An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis

Yun Wan
Faculty of Computer Science
Dalhousie University
Halifax, NS B3H 4R2, Canada
yn378100@dal.ca

Dr. Qigang Gao
Faculty of Computer Science
Dalhousie University
Halifax, NS B3H 4R2, Canada
qggao@cs.dal.ca

***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, an ensemble sentiment classification strategy was applied based on Majority Vote principle of multiple classification methods, including Naive Bayes, SVM, Bayesian Network, C4.5 Decision Tree and Random Forest algorithms. In our experiments, six individual classification approaches, and the proposed ensemble approach were all trained and tested using the same dataset of 12864 tweets, in which 10 fold evaluation is used to validate the classifiers. The results show that the proposed ensemble approach outperforms these individual classifiers in this airline service Twitter dataset. Based on our observations, the ensemble approach could improve the overall accuracy in twitter sentiment classification for other services as well.**

Keywords— Twitter data mining; Sentiment classification; Airline services analysis; Ensemble classification

I. INTRODUCTION

Customer feedback analysis is one of the essential components 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. It needs 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 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. 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.

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. In this paper, six popular classification methods are compared against the proposed ensemble approach, which integrates the five individual algorithms into an ensemble based classification decision making, including Naïve Bayesian classifier, Support Vector Machine (SVM) classifier, Bayesian Network classifier, C4.5 Decision Tree classifier and Random Forest classifier. The ensemble classification takes into account classification results of the five classifiers and uses the Majority-Vote method to determine the final sentiment class prediction. The detail experimental comparisons of the six different sentiment classification methods and analytics comments are provided.

The paper is organized as follows. In section 1, the motivation are explained and the objective is introduced. In section 2 the relevant work is briefly surveyed and summarized. Section 3 presents the data preparation process, including the data collection, data pre-processing and feature selection procedure. In section 4, the seven classifiers are briefly presented, including the proposed ensemble method on sentiment classification. In section 5, experiment design and evaluation results are presented, analysis comments are also provided., Section 6 provides the conclusion, which summarizes our findings from this research and some future research directions..

II. RELATED WORK

Sentiment classification 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 [1]. The

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.7

most important application of opinion mining and sentiment classification has been customer review mining. There have been many studies recorded on different review sites.

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 weakness of this approach is poor recognition of affect when negation is involved [2]. Many supervised methods have been applied to sentiment classification and these systems were mainly based on supervised learning relying on manually labelled samples [3]. The Naïve Bayes method has been a very popular method in text categorization because its simplicity and efficiency [4]. 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 Naïve Bayesian method is that the attributes in classification are independent to each other, which considerably reduces the computing complexity of the classification algorithm. The Support Vector Machine (SVM) method was considered the best text classification method [5]. The Support Vector Machine method is a statistical classification approach which is based on the maximization of the margin between the instances and the separation hyper-plane. This method is proposed by Vapnik. Different from other machine learning methods, the K-nearest neighbors (KNN) method does not extract any features from the training dataset but compare the similarity of the document with its neighbors [6]. In feature selection part, Songbo Tan [7] compared four feature selection approaches and five machine learning methods on Chinese texts. He concluded that the Information Gain algorithm outperforms other feature selection approaches and the Support Vector Machine approach works best in sentiment classification. Yi et al [8] also discovered that the Support Vector Machine approach performs better than the Naïve Bayesian approach and an N-gram model do.

Big Data social data analysis has been very popular [9]. 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 Naïve Bayesian approach and the Support Vector Machine approach, both of which reached accuracy up to 70% [10]. 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 [11]. 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 [12]. 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 didn't contain emoticons, which are the majority of tweets. In this work, they didn't 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 Naïve Bayesian approach, have been applied to some tweet data and the performance was not bad [12]. Lee et al used twitter as the data source to analyze consumers' communications about airline services [13]. 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 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 [14]. 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 labeled data. It can yield inaccurate testing results because sentiment classifications are highly domain specific. Adeborna et al adopted Correlated Topics Models (CTM) with Variational Expectation-Maximization (VEM) algorithm [15]. Their lexicons for classification were developed with Airline Quality Rating (AQR) criteria. In Sentiment detection process, the performances of the SVM classifier, the Maximum Entropy classifier and Naive Bayesian classifier were compared and Naive Bayesian classifier was adopted. Besides that, tweets are categorized by topics using the CTM with the VEM algorithm. In this research, the overall dataset they used contains only 1146 tweets, which includes only three airline companies. Besides, the author only used unigrams as sentiment classification features in the Naive Bayesian classifier, which can cause problems because phrases and negation terms can change sentiment orientation of the unigrams in sentences. 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. However, our work not only contains the analysis of tweets with different sentiments but also includes the comparison of the performance of different approaches.

