International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 02 - Issue 08 || August 2017 || PP. 67-73

Opinion Mining on News Articles Using Feature Reduction Method

N. Sudha
Research Scholar
Department of Computer Science and Engineering
Annamalai University

Dr.M.Govindarajan
Assistant Professor
Department of Computer Science and Engineering
Annamalai University

Abstract: Sentiment analysis is the process of extracting opinions and emotions expressed in a text shared by the users on social media such as blogs, web forum, online news and web pages etc. Sentiment can be determined using a data mining technique called machine learning. In this paper, used Naïve Bayes and Support vector machine are the most common classifiers. Currently large number of features is available over the social media. Due to this, may lower the classification performance in terms of accuracy and increase the computational cost too. Additionally, the feature reduction technique is required by selecting an optimal sub set of features and that to eliminate the irrelevant features. The feature reduction technique is gain ratio and ranker search algorithm is to rank all the attributes over two types of datasets such as email and Usenet So this experiments did with feature reduction method and their accuracies is compared with the two individual classifiers.

General Terms: Data mining, Classification, Data sets, Social network feature reduction and Algorithms
Keywords: Sentiment analysis, opinion extraction, news articles, Support vector machine, Naïve Bayes

1. Introduction

In recent times, immensely large amount of reviews (data) such as product, movie, etc. can be uploaded in the internet. Currently many people can post their opinion through various social media like forums, blogs or news articles sites and discussing current issues of finance, education, religion, politics and a host of general social issues. The web discourse domain of sentiment analysis which includes rating of web forums, newsgroups, and blogs [1]. By this domain, users can communicate with others in different parts of the world to come and share their feelings, opinion and discuss issues of common interest. Thousands of web forums devoted to multiple of topics exist where millions of users often participate in various discussions. People use web forums and web pages to discuss and ask questions about various topics such as news, sports, technology, health, etc. The archives of web forums contain millions of such discussion threads and act as a valuable repository of human generated information that needs to be efficiently managed [2]. To evaluate the opinion of the users is not easy since the user generate content can be numerical or text file, can be structured, semi-structured or non-structured. Text mining is an approach used several kinds of fields such as machine learning, information retrieval, statistics, and computational linguistics for opinion mining. Web mining is a subset of text mining used to mine the unstructured web data. Therefore there is a need of a automation system will automatically extract the opinion and categorize the emotions called as Sentiment analysis which is also known as opinion mining [3, 4]. It is a type of natural language processing that analyzes people's opinions, sentiments and emotions towards the entities (products, services, organizations, individuals, issues, events, topics and their attributes) and classify the polarity of a given text at the document, sentence whether the expressed opinion in a document, a sentence or an entity feature is positive, negative, or neutral. Such sentiments are useful for the Consumers can obtain opinion to research products or services before making a purchase and Marketers can obtain public opinion about their company and products, or to analyze customer feedback. Finally, organizations can also obtain critical feedback about problems in newly released products. The main goal of sentiment analysis is to analyze the reviews and examine the scores of sentiments. Sentiment analysis is also more important in social media, since number of people accessing the social network from anywhere [5]. Through the network they can share their feelings and post about products and services or express their political and religious views especially in micro blogging web-sites. Such data can be efficiently used for marketing or social studies. Politicians may be interested to know if people support their program or not. Social organizations

may ask people's opinion on current debates. All this information can be obtained from micro blogging services, as their users post everyday what they like/dislike, and their opinions on many aspects of their life [6].

There are multiple methods for measuring sentiments, including lexical based approaches and supervised machine learning methods. Machine learning based approach uses classification technique to classify text. Two sets of documents are needed: training and a test set. A training set is used by an automatic classifier to learn the differentiating characteristics of documents, and a test set is used to check how well the classifier performs. Lexicon based method uses sentiment dictionary with opinion words and match them with the data to determine polarity [7]. Subsequent research focused on supervised learning techniques that are common in text classification includes Support Vector Machine (SVM) and Naive Bayes (NB) classifiers. Though these techniques are far more accurate than the earlier text-based approaches, they are a lot more computationally expensive to run due to the large number of features. Very few of these features actually provide useful information to the classifier, so feature reduction can be used to reduce the number of features. One major difficulty of the sentiment classification problem is the high dimensionality of the features used to describe texts, which raises problems in applying many sophisticated learning algorithms to text sentiment classification. The aim of feature reduction methods is to obtain a reduction of the original feature set by removing some features that are considered irrelevant for sentiment classification to yield improved classification accuracy and decrease the running time of learning algorithms [8]. In this research , the existing techniques are analyzed and the experimental study is carried out with different web forums datasets applied on supervised machine learning methods in order to classify the sentiments. The proposed work is mainly focus on feature reduction technique is to reduce the feature size that perform well in sentiment

