International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 3 Issue 3, March 2015

A Survey of Sentiment Classsifiction Techniques

Sangita N. Patel[1], Jignya B. Choksi[2]

[1]ME Student, Computer Engineering, Sardar Vallabhbhai Institute of Technology, Vasad-388306, Gujarat-India

[2]Assistant Professor, Computer Engineering, Sardar Vallabhbhai Institute of Technology, Vasad-388306, Gujarat-India

Abstract: ***Sentiment classification is an ongoing field and interesting area of research because of its application in various fields collecting review from people about products and social and political events through the web. Currently, Sentiment Analysis concentrates for subjective statements or on subjectivity and overlook objective statements which carry sentiment(s). During the sentiment classification more challenging problem are faced due to the ambiguous sense of words, negation words and intensifier. Due to its importance the correct sense of target word is extracted and determined for which the similarity arise in WordNet Glosses. This paper presents a survey covering the techniques and methods in sentiment analysis and challenges appear in the field.***

Keywords: Sentiment classification, Word sense disambiguation, Intensifier, SentiWordNet, WordNet

1. Introduction

Large datasets are available on-line today, they can be numerical or text file and they can be structured, semi-structured or non-structured. Sentiment classification is tracking the mood of public about particular product or event or topic. Many different information retrieval techniques and tools have been proposed according to different data types. Sentiment classification, also known as opinion mining, is to identify and extract subjective information in source materials, which can be positive or negative. Using appropriate mechanisms and techniques, this large amount of data can be processed into information to support decision making.

During the sentiment analysis people require fast ans accurate information so that they can make quick and accurate decisions. People often ask their friends, family members for decisions making. Researchers in sentiment analysis have focused mainly on two problems– detecting whether the text is subjective or objective, and determining whether the subjective text is positive or negative, and the objective text in SentiWordNet by considering the sentimental relevance of objective text and their associated sentiment sentences.

The main task in sentiment classification is to determine the polarity of the comments as positive, negative or objective. It can be done at different levels such as word/phrase levels, sentence level and document level. Sentiment can be expressed in text, in different ways. The following are examples of them:

- I read this book later.
- The book is good.
- I like to read this book.
- The book is very good.

Opinion can be collected from different sources, e.g. Newspaper, Television, Internet etc.The web has become the largest source of opinion. Before web opinion collected manually.

Word Sense Disambiguation (WSD) refers to a task that automatically assigns a sense, selected from a set of pre-defined word senses to an instance of a polysemous word in a particular context. WSD is an important but challenging technique in the area of natural language processing (NLP). It is necessary for many real world applications such as machine translation (MT), semantic mapping (SM), semantic annotation (SA), and ontology learning (OL). It is also believed to be helpful in improving the performance of many applications such as information retrieval (IR), information extraction (IE), and speech recognition (SR)[1].

WSD in text, the following are examples of them:

- I went fishing for some sea bass.
- The bass line of the song is too weak.

To a human, it is obvious that the first sentence is using the word "bass (fish)", as in the former sense above and in the second sentence, the word "bass (instrument)" is being used as in the latter sense below. Developing algorithms to replicate this human ability can often be a difficult task, as is further exemplified by the implicit equivocation between "bass (sound)" and "bass" (musical instrument)[1].

2. Classification Algorithm

Classification techniques are widely used to classify data among various classes. There are many algorithm used for Sentiment classification. There are mainly two types of Sentiment classification algorithms Machine Learning Approach and Lexicon-based Approach. Sentiment classification techniques can be divided into following approach.

Machine Learning Approach:
a) Supervised Learning
- Decision tree classifier
- Rule-based classifier
- Support vector machine
- Neural Network
- Naive Bayes
- Bayesian classifier
- Maximum Entropy

b) Unsupervised Learning

Lexicon- Based Approach:
c) Corpus –based Approach
- Statistical
- Semantic

d) Dictionary –based Approach

Here briefly discuss about classification techniques. Supervised machine learning techniques are used for classified document or sentences into finite set of class i.e into positive, negative and objective. Training data set is available for all kind of classes. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way. We are using Support Vector Machine (SVM), Naive-Bayes, Maximum Entropy for classification purpose.SVM efficiently classifies Moive Review dataset into positive, negative category [2].

