International Journal of Computer Science and Business Informatics

IJCSBI.ORG

Performance Evaluation of Sentiment Mining Classifiers on Balanced and Imbalanced Dataset

G.Vinodhini

Department of Computer science and Engineering,
Annamalai University, Annamalai Nagar -608002.

R M. Chandrasekaran

Department of Computer science and Engineering,
Annamalai University, Annamalai Nagar -608002.

ABSTRACT

The transition from Web 2.0 to Web 3.0 has resulted in creating the dissemination of social communication without limits in space and time. Sentiment analysis has really come into its own in the past couple of years. It's been a part of text mining technology for some time, but with the rise in social media popularity, the amount of unstructured textual data that can be used as a machine learning data source, is enormous. Marketers use this data as an intelligent indicator for customer preferences. This paper aims to evaluate the performance of sentiment mining classifiers for problems of unbalanced and balanced large data sets for three different products. The classifiers used for sentiment mining in this paper are Support Vector Machine (SVM), Naïve bayes and C5.The results shows that the performance of the classifiers depends on the class distribution in the dataset . Also balanced data sets achieve better results than unbalanced datasets in terms of overall misclassification rate.

KEYWORDS

Sentiment, opinion, SVM, classifiers, balanced, imbalanced.

1. INTRODUCTION

Sentiment analysis is a part of text mining technology, but with the rise in social media popularity, the amount of unstructured textual data that can be used as a machine learning data source, is enormous. Sentiment analysis is understanding the meanings and feelings behind statements made in social media and other forums (Pang, Bo., & Lee, L.,2004, Kunpeng Zhang et al, 2010, X. Fu et al, 2013). Public opinions and sentiments can have major impact on our society. They can affect the sales of products, the change of government policy, and even people's vote in elections. Thus, it is of high significance to study sentiment analysis also known as opinion mining. In the age of the Web, more and more

people choose to express their opinions on a wide range of topics on the Web in the forms of blogs, product/service reviews, and comments (A. Balahur et al, 2012). The amount of data exchanged over social media is witnessing a major growth in the last few years. Opinion mining at both the document level and sentence level has been too coarse to determine precisely what users like or dislike (Turney, P. D. 2002). In order to address this problem, sentiment mining at the attribute level is aimed at extracting opinions on products specific attributes from reviews in this work (Magdalini et al, 2012). Various studies in different domains investigated extracting sentiment information from this exchanged data. Less attention was directed toward studying the effect of class imbalance problem in sentiment mining. In recent years, class imbalance problem has emerged as one of the challenges in data mining community. This situation is significant since it is present in many real-world classification problems.

Previous studies have used a balanced dataset, however in the product domain it is commonly the case that the ratio of positive and negative reviews is unbalanced, therefore this paper focuses on and investigating the effects of the size and ratio of a dataset. The proposed system architecture takes customer reviews as input to each of the classifiers and outputs the dataset split into positive and negative reviews.

In this work, we analyze the performance of three different classifiers like SVM, Naive Bayes (NB) and C5 for sentiment mining. The classification model uses product attributes as features. The models are empirically validated using review data sets of nokia, ipod and nipon camera.To analyse the effect of class distribution two data models are developed. Model A using balanced class distribution i.e. equal number of positive and negative classes. Model B using unbalanced class distribution i.e. unequal number of positive and negative classes .The results of three different classifiers are compared for both Model A and Model B.

This paper is outlined as follows. Section 2 discusses about the related work. Section 3 describes the proposed work used.. The various classification methods used to model the prediction system are introduced in Section 4. The Experimental analysis done is reported in Section 5. Section 6 summarizes the results and Section 7 concludes our work.

2. RELATED WORK

The area of sentiment mining has seen a large increase in academic interest in the last few years. Researchers in the areas of natural language processing, data mining, machine learning, and others have tested a variety of methods of automating the sentiment analysis process. A number of machine learning techniques have been adopted to classify the reviews based on sentiment. Various

machine learning methods like Support vector machines (SVM), Naive Bayes (NB), Maximum Entropy (ME), K-Nearest neighbourhood, ID3, C5 and centroid classifier classification have been already applied in sentiment classification. (Songho tan et al., 2008 ; Qingliang et al., 2009; Rui Xia et al., 2011, Hassan Saif et al, 2012). Various comparative studies have been done to find the best choice of machine learning method for sentiment classification. As the result of a sentiment analysis varies according to the composition method of a domain and feature and the type of learning algorithm, a need to perform comparative analysis arises.

Inspite of using various single classifiers, many works has been done in recent years focussing on the combination of classifier like hybrid and ensemble methods to improve the classification accuracy (Rudy Prabowo et al.,2009; Whitehead et.al., 2008) . From the literature review done, it is also observed that only a very few studies has been conducted in analysing the performance of classifiers on class imbalanced condition.Most of the existing works are based on product review datasets because a review usually focuses on a specific product and contains little irrelevant information. These datasets have an even number of positive and negative reviews, however in the product domain it is typical that there are substantially more positive reviews compared to negative reviews. Our work will therefore compare the effects of a balanced and unbalanced dataset.

