Survey on Sentiment Analysis Techniques

Vijay More[1], Vishal Hulawale[2], Shubham Gotal[3], Rushikesh Killedar[4], Mahesh Shelar[5], Jayesh Choudhari[6]
[1, 2, 3, 4, 5, 6] Department of Information Technology
[1, 2, 3, 4, 5, 6] SPPU AISSMS IOIT, Pune, India

Abstract- *Nowadays the use of e-commerce is increased. People buy products and post their opinions, suggestions about the topic or product also before buying a product they may need to go through those reviews before making any decision. This will help them finalize their choice and take a to purchase the product. If the number of reviews is more, then going through all reviews will become a time-consuming process for users. They will not able to interpret all product reviews and might get confused. The purpose of this study is to explore the methods which work on the sentiment analysis that can be used to interpret reviews and summarization of reviews in the user suitable format. It focuses on specific words or attributes that the customer will be interested in a product. Product reviews will be classified based on emotions extracted from reviews. Our goal is to study various aspects of the classification technique used for sentiment classification.*

Keywords- Sentiment Analysis, Classifiers, SentiWordNet, SVM

I. INTRODUCTION

Nowadays, everything can be done using the internet, like buying, booking etc. Every transaction can be completed on the internet. This improvement in e-commerce makes the incredible or revolutionary change in the financial process. People also have changed their perceptive about e-commerce and started to rely on it. So, to increase in sales merchants enabled the customer to share their reviews about the product on their site to increase the interaction with the customer. If the customer wants to check the product all he must do is to go through reviews but going through all these reviews it can be problematic in terms of an amount of time that customer has to spend to read all reviews, will be large.

The merchandise provides the rating system to reduce the time required for selecting the product. The review used to generate the ratings. Data mining technique can be applied to assess the information and their classification. Text mining consists of techniques to analyze human language. So, sentiment analysis can automate the process of rating based on summarization of reviews. Sentiment analysis is a process of computationally identifying and categorizing reviews expressed in the form of text, to determine user's attitude towards the product. Sentiment analysis aims to detect polarities regarding the product. The main problem is to distinguish it from topic based classification because topics can be identified by keywords and sentiments can be expressed in more delicately. For classification of reviews written by the single user or of one product are considered as a temporal ordered sequence. By assessing the reviews of the same it person is easy to classify them because reviews from the same person are more consistent than that of the others. The neural network has been used for distributed representation learning and can be used in sentiment analysis. In this, it is possible to learn distributed representation of a product, which captures the semantic information contained in a review posted by the user

II. LITERATURE REVIEW

Several social media analysis and text mining used for sentiment analysis have already done. Some of them include:

- SentiWordNet
- Naive Bayes
- Maximum Entropy
- J48

Sentiment classification is done on data using SentiWordNet [1][7]. It is a two-step process: first, WordNet is a lexical database of the English language that groups the word into a set of synonyms called the sunset. It assigns each word or synset three numerical scores Obj(s), Pos(s) and Neg(s), describing the objectivity, positivity and negativity. In the second step, after a fixed number of iterations, a subset of WordNet terms is obtained with either positive or negative label. These term's glosses are then used to train committee of machine learning classifiers. SentiWordNet classifier uses opinion lexical resource SentiWordNet and WordNet along with Word Sense Disambiguation for accurate classification of tweets which is extracted in real time manner.

Naive Bayes [1][2][3][4][5][6][8] classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. Naive Bayes is the simplest and the most commonly used classifier. Naive Bayes estimates prior probabilities from training data and generates the posterior probability of the document based on the prior probability. Naive Bayes computes the posterior probability of class based on a distribution of the words in the document. It uses a bag of word feature extraction which ignores the position of the word. Naive Bayes is optimum for certain problem classes with high dependent features. Naive Bayes (NB) classifier Bayes' rule,

$$p(c|x) = \frac{p(c)p(x|c)}{p(x)}$$

Maximum Entropy (ME) [2][3][4] is a classification technique which has proven effective in several natural language processing applications. In ME, no assumptions are taken regarding the relationships between features. This classifier always tries to maximize the entropy of the system by estimating conditional distribution of the class label. Maximum Entropy converts labeled feature sets into vectors by using encoding, which can be used to calculate weights for each feature that can be combined to determine the most likely label for the feature set. It has proved effectiveness in several natural processing applications.

$$p(fj) \equiv \sum_{x,y} p(x,y)fj(x,y)$$

J48 [8][9] classifier is a Decision tree induction based algorithm. Decision tree based on top-down recursive divide and conquer technique is constructed. It classifies the training data by sorting samples in the dataset depending on features. It provides hierarchical decomposition of the training data using features. Initially, a root node is assigned with all the training samples and the categorical attributes. Based on data divergence where the data with the highest information gain is used to make the split decision on each node, samples are partitioned recursively. The test attributes are selected based on heuristic or statistical measure. When all the samples are classified and no more attributes for further partitioning, the process is stopped. The leaf nodes will give the classification results.