Unsupervised machine learning techniques don't use training data set for classification. Semantic Orientation also provides to generate accurate result for classification. Point wise mutual information (PMI) is also one of the unsupervised classification methods for sentiment analysis [3].

The corpus-based techniques try to find co-occurrence patterns of words to determine their sentiments. Turnery(2002) calculated a phrase's semantic orientation to be the mutual information between the phrase and the word "excellent"(as positive polarity)minus the mutual information between the phrase and the word "poor"(as negative polarity).The overall polarity of an entire text was predicted as the average semantic orientation of all the phrases that contained adjectives or adverbs.[3]

Dictionary based techniques are synonyms, antonyms and hierarchies in WordNet (or other lexicons with sentiment information) to determine word sentiment.

3. Literature Survey

This section describes literature review or the studies which give an idea that for our research done in direction of sentiment classification.

A. Yan Dang, Yulei Zhang proposed lexicon enhanced method for sentiment classification combines machine learning and semantic-orientation approaches into one framework that significantly improves sentiment classification performance. We also found that conducting feature selection can further improve the performance, especially for large data sets. They compared Naïve Bayes, Maximum Entropy, and SVM and achieved the highest classification accuracy (82.9 percent) using SVM.[4]

The semantic-orientation approach, on the other hand, performs classification based on positive and negative sentiment words and phrases contained in each evaluation text and mining the data requires no prior training.[4]

Advantages
- With introduction of sentiment features this approach provides better performance.

Disadvantages
- This method requires further refinement in the direction of lexicon extraction process.

For further study in this area is to refine the lexicon and extend the sentiment feature-extraction procedure.
Further research can also explore other sentiment feature-generation methods, such as corpus-based techniques, and compare their performance.

B. Chihli Hung,Hao-Kai Lin proposed approach for mine sentiments of opinions from word-of-mouth (WOM) to improve the performance of word-of-mouth Sentiment classification by re-evaluates objective sentiment words in the SentiWordNet sentiment lexicon with the help of SVM classifier.[5]

WordNet is a public sentiment lexicon that's used to extract sentiments of WOM for sentiment classification. However, most existing sentiment mining models ignore objective words, which comprise more than 90 percent of the words in SentiWordNet. These objective words are often considered useless. Research reevaluates objective words in SentiWordNet by assessing the sentimental relevance of objective words and their associated sentiment sentences. In this paper two sampling strategies and integrate them with the support vector machines (SVMs) for sentiment classification.[5]

As an example, we'll use two sentences wherein each word contains three sentiment values in brackets—that is, Positive, objective, and negative—while looking up SentiWordNet as follows:
• Sentence 1: I (p:0, o:1, n:0) will (p:0, o:1, n:0) read (p:0, o:1, n:0) this (n/a) book (p:0, o:1, n:0) later (p:0, o:1, n:0).
• Sentence 2: Reading (p:0, o:1, n:0) this (n/a) book (p:0, o:1, n:0) is (n/a) happy (p:0.875, o:0.125, n:0).

Calculation:
A word whose sentiment value is the greatest in positive, negative, or objective orientation is defined as a positive, negative, or objective word, respectively.

Advantages
- Based on the average accuracy and standard deviation, the proposed, revised SentiWordNet model achieves a higher and more stable classification performance.

Disadvantages
- This method, extracts the first sense of a word from assigned POS tag in SentiWordNet because this usage is generally the most common. But it can cause word sense disambiguation.

The technique of word sense disambiguation could be applied before the extraction of SentiWordNet. Sentiment extraction from linguistic or semantic viewpoints is another possible direction. This work uses SVM techniques; a further research direction might focus on using various

classification algorithms such as ensemble learning for sentiment classification.

C. Jasmine Bhaskar, Sruthi K.,Prema Nedungadi proposed an enhanced technique for sentiment classification of online reviews by considering the objective words [5] and intensifiers[6].

