Perform Sentiment Analysis with Machine Learning Techniques

Ruchika Sharma

***Abstract*—Sentiment Analysis has become an indispensible part of product reviews in present scenario. We consider the problem of analyzing the overall sentiment of a document using Machine learning techniques. Sentiment Analysis is a very well studied field, but the scale remains limited to not more than a few hundred researchers. We improve the results using SVM kernel approach and compare the same with previously used techniques. The present research is a comparison and extension of the work proposed by Mullen and Collier (2003). Our system consists of a feature Extraction phase and a learning phase; on the basis of which the overall sentiment of the document is analyzed. Our present work uses the movie review data set used by Pang (2002). The present work shows that SVM Kernel approach outperforms the Naïve bayes approach.**

***Index Terms*—Sentiment Analysis, Classifier, SVM, td-idf, Naïve Bayes, PCA.**

I. INTRODUCTION

Highlight a section that you want to designate with a certain style, and then select the appropriate name on the style menu. The style will adjust your fonts and line spacing. **Do not change the font sizes or line spacing to squeeze more text into a limited number of pages.** Use italics for emphasis; do not underline Sentiment Analysis aims at analyzing the overall sentiment of the text, whether it is positive or neutral or negative. It is a combination of Natural Language Processing and Information Retrieval methods. The analysis can be performed word/sentence/paragraph/document wise. With the growing data on web, a large amount of data is available online, which can be manipulated for finding reviews about a particular product. It has been used to generate the list of people who were regarded as positive and negative characters in newspapers and blogs [1]. Some others areas where its use has been marked are: Business and Government Intelligence for knowing consumer attitudes and trends, knowing public opinions for political leaders or their notions about rules and regulations in place, for detecting heating language in mails etc [10]. It is often confused with text categorization task, which is quite not the case. It faces many challenges like: implicit meaning of the sentence, entity identification, negation, subjectivity detection, pragmatics etc. Sentiment140.com collects the tweets from twitter for the keyword entered and represents the sentiment in the form of a pie chart for it. Most related work has been partially knowledge based. Some of this work is based on determining the semantic orientation of words. Machine Learning Techniques have their reach beyond expectations. Srinivasaiah and Skiena (2007) performed sentiment analysis for news and blogs. It incorporated the use of a system which assigns scores indicating positive or negative opinion in the document [1].

The system consists of two phases: sentiment identification phase and sentiment aggregation and scoring phase. They generated a list of top positive and negative entities in news and blogs.Feature selection has been described by Ikonomakis, Kotsiantis and Tampakas (2005) using machine learning techniques. The process followed by them is as: tokenization results in stemming. The vector representation of text is done followed by the feature selection and transformation so as to delete redundant words. Then the features are put into a learning algorithm. There is work which contrasts the various machine learning methods. Pang, Lee and Vaithyanathan (2002) [10] show that SVMs outperform Navies Bayes and Maximum Entropy in terms of performance. The motivation of present research is to incorporate methods to perform sentiment analysis using machine learning techniques..

II. METHOD

The system consists of two phases: Processing and Validation Phase. The text is passed into the Processing Phase which provides us with the features from which the sentiment would be analyzed. These features are then passed into learning phase, which uses an approach to learn from previous examples (Machine learning). The document, which is a combination of words, is passed into the system and the text is converted into tokens, termed as tokenization. The document is represented by a binary vector [3]. After this, stopwords are deleted. These are the words which are of hardly any significance to us. Another preprocessing step is Stemming. It refers to replacement of the words which originate from the same stem with a root word. For e.g. the words like: play, playing, played, etc can be replaced with a single word: play. This step reduces ambiguity. The above stated steps become necessary to be implemented as the number of features can reach orders of tens of thousands without these steps. The ultimate aim remains to reduce the size of the feature set [5]. After feature selection, Feature transformation is done. This is done using PCA (Principal Component Analysis) [7]. The aim of using PCA is to learn a discriminative transformation matrix in order to reduce the complexity of the feature set so obtained. After the feature set is obtained, a machine learning algorithm can be applied. The algorithm varies in the approach adopted for its implementation. It allows the use of Naives bayes, minimum entropy, support vector machines, neural networks, nearest neighbors, etc. we use support vector machines in collaboration with kernel approach for our purpose of research.

Manuscript received December 15, 2013.

Ruchika Sharma, Computer Science Department, Chitkara University/ Baddi, Himachal Pradesh, India.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

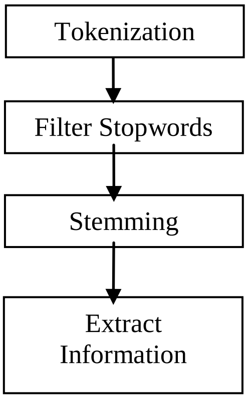

Figure 1: Processing of documents from text.

