Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis

Bhumika M. Jadav
M.E. Scholar,
L. D. College of Engineering
Ahmedabad, India

Vimalkumar B. Vaghela, PhD
Assistant Professor,
L. D. College of Engineering
Ahmedabad, India

ABSTRACT
Social media is a popular network through which user can share their reviews about various topics, news, products etc. People use internet to access or update reviews so it is necessary to express opinion. Sentiment analysis is to classify these reviews based on its opinion as either positive or negative category. First we have preprocessed the dataset to convert unstructured reviews into structured form. Then we have used lexicon based approach to convert structured review into numerical score value. In lexicon based approach we have preprocessed dataset using feature selection and semantic analysis. Stop word removal, stemming, POS tagging and calculating sentiment score with help of SentiWordNet dictionary have been done in preprocessing part. Then we have applied classification algorithm to classify opinion as either positive or negative. Support vector machine algorithm is used to classify reviews where RBF kernel SVM is modified by its hyper parameters which are soft margin constant C , Gamma γ. So optimized SVM gives good result than SVM and naïve bayes. At last we have compared performance of all classifier with respect to accuracy.

Keywords
Sentiment analysis, Text mining, SentiWordNet, SVM, Naïve Bayes, RBF kernel SVM

1. INTRODUCTION
Sentiment analysis is an ongoing research area which is growing due to use of various applications. Sentiment analysis is also called as opinion mining. People give their reviews in form of unstructured format via blogs, forums etc. These unstructured reviews are preprocessed to extract opinion from it and this opinion is positive, negative or neutral.

Sentiment analysis is done by using classification approaches which are lexicon and machine learning based approaches. Lexicon based approach is of dictionary based approach and corpus based approach. Machine learning techniques are most widely used to classify and to predict sentiment as either positive or negative sentiment. Machine learning algorithms are mainly classified as either supervised or unsupervised approach. Supervised approach takes labeled dataset where each training set has already assigned its sentiment. Unsupervised approach takes unlabeled dataset where review is not defined with its label [2].

Sentiment analysis refers to the task of identifying opinion from reviews. Sentiment analysis is categorized into three different levels which are document level, sentence level and entity-aspect level. Overall opinion is to be identified in document level analysis. Only opinion of sentence is to be identified in sentence level analysis. Focus is directly on opinion itself in entity-aspect level analysis [1].

In this study, First unstructured movie review is converted into structured form. Then score is calculated to tagged structured word. This score is given as input to support vector machine with kernel hyper parameter to classify reviews as either positive or negative.

2. RELATED WORK
Different features techniques like unigrams, bigrams, unigrams + bigrams, unigrams + POS tagging, Position and unigrams + Position are used and then machine learning techniques like Naïve bayes, Maximum Entropy and support vector machine classification algorithms are applied on preprocessed dataset. Classification algorithms perform better than human based classifier [3]. Tokenization, stop word removal, TF-IDF and POS tagging is used as part of preprocessing then new approach sentiment fuzzy classification algorithm is used to improve result of movie review dataset and POS makes accurate classification in sentiment analysis [4]. Feature extractors which are unigram, bigram, unigram with bigram combination and unigram with POS tagging is used. Machine learning algorithms like Naïve bayes, Maximum Entropy and SVM etc are used to classify tweets and proved that their accuracy is above 80% though they are trained with emotions dataset [5]. Movie and Twitter review are classified using natural language techniques which are synonym using WordNet and word sense disambiguation. The results show that accuracy is increased by 5% using machine learning ensemble classifiers consist from Random Forest, Decision Tree, Extremely Randomized Trees and Ada Boost. WordNet synset increases accuracy [6]. Machine learning approach is supervised learning approach because classifier is trained on dataset whereas semantic approach is unsupervised because it measures how far a word is related to positive or negative. Both approaches have its pros and cons. In this paper supervised machine learning approach is more accurate then semantic orientation but it is time consuming process to train model. Semantic orientation approach is less accurate but it is efficient [7]. Tweet sentiment analysis model consists of feature selection module, sentiment identification module and sentiment aggregation & scoring module. In feature extraction, they are extracted opinion words using lexicon list. They have used Wilson opinion lexicon list for semantic orientation. It is used to predict prior public opinion [8]. The aim of paper is to find best effective features which provide better result and also provide better feature selection method. They have also express that how unigram feature set can be reduced to get better result. As a preprocessing sop word removal, stemming, pos tagging is performed. Mutual information (MI), Information gain (IG), Chi-square ($\lambda 2$) and TF-IDF feature selection method is used to extract feature [9].