Intensifier Handling: People usually use intensifiers in reviews to express their emotion deeply. Presence of the words like 'very " 'really 'and 'extremely ' in negative and positive sentences make the adjective and adverb stronger. But this effect is not considered during the score calculation in existing method.

Table 1: Intensifier Handling in Positive and Negative sentence[6]

Previous Word	*Next word*	*Score*
Intensifier	Adjective[Negative]	High Negative
Intensifier	Adjective[Positive]	High Positive
Intensifier	Adverb[Negative]	High Negative
Intensifier	Adverb[Positive]	High Positive

The polarity of the sentence can be obtained by following equation. Sentence Score= $\sum_{I=1}$ Score(i)
Score (i) is the positive and negative score of the words and n is the number of words in the sentence. If Sentence Score is greater than 0, then we can say that the sentence is positive otherwise sentence is negative.

Advantages

- Prediction accuracy of this method is much better than the traditional and existing methods.
- Though the existing method out performs the traditional method its accuracy is less compared to the presented method. This is because miss-classification is less in the proposed method related to the negative sentences as compared to the existing method. This improvement is due to the proper handling of intensifiers.

Disadvantages

- This method can effectively handle intensifier but they doesn't present effective approach for negation modifier.

In further direction research lies in applying Word sense disambiguation and identification of the product feature about which the sentiment is expressed.

D. M. Govindarajan proposed new hybrid classification method is proposed based on coupling classification methods using arcing classifier and their performances are analyzed in terms of accuracy.[7] A Classifier ensemble was designed using Naïve Bayes (NB), Support Vector Machine (SVM). In the proposed work, a comparative study of the effectiveness of ensemble technique is made for sentiment classification. The ensemble framework is applied to sentiment classification tasks, with the aim of efficiently integrating different feature sets and classification algorithms to synthesize a more accurate classification procedure.[7]

Advantages

- A comparison between Naive bayes and SVM classifier and SVM provides better performance.
- Other comparisons between SVM and ensemble Naive bayes SVM classifier. Hybrid classifier show the significant improvement over the single classifiers.

In Future direction this method requires further refinement in the direction of various classification algorithms.

E. In this work Muhammad Faheem Khan, Aurangzeb Khan and Khairullah Khan proposed a new method of word sense disambiguation (WSD) using matrix map of the semantic scores extracted from SentiWordNet of WordNet glosses terms.[8] The correct sense of the target word is extracted and determined for which the similarity between WordNet gloss and context matrix is greatest. Experiment results have shown that the proposed method improves the result of sentence level sentiment classification as evaluated on different domain datasets. From the result it is clear that the propose method achieves an accuracy of 90.71% at sentence level sentiment classification of online reviews.[8]

In future direction research lies in applying Word sense disambiguation using matrix map for semantic orientation at document level and feedback level and Word sense disambiguation matrix map will applied for the improvement of sentence clustering which may in turn be based on improved sentence similarity measures. We are currently exploring the feasibility f using the matrix map technique in other text mining task.

4. Cross-Domain Sentiment Classification[9]

Cross domain sentiment analysis is introduced to reduce the manual effort in training the machine using labeled data. Instead the machine learns from a particular domain and analyze the sentiment polarities of texts in another domain. This is a very challenging problem because the kind of words used to express emotions in two different domains may be very different. A paper [9] approaches this topic vastly covering all the difficulties evolved in the problem. A sentiment sensitive distributional thesaurus is created using labeled data for the source domains and unlabelled data for both source and target domains. Sentiment sensitivity is achieved in the thesaurus by incorporating document level sentiment labels in the context vectors used as the basis for measuring the distributional similarity between words. The created thesaurus is used to expand feature vectors during train and test times in a binary classifier.

Advantages

- This approach overcome feature mismatch problem arise in cross-domain sentiment classification, by using labelled data from multiple source domains and unlabeled data from source and target domains to compute the relatedness of features and construct a sentiment sensitive thesaurus**.**

Disadvantages

- This method restricted to semi-supervised domain adaptation category. For fully supervised category, this method doesn't provide desirable result.

This method can be extended for fully supervised category, in order to determine cross-domain sentiment classification.