2. Literature Review

The vast research has been conducted on text sentiment analysis, a large amount of datasets are accessible on the Web. These datasets are usually created for particular domains. For instance, 50,000 movie reviews with comments for positive and negative sentiments are provided [9]. Currently, as many researchers turned their investigations to more timely and convenient social data, some corresponding datasets are put forward for consideration. In [10], a large dataset including manually labeled 20K tweets is constructed for the annually organized competition in SemEval challenge. In these datasets, each message is joined with one label. However, each tweet may have positive and negative sentiments. In STS-Gold [11], both message-level and entity-level sentiments are reserved to 2,206 tweets and 58 entities. In addition to this datasets for general sentiment analysis, there are some other datasets are constructed for specific domains or topics, like Health Care Reform (HCR) dataset [12] including eight subjects and Sanders dataset for four topics. As Compare with textual data, only a few datasets have been generated for sentiment analysis on visual instances.

The author studied the problem of identifying intention posts in discussion forums using new transfer learning method, called CoClass, Unlike a general transfer learning method, Co-Class can deal with two specific difficulties of the problem to produce more accurate classifiers[13]. Pfitzer, Garas, and Schweitzer 14] classified twitter posts as serving distinctly one of two functions. A tweet is either information creation or the pure distribution of information where a user reposts another's original idea or thought (retweet). It was found that emotional divergence has an impact on the probability of a piece of information being retweeted; tweets with higher emotional divergence have a higher probability of being retweeted. Previous work in sentiment analysis on Twitter has revealed that there is a distinct correlation of the swing of the collective mood due to the ability to retweet another user's post. There are several classification models such as Support Vector Machine (SVM), Multinomial Naïve Bayes (MNB), and Maximum Entropy (ME) to analyze sentiments from comments in blogs, reviews and forums in English, Dutch and French. The experiments had achieved reasonable performance for English (accuracy = 83%), but poor performance for Dutch (accuracy = 70%) and French (accuracy = 68%) [15]. The commonly employed supervised machine-learning techniques such as support vector machines (SVM), Naive Bayes, maximum entropy and artificial neural networks to classify the Twitter data using unigram, bigram and unigram + bigram (hybrid) feature extraction model for the case study of US Presidential Elections 2012 and Karnataka State Assembly Elections (India) 2013[16]. The feature reduction method such as Information Gain (IG) used as a feature ranker to select between 42 and 34,855 features (consisting of a combination of unigrams and either sentiment-topic features or semantic features) used to describe 1000 instances from the Stanford Twitter Corpus. They conclude that using more than 500 features yielded no significant improvement in classification performance; however, they only tested a single ranker and learner which includes information gain and Naive Bayes[17]. So, our research work is to make an intensive study of the effectiveness of reduced features using gain ratio for sentiment classification of social networking sites. The effectiveness of the features thus selected is evaluated using SVM and Naive Bayes classifier.

3. Methodology

3.1 Feature Reduction

To improve the accuracy, number of features can be reduced for sentiment classification. Select the favorable attributes from a large features space. There are different Feature reduction methods for text classification: Information Gain, Document Frequency, Gain Ratio, Chi-Squared and Relief-f. In this work Gain ratio is used for feature reduction.

3.1.1 Gain Ratio

This is magnified IG as it normalizes distribution of all features to final classification decision [12]. Problem with using gain ratio, in some cases the gain ratio modification overcompensates and can lead to preferring an attribute just since its intrinsic information is much lower than other attributes. A permanent fix is to choose the attribute that maximizes the GR, provided that the information gain or that attribute is at least as great as the average information gain for all the attributes examined. This technique especially used to improve the accuracy as well as focused on selected subset of sentiment discriminators.

It applies normalization to information gain score by utilizing a split information value.

The split information value corresponds to the important information obtained by dividing the training dataset D into v divisions, resulting to v outcomes on attribute A

$$\text{SplitInfo}_A(D) = -\sum_{j=l}^{v} \frac{|D_j|}{|D|} \text{x} log_2 \frac{|D_j|}{|D|}$$

Gain ratio is the ratio between the information gain and intrinsic value can be defined by the following equation, Gain Ratio (A) = Information Gain(A)/Split Info(A)

3.1.2 Ranker Search

Ranker method is ranked attributes by their individual evaluations Use in conjunction with attribute evaluators (ReliefF, Gain Ratio, Entropy etc.) with the parameter generate ranking (true or false), number to select, and threshold values is set threshold by which attributes can be discarded. Default value results in no attributes are discarded. Use either this option or number to select to reduce the attribute set. The classification, variable ranking is a filter method: it is a preprocessing step, independent of the choice of the predictor]. The ranker method generally performs the rank which attributes should be obtain high or low rank according to the selected attribute in the given datasets. Ranker is providing a rating of the attributes, orderly by their score to the evaluator.