Twitter dataset is used and analyzed using unigram feature extraction technique. Further, the different machine learning techniques trains the dataset with feature vectors and then the semantic analysis offers a large set of synonyms and similarity which provides the polarity of the content. WordNet increases accuracy [10]. As a part of preprocessing they have removed vague information and unnecessary blank spaces. After preprocessing, this preprocessed data is converted into numerical vector using TF-IDF and CountVectorizer. Support vector machine and naïve Bayesian classifiers are used to classify numerical vector [11].

3. PROPOSED METHODOLOGY

Step 1: Review Dataset

Here we have used Polarity movie review dataset. Separate text file is maintained for each review. Other Twitter and Gold dataset is also taken to show effect of proposed method on different dataset. Twitter dataset is taken from twitter API and gold dataset is taken from amazon.com

Step 2: Preprocessing

Reviews contain information which are not clearly expressive or say meaning and need to be removed.

- Remove unwanted punctuations: All punctuations which are not necessary, it has been removed.
- Stop Word Removal: Some words used more and more time such words are called stop word. This pronouns, prepositions, conjunctions have no specific meaning.
- “i","a","an",”is”,"are","as","at","from","in",”this”,"on","or","to","was","what","will","with” etc are example of stop word, so these types of words has been vanished.
- Stemming : It converts word into its grammatical root form. Stemming technique converts word like “teach”,”teacher”,”teaching”,”teached”,”teaches” to root word teach.
- Among many available algorithms, we have used M.F Porter stemming algorithm.
- It minimizes the feature set and makes efficient classification performance by using java language.
- Part of Speech Tagging :
- The Part-Of-Speech of a word is a linguistic category that is defined by its syntactic or morphological behavior. Noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection are POS common categories.
- POS tagging is the task of marking each word in a sentence with its appropriate POS. we have used the Stanford tagger to tag the words. We have assign tag as verb, adjective, noun and adverb category.
- SentiWordNet dictionary calculates score to tagged words and score is given to Proposed SVM to classify Reviews. Every word has positive and negative score already defined in the SentiWordNet dictionary so with help of that score, weighted score is assigned to tagged word to calculate its sentiment score.

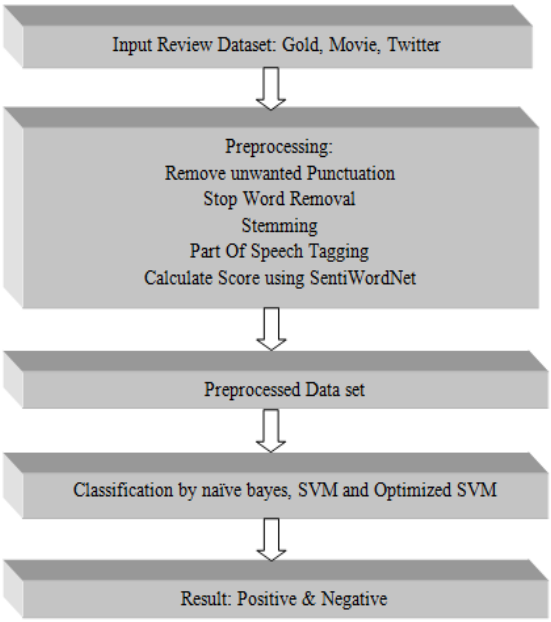

Figure 1. Proposed System Architecture

Step 3 : Classify by Optimized SVM

The preprocessed dataset is given as input to the classification algorithm. The Naïve Bayes and Support Vector Machine classification algorithm is also used to classify dataset because of comparison with optimized SVM. Here we have changed value of kernel hyper parameters which are gamma and margin constant.