2.1 Document Processing Phase

We consider the problem of analyzing the overall sentiment of a document using Machine Learning techniques. Cornell Movie Review dataset has been used to show that machine learning techniques outperforms the traditional cognitive sentiment classification methods. This is the dataset which was presented in Pang et. al. (2002) and can obtained from www.cs.cornell.edu/people/pabo/movie-review-data/. It comprises of approximately 700 positive and 700 negative reviews. 300 reviews in each category are passed into the system as training data set and the remaining 400 reviews are checked for the results.

Tokenization

In this step the text of the document is spilt into a sequence of tokens. A 'token' refers to each word which has an independent existence in the sentence. For example, in the sentence- "The movie was really good", the tokens would be- 'the', 'movie', 'was', 'really' and 'good'. Tokenization is the most primitive step while processing any document in natural language processing. It is also the basic step performed by compilers while converting a program from high level language to machine level language.

Filter Stopwords

Stopwords are those words which are of hardly any importance to us during this research. It comprises of punctuations, articles, conjunctions, connectors, etc. These words should be detected in order to reduce a precise and smaller data set. These words are incurred in almost all the documents and are insignificant while analyzing the overall sentiment of the document.

Stemming

Stemming refers to the processing of words in order to reduce the feature set size. The features with the same stem are replaced e.g. the words like trainer, trains, trained can be replaced by a single word- "train". Stemming is useful until it does not turn out to be aggressive in nature. Aggressive stemmers such as Porter Stemmer [13] sometimes tend to lose some important words. Thus aggressive Stemming remains a topic of controversy. We prefer moderate level stemming for our work.

Extract Information

The information collected so far, is in the form of words. These words are in the form of tokens, free from ambiguity. This collection of words has also been freed from words and symbols (like punctuations, conjunctions, and etcetera) of zero importance to us. Now, we need to collect important features which could be used to recognize the sentiment of the text.

2.2 Validation

This phase uses a machine learning approach to learn from previous examples to perform sentiment analysis. Following is a list of some machine learning approaches being used in this paper.

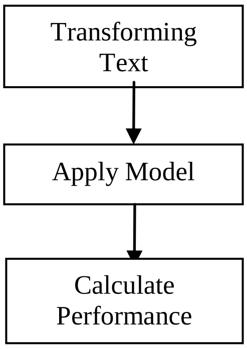

Figure 2: Processing of documents from text.

Transforming Text

Since a machine cannot understand high level languages (English, in this case), so it needs to be converted into the form which is understood by the machine. For our research purpose, we use Td-idf (Term Frequency- inverse document frequency) which evaluates the importance of a word in the document. It is a way to convert textual representation of information to VSM (Vector Space Model). VSM is an algebraic model which represents text as a vector. The sole purpose of this preprocessing step is to reduce the size of feature set. PCA (Principal Component Analysis) is the most commonly used technique for feature selection. It is a way of highlighting similarities and differences in pattern recognition. It works where the luxury of graphical representation of data fails. With the use of PCA, we can compress the data, without the loss of any information. This can be done by reducing the number of dimensions.

Apply Model

In this step, various algorithms are applied on the data transformed in the previous steps. In our work, we perform experiments with four models namely- Naïve Bayes, Naïve bayes (kernel), SVM (linear) and SVM (Kernel). All these methods have been explained in the following section.

Calculate Performance

In this step, actual sentiment is compared with the results so obtained by the respective models. This comparison is then translated into a percentage vector in the form of accuracy obtained by the model. SVM Kernel approach obtained an accuracy of 76.43%, followed by SVM (Linear) which attained an accuracy of 74.43% only.

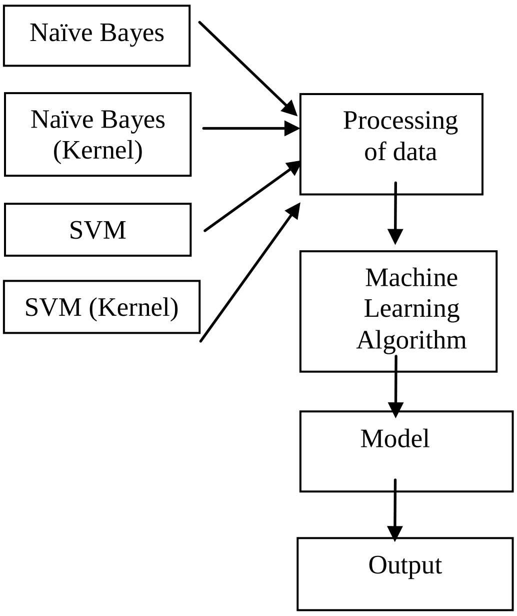

Figure 3: Flowchart showing the steps for Learning Phase.

