ISSN 2229-5518

Comparative Study of Sentiment Analysis Techniques in Web

Sasikala P, L.Mary Immaculate Sheela

Abstract— The development of web contributes a huge quantity of user created content such as customer feedbacks, opinions and reviews. Sentiment analysis in web embraces the problem of aggregating data in the web and extraction about opinions. Studying the opinions of customers helps to determine the people feeling about a product and how it is received in the market. Various commercial tools are available for sentiment analysis. In this paper, we are going to compare and analyze the techniques for sentiment analysis in natural language processing field.

Keywords—Machine learning, Natural Language Processing, Opinion Mining, Sentiment Analysis.

---

1 Introduction

"What others think?" is always important information in a decision-making process. Every day people discuss various products on social media sites. Companies want a piece of that pie to determine how their audience communicates to find the important information that drives business. Sentiment analysis is the robotic mining of opinions and feelings from content through Natural Language Processing (NLP). Sentiment analysis is nothing but categorizing opinions in the given content or documents into "positive" or "negative" or "neutral"[1].

Sentiment analysis can be processed in three levels: aspect-level, document-level, and sentence-level. Sentiment analysis in document level considers the entire document as a single topic and classifies positive or negative sentiment. The sentiment expressed in each and every sentence is classified in sentence level. Since sentences are part of the documents does not make much difference between sentence and document level. To get the detail opinion or sentiments have to process the document to aspect level. Aspect level sentiment analysis is to determine the features of the sentiment conveyed towards each aspect and the given target entities.

The first step in Aspect level sentiment analysis is aspect term extraction identifies the terms in the sentence and list all the distinct aspect terms. For example, "I liked the *movie* but not the *music*", Multi-word aspect terms (e.g., "hard disk") should be treated as single terms (e.g., in "The *hard disk* is very heavy" the only aspect term is "hard disk"). The next step is aspect term polarity which classifies the term as positive, negative and conflict or neutral. The third step is aspect category detection and aspect category polarity [2]. For example, given the set of aspect categories {*food, service, price, ambiance, anecdotes /miscellaneous*}:"The restaurant was too expensive" → {*price*} "The restaurant was expensive, but the menu was great" → {*price, food*}

- *Sasikala. P , Research Scholar, Mother Teresa Women's University, Kodaikanal, Tamilnadu , India. E-mail: shakthesasi@gmail.com*
- *L.Mary Immaculate Sheela, Supervisor, Dilla University, Ethiopia. E-mail: drsheela09@gmail.com*

2 Sentiment classification and feature selection

Sentiment classification is to select and extract the text features. Feature selection in sentiment analysis is collecting the information from reviews in web and performing the following steps.

Data Preparation: The data preparation step will pre-process the data and removes all the non-textual information and tags. Data pre-processing performs cleaning of data by removing the information like review date and name of the reviewer which is not required for sentiment analysis.

Review Analysis: finding parts of speech (POS) adjectives and counting the presence and frequency.

Sentiment Classification: Classifies the extracted words as positive or negative.

2.1 Feature selection methods

Feature selection method is divided into lexicon based and statistical methods. Statistical methods are fully automated where lexicon based starts with a small set of words. Three methods are included in this study are mutual information (MI), Chi-Square χ^2 and information gain (IG).

2.1.1. Mutual Information

The Mutual Information is the difference between expected and frequency of co-occurrence. In statistical terms, this is a measure of the strength of association between words s and t. In a given finite corpus where K is the number of times s and t co-occurs, D is the domain count, Y is the number of times t occurs without s and Z is the number of times s occurs without t[18]. Then mutual information of s and t is calculated as

$$Mi(s,t) \approx \log ((K \times D))/((K+Z) \times (K+Y)) \quad (1)$$

2.1.2. Chi-Square χ^2

Chi-square (χ^2) is the test of independence compares two categories s and t in a cross-tabulation fashion to determine the amount of association or difference. In a given finite corpus where K is the number of times s and t co-occurs, Y is the number of times t occurs without s, L is the number of times neither s nor t appears and Z is the number of times s occurs without t[18].

$$\chi^2(s,t)=\frac{D \times (KL-ZY)^2}{(K+Z)\times(Y+L)\times(K+Y)\times(Z+L)} \qquad (2)$$

2.1.3. Information gain

Information gain measures the information in bits by predicting the presence and absence of the information. Given the training set, for each term information gain can be computed and removed from the feature selection whose gain value is less than the estimated threshold.