3.2 Sentiment Classification

Sentiment analysis deals with analysis of text containing opinions and emotions. We consider sentiment classification studies that attempts to determine whether a text is subjective or objective or a subjective text contains positive or negative sentiments. Sentiment classification has several important characteristics such as various tasks, features, techniques, and application domains.

a) Tasks

Three important characteristics for polarity classification tasks are the classes, classification levels, and assumptions about sentiment source and target (topic). The classes which involves classifying sentiments as positive or negative. Additionally which include classifying messages as opinionated/subjective or factual/objective. Examples are happiness, sadness, anger, horror etc. Sentiments can be classified at document, sentence, or phrase (part of sentence) level. Document level aims to classify sentiments in movie reviews, news articles, or web forum postings Sentence level aims to classify positive and negative sentiments for each sentence or whether a sentence is subjective or objective. The third level is phrase level aims to capture multiple sentiments which may be present within a single sentence

In addition to sentiment classes and categorization levels, different assumptions have also been made about the sentiment sources and targets. In this work, we focus on document level sentiment polarity categorized into positive and negative sentiment texts.

b) Features

There are four features that have been used in sentiment analysis are syntactic, semantic, link-based, and stylistic features. Semantic and syntactic attributes are the most commonly used for sentiment analysis. They are word n-grams, part of-speech (POS) tags and punctuation. Link-based features are link/citation

analysis which determine sentiments for web artifacts and documents. Due to the limited usage of link-based features, it's not clear that how it will be effective for sentiment classification.

Stylistic features are lexical and structural attributes are built in numerous authorship studies. But they are limited usage in sentiment analysis research. Lexical style markers like words per message, and words per sentence for affective analysis of web blogs. But it is not clear that make sure it is stylistic features are effective sentiment discriminators for movie/product reviews and web discourse.

c) Techniques

Sentiment classification can be classified into three categories. These include machine learning algorithms, link analysis methods, and score based approaches. In this work, machine learning algorithms have been used. The support vector machines and Naïve Bayes being the most commonly used technique. SVM has been used extensively for movie reviews [Pang et al, 2002; Pang and Lee, 2004; Whitelaw et al., 2005] while Naïve Bayes has been applied to reviews and web discourse [Pang et al, 2002; Pang and Lee, 2004; Efron, 2004]. In comparisons, SVM has outperformed other classifiers such as NB [Pang et al., 2002]. While SVM has become a dominant technique for text classification, other algorithms such as Winnow [Nigam and Hurst, 2004] and AdaBoost [Wilson et al., 2005] have also been used in previous sentiment classification studies.

Machine learning approaches initially collecting training dataset, then to train a classifier based upon training data. Once a supervised classification technique is selected, then to make a decision is feature selection. So that, supervised classifier tells us how documents are represented.

d) Sentiment analysis domain

Sentiment classification techniques applied on different data set types such as Reviews, Web Discourse and News Articles. The reviews are movie reviews, product reviews and music reviews Product reviews are complex, because a review of a single person can have both positive and negative sentiment about a specific feature of the product. Another review of movie is very interesting, because movie reviewers can gave their opinion in large summaries and use some literary devices like rhetoric and sarcasm. Web discourses are social website like News groups, Web forums and blogs (face book, twitter). In this domain usually sentiment can be extracted on particular topics or issues like global warming, gun control and politics. Some authors have performed sentiment analysis on news articles [18].

4. Experimental Results

4.1 Dataset description

All the experiments carried out in this section are computed using open source tool Weka. Two various domain in the experiments are used to investigate the performance of the proposed method using machine learning algorithms such as Naïve Bayes and Support Vector Machine.

Use net Dataset

The Usenet1 and Usenet2 Datasets are based on the 20 newsgroups collection. This dataset consists of 1500 instances and 100 attributes. It is available at http://qwone.com/~jason/20Newsgroups/. They simulate a stream of messages from different newsgroups that are sequentially presented to a user, who then labels them as interesting or junk, according to his/her personal interest.

Email Dataset The dataset is a stream of 1500 examples and 913 attributes which are words that appeared at least 10 times in the corpus (Boolean bag-of-words representation).

