International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Sentiment Analysis on Movie Reviews Based on Combined Approach

Anurag Mulkalwar[1], Kavita Kelkar[2]

[1]Research Scholar, Dept. of Computer Engineering, K.J.Somaiya College of Engineering, University of Mumbai VidyaVihar, Mumbai, India

[2]Associate Professor, Dept. of Computer Engineering, K.J.Somaiya College of Engineering, University of Mumbai VidyaVihar, Mumbai, India

Abstract: ***Sentiment Analysis is one of the most important area of concern towards classifying sentiments from any given textual information. Sentiment analysis can be performed on textual data by using various machine learning methods like supervised learning or unsupervised learning. But no any individual method is sufficient to classify sentiments with that much precision. In this paper we are going to propose new approach to perform sentiment analysis on movie review called as Combined Approach. It uses two separate classifier Support Vector Machine (SVM) and Hidden Markov Model (HMM). Then it combines results of these classifier using classifier combine rule.***

Keywords: Sentiment Analysis Combined Approach, HMM, SVM.

1. Introduction

Sentiment analysis is sub division of Data Mining for detection of human behavioral contain from textual data. Now today World is moving toward complete digitalization. We can see that massive growth in user of internet from all age group. By considering this as a benefit for business promotion each business wants to take command on internet platform. Looking at growth of users and their attraction towards internet so many business start promoting their products facilities on internet. It is not stick to only single business but varies from promotion of movies, electronic products, online marketing and also for social networking like twitter, facebook. Each of this marketing portal start providing some space for user to leave their reviews regarding particular thing. If we look at any site which is providing information related movies like IMDB we can see lots of comments from visitors over there. It's not only about movie reviews but it is all about each and every site which wants take reviews from their visitors. Reviews are helpful contain for any business to know what is exact response of their reviewer. But if we what is exact response of their reviewer. This is because of huge amount of reviews. Recent days film industry is in boom. Turnover of this industry is in millions or billions. This is because of huge amount of reviews.

Recent day film industry is in boom. Turnover of this industry is in millions or billions. Each weak number of movies are released. Numerous websites provide information about movies. Movie lovers always want watch good movie for which they spend money. After watching movie they always want to leave reviews or before going for movie wants to know comments on that movie. Beside of reading all reviews it will be always better to know how many there are positive or negative reviews. Movie review contains talk about various aspect like acting, direction, story, film quality. But if we look carefully it is not possible for any human analyzer to read each and every review and to decide what reviewer saying either positive or negative. To know about each aspect without going throw each review process like sentiment analysis is really useful.

Sentiment analysis techniques take care of such situations. Sentiment classifier looks for sentiment from text data and classify that in their appropriate class positive, negative or neutral. Sentiment analysis process get completed in various steps. This process start with gathering of reviews in text form., then collected data get cleaned and form unit for processing called as token. Required features are get extracted from tokens and all features are pass to the sentiment classifier. Sentiment classifier is a core part of sentiment analysis process.

Sentiment Analysis is possible at various levels like Aspect level, sentence level and document level.

- **Aspect level:** classification based on sentiment related to aspect.
- **Sentence level:** classification based on sentiment present in particular sentence.
- **Document level:** classification based on sentiment present in whole document.

It is most important thing in sentiment analysis processes which classifier you are using for sentiment classification. If classifier is not working properly then main intension of sentiment analysis will never get achieved. Mainly for sentiment classification machine learning techniques are used. Supervised learning or unsupervised learning methods are preferably used for classification. But no single classifiers efficiently manage to classify sentiment correctly. So to improve classification technique we are going to use two classifiers SVM and HMM.

2. Related Work

As per mention in [1] Sentiment analysis applies natural language processing techniques and computational linguistics to extract information about sentiments expressed by users about any subject. It shows that there are number of number of approaches are present to do sentiment analysis.