Naïve Bayes

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be 'independent feature model'. In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class (i.e. attribute) is unrelated to the presence (or absence) of any other feature.

Naïve Bayes (Kernel)

A kernel is a weighting function used in non-parametric estimation techniques. Kernels are used in kernel density estimation to estimate random variables' density functions, or in kernel regression to estimate the conditional expectation of a random variable. We have used kernel approach with Naïve bayes classifier (used in the previous step) to the transform data. It attained an accuracy of 54.40%.

SVM

This concept was given by Vapnik (Vapnik, 1979), and since then it has become the most widely used approach in the field of machine learning. The use of SVM is highly dependent on Model selection. It has the capability to produce better results than many other models. The sole purpose of SVM is pattern recognition and results obtained using this model has been spotted as remarkable. We use Libsvm package for training and testing.

SVM (Kernel)

A string kernel is a mathematical tool, where sequence data are to be clustered or classified. We have used kernels with support vector machines to transform data from its original space to one where it can be more easily separated and grouped [10].Then the inner product of those vectors is taken. There is no need to explicitly map the data into high dimensional space for optimizing the results. It has proved to produce best results with text related operations.

We use Libsvm package for training and testing. Mercer's theorem defines:

$$K(x,y)=\varphi(x).\varphi(y)$$

Where K represents Kernel and φ is the mapping function which maps the arguments into an inner space.

III. RESULTS

To compare the performances of the above stated models, we will perform cross validation on the data set. Since the dataset is large, so we test the vales with 10, 15 and 20 fold cross validation on the features extracted by the system. In 10 fold cross validation, we divide the data set into 10 subsets of equal size and follow the following algorithm:

Train classifier on folds: 2 3 4 5 6 7 8 9 10; test against fold: 1
Train classifier on folds: 1 3 4 5 6 7 8 9 10; test against fold: 2
Train classifier on folds: 1 2 4 5 6 7 8 9 10; test against fold: 3
Train classifier on folds: 1 2 3 5 6 7 8 9 10; test against fold: 4
Train classifier on folds: 1 2 3 4 6 7 8 9 10; test against fold: 5
Train classifier on folds: 1 2 3 4 5 7 8 9 10; test against fold: 6
Train classifier on folds: 1 2 3 4 5 6 8 9 10; test against fold: 7
Train classifier on folds: 1 2 3 4 5 6 7 9 10; test against fold: 8
Train classifier on folds: 1 2 3 4 5 6 7 8 10; test against fold: 9
Train classifier on folds: 1 2 3 4 5 6 7 8 9; test against fold: 10

S.No.	MODEL	10 FOLDS	15 FOLDS	20 FOLDS
1	Naïve Bayes	62.86%	63.01%	62.86%
2	Naïve Bayes (Kernel)	54.36%	54.65%	54.40%
3	SVM (Linear)	75.71%	55.03%	75.43%
4	SVM (Kernel)	75.36%	75.79%	76.43%

Table 1: Accuracy result comparison for 10, 15and 20 fold cross validation on Movie review dataset.

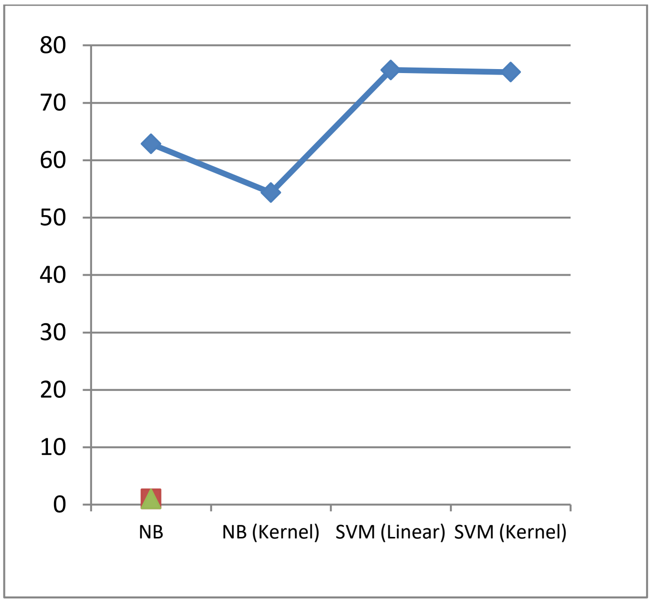

Figure 4: Graph showing accuracy obtained various models with 10 fold cross validation.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