3 Sentiment Analysis techniques

Sentiment analysis techniques are machine learning, lexicon based and hybrid techniques. Machine learning techniques are implemented in supervised classification. Lexicon based approach depends on the collection of opinion terms. Combined approach of lexicon and machine learning is hybrid approach.

3.1 Machine learning technique

Machine learning approach is the design and development of algorithms which infers data from datasets or databases. Two types of datasets used are training and test set. Training set used by separator categorizes documents based on characteristics and performance is validated by the test set. Supervised learning method finds the relationship between input and target attributes. Input attributes are nothing but independent variables and target attributes are dependent variables. A model is the structure of relationship which is used for predicting the relationship in attributes.

The various classifiers in supervised learning methods are:

3.1.1. Naïve Bayes

The Bayesian Classification is a supervised statistical method for classification and contains practical learning algorithms. The posterior probability of a class can be computed using Naive Bayes model. This model works is suitable for a large data set. The use of the Bayes Theorem is to presume the chance of the inclined feature set matches to specified label.

Bayes theorem provides a way of calculating the posterior probability, P(L|F), from P(L), P(F), and P(L|F). Naive Bayes classifier assumes that the effect of the value of a predictor (F) on a given class (L) is independent of the values of other predictors. The equation can be written as follows [9]:

$$P(L|F) = \frac{P(L|F) * P(L)}{P(F)} \qquad (3)$$

3.1.2. Bayesian Network

A Bayesian network is part of probabilistic graphical models (GMs). An ambiguous domain can be represented using these structures. Every node in the graph points to a arbitrary variable and the edges represents their chance of dependence. These conditional dependencies in the graph are calculated by using known statistical and computational methods[18].

3.1.3. Maximum Entropy Classifier

Maximum entropy is a probabilistic approach used for natural language processing, segmentation, modeling and POS (part-of-speech) tagging. Maximum Entropy (MaxEnt) uses search based optimization to find the features. The general form of MaxEnt classifier uses word-level features can be described as: The probability of class c in document d and weight W is

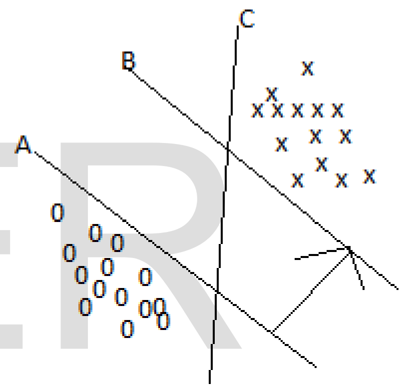
$$p(c \mid d,w)=\exp\sum_i f_{iWi(c,d)} / \sum_{c'\in c}\sum_i f_{iWi(c',d)} \qquad (4)$$

For each word S and class c, features of (S, c) =T, where T is the total number of times S occurs in a document in class c.

3.1.4. Support Vector Machine Classifiers

Support vector machines (SVM) are supervised learning models with associated learning algorithms.SVM are effective approaches for non linear separation and regression analysis.

Fig. 1. Using SVM for classification



In fig1: There are two classes 0 and x, there are three hyperplanes A, B and C. Maximum margin of separation is represented due to the regular interval of one of the data point is leading which margins to hyperplane A to provide high quality separation between classes. Text data is suitable for SVM because of the scattered quality of the text. The features are unrelated but they have a impulse to equate with each other and formed into linearly separable categories [3]. SVM is used for sentiment polarity classification.

3.1.4. Unsupervised Learning

Unsupervised machine learning is a machine learning function to depict unseen structure from unlabeled data. In supervised learning, a large number of training documents are provided which are used to train the machine and get the desired outputs. Sometimes it is difficult to generate a training set and it is easy to assemble unlabeled documents. Unsupervised learning prevails over these complexities [14]. Two classes of the method have been suggested for unsupervised learning is density estimation and feature extraction. Statistical models are built by density estimation technique. Feature extraction techniques try to dig out statistical regularities (or sometimes irregularities) straight from the inputs. The most widespread unsupervised learning method is cluster analysis, which is used to find hidden patterns or grouping in data.

3.1.5. Rule-based classifiers

The rule-based classifier is a machine learning methods that learn 'rules' to apply, knowledge and store. The characteristics of the rule-based model are an identification of the set of rules that represents the knowledge. Association rule mining is part of the rule-based approach. The training phase generates the rule based on certain constraints. Support and confidence is the general constraints. Support indicates the frequency of item-set in the database and confidence refers success or truth of the rule.