Step 4 : Result

Confusion Matrix is generated which shows classified positive and negative reviews. Accuracy is calculated based on confusion matrix. Then it has been compared with values of same with naïve bayes and support vector machine.

3.1 SVM & RBF kernel

Support vector machine is non probabilistic algorithm which is used to separate data linearly and nonlinearly. Here dataset D = {Xi, yi} where Xi is set of tuples and yi is associated class label of tuples. Class labels are -1 and +1 for no and yes category respectively. The goal of SVM is to separate negative and positive training example by finding n-1 hyper plane.

Quadratic Programming (QP) problem is needed to be solved in linear data. This problem is transformed using the Lagrange Multipliers theory and Optimal Lagrange coefficients sets are obtained. A separating hyper plane is written as:

$$W*X + b = 0 \quad (1)$$

where $W = \{ w_1, w_2, w_3, \dots , w_n\}$, w_n is weight vector of n attributes and b is bias. Distance from separating hyper plane to any point on H1 is 1/|W| and Distance from separating hyper plane to any point on H2 is 1/|W| so maximum margin is 2/|W|. The MMH is rewritten as the decision boundary according to Lagrangian formulation.

$$D(X^T) = \sum_{i=1}^{l} y_i a_i X_i X^T + b_0 \quad (2)$$

Where X^T is test tuple, α_i and b_0 are numeric parameters, y_i is class label of support vector X_i. so If sign is positive of MMH equation then X^T comes in positive category. If sign is negative of MMH equation then X^T comes in negative category. SVM classifier formula is defined as following.

$$f(x)= \sum_{i=1}^{n} a_i k(x, x_i) + b \quad (3)$$

3.2 Naïve Bayes

It is probabilistic classifier which requires small set of training data to determine parameter prediction. Only variance of feature is calculated because of independence of features instead of calculating full covariance matrix. Bayes theorem is defined as.

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (4)$$

d is review and c is class. For a given textual review 'd' and for a class 'c' (positive, negative), the conditional probability for each class given a review is P(c|d) .

3.3 SVM RBF kernel

SVM classifier formula is defined as following.

$$f(x)= \sum_{i=1}^{n} a_i k(x, x_i) + b \quad (5)$$

$$\text{where } k(x,x_i) = \exp\left(-\frac{||x-x'||^2}{2\sigma}\right) \quad (6)$$

In nonlinear data, Dimension is transformed to higher dimension so cost is increased due to multiplication of test tuple with every support vectors. So due to higher dimension space, training tuples are in form of $\phi(X_i)^* \phi(X_j)$, which are replaced by kernel $K(X_i, X_j) = \phi(X_i) * \phi(X_j)$. $K(X_i, X_j)$ is kernel function and it is Mathematically equivalent to product of $\phi(X_i)$ and $\phi(X_j)$ So there is no need of nonlinear mapping. Further process is same as linear data case.

SVM can able to handle linear separation on the high dimension non linear input data, and this is gained by using an appropriate kernel function. There are many kernel functions of SVM which are Polynomial kernel of degree, Gaussian radial basic kernel function and sigmoid kernel function. Gaussian Radial Basic Kernel (RBF) function has been chosen which has kernel hyper parameter γ (gamma) and soft margin constant C.

Hyper parameters are modified with different combination of regularization Constant (Soft Margin) C , kernel hyper parameter γ (gamma). The aim is to identify good (C, γ) by comparing different combination of (C, γ). SVM has been implemented on R tool.

4. EXPERIMENTAL RESULT ANALYSIS

Table 1. Confusion Matrix

	Positive	**Negative**
Positive	True Positive(TP)	False Positive(FP)
Negative	False Negative(FN)	True Negative(FN)

The performance matrix is used to calculate classification accuracy.

Accuracy: It is one of the most common performance evaluation parameter and it is calculated as the ratio of number of correctly predicted reviews to the number of total number of reviews present in the corpus. The formula for calculating accuracy is given as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \quad (7)$$

Proposed approach gives high accuracy than existing system.