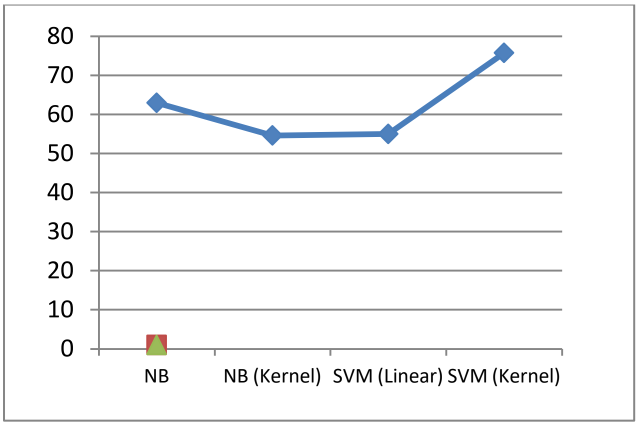

Figure 6: Graph showing accuracy obtained various models with 15 fold cross validation.

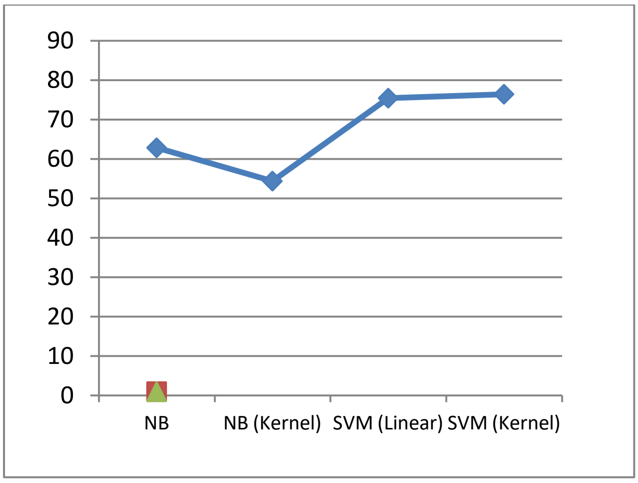

Figure 6: Graph showing accuracy obtained various models with 20 fold cross validation.

IV. CONCLUSION

Some of the machine learning approaches namely Naives Bayes, Maximum Entropy, SVM and Kernels were explored. SVM Kernel produces an accuracy of 75.79% and 76.43% for cross validation in 15 fold and 20 fold respectively. However the combination of multiple kernels with other machine learning approaches remains untouched and can be worked upon in future. Polysemy (words with more than one meaning) and Synonymy (different words with same meaning) in case of feature extraction are the areas which need special attention.

ACKNOWLEDGMENT

I am grateful to my institution for providing me the opportunity to carry out the present work. I would like to express a deep sense of gratitude and thanks to all the faculty members of Department Of Computer Science and Engineering, Chitkara University for their counsel and guidance.

REFERENCES

1. Godbole, N., Srinivasaiah, M., Skiena, S.: Large-scale sentiment analysis foe news and blogs. In: Proc. Int. Conf. Weblogs and Social Media (ICWSM 07). (2007)
2. Benjamin Snyder; Regina Barzilay (2007). "Multiple Aspect Ranking using the Good Grief Algorithm". Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL). pp. 300–307.
3. M. Ikonomakis, S. Kotsiantis, V. Tampakas: Text classification using machine learning techniques. In: WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, August 2005, pp. 966-974
4. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL. (2004) 271-278
5. Han X., Zu G., Ohyama W., Wakabayashi T., kimura F., Accuracy improvement of automatic Text Classification Based on feature Transformation and multi-classifier combination, LNCS, Volume 3309, Jan 2004, pp. 463-468
6. Nasukawa, T., Yi, J.: Sentiment Analysis: Capturing favorability using natural language processing. In: the Second International Conferences on Knowledge Capture. (2003) 70-77
7. Zu G., Ohyama W., Wakabayashi T., Kimura F., "Accuracy improvement of automatic text classification based on feature transformation": Proc: the 2003 ACM Symposium on Document Engineering, November 20-22, 2003, pp. 118-120
8. Tony Mullen and Nigel Collier, Sentiment analysis using support vector machines with diverse information sources. (2003)
9. J. Yi, T. Nasukawa, R.B., Niblack, W.: Sentiment analyzer: Extracting sentiments about a given topic usin natural language processing techniques. In: 3rd IEEE Conf. on Data Mining (ICDM'03). (2003) 423-434
10. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in natural Language Processing (EMNLP). (2002) 79-86
11. Bao Y. and Ishii N., "Combining Mutilple kNN Classifiers for Text Categorization by reducts", LNCS 2534, 2002, pp.340-347
12. Huma lodhi, Criag Saunders, John Shawe-Taylor, Nello Cristianini, Chris Watkins, "Text Classification using string kernels", Jornal Of Machine Learning Research, 2002, pp. 419-444
13. Sebastiani F., "Machine Learning in Automated Text Categorization", ACM Computing Surveys, vol. 34 (1), 2002, pp 1-47