Sentiment classification can be possible at three levels like document; sentence and feature level complete discussion is given in [2]. Sentiment analysis has their sentiment classification techniques lexical approach and machine learning approach. Lexical approach and its improved approach is explained in [3] which shows improvement of basic lexical approach for topic detection. Same ways there are machine learning approaches are also used for sentiment classification [4] like SVM, Naïve Bayes, Maximum Entropy Approach. It also shows comparison between these approaches it concludes that SVM gives batter classification of sentiment. In 2009 FAN Na , CAI Wan-dong, ZHAO Yu propose A Method based on Generation Models for Analyzing Sentiment-Topic in Texts[5]. As per this method it firstly sentiment and topic of training texts are labeled by hand and sentiment models and topic models are established and secondly compute the Kullback-Leibler divergence between a testing text and sentiment models in order to determine sentiment of the text. Traditional approaches of text sentiment analysis typically work at a particular level, such as phrase, sentence or document level, which might not be suitable for the documents with too few or too many words. Considering every level analysis has its own advantages, a combination model may achieve better performance. In this [6], a novel combined model based on phrase and sentence level's analyses and a discussion on the complementation of different levels' analyses are presented. Sentiment analysis seeks to characterize opinionated or evaluative aspects of natural language text thus helping people to discover valuable information from large amounts of unstructured data. A new methodology for sentiment analysis called proximity-based sentiment analysis. It is a different approach, by considering a new set of features based on word proximities in a written text. It [7] propose three proximity-based features, namely, proximity distribution, mutual information between proximity types, and proximity patterns. Experimental results show that proximity based sentiment analysis is able to extract sentiments from a specific domain, with performance comparable to the state-of-the-art. Sentiment analysis of product reviews has recently become very popular in Web text mining, natural language processing and computational linguistics research. Automated analysis of the sentiments presented in online consumer feedbacks can facilitate both organizations' business strategy development and individual consumers' comparison shopping. The main contribution of this paper [8] is the illustration of a novel feature-level sentiment analysis mechanism which is underpinned by a fuzzy domain sentiment ontology tree extraction algorithm. The proposed mechanism can automatically construct fuzzy domain ontology tree (FDSOT) based on the product reviews, including the extraction of sentiment words, product features and the relations among features. Here product features (or features) mean product components and attributes. [9]This paper goes beyond sentiment classification by focusing on techniques that could detect the topics that are highly correlated with the positive and negative opinions. Such techniques, when coupled with sentiment classification, can help the business analysts to understand both the overall sentiment scope as well as the drivers behind the sentiment. [10] This paper shows an empirical study to apply classification-based sentiment analysis on online reviews with multiple dimensions using natural language processing techniques. The aim of this study is to find the most influential part-of-speech on the sentimental analysis and the performance of the multi-dimensional classification methods. [11] this paper focus on various feature selection methods currently present in action. This paper also propose new approach of feature selection it also shows how feature selection is important in sentiment analysis. [12] Paper shows result of sentiment classification by using MI as a feature extraction method and various machine learning classifiers.

3. Methodology

3.1 Preprocessing

In this module we clean incoming textual review for ease of classification. This module consists of various processing tasks like tokenization, stop-word removal and lemmatization.

Tokenization: Tokenization process start with taking input as raw text stream which is in our case one of the review regarding movie. Task of this step is to form tokens of continued text stream. Logic behind implementation of this sub module is firstly convert whole uppercase in to lowercases and then process each word until found whitespace if whitespace found then consider it as a token.

- **Stop-word Removal:** Fully tokenized input of Tokenization sub module taken for finding unwanted words. Stop word are those words which are the parts of sentence but doesn't contain any sense.
- **Lemmatization:** In English language single word has many forms. It causes increase no. of features. To avoid needs to cut down each word in to its root form. Lemmatization processes transform each word in to its original form.
- **Synonyms Finder:** This step finds synonyms of each word and generate proper set of tokens. It helps to reduce number of tokens.

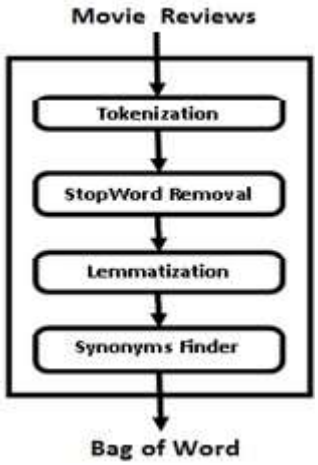

Figure 1: Preprocessing

3.2 Slang Words and Smiley Processing Unit

Smiley is mostly preferable type for giving comments. Wide range of users or viewers use smiley along with text for commenting. And it is really necessary task to find out such

smiley from textual review and analyze it. Most of the time users write comment by using slang words. By simple looking at the slang word human can understand its meaning but it is complicated task for classifier to understand such tokens. So slangs are another challenge for proper sentiment classification. This Slang words and Smiley processing module take care of find out such slang words and return its appropriate meaning to classifier. Same way slang words are getting converted in to its meaningful form. For doing that Slang words and preprocessing module use predefined database to recover exact meaning of that smiley or slang.