4.1 SVM

The confusion matrix is obtained after implementing SVM algorithm. Table 2, 3 and 4 shows confusion matrix for respectively gold, small movie and twitter dataset.

Table 2. Confusion Matrix for Gold Dataset

	Correct Labels	
	Positive	**Negative**
Positive	0	600
Negative	0	1601

Table 3. Confusion Matrix for Movie Dataset

	Correct Labels	
	Positive	**Negative**
Positive	5331	0
Negative	1802	0

Table 4. Confusion Matrix for Twitter Dataset

	Correct Labels	
	Positive	**Negative**
Positive	3000	0
Negative	900	0

4.2 Naïve Bayes

The confusion matrix is obtained after implementing naïve bayes algorithm. Table 5, 6 and 7 shows confusion matrix for respectively gold, small movie and twitter dataset.

Table 5. Confusion Matrix for Gold Dataset

	Correct Labels	
	Positive	**Negative**
Positive	143	457
Negative	223	1378

Table 6. Confusion Matrix for Movie Dataset

	Correct Labels	
	Positive	**Negative**
Positive	5300	31
Negative	1784	18

Table 7. Confusion Matrix for Twitter Dataset

	Correct Labels	
	Positive	**Negative**
Positive	2958	42
Negative	868	32

4.3 SVM with hyper parameter

The confusion matrix is obtained after implementing SVM with RBF kernel hyper parameter (C,γ). Table 5, 6 and 7 shows confusion matrix for respectively gold, small movie and twitter dataset. This dataset find (10,1) value for (C,γ).

Table 8. Confusion Matrix for Gold Dataset

	Correct Labels	
	Positive	**Negative**
Positive	83	517
Negative	65	1536

Table 9. Confusion Matrix for Movie Dataset

	Correct Labels	
	Positive	**Negative**
Positive	5331	0
Negative	1802	0

Table 10. Confusion Matrix for Twitter Dataset

	Correct Labels	
	Positive	**Negative**
Positive	2977	23
Negative	828	72

4.4 Graphical Representation

This section represents comparison of proposed and existing method. As shown in Table 11 and Figure 2, proposed method improves accuracy compare to existing method.

Table 11. Accuracy with dataset and classifiers

Dataset	Naïve Bayes	SVM	RBF kernel SVM
Gold	69.10	72.74	73.56
Movie	74.55	74.73	74.74
Twitter	76.67	76.92	78.18