The main objective of the work is to perform feature based sentiment mining to decide whether the opinions are positive or negative. Moreover the main focus in on evaluating the performance of various classifiers in two different data distributions i.e. class balanced and class imbalanced.

3. METHOD

The following list describes the methodology of the proposed work

i. Identify the data sources.
ii. Create two datasets i.e balanced and unbalanced for each product.
iii. Preprocess the data to remove noise and redundancy.
iv. Identify the features for creating a word vector model.
v. Develop two word vector model
    a. Model A using balanced dataset with term presence method.
    b. Model B using unbalanced dataset with term presence method.
vi. Develop the classification models
    a. Naïve Bayes
    b. Support Vector machine
    c. C5
vii. Predict the result for classification and compare with the actual results.

viii. Evaluate the performance of classifiers using overall misclassification rate.

a. Classification Methods

The following section describes about the various classification methods used in this work. Most of the literatures showed that SVM and Naive Bayes and C5 are perfect methods in sentiment classification.

Naïve Bayes Classifier

Bayesian learning algorithms use probability theory as an approach to concept classification. Bayesian classifiers produce probabilities for class assignments, rather than a single definite classification. Naïve Bayes classifier (NBC) is perhaps the simplest and most widely studied probabilistic learning method. It learns from the training data, the conditional probability of each attribute *Ai*, given the class label *C*. The strong major assumption is that all attributes *Ai* are independent given the value of the class *C*. Classification is therefore done applying Bayes rule to compute the probability of *C* and then predicting the class with the highest posterior probability. The assumption of conditional independence of a collection of random attributes is very critical.

Support Vector Machines

Support Vector Machines (SVMs) are pattern classifiers that can be expressed in the form of hyper-planes to discriminate positive instances from negative instances. SVMs have successfully been applied to numerical tasks, including classification. They perform structural risk minimization and identify key "support vectors". Risk minimization measures the expected error on an arbitrarily large test set with the given training set and SVMs non-linearly map their *n*-dimensional input space into a high dimensional feature space. In this high dimensional feature space a non-linear classifier is constructed. Given a set of points which belong to either of two classes, a linear SVM finds the hyper-plane leaving the largest possible fraction of points of the same class on the same side, while maximizing the distance of either class from the hyper-plane. The hyper plane is determined by a subset of the points of the two classes, named support vectors, and has a number of interesting theoretical properties.

C5

C5 is one of the simplest forms of supervised learning algorithm. It has been extensively used in many areas such as statistics and machine learning for the purposes of classification and prediction. C5 classifiers can be generalize beyond the training sample so that unseen samples could be classified with as high accuracy as possible. C5s are non-parametric and a useful means of representing

the logic embodied in software routines. C5 takes as input a case or example described by a set of attribute values, and outputs a Boolean decision . In the classification case, when the response variable takes value in a set of previously defined classes the node is assigned to the class which represents the highest proportion of observations

b. Data Preparation

In order to compare the two classifiers, naive Bayes and language model, we looked at the results based on a balanced and an unbalanced dataset and also the consistency of results when the dataset was different sizes. To conduct our experiments we created the following datasets;

- Unbalanced dataset - all reviews extracted, a realistic representation of the ratio of positive and negative reviews (Model B)
- Balanced dataset - all negative reviews and the same number of positive reviews.(Model A)

We used the publicly available customer review datasets (http://www.cs.uic.edu /~liub/FBS/sentiment-analysis.html). These dataset contains annotated customer reviews of various products. We have selected reviews of 3 different products like Nokia 6600, iPod, Nikon coolpix. There reviews are presented in plain text format. The dataset consists of negative, positive and neutral reviews. In this binary classification problem, we have considered only positive and negative reviews. The product attribute discussed in the review sentences are collected for each review sentences. Unique product features are grouped, which results in a final list of product attributes (features). A word vector representation of review sentences is created for Model A and B. The word vector set can then be reused and applied to different classifiers. To create the word vector list, the review sentences are pre-processed. The descriptions of review dataset models to be used in the experiment are given in Table 1. For our investigation we created two data model, one balanced and other unbalanced. The dataset is made balanced by random sampling.

Table 1. Descriptions of review dataset

Product	Model A (Balanced)		Model B (Unbalanced)	
	Positive	Negative	Positive	Negative
Nokia 6600	175	175	414	186
Ipod	120	120	328	122
Nikon coolpix	98	98	176	98

4. RESULTS & DISCUSSION

The classification model used is employed using Weka tool. The parameters for classifiers use the default values available in the tool. Experiments used a 10-fold cross validation. Each dataset was randomly spilt into 10 folds, 9 folds used for training and 1 fold used for testing. The average of the 10-folds was then used for performance analysis. In order to evaluate the accuracy of the classification model, overall misclassification rate is used as a metric. Misclassification rate is defined as the ratio of number of wrongly classified reviews to the total number of reviews classified by the prediction system. Misclassification rate considers both positive and negative reviews in formula. We first focus on the commonly used balanced dataset. Table 2. and fig 1 show the overall misclassification rate of each classifier. Then we focus on the unbalanced dataset. Table 3. and fig 1. shows the overall misclassification rate of each classifier.