5. SentiWordNet and WordNet

SentiWordNet is sentiment analysis lexical resource made up of synset from WordNet, a thesaurus-like resource; they are allocated a sentiment score of positive, negative or objective. These scores are automatically generated using the semi-supervised method which is described in [10]. It is also available freely for research purpose on web. SentiWordNet is one of the sources of sentiment analyses. It is a semi-automatic way of providing word/term level information on sentiment polarity by utilizing WordNet database of English terms and relations. WordNet is is a very rich source of lexical knowledge Since most entries have multiple senses. Each term in WordNet database is assigned a score of 0 to 1 in SentiWordNet which indicates its polarity. Strong partiality information terms are assigned with higher scores whereas less bias/subjective terms carry low scores. SentiWordNet is made up of a semi-supervised method which refers to a subset of seed terms to obtain semantic polarity. Each set of synonymous terms is assigned with three numerical scores ranging from 0 to 1 which indicates its objectiveness i.e. positive and negative bias [11]. One of the key features of SentiWordNet is that it assigns both positive and negative scores for a given term according to the following rule [10]: For a synset *s*, we define [11].

- *Pos(s) Positive score for synsets.*
- *Neg(s) Negative score for synsets.*
- *Obj(s) Objectiveness scores for synsets.*

Then the following scoring rule applies:
Pos(s) + Neg(s) + Obj(s) = 1 ; The positive and negative scores are always given, and objectiveness can be implied by the relation: Obj(s) = 1 – (Pos(s) + Neg(s)). Polarity scores according to synset and relevant part of speech are grouped by SentiWordNet database as a text file. The table below describes the columns for one entry in the database reflecting opinion information of a synset.

Table 2: SentiWordNet database structure

Fields	*Description*
POS	Part of speech linked with synset. This can take four possible values: a = adjective=jj n = noun=nn v = verb=vb r = adverb=rb
Offset	Numerical ID which associated with part of speech uniquely Identifies a synset in the database.
Positive Score	Positive score for this synset. This is a numerical value ranging from 0 to 1.
Negative score	Negative score for this synset. This is a numerical value ranging from 0 to 1.
Synset terms	List of all terms included in this synset.

6. Comparative Analysis

A. Discussion

Sentiment classification plays vital role in Businesses and Organizations.

- Product and service bench marking.
- Market intelligence.

People get the other's opinion to make some decision about Product or services.

- Finding opinions while purchasing a new product.
- Finding opinions on political topics.
- In Advertisement (ads) opinion mining helps to Display the product based on the stake holders view.
- Placing ads in the user-generated content.
- Place ads when one praises a product.
- Place ads from a competitor if one criticizes a product.

Finally sentiment can be served in the field of Information search & Retrieval. In opinion mining Determining sentiments seems to be easier, determining objects and their corresponding features is harder.

B. Comparision

Table 3: Comparative study of papers

Sr.no	Author name	Technique/ Approach/ Method and [Dataset]	Advantage	Limitation
1	Yan Dang, Yulei Zhang	Combine Machine Learning and semantic orientation(naïve bayes, Maximum entropy, SVM) [Product Review Dataset]	Introduce sentiment features to provide better performance	Requires refinement in direction of lexicon extraction process
2	Chihli Hung,Hao-kai,chung yung	SVM, Reevaluate objective word (add threshold value) as pos,neg. and negation handling [Movie Review Dataset]	Revised method achieves high and more stable class performance	To extract first sense of word from SentiWordNet. so Word sense Disambiguation problem.
3	Jasmine Bhaskar, Sruthi K.,Prema Nedungadi	SVM, add sentiment threshold value to Objective words(as Pos,neg)& considering intensifier, negation handling [Product Review Dataset]	Prediction accuracy much better and proper handling of intensifier	To extract first sense of word from SentiWordNet. so Word sense Disambiguation problem.
4	Govindraj an M.	Compare naïve bayes, SVM and NBSVM [Movie Review Dataset]	Ensemble arcing technique gives much better performance	Not applying negation handling,intensifier ,not evaluate objective word
5	Muhammad Faheem khan,aurangzeb khan	WSD technique using Matrix map [Twitter,Airlines,Election Dataset]	Proper handling of WSD and apply for different domain	Further WSD apply for Document level

From literature survey of different techniques we conclude that a new method of word sense disambiguation (WSD) using semantic scores extracted from SentiWordNet of WordNet glosses terms. Along with handling negation scope

and intensifier by considering positive, negative and objective orientation of sentiment.