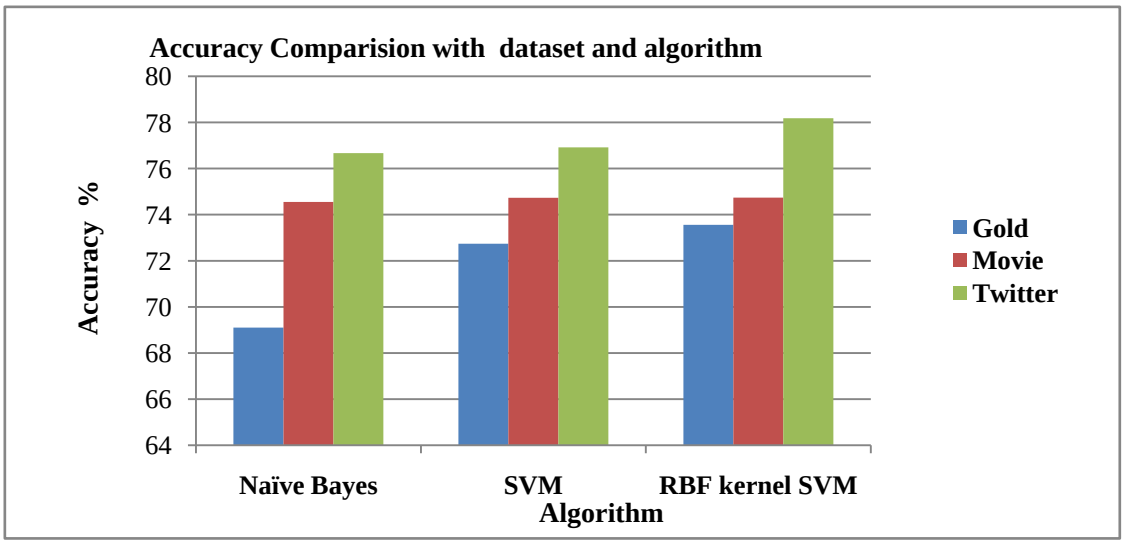

Figure 2. Accuracy comparison with classifier and dataset

5 CONCLUSION

Sentiment analysis has been done for movie Review, Twitter and Gold dataset using optimized SVM. Here Comparison is made between Optimized Support Vector Machine towards Support Vector Machine and naïve bayes classifier. Modifying hyper parameter value of RBF kernel SVM gives better result compare to Support Vector Machine and Naïve Bayes algorithm. Hyper parameters are soft margin constant C and Gamma γ. Proposed approach has found optimal value for hyper parameter which classifies dataset with more accuracy than existing system.

There are many SVM kernel functions available with many hyper parameters. These values can be modified to improve accuracy.

6 REFERENCES

[1] Liu, B., 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), pp.1-167.

[2] Y. Singh, P. K. Bhatia, and O. Sangwan, "A review of studies on machine learning techniques," International Journal of Computer Science and Security, vol. 1, no. 1, pp. 70–84, 2007.

[3] Pang, B., Lee, L. and Vaithyanathan, S., 2002, July. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

[4] Mouthami, K., Devi, K.N. and Bhaskaran, V.M., 2013, February. Sentiment analysis and classification based on textual reviews. In Information Communication and Embedded Systems (ICICES), 2013 International Conference on (pp. 271-276). IEEE.

[5] Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1, p.12.

[6] Kanakaraj, M. and Guddeti, R.M.R., 2015, February. Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques. In Semantic Computing (ICSC), 2015 IEEE International Conference on (pp. 169-170). IEEE.

[7] Chaovalit, P. and Zhou, L., 2005, January. Movie review mining: A comparison between supervised and unsupervised classification approaches. In System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on (pp. 112c-112c). IEEE.

[8] Zhou, X., Tao, X., Yong, J. and Yang, Z., 2013, June. Sentiment analysis on tweets for social events. In Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on (pp. 557-562). IEEE.

[9] Shahana, P.H. and Omman, B., 2015. Evaluation of Features on Sentimental Analysis. Procedia Computer Science, 46, pp.1585-1592.

[10] Gautam, G. and Yadav, D., 2014, August. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In Contemporary Computing (IC3), 2014 Seventh International Conference on (pp. 437-442).IEEE.

[11] Tripathy, A., Agrawal, A. and Rath, S.K., 2015. Classification of Sentimental Reviews Using Machine Learning Techniques. Procedia Computer Science,57, pp.821-829.

[12] Miller, G.A., 1995. WordNet: a lexical database for English. Communications of the ACM, 38(11), pp.39 41.

7 AUTHOR PROFILE

Bhumika M. Jadav, is currently pursuing her M.E. in Computer Science & Engineering Department at L.D. College of Engineering, Ahmedabad. She has received the degree of B.E. from G.E.C. Rajkot in 2010.

Prof. (Dr.) Vimalkumar B Vaghela, holds Ph.D. in Computer Science & Engineering. This author is Young Scientist awarded from Who's Science & Engineering 2010-2011 & also his biography is included in American Biographical Institute in 2011. His publication is also available in ieeeexplorer and also in spocus online database. He is currently working as Assistant Professor in Computer Engineering Department at L.D. College of Engineering, Ahmedabad, Gujarat, India. He received the

B.E. degree in Computer Engineering from C. U. Shah College of Engineering and Technology, in 2002 & M.E. degree in Computer Engineering from Dharmsinh Desai University, in 2009. He has published book titled "Ensemble Classifier in Data Mining" in LABERT Academic publisher, Germany, 2012 & "Operating System" in Dreamtech-India 2015. His research areas are Relational Data Mining, Ensemble Classifier, and Pattern Mining. Author has published / presented more than 23international papers and 5 national papers.