3.1.6. Decision tree classifier

Decision tree classifier is a statistical model uses a decision tree which represents the class labels in the leaves and features of those labels in branches. Algorithm for decision tree works by picking the suitable attribute to split the data and expanding the leaf nodes of the tree until it satisfies the required condition.Decision tree classifier is all about finding attributes that return the highest information gain[7].

3.1.7. k-Nearest Neighbour

K-Nearest Neighbour is an Instance-based classifier works on unknown instances. It relates the known to unknown instances by distance or similarity. It does not involve prior assumptions about the distribution of data taken from set of positive and negative samples. A new sample is categorized by computing the interval to the nearest training pattern. Positive or negative sign of that point decides the classification of sample. This approach of locating nearest neighbour and marking the unfamiliar item with the located instance as that of the known neighbour is referred as nearest neighbour classifier.

3.1.8. Lexicon-based approach

A sentiment classification task uses opinion words. Positive and negative opinions are implemented to precise the desired and undesired states. The collection of opinion idioms and phrases are called as opinion lexicons. There are three main approaches to collect the opinion word list. The first approach is manual and can be combined with the other automated approaches to obtain the desired results without mistakes.

The automated approaches are Dictionary-based (DB) approach and Corpus-based (CB) approach. The basic steps of lexicon based techniques are [4]:

i. Pre-process the text (remove the noisy data)
ii. Initialize the sentiment score: sum $\leftarrow$ 0
iii. Tokenize text: For each token present in the dictionary,
   a. If the token is positive then sum $\leftarrow$ sum + N
   b. If the token is negative then sum $\leftarrow$ sum - N
iv. T is the threshold value, if the score sum is
   a. sum > T , identify the text as positive
   b. sum < T , identify the text as negative

A group of handpicked sentiment words with known differences are collected in DB approach. This collection is extended further by finding thesaurus and WordNet corpora for meanings and conflicts. In iteration, freshly identified terms are updated to the archive. Iteration will proceed till no new words found. Corpus-based approaches rely on context specific orientation opinion words.

3.1.9. Hybrid Techniques

In hybrid techniques, both lexicon and machine learning approaches are used. The entropy weighted genetic algorithm (EWGA) uses the information gain searching to enhance the distinct sentiment properties. Existing techniques are merged to prevail over their limitations and by utilising their benefits to increase opinion classification performance. Malandrakis et al. [21] had proposed a hierarchical model based for Twitter sentiment analysis. The hierarchical lexicon-based model proved very successful in spite of using part-of-speech information and n-gram ratings. The suggested model did not perform well independently, but confirmed a visible progress to the lexicon-based model. In general these models concluded a superior accomplishment.

4 Sentiment Analysis tools

There are many tools used for sentiment analysis for detecting the opinions of reviews, blogs or forums in a web which include text, star rating and emoticons. The popular lexicon-based tools available in market are SentiWordNet, Panas-t, NRC and SentiStrength. A new sentiment analysis method that combines various approaches is SASA0.1.3 (PYTHON package).The tool that explores Artificial Intelligence techniques is SenticNet. The tool which uses a wrapper model based entropy weighted generic algorithm is EWGA. The other java based tools are LingPipe, OpenNLP, MALLET and Weka. Opinion Observer is a sentiment analysis tool used to compare the reviews and presents the results in a graph [19].

5 Discussion and Analysis

The trend of research shows a general cataloguing of sentiments relatively than building positive or negative classifications. The increase in the number of articles for general classification shows that sentiment analysis is growing. Study of the papers concluded that rule oriented data accuracy is better than non reliant data [5, 6]. Unsupervised methods are used because of the straightforward availability of unlabelled data.

Most of the research proved that Support Vector Machine (SVM) has high precision. Main constraint of the supervised learning is the creation of expert-annotated training set, and may not succeed when training data is inadequate. The following table shows the result of the comparative study of sentiment analysis techniques in web based on various techniques.