Table 4: Comparison of classification techniques

Sentiment classification Method	*Pros*	*Cons*
Support Vector Machine	Kernel-based framework is very powerful, flexible ,SVMs work very well in practice, even with very small training sample sizes	No "direct" multi-class SVM, must combine two-class SVMs , Computation, memory - During training time, must compute matrix of kernel values for every pair of examples and Learning can take a very long time for large-scale problems
Naïve Bayes Classifier	Fast and Good performance, Induced classifiers are easy to interpret, Uses evidence from many attributes, handle missing data and Robust to irrelevant attributes, good computational complexity, incremental updates	Assumes independence of attributes, Low performance ceiling on large databases
Decision tree	Fast, Segmentation of data	Fragmentation as number of splits becomes large, Interpretability goes down as number of splits increase
Neural Network	A neural network can perform tasks that a linear program cannot., When an element of the neural network fails, it can continue without any problem by their parallel nature. A neural network learns and does not need to be re-programmed,It can be implemented in any application.	The neural network needs training to operate. ,The architecture of a neural network is different from the architecture of microprocessors therefore needs to be emulated. Requires high processing time for large neural networks.

From the survey of different techniques of sentiment classification. it is concluded that the SVM technique give high performance if data set is small.

7. Conclusion

Sentiment Analysis problem is a machine learning problem that has been a research interest for recent years. Through this literature survey, the relevant works done to solve this problem could be studied. Although several works have come in this field, a fully automated and highly efficient and all problems combine together in single system has not been introduced till now. Because of the unstructured nature of natural language. The vocabulary of natural language is very large that things become even hard.In future, extraction of the acute sense of sentence and remove noisy text for an efficient semantic orientation. Furthermore, the knowledgebase need to improve for the semantic scores of all parts of speech.

8. Acknowledgment

We take the immense pleasure in expressing our humble note of gratitude to our project guide, Ms. Jignya B. Choksi, Assistant Professor, Department of Computer Engineer, Sardar Vallabhbhai Institute of Technology, for this remarkable guidance and suggestions, which helped us in completion of paper.

References

[1] http://en.wikipedia.org/wiki/Word-sense_disambiguation
[2] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002.
[3] Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002.
[4] Yan Dang; Yulei Zhang; Hsin chun Chen, "A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews, "Intelligent Systems, IEEE , vol.25, no.4, pp.46,53, July-Aug.2010 doi: 10.1109/MIS.2009.105
[5] Hung, Chihli, and Hao-Kai Lin. "Using objective words in SentiWordNet to improve word-of-mouth sentiment classification." *IEEE Intelligent Systems* 28.2 (2013): 0047-54.
[6] Bhaskar, J.; Sruthi, K.; Nedungadi, P., "Enhanced sentiment analysis of informal textual communication in social media by considering objective words and intensifiers," *Recent Advances and Innovations in Engineering (ICRAIE), 2014* , vol., no., pp.1,6, 9-11 May 2014 doi: 10.1109/ICRAIE.2014.6909220
[7] Govindrajan M. "Sentiment classification of movie review using hybrid method."
[8] Muhammad faheem Khan, Aurangzeb and khairullah khan "efficient word sense disambigution teqnique for sentence level sentiment classification of online review" Sci.Int(Lahore).25(4),2013
[9] Bollegala, D.; Weir, D.; Carroll, J., "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus," Knowledge and Data Engineering, IEEE Transactions on , vol.25, no.8, pp.1719,1731, Aug. 2013 doi: 10.1109/TKDE.2012.103
[10] A. Esuli and F. Sebastiani, "SentiWordNet: A High-Coverage Lexical Resource for Opinion Mining" Proceedings of LREC, 2006, pp. 417-422.
[11] B. Ohana, "Opinion mining with the SentWordNet lexical resource," Dublin Institute of Technology, 2009.