III. COMPARISON

[1] SentiWordNet

SentiWordNet has a disadvantage which is, it does not include information about the pronunciation of words and it contains only limited information about usage. WordNet aims to cover most of the everyday English and does not include much domain-specific terminology.

[2] Naive Bayes

An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification. But the problem happens in Naive Bayes is data scarcity. For any possible value of a feature, you need to estimate a likelihood value by a frequentist approach. This can result in the probabilities going towards 0 or 1, which in turn leads to numerical instabilities and worse results.

[3] Maximum Entropy

ME work well with depended features. But it has a low performance with independent features. Also, the feature selection could become a complex procedure

[4] J48

J48 classification trees to yield the highest classification performance. But J48 may suffer from overfitting and can also get stuck in local minima, so it needs ensembles to reduce the variance.

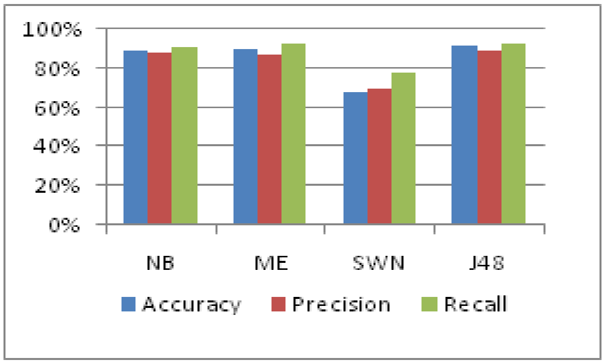

Figure 1 Working Comparison of Techniques

Table 1 Comparison of Various Techniques

	NB	**ME**	**J48**	**SWN**
Characteristic	-Based on assumption. -Perform well with independent features.	-Estimate the probability distribution from data. -Works well with independent features.	-Easy to adopt from non-expert users.	-Based on Bag of synset model.
Supervised	Yes	Yes	Yes	No
Semi-Supervised	Yes	Yes	Yes	Yes
Un-supervised	No	No	No	Yes
Advantages	- Assumptions are not dependent. -Small training data set can be used.	-Use algorithm like GIS and IIS to apply features.	-Easy to understand if few decisions and outcomes includes in the tree.	- Training data is not needed. -Fast
Disadvantages	-Limited applicability -We should lose performance with assumptions.	-Low performance with independent features -Feature selection could become a complex procedure.	-When the actual decisions are made, the payoffs & resulting decisions may not be same.	-Does not include domain words. -Not suitable for words of multiple senses.

IV. CONCLUSION

In this paper, we had a comparative study of various text mining techniques for extracting relevant data. Also, illustrated case about which technique will be useful in which scenario and the effectiveness of techniques in terms of accuracy, precision and recall.

REFERENCES

[1] R. Jose, V. Chooralil, “Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Classifier Ensemble Approach,” IEEE International Conference on Data Mining and Advanced Computing, 2016.

[2] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” in Proc. 2002 Conf. Empirical Methods in Natural Language Processing, 2002.

[3] Neethu S., Rajashree R.,” Sentiment Analysis in Twitter using Machine Learning Techniques,” Fourth International Conference on Computing, Communications and Networking Technologies, 2013.

[4] E. Aydogan, M. Akcayol,” A Comprehensive Survey for Sentiment Analysis Tasks Using Machine Learning Techniques,” International Symposium on INnovations in Intelligent SysTems and Applications, 2016.

[5] Kamal, M. Abulaish, “Statistical Features Identification for Sentiment Analysis using Machine Learning Techniques,” International Symposium on Computational and Business Intelligent, 2013.

[6] S. Saini, S. Kohli, “Machine Learning techniques for Effective Text Analysis of Social Network E-health Data,” 3rd International Conference on Computing for Sustainable Global Development, 2016.

[7] B. Ohana, B. Tiernay, “Sentiment Classification of Reviews Using SentiWordNet,” 9th IT&T Conference, 2009.

[8] K. Krishnaveni, E. Radhamani, “Diagnosis and Evaluation of ADHD using Naïve Bayes and J48 Classifiers,” 3rd International Conference on Computing for Sustainable Global Development, 2016.

[9] Pål Christian S. Njolstad, Lars S. Hoysæter, Wei Wei, Jon Atle Gulla, “Evaluating Feature Sets and Classifiers for Sentiment Analysis of Financial News,” International Joint Conference on Web Intelligence and Intelligent Agent Technologies,2014