Improve Sentiment Analysis Accuracy using Multiple Kernel Approach

Ruchika Sharma
M.E.
Chitkara University,
Himachal Pradesh

Amit Arora
Assistant Professor
Chitkara University
Himachal Pradesh

ABSTRACT
Sentiment Analysis has become an indispensible part of product reviews in present scenario. Sentiment Analysis is a very well studied field, but the scale remains limited to not more than a few hundred researchers. The problem of analyzing the overall sentiment of a document using Machine learning techniques has been considered. Results have been improved using multiple kernel approach and compared with previously used techniques..The present research is a comparison and extension of the work proposed by Mullen and Collier (2003). The system consists of a feature Extraction phase and a learning phase; on the basis of which the overall sentiment of the document is analyzed. The present work uses the movie review data set used by Pang (2002). The approach significantly outperforms the previous methods attaining 90% and 92% accuracy using 5 fold cross validation 10 fold cross validation respectively.

General Terms
Machine Learning, Information Retrieval.

Keywords
Sentiment Analysis, Feature Extraction, SVM, PCA, Kernel, Multiple kernel.

1. INTRODUCTION
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 [3]. 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 [12]. Vast work has been done on analyzing the citations of various papers i.e. the number of times a paper is cited and the sentiments associated with that paper, whether it has positive or negative reviews [1]. It is a tool widely used by CISCO to know their product reviews [2]. 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.

Mullen and Collier (2003) have described various methods together to assign values to the words selected (using feature extraction and selection) [10]. They show that hybrid SVM (PMI/Osgood and Lemmas) produce the best results for sentiment analysis. Simple lemmas obtain an average score of 84% whereas simple unigrams model produces 79.75% accurate results. Hybrid SVM outperforms the former by producing 86.5% accurate results. These have been obtained over the four n fold cross validation experiments. Godbole, Srinivasaiah and Skiena (2007) performed sentiment analysis for news and blogs [3]. It incorporated the use of a system which assigns scores indicating positive or negative opinion in the document. 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) [12] 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.

2. METHOD
The system consists of two phases: Feature Extraction and Learning Phase. The text is passed into the Feature Extraction Phase which provides 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. The document which is combination of words is passed into the system, following which it is represented in an array of words. The document is represented by a binary vector [5]. After this, stop words 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. 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 [7]. After feature selection, Feature transformation is done. This is done using PCA (Principal Component

Analysis) [9]. 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 implemented. It allows the use of Naives bayes, minimum entropy, support vector machines, neural networks, nearest neighbors, etc. we use support vector machines for our purpose of research.

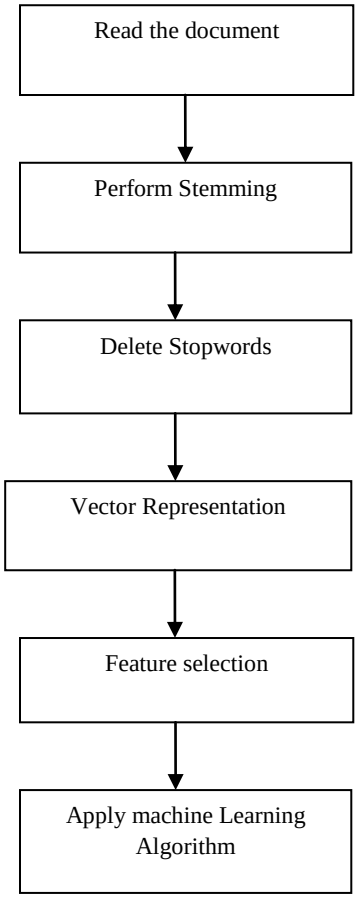


Figure 1: Flowchart showing the steps for feature selection.

2.1 Feature Selection Phase

2.1.1 Read the document
The problem of analyzing the overall sentiment of a document using Machine Learning techniques has been considered. 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.

2.1.2 Perform 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 [15] sometimes tend to lose some important words. Thus aggressive Stemming remains a topic of controversy. Moderate level stemming has been preferred for the present work.

2.1.3 Delete Stopwords
Stopwords are those words which are of hardly any importance for analyzing the text for present work. 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.

2.1.4 Vector Representation
Document is represented in the form of an array. It becomes easy to manipulate the data in an array format. The document can be represented as a vector. For present research purpose, binary vector has been used. If a feature is contained in the document, it is assigned the value 1 and 0 otherwise. The values as stated above are assigned by the placing the document in a $R^{|V|}$ space where |V| refers to the size of the vocabulary. Vocabulary refers to the number of words of training set. It is generally called the feature set of the document. Td-idf (Term Frequency- inverse document frequency) 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.