III. DATA PREPARATION

We use the Twitter Search API to retrieve tweets data about 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 the word 'flight' to collect tweets that convey airline services feedback. To get a full and comprehensive coverage of English tweets about 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 Airlines, JetBlue Airways, United Airlines, Air Canada, SouthWest Airlines, AirTran Airways, WestJet, American Airlines, Frontier Airlines, Virgin Airlines, Allegiant Air, Spirit Airlines, US Airways, Hawaiian Airlines, Skywest Airline, Alaska Air Group [16]. Retrieving tweets about those brands can build the best dataset for sentiment analysis of airline services.

These tweets include original tweets and retweets. We discard the irrelevant tweets and label each relevant tweet in the dataset as positive sentiment, negative sentiment or neutral sentiment manually. In total, there is a dataset containing 107866 tweets in our work. In the dataset, 4288 tweets are labeled positive, 35876 tweets are labeled negative, 40987 tweets are labeled neutral and 26715 tweets are discarded for being irrelevant.

Table 1 Class distribution

class	positive	negative	neutral	irrelevant
tweets	4288	35876	40987	26715

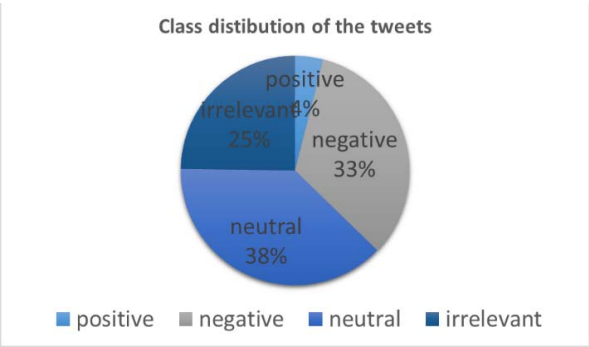

Figure 1 Class distribution

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. For example, in the Naïve Bayesian classification model training, as shown in (1).

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

The probability of the document D being classified as the sentiment class S is $p(S|D)$, which is determined by $p(S)$, $p(D)$ and $p(w_i|S)$. If the class distribution in the training data is not balanced, then $p(S|D)$ will be biased because $p(S)$ are different for different classes. We randomly resample a dataset with exact same number of documents for each sentiment class. We get a dataset with 4288 documents for each sentiment class and 12864 documents in total. We remove the punctuations, symbols, emoticons and all other non-alphabet characters from the tweets. Besides that, we also remove web links and decapitalized all of tweets because these features provides little information in sentiment classifications.

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, 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. Nevertheless, stemming techniques still can considerably reduce the sparsity of the features.

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 [17]. As shown in Figure 2 the Information Gain decreases as the feature length increases.

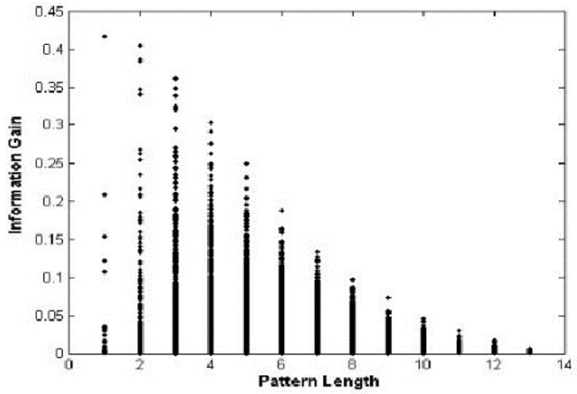

Figure 2 IG for different length features[14]

We use Weka to compute the Information Gain for each attribute and rank them in decreasing order. In Weka, We select the supervised filter, Attribute Selection to implement feature selections. In the Attribute Selection filter, we select the InforGainAttributeEval algorithm for the evaluator option and the Ranker algorithm for the search option. We keep the default value of the threshold for the Ranker algorithm, which is -1.7976931348623157E308. By keeping the threshold default value, the algorithm ranks all of the attributes decreasingly without removing any attributes. We export the ranking results and plot them in a line chart to see the rates of information gain decreasing. As shown in Figure 3, the x coordinate is the ranks for the attributes and the y coordinate is the information gain for each attribute.