3.3 Feature Extraction

Before sending all words to classifier minimize the whole bag of word. Because when we use supervised learning method for classification and if the size of feature set gets increased then there are chances of accurate classification of particular review by classifier get decreases. This violation of supervised classifier from its goal is because of over fitting. To avoid this situation it is compulsory task to minimize size of feature set. Appropriate feature extraction method can handle this case in efficient manner. We are going to use Mutual Information (MI) feature extraction method to extract useful features that will helps in accurate classification to classifier. MI calculates mutual independence of two random variables. In our case random variables are nothing but features. W is a token from bag of word and C is a desired class of classification from class set.

$$MI(W,C)= \log \frac{P(W,C)}{P(W)P(C)}$$

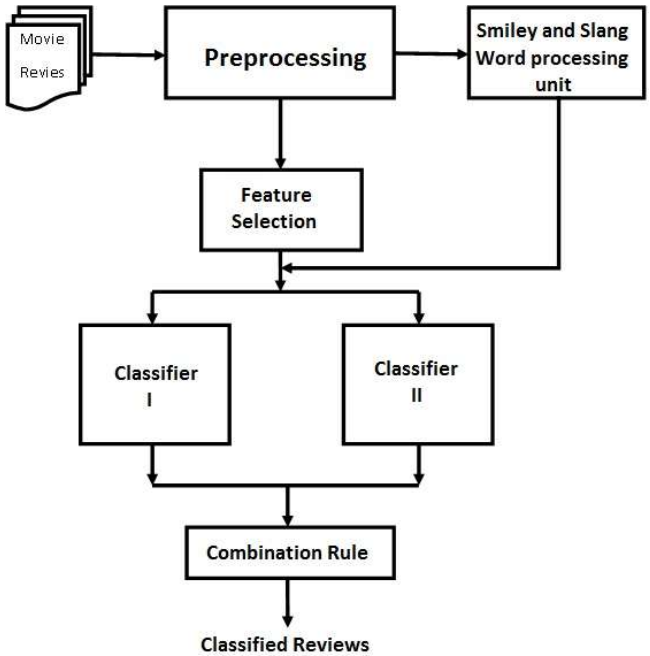

Figure 2: Architecture

3.4 Hidden Markov Model (HMM: Classifier I)

Hidden Markov Model is a well known structure of probabilistic automata. It deals with number of states, and associated transition probability. In our case number of states are nothing but set of class {Positive and Negative}. Transition probabilities are probability of observation by which transitions are takes place. Hidden Markov Model for text classification is depend on vector

$$M= (I, E, T, O, S)$$

where I initial probabilities, E output symbol emission probabilities, T states transition probabilities, O set of output symbols and S set of states. So vector representation of HMM.

3.5 Support Vector Machine (SVM : Classifier II)

Support Vector Machines are highly effective rather than traditional test classification methods. SVMs are proven best technique over Maximum Entropy and Naïve Bayes Method. SVMs has various types but of which Linear SVM is consider as a suitable type for text classification. Simple idea behind the SVM in two-category classification is to find hyperplane which separates two classes with the help of vector $\vec{w}$. It not only separates the document vectors but maximize the margin that separates document vector.

$$\vec{w} = \sum_{j}^{n} \propto_i y_i \vec{d}_i \propto_i \geq 0$$

Where $\propto_i$ is given by solving dual problem. $\vec{d}_i$ with value of $\propto_i$ is greater than zero are consider as Support Vectors. SVM finds proper support vector which will separate two categories exactly.

3.6. Classifier Combination Rules

As above mention our two classifier, one is SVM and HMM are combine with classification rule as describe in following section. There are many possible method for combination of two different classifier. Consider, text classification problem where text is assign to one of the possible class label L_k ($L_k = f_1, f_2, \ldots\ldots f_m$). Let assume that we have N classifiers denoted as C_k(k=1, 2, ……N). Input sample X_k is assign to every classifier C_k and output measure in form of posterior probability vector represent as follows,

$$p_k = [\, p(f_i|X_k), \ldots\ldots\ldots, p(f_m|X_k)]$$

Where, $p(f_i|X_k)$ denote the probability of classifier when X labeled as f_i. Here, consider j type of label is assign to z test data in given combination rules.