Table 2: Results of balanced dataset

Product	Overall misclassification rate		
	Naive Bayes	SVM	C5
Nokia 6600	12.6	11.1	12.1
Ipod	11.3	9.8	10.5
Nikon coolpix	10.5	9.2	9.9

Table 3: Results of Unbalanced dataset

Product	Overall misclassification rate		
	Naive Bayes	SVM	C5
Nokia 6600	22.8	20.8	21.7
Ipod	21.4	19.6	20.8
Nikon coolpix	19.3	18.5	19.1

The overall misclassification rate is reduced considerably for Model A than Model B for all three methods used. This is due to the class imbalance nature of Model B. Model B has nearly 50% more positive reviews than negative reviews. This results in higher Type I error (number of negative reviews wrongly classified as positive). Hence increases the overall misclassification rate.

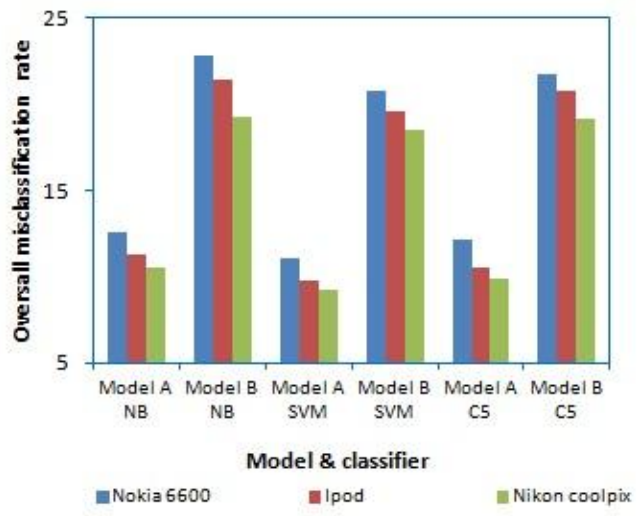

Figure 1. Overall misclassification of Model A and Model B

5. CONCLUSION

The major contribution of the paper has been the application of three different machine learning algorithms to predict sentiment orientation of the review sentences and to evaluate the effect of class distribution on classifier performance. Three different product review datasets were utilized for this task. The results suggest that the machine learning algorithms can be successfully applied in sentiment mining under balanced distribution of classes. Though classifiers perform better in balanced distribution, it has been found that among all classifiers (c5, NB and SVM), SVM performs better in balanced and imbalanced conditions. While many researches continue, practitioners and researchers may apply various sampling methods for under sampling and over sampling to construct a balanced model from an imbalanced model. We plan to replicate our study to predict the models based on hybrid machine learning algorithms under data imbalanced condition.

REFERENCES

[1]. A. Balahur, J.M. Hermida, A. Montoyo, Detecting implicit expressions of emotion in text: a comparative analysis, Decision Support Systems 53 (2012) 742–753.

[2]. Hassan Saif, Yulan He and Harith Alani, Semantic Sentiment Analysis of Twitter Knowledge Media Institute, The Open University, United Kingdom, 2012.

[3]. Kunpeng Zhang, Ramanathan Narayanan, "Voice of the Customers: Mining Online Customer Reviews for Product", 2010.

[4]. Magdalini Eirinaki, Shamita Pisal , Japinder Singh , "Feature-based opinion mining and ranking" , Journal of Computer and System Sciences 78 (2012) 1175–1184.

[5]. M. Whitehead, L. Yaeger, Opinion mining using ensemble classification models, in: International Conference on Systems, Computing Sciences and Software Engineering (SCSS 08), Springer, 2008.

[6]. Pang, Bo., & Lee, L. (2004). A opinional education: Opinion analysis using subjectivity summarization based on minimum cuts. In Proceedings 42nd ACL.

[7]. Popescu, A. M., Etzioni, O.: Extracting Product Features and Opinions from Reviews, In Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, 2005, 339–346.

[8]. Qingliang Miao, Qiudan Li, Ruwei Dai , "AMAZING: A sentiment mining and retrieval system", Expert Systems with Applications 36 (2009) 7192–7198.

[9]. Sujan Kumar Saha, Sudeshna Sarkar, Pabitra Mitra , " Feature selection techniques for maximum entropy based biomedical named entity recognition" , Journal of Biomedical Informatics 42 (2009) 905–911

[10]. Rudy Prabowo, Mike Thelwall, "Sentiment analysis: A combined approach .", Journal of Informatics , (2009) 143–157.

[11]. Rui Xia , Chengqing Zong, Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences 181 (2011) 1138–1152.

[12]. Songbo Tan , Jin Zhang, "An empirical study of sentiment analysis for chinese documents ", Expert Systems with Applications 34 (2008) 2622–2629.

[13]. Turney, P. D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, In Proceedings of the 40th Annual Meetings of the Association for Computational Linguistics.

[14]. X. Fu, G. Liu, Y. Guo, Z. Wang, Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon, Knowledge-Based Systems 37 (2013) 186–195.