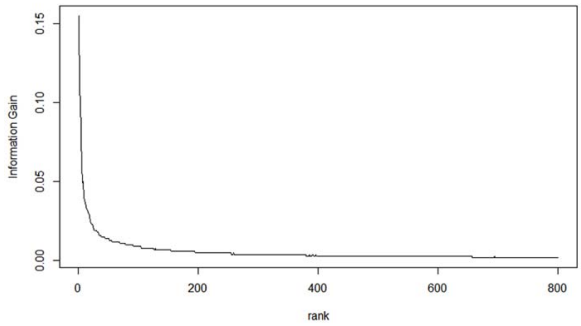

Figure 3 Information Gain for features

There is a cutoff of the Information Gain between the feature which ranked 656 and the features ranked below. So for our experiment, the features that ranks above the 656th feature are selected.

In data transformation, the first step is to make all distinct features appearing in the tweet data a set. This set contains all the distinct features appearing in the tweet dataset and no duplicate features exist in this set. The second step of data transformation is to make a matrix and in this matrix, each column represents a feature appearing in the feature set from the previous step. By doing this, we can convert each of the tweet documents into a binary row of the matrix. For any feature appearing in a tweet document, there must be a column in the matrix representing it. When converting tweet documents to binary matrix rows, for each column in the row, if the feature it represents appears in this tweet document then its value is set to 1, and if the feature it represents does not appears in this tweet document then the value is set to 0.

IV. Methodology and System Design

Here we describe seven different classifiers using different classification methods. They are the Lexicon-based classifier, the SVM classifier, the Naive Bayesian classifier, the Bayesian Network classifier, the C4.5 Decision Tree classifier, the Random Forest classifier and the ensemble classifier.

A. 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 scans 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 adopt the word lists produced by Hu and Liu in their work *Mining and Summarizing Customer Reviews* [14] and we add four words including 'delayed', 'late', 'oversold' and 'bumped' into the negative word list because those words indicate strong negative sentiment in the airline services domain.

B. 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. The Naïve 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 Weka to implement experiments and tests.

C. Bayesian Network classifier

Like Naïve Bayes method, Bayesian Network also derives from Bayes' theorem, 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.

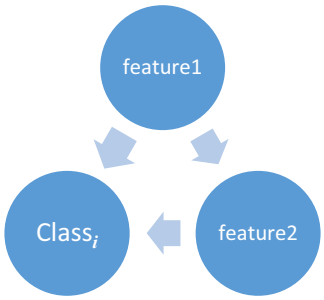

Figure 4 Bayesian Network Model

As illustrated in Figure 4, 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:

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

In (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.

D. 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. In the pairwise classification method, each pair of classes will have one SVM classifier trained to separate the classes. The accuracy of the classification will be the overall accuracy of every SVM classification included. We adopt pairwise classification approach in the SVM classification method. We utilized the libSVM algorithm in Weka, which use pairwise classification for multiclass SVM classification, in Weka to train the SVM classifier and implement experiments and tests.

E. C4.5 Decision Tree

A Decision Tree is a flowchart-like tree structure, in which each internal node represents a test on an attribute and each branch represents an outcome of the test, and each leaf node represents a class. The first popular Decision Tree algorithm was Iterative Dichotomiser 3 (ID3), developed by J. Ross Quinlan. Because the ID3 algorithm keeps iterating the process of splitting subset data, it can cause the over-fitting problem. Besides, ID3 algorithm cannot deal with continues attributes or attributes containing missing values. As an extension of ID3, C4.5 was developed by Ross Quinlan to solve these problems.

C4.5 discretizes the continuous attribute by setting a threshold and splitting the data to a group whose attribute value is above the threshold and another group whose attribute value is below or equal to the threshold. C4.5 handle missing values in attribute by just not using the missing values in Information Gain calculations. C4.5 handles over-fitting problems by using the post-pruning approach. C4.5 uses a post-pruning approach called pessimistic pruning, which uses the Cost Complexity pruning algorithm and uses the training data to estimate the error rate. Error rate of the tree is the percentage of misclassified instances in the tree. node and its subtree are calculated and compared. We adopt the J48 C 4.5 algorithm in Weka.