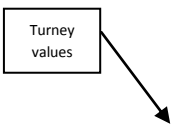
2.1.5 Feature selection
The sole purpose of this preprocessing step is to reduce the size of feature set. PCA (Principal Component Analysis) is the most commonly used technique for feature selection. It is a way of highlighting similarities and differences in pattern recognition. It works where the luxury of graphical representation of data fails. With the use of PCA, data can be compressed without the loss of any information. This can be done by reducing the number of dimensions.

2.2 Learning Phase

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.

2.2.1 Turney Values
It refers to the average of all SO (Semantic Orientation) values for the text. In the present work, the calculation of SO values involves the approach used by Mullen and Collier. The SO value is the difference between PMI (Pointwise Mutual Information) with the word "best" and the PMI with the word "worst".

2.2.2 Osgood values
In this approach, three values are obtained, namely: potency, activity and evaluative. These were introduced in Charles Osgood's Theory of Semantic Differentiation. Potency defines whether the word is strong or weak. Activity defines the active or passive nature. Evaluative defines the overall idea of the word, if it is good or bad. These values can be derives in WordNet.



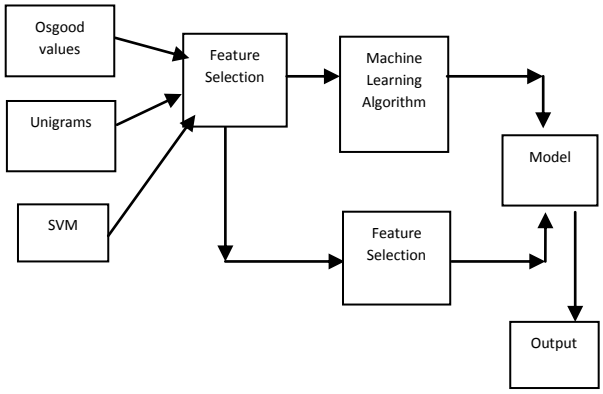


Figure 2: Flowchart showing the steps for Learning Phase.

2.2.3 Unigrams
It is a kind of probabilistic language model. The main advantage of using unigrams is its simplicity for usage. It refers to making a continuous sequence of one item from the given data set. The features as extracted by the system are treated as a single entity and the sentiment is extracted from the individual word (feature). It has reported to produce 82.8% and 83.5% accuracy with 5 fold and 10 fold respectively. But tend to improve the results when used with Osgood values.

2.2.4 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. Libsvm package has been used for training and testing.

2.2.5 SVM String 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 [12].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. Libsvm package has been used 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.

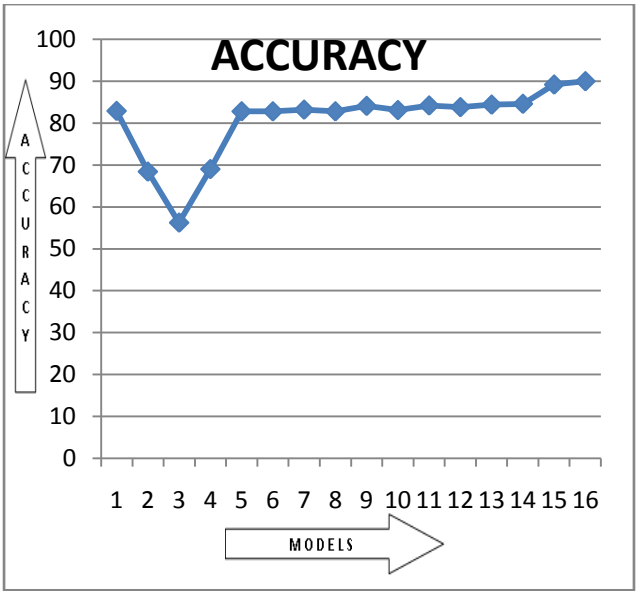
2.2.5 Multiple Kernel
It refers to the combination of two or more kernels in such a way that performance measure of the system is enhanced. Measures for studying multiple kernel learning include maximum margin classification errors, kernel alignment, Fisher discriminative analysis etc. [13] it has been regarded as the most promising approach during our work. It is capable of producing best results amongst all other approaches used. Shogun package has been used for training and testing.
.