TABLE I. COMPARISON OF SENTIMENT ANALYSIS TECHNIQUES

Paper	Dataset	Technique (Accuracy, %)
Nan Li [7]	Sino-Sports forum	SVM (80%) Decision Tree (58.2%)

Paper	Dataset	Technique (Accuracy, %)
K Nirmala Devi et al .[8]	Forums.digital.point.com	SVM (60%) Naïve Bayes (48.6%)
Evandro et al.[9]	Online blogs	Naïve Bayes (79.67%) SVM (85.50%)
A.Ortigosa et al.[10]	Facebook	SVM (83.27%)
Qiang Ye [11]	Yahoo.com(Tourism Review)	Naïve Bayes (80.71%) SVM(85.14%)
Turney [12]	Epinions	PMI (66%)
Ziqiong et al. [13]	Cantonese Reviews	Naïve Bayes (93%) SVM (90%)
Zang et al.[14]	Twitter	Machine Learning and Lexicon based techniques (82.62%)
Kaiquan Xu et al. [15]	Amazon Reviews	SVM (61%)
Pang et al.[16]	IMDB	Naïve Bayes (81.5%) SVM (82.9%)
K Dashtipour et al. [17]	Blitzer (Books and DVD reviews)	Naïve Bayes and SVM (65%)
M.Govindarajan	Movie-Review Data	Naïve Bayes (NB) 91.15 % Genetic Algorithm (GA) 91.25 % Proposed Hybrid NB-GA Method (93.80%)

6 Conclusion

This comparative study paper conferred an outline on the current updates in sentiment analysis and its techniques. After analyzing the articles, it is apparent that applying sentiment analysis to excavate the vast quantity of data has become a significant research problem. Most of the techniques accorded excellent conclusion, but nothing resolved all the challenges. SVM and Naive Bayes have high precision than other algorithms. We can conclude that SVM has ascendancy with performance and accuracy, but still it is not suitable for imbalanced data sets.

Naïve Bayes is selected for less memory and preprocessing power requirements. More research on context-based sentiment is required. Using Natural Language processing tools in sentiment analysis has attracted researchers and still needs some enhancement. Hybrid techniques with improvement had shown good performance. The right selection of a classification model plays a vital role in sentiment analysis since the result influences the correctness of the system and the end product.

There is an immense need in the market for sentiment analysis tools and applications because every company wants the opinion of customers about their products and services for further enhancements and to compete with their opponents.

References

[1] Patil, Pratibha Yalagi, “Sentiment Analysis Levels and Techniques: A Survey’, International Journal of Innovations in Engineering and Technology (IJIET), pp.523-528, 2016.

[2] Steinberger, Tomas Hercig, Michal Konkol, ”Aspect-Level Sentiment Analysis in Czech”, Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis ,pp.365-375,2014.

[3] Joachims, “A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization” , ICML conference, pp. 143-151 , 1997.

[4] Annett,G.Kondrak, ”A comparison of sentiment analysis techniques : Polarizing movie blogs” , Canadian Conference , pp.25-35,2008.

[5] Luigi Di Caro, Matteo Grella ,”Sentiment analysis via dependency parsing “, CSI ,pp. 442–453 , 2013.

[6] Fermín L. Cruz, José A. Troyano, Fernando Enríquez, F. Javier Ortega, Carlos G. Vallejo,”Long autonomy or long delay?’ The importance of domain in opinion mining”,Expert Syst Appl, 40 (2013), pp. 3174–3184,2013.

[7] Nan Li ,Desheng Dash Wu,” Using text mining and sentiment analysis for online forums hotspot detection and forecast”, Decision Support Systems archive,Volume 48 Issue 2, Pages 354-368,January, 2010.

[8] K. Nirmala Devi ,and Dr. V. Murali Bhaskarn,” Online Forums Hotspot Prediction Based on Sentiment Analysis”, Journal of Computer Science 8(8),pp.1219-1224,2012 .

[9] Evandro Costa, Rafael Ferreira, Patrick Brito, Ig Ibert Bittencourt, Olavo Holanda, Aydano Machado,Tarsis Marinho “A framework for building web mining applications in the world of blogs: A case study in product sentiment analysis”,Expert Systems with Applications (39),pp.4813-4834,2012.

[10] Alvaro Ortigosa, José M. Martín, Rosa M. Carro,”Sentiment analysis in Facebook and its application to e-learning”,Computers in Human Behavior 31 ,pp.527-541,2014.

[11] Qiang Ye,Ziqiong Zhang,Rob Law,” Sentiment classification of online reviews to travel destinations by supervised machine learning approaches”, Expert Systems with Applications: Volume 36 Issue 3,pp.6527-6535,2009.

[12] Peter D. Turney,” Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”, (ACL), 40, July 2002, pp. 417-424, 2002.

[13] Zhang, Qiang Ye, Zili