F. Random Forest

Because the Decision Tree generated by ID3 algorithm and C4.5 algorithm are not necessarily the best decision tree for classification, Random Forest was developed as an ensemble approach based on many decision trees. Random Forest uses the Majority Vote method and returns the class with most votes. Random Forest uses the Bagging approach in building classification models. For a dataset, D, with N instances and A attributes, the general procedure to build a Random Forest ensemble classifier is as follows. For each time of building a candidate Decision Tree, a subset of the dataset D, d, is sampled with replacement as the training dataset. In each decision tree, for each node a random subset of the attributes A, a, is selected as the candidate attributes to split the node. By building K Decision Trees in this way, a Random Forest classifier is built. In classification procedure, each Decision Tree in the Random Forest classifiers classifies an instance and the Random Forest classifier assigns it to the class with most votes from the individual Decision Trees. In our experiment, the Random Forest algorithm in Weka is adopted.

G. The Ensemble Classifier

We conducted experiments with the six classification models. We used the 10-fold validation plan to evaluate the machine learning classification approaches including: the Naïve Bayesian classifier, the SVM classifier, the Bayesian Network classifier, the C4.5 Decision Tree and the Random Forest classifier. Test results for the three-class classification experiment are shown in table 2. The Lexicon-based classifier got the lowest accuracy, which is 60.5%. The accuracy of the Naïve Bayesian model classification reached 81.8%. The Bayesian Network classifier outperformed the Lexicon-based classifier and the Naïve Bayesian classifier by reaching an accuracy of 85.1%. The SVM classifier got an accuracy of 74.7%, the C4.5 Decision Tree got an accuracy of 82.9% and the Random Forest classifier got an accuracy of 82.4%.

Table 2 Accuracy in three-class dataset

Classifier	Lexicon Based	Naïve Bayesian	Bayesian Network	SVM	C4.5	Random Forest
Accuracy	60.5%	81.8%	85.1%	74.7%	82.9%	82.4%

Table 3 Accuracy in two-class dataset

Classifier	Lexicon Based	Naïve Bayesian	Bayesian Network	SVM	C4.5	Random Forest
Accuracy	67.9%	90.0%	91.4%	84.6%	86.0%	89.8%

Test results in two-class dataset are pretty the same, which shows the Lexicon Based classifier is much worse than other classifiers.

After comparing the experiment results of the six sentiment classifiers. We select Naïve Bayesian method, Bayesian Network method, Support Vector Machine (SVM) method, C4.5 Decision Tree method and Random Forest method to build the ensemble classifier.

Figure 5 present the ensemble system built by the five independent classifiers. They are all trained independently with sub-sample training datasets from the overall training set.

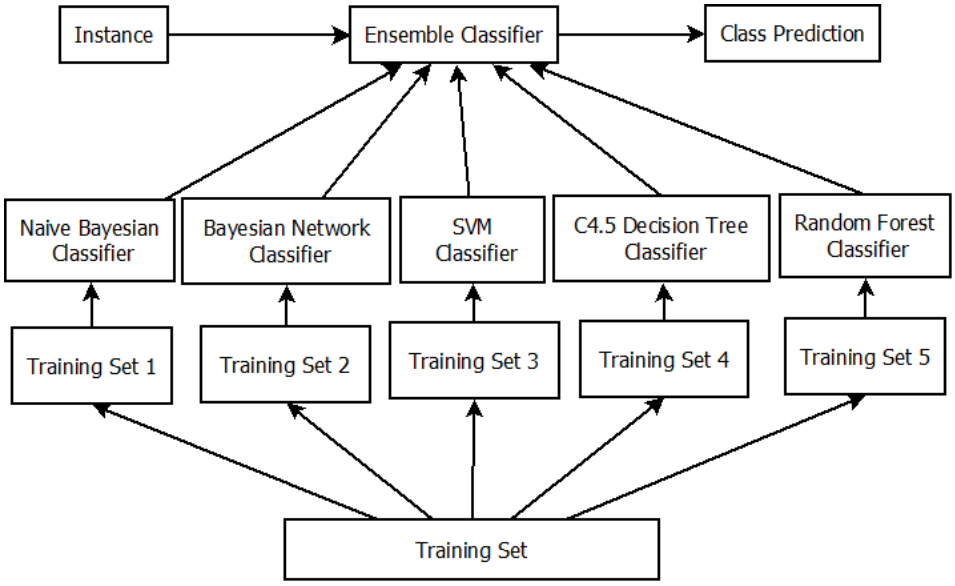

Figure 5 Ensemble System

The ensemble classifier uses the Majority Vote method to classify each document's class. The five classifiers have the same weights in the majority vote process. In the sentiment classification procedure, each tweet instance is classified independently by each of the five classifiers, and the final class prediction result is classified to the class which has the most votes in the five class prediction results.