Sum Rule:

Label f_i is assign to z when following condition occur.

$$assign\ Z \rightarrow f_j$$

$$j = \arg\max \sum_{k=1}^{N} p(f_i/x_k)$$

Major Voting Rule:

Shortly, each classifier directly assign label to test data then major voting rule assign the final label to test data by selecting major vote given by all the classifier for that test data. When different label got the same number of maximum counts, then randomly class label is selected among them. The mathematical formula for major voting rule is given as follow.

$$assign\ Z \rightarrow f_i$$

$$j = \max_{f_i} \sum_{k=1}^{N} \Delta_k\, p(\frac{f_i}{x_k})$$

Where,

$$\Delta_k p\left(\frac{f_i}{x_k}\right) = \begin{cases} 1\ L_k = f_i \\ 0\ L_k \neq f_i \end{cases}$$

Max Rule:
Max rule is applied over information provided by probability of $P_j(f_i|x_k)$. Max rule is always winner in major voting class. It is expressed by following function.

$$assign\ Z \rightarrow f_j$$

$$j = \arg\max\{\max p_k(\frac{f_i}{x_k})\}$$

Min and mean rule also use for combination of two classifier. Min rule can be derive by the last result of max rule. Mean rule can be obtain by max rule also by taking mean of probability function{ mean $p_k(\frac{f_i}{x_k})$}.

Based on this classifier combination rules HMM and SVM model greatly outperform. Sum rule give the more accuracy than any other combination rule.

4. Conclusion

In this paper we propose new approach called combined approach to classify text reviews based on sentiment present in that reviews. With the help of two classifier and classifier combination rules it is possible to improve expected classification results. We also propose way of handling slang words and smiley. It will overall causes good sentiment classification with higher accuracy.

5. Future Work

Currently we are working towards development of such system which will parse through all reviews regarding particular movie from various sites and it will classify all the reviews based on sentiments present in reviews. We plan to design such a model which will perform best classification on reviews from all domains. We plan to apply this technique on audio sentiment classification.

References

[1] Sowmya Kamath S, Anusha Bagalkotkar, Ashesh Khandelwal, Shivam Pandey, Kumari Poornima, "Sentiment Analysis Based Approaches for Understanding User Context in Web Content", International Conference on Communication Systems and Network Technologies, 2013
[2] Bing Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.
[3] Keke Cai*, Scott Spangler!, Ying Chen!, Li Zhang, **"**Leveraging Sentiment Analysis for Topic Detection" IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology ,2008.
[4] Bo Pang and Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002.
[5] FAN Na , CAI Wan-dong, ZHAO Yu, "A Method based on Generation Models for Analyzing Sentiment-Topic in Texts" IEEE, 2009.
[6] Si Li, Hao Zhang, Weiran Xu, Guang Chen and Jun Guo, " Exploiting Combined Multi-level Model for Document Sentiment Analysis", International Conference on Pattern Recognition , 2010
[7] S.M.Shamimul Hasan, Donald A. Adjeroh, "Proximity-Based Sentiment Analysis", IEEE , 2011.
[8] Lizhen Liu, Xinhui Nie, Hanshi Wang, "Toward a Fuzzy Domain Sentiment Ontology Tree for Sentiment Analysis" 5th International Congress on Image and Signal Processing (CISP) ,2012.
[9] Mostafa Karamibekr, Ali A. Ghorbani, "Sentiment Analysis of Social Issues", International Conference on Social Informatics , 2012
[10] Samatcha Thanangthanakij, Eakasit Pacharawongsakda, Nattapong Tongtep, Pakinee immanee, Thanaruk Theeramunkong, "An Empirical Study on Multi-Dimensional Sentiment Analysis from User Service Reviews", Seventh International Conference on Knowledge, Information and Creativity Support Systems, 2012.
[11] Peter Koncz and Jan Paralic, "An approach to feature selection for sentiment analysis", 15th International Conference on Intelligent Engineering Systems, 2011.
[12] Wang Zuhui Jiang Wei, "Online Reviews Sentiment Analysis Applying Mutual Information", 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2012.

Author Profile

Mr Anurag V. Mulkalwar is B.E.(Computer) From University of Mumbai. Currently pursuing his M.E.(computer) from University of Mumbai, Maharashtra, India.

Mrs Kavita M Kelkar is working as Associate Professor in Department of Computer Engineering, KJSCE, Vidya Vihar, Mumbai. She has been working as faculty at KJSCE since June 2000. Mrs Kavita is B E (Computer) from University of Pune and M E(IT) from University of Mumbai. She is currently pursuing her Ph.D from University of Mumbai, Maharashtra, India