It is downloaded athttps://www.cs.cmu.edu/~./enron/The emailing list (elist) dataset simulates a stream of email messages from different topics that are sequentially presented to a user who then labels them as interesting or junk according to his/her personal interests.

Table 1: Classification accuracy for NB and SVM classifiers

Approach		Before	After
NB	E-Mail	82 %	94.5 %
	Usenet-1	66.4 %	98.4 %
	Usenet-2	75 %	99 %
SVM	E-Mail	92.6 %	98 %
	Usenet-1	66.1 %	74.2 %
	Usenet-2	74.2 %	99.6 %

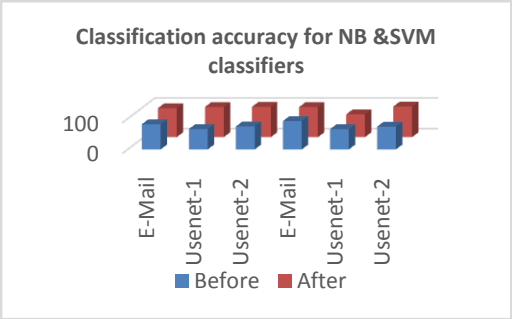

Figure1: Accuracy for E-mail, Usenet-1 and Usenet-2 review.

Table 2: Precision, Recall and F-Measure using NB

NB			
Before			
	Precision	**Recall**	**F-measure**
Email	0.12	0.431	0.187
Usenet-1	0.729	0.446	0.553
Usenet-2	0.717	0.416	0.527
After			
	Precision	**Recall**	**F-measure**
Email	0.988	0.945	0.956
Usenet-1	0.988	0.985	0.986
Usenet-2	0.933	0.99	0.992

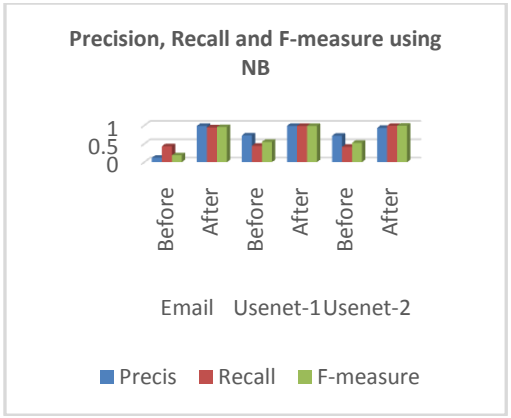

Figure 2: Naïve Bayes Performance Measures

In this research, different datasets are taken such as twitter, blogs, Usenet, email which are stored in unstructured textual format. Then this data need to be converted into meaningful text using machine learning supervised algorithm, since it uses a labeled dataset where each document of training set is labeled with appropriate sentiment. First do preprocessing by removal of stop words and blank spaces. After removing, potential features (words) are extracted. These words have been converted into numeric vector. For that vectorization technique is applied where a matrix is created in which each row denotes an individual review and each column represents a feature. Then this matrix is given as input to the classification algorithm. Before applying classifiers, the data reduction technique such as gain ratio is used as an attribute evaluator. By default the gain ratio uses ranker search method for ranking all the features using parameter values. In order to identifying the optimum number of features for the dataset used, the number of features are varied from min n= 13 to max n= 50. Among the different values, the number of features used and it is observed that the accuracy is max when the value of n is more than 35. Then SVM and NB algorithms are used for 10 fold cross –validation. By this the results are calculated in terms of accuracy, precision, recall and F-measure which is shown in table. The performance of SVM and Naive bayes classifier before feature reduction and after feature reduction is shown in Figure 1, 2 and 3. Thus the classification accuracy obtained through Gain ratio based feature reduction method with classifiers is better than individual classifiers on different review datasets

Table 3: Precision Recall and F-measure using SVM

SVM			
Before			
	Precision	**Recall**	**F-measure**
Email	0.209	0.194	0.201
Usenet-1	0.678	0.521	0.59
Usenet-2	0.717	0.416	0.527
After			
	Precision	**Recall**	**F-measure**
Email	0.96	0.98	0.97
Usenet-1	0.988	0.994	0.991
Usenet-2	0.933	0.997	0.995