Require: C (five classification results for one instance), S (sentiment classes: Positive, Negative and Neutral)
if for any class i in S, C_i in C
count(C_i) >=3
then Class=C_i
else if for any two class j, k in S, C_j and C_k in C
count(C_j) = count(C_k)=2
then Class=C_j or Class= C_k
return Class

Figure 6 The Majority Vote combination rule

As shown in the pseudo code in Figure 6, a tweet document is assigned an arbitrary class if the classification results of the five classifier cannot be determined by the Majority Vote method. For example, if a tweet document is classified to negative by the Naïve Bayesian classifier, classified to neutral by the Bayesian Network classifier, classified to positive by the Support Vector Machine (SVM) classifier, classified to positive by the Decision Tree classifier and classified to negative by the Random Forest. Then this tweet document is assigned arbitrarily to either of the two classes, which are positive or negative, because this tweet document has same probabilities of being either of the two classes.

In our paper, We use 10-fold validation. We use the same dataset to train the Naïve Bayesian classifier, the Bayesian Network classifier, the Support Vector Machine (SVM) classifier, the C4.5 Decision Tree classifier and the Random Forest classifier individually. The process of the model training and classification are implemented with Weka.

V. EXPERIMENT AND EVALUATION

A. 5.1 Evaluation Plan

1) Classification Validation

Generally speaking, over-fitting happens when the training data is relative small, and cross-validation is a good solution to avoid this. In our research, We take 12,864 tweets data, which is relatively not small, but it is still a good choice to implement cross-validation. Cross-validation is a method for model validation, which samples a subset of data to do model training and another subset of data to do model validation. 10-fold validation is one cross-validation method. In 10-fold validation, the dataset is randomly partitioned into 10 subsets with equal sizes. In the model training and validation process, each 9 subsets of data is used as a training dataset to train a model and the remaining 1 subset is used to validate the model. After repeating 10 times, each 9 subsets have been used as a training dataset to train a model and 10 classification validation results are produced. The overall validation result of the 10-fold validation is the average validation result of the 10 models. In the data mining research area, 10-fold validation is a popular validation method and it is used in our experiment.

2) Two-Class Dataset and Three-Class Dataset

The experiment is implemented in both the three-class dataset, which includes the positive sentiment, the negative sentiment and the neutral sentiment, and in the two-class dataset, which only includes the positive and the negative sentiments. The two-class dataset is from the three-class dataset by just deleting the neutral tweets.

3) 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 [4]. For example, if 900 tweets are classified to positive and 800 of them are correct, and in the dataset there are 1000 tweets which are positive, then the recall for the positive class is 800/1000, which equals to 0.8.

b) Precision

Precision is the fraction of the correctly classified instances for one class of the overall instances which are classified to this class [4]. For example, if 900 tweets are classified to positive and 800 of them are correct, and in the dataset there are 1000 tweets which are positive, then the Precision for the positive class is 800/900, which equals to 0.89.

c) F-measure

To get a comprehensive evaluation of the classification, F-measure is developed to integrate the Recall and the Precision. The F-measure can be expressed as

$$F_\beta = (1+\beta^2) \times \frac{precsion \times recall}{\beta^2 \times precsion + recall} \qquad (3)$$

This is a general form of F-measure and the parameter β is used to change the weights for Precision and Recall in calculating the F-measure value. In our work, because recall and precision are equally important [4]. We set β to 1, and it is called the harmonic mean of precision and recall. The formula can be rewritten as:

$$F_{0.5} = 2 \times \frac{precsion \times recall}{precsion + recall} \qquad (4)$$

d) Error Rate

To illustrate the accuracy of different classifiers, we use the error rate on bar chart. Error rate is the (1-F-value).

B. Experiment Result Evaluation

As shown in Table 4, in the experiment with three classes, the ensemble classifier gets the highest accuracy in terms of Precision, Recall and F-measure, which is 84.2%.