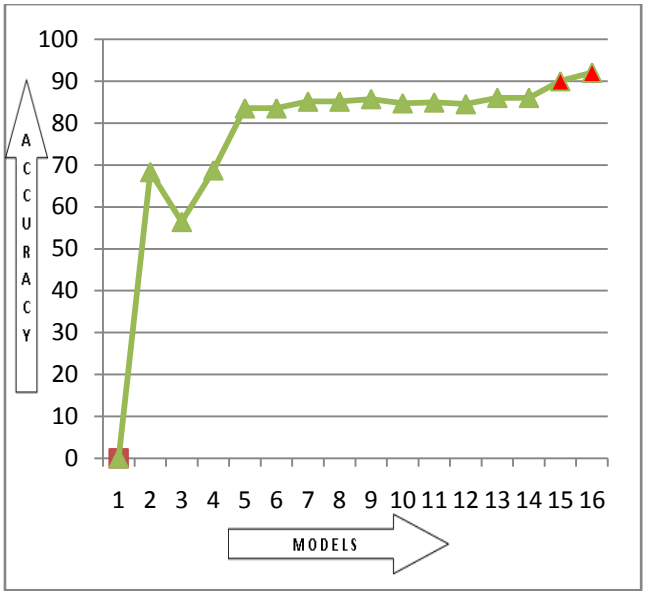
3. RESULTS

To compare the performances of the above stated models, cross validation has been performed on the data set. Since the dataset is large, so testing of the values has been done with 5 fold and 10 fold cross validation on the features extracted by the system. In 5 fold cross validation, the data set is divided into 5 subsets of equal size and the following algorithm is followed:
Train classifier on folds: 2 3 4 5; test against fold: 1
Train classifier on folds: 1 3 4 5; test against fold: 2
Train classifier on folds: 1 2 4 5; test against fold: 3
Train classifier on folds: 1 2 3 5; test against fold: 4
Train classifier on folds: 1 2 3 4; test against fold: 5
In case of 10 fold cross validation:
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

Table 1. Accuracy result comparison for 5 fold and 10 fold cross validation on Movie review dataset.

S.No.	Model	5 folds	10 folds
1	Pang et al. 2002	82.9%	NA
2	Turney Values only	68.4%	68.3%
3	Osgood only	56.2%	56.4%
4	Turney Values and Osgood	69.0%	68.7%
5	Unigrams	82.8%	83.5%
6	Unigrams and Osgood	82.8%	83.5%
7	Unigrams and Turney	83.2%	85.1%
8	Unigrams, Turney, Osgood	82.8%	85.1%
9	Lemmas	84.1%	85.7%
10	Lemmas and Osgood	83.1%	84.7%
11	Lemmas and Turney	84.2%	84.9%
12	Lemmas , Turney, Osgood	83.8%	84.5%
13	Hybrid SVM (Turney and Lemmas)	84.4%	86.0%
14	Hybrid SVM (Turney/ Osgood and Lemmas)	84.6%	86.0%
15	**SVM String Kernel(Turney and Lemmas)**	**89.2%**	**90%**
16	**Multiple kernel**	**90%**	**92%**





4. CONCLUSION

Some of the machine learning approaches namely Naives Bayes, Maximum Entropy, SVM and Kernels were explored and Multiple Kernel outperforms them all. Multiple Kernel produces an accuracy of 90% and 92% for cross validation in 5 fold and 10 fold respectively. However the combination of Multiple kernel 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.

5. ACKNOWLEDGMENTS

Our sincere thanks to all the experts who have contributed towards completion of the project. We would fail in our duty if we do not thank our institution which provided us the platform and resources to develop the project.

6. REFERENCES

[1] Prudêncio, R.B.C. ; Pradhan, S.S. ; Shah, J.Y. ; Pietrobon, R.S.: Good to be Bad? Distinguishing between Positive and Negative Citations in Scientific Impact. In: Centro de Inf., Univ. Fed. de Pernambuco, Recife, Brazil. (2012)

[2] Rotella, P., Chulani, Sunita: Analysis of customer satisfaction survey data. In: Cisco Syst., Inc., Research Triangle Park, NC, USA. (2011)

[3] 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)

[4] 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.

[5] 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

[6] Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL. (2004) 271-278

[7] 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

[8] Nasukawa, T., Yi, J.: Sentiment Analysis: Capturing favorability using natural language processing. In: the Second International Conferences on Knowledge Capture. (2003) 70-77

[9] 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

[10] Tony mullen and Nigel Collier, Sentiment analysis using support vector machines with diverse information sources. (2003)

[11] 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

[12] 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

[13] Bao Y. and Ishii N., "Combining Mutilple kNN Classifiers for Text Categorization by reducts", LNCS 2534, 2002, pp.340-347

[14] 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

[15] Sebastiani F., "Machine Learning in Automated Text Categorization", ACM Computing Surveys, vol. 34 (1), 2002, pp 1-47