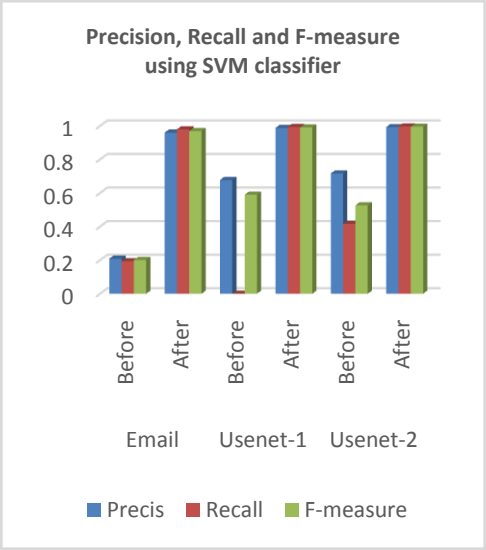

Figure 3: SVM Performance Measure

Then the accuracy is measured using the 2 classifiers. For all domain we are getting accuracy such as 91%, 93% and 100% respectively. The performance of SVM and Naive Bayes classifier before feature reduction and after feature reduction is shown in Figure 1, 2 and 3. Thus the classification accuracy obtained through Gain ratio based feature reduction method with classifiers is better than existing classifiers on news articles datasets.

5. Conclusion and Future work

This research work focused on feature reduction method with Naïve Bayes and SVM as classifiers used to carry out the sentiment analysis over datasets such as E-mail and Use net articles. Extracting features from the document (review) through the attribute evaluator gain ratio and ranker search method. The extracted features and the classification algorithms obtained the accuracy greater.

We increased the accuracy using ranker method to rank all the features on the dataset, we are having list of features that have some ranks given by ranker algorithm in association with attribute evaluator. In future, the experiments can be performed with other domain such as web discourse and other sites with a combination of different feature reductions can be implemented.

References

[1]. Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien and Jun Long, "A Lexicon-based Approach for Hate Speech Detection", International Journal of Multimedia and Ubiquitous Engineering , Vol.10, No.4, pp.215-230 , 2015.

[2]. S. Bhatia,P. Mitra, "Adopting inference networks for online thread retrieval", In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, pp. 1300–1305, July 2010.

[3]. R. M. Elawady, S. Barakat H. M. El-Bakry,N. M. Elrashidy, "Sentiment analysis For Arabic And English Datasets", International Journal of Intelligent Computing and Information Science, IJICIS, Vol.15, No. 1, pp.55-69, January2015.

[4]. Bing Liu. Sentiment analysis and subjectivity. Handbook of natural language processing, 2:568, 2010.

[5]. Ms. GaurangiPatil, "Sentiment Analysis Using Support Vector Machine", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, January 2014.

[6]. Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", pp. 1320-1326.

[7]. Mr. S. M. Vohra, Prof. J. B. Teraiya, "A Comparative Study of Sentiment Analysis Techniques", Journal of Information, Knowledge and Research in Computer Engineering, Volume – 02, Issue – 02.Pp. 313-316, Oct 20 13.

[8]. G.Vinodhini, RM.Chandrasekaran, "Effect of Feature Reduction in Sentiment analysis of online reviews ",International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 6, Pp.2165-2172, June 2013.

[9]. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: ACL (2011.

[10]. Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S.M., Ritter, A., Stoyanov, V.: SemEval-2015 Task 10: sentiment analysis in twitter. In: SemEval 2015 Workshop (2015.

[11]. Saif, H., Fernandez, M., He, Y., Alani, H.: Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the STS-Gold. In: ESSEM Workshop (2013.

[12]. Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with label propagation over lexical links and the follower graph. In: EMNLP Workshop (2011.

[13]. Zhiyuan Chen, Bing Liu Meichun Hsu, Malu Castellanos, Riddhiman Ghosh, "Identifying Intention Posts in Discussion Forums".

[14]. Pfitzner, R., Garas, A. and Schweitzer, F. (2012) Emotional Divergence Influences Information Spreading in Twitter. ICWSM-12.

[15]. E.Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," Information Retrieval, vol. 12, no. 5, pp. 526-558, Sep. 2009.

[16]. Malhar Anjariar,Ram Mohana Reddy Guddeti, "A novel sentiment analysis of social networks using supervised learning", social Network analysis and mining. Springer link, Dec 2014.

[17]. Saif, H.; He, Y.; and Alani, H. 2012. Alleviating data sparsity for twitter sentiment analysis. CEUR Workshop Proceedings (CEUR-WS. org).

[18]. Shailendra Kumar Singh, Sanchita Paul Dhananjay Kumar, "Sentiment Analysis Approaches on Different Data set Domain: Survey", International Journal of Database Theory and Application, Vol.7, No.5, pp.39-50, 2014.

[19]. Sharma and S. Dey, "A comparative study of feature selection and machine learning techniques for sentiment analysis", presented at the ACM Research in Applied Computation Symposium, 2012.