Table 4 Accuracy in the three-class dataset

Classifiers	Precision	Recall	F-measure
Lexicon-based	60.5%	61.5%	61.0%
Naïve Bayesian	82.4%	82.2%	82.3%
Bayesian Network	82.2%	82.3%	82.2%
SVM	77.8%	77.0%	77.4%
C4.5 Decision Tree	83.7%	83.6%	83.6%
Random Forest	83.5%	83.4%	83.4%
Ensemble	84.2%	84.2%	84.2%

As shown in Figure 5, the error rate for the ensemble classifier is the lowest, which is 15.8%

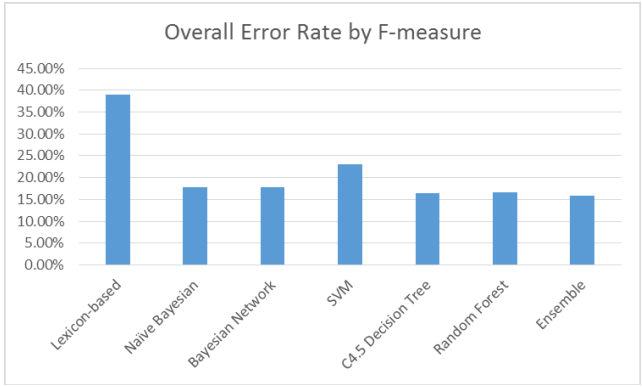

Figure 7 Three-class error rate by F-measure

Table 5 Accuracy in the two-class dataset

Classifiers	Precision	Recall	F-measure
Lexicon-based	67.3%	67.9%	67.6%
Naïve Bayesian	90.4%	90.3%	90.4%
Bayesian Network	90.5%	90.5%	90.5%
SVM	87.2%	87.0%	87.1%
C4.5 Decision Tree	87.4%	87.3%	87.3%
Random Forest	90.8%	90.8%	90.8%
Ensemble	91.7%	91.7%	91.7%

As shown in Table 5, in the experiment with only two classes, positive class and negative class, the ensemble classifier also gets the highest accuracy, which is 91.7%.

As shown in Figure 6, the error rate for the ensemble classifier is still the lowest in the two-class experiment, which is 8.3%

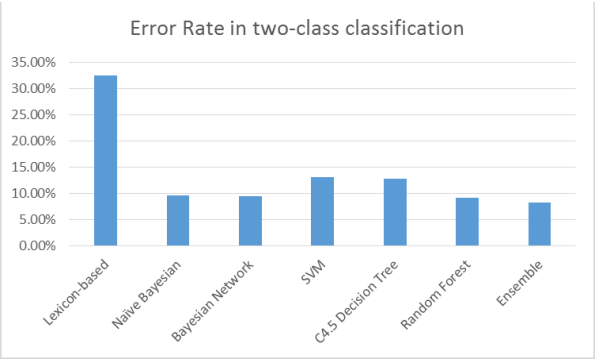

Figure 8 Two-class error rate by F-measure

VI. CONCLUSION

This paper makes empirical contributions to this research area by comparing the performance of different popular sentiment classification approaches and developing an ensemble approach, which further improves the sentiment classification performance. In the domain of twitter sentiment analysis about airline services, little work has been done. These past work compares several different traditional classification methods and selects the most accurate individual classification method to implement sentiment classification. However, the ensemble approach we present improves the accuracy by combining these sentiment classifiers. 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 paper, 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.

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] Cambria, Erik, Bjorn Schuller, Bing Liu, Haixun Wang, and Catherine Havasi. "Knowledge-based approaches to concept-level sentiment analysis." IEEE Intelligent Systems 2 (2013): 12-14

[3] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2.1-2 (2008): 1-135.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] 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.

[5] Xia, Rui, Chengqing Zong, and Shoushan Li. "Ensemble of feature sets and classification algorithms for sentiment classification." Information Sciences 181.6 (2011): 1138-1152.

[6] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining, southeast asia edition: Concepts and techniques. Morgan kaufmann, 2006.

[7] Tan, Songbo, and Jin Zhang. "An empirical study of sentiment analysis for chinese documents." Expert Systems with Applications 34.4 (2008): 2622-2629.

[8] Yi, Jeonghee, and Wayne Niblack. "Sentiment mining in WebFountain." Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on. IEEE, 2005.

[9] E. Cambria, H. Wang, and B. White, "Guest editorial: Big social data analysis," Knowledge-Based Systems, vol. 69, pp. 1–2, 2014.

[10] 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.

[11] 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.

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

[13] 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.

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

[15] Adeborna, Esi, and Keng Siau. "An approach to sentiment analysis–the case of airline quality rating." (2014).

[16] 2014. Major Canadian and US Airlines. http://www.nationsonline.org/oneworld/Airlines/airlines_north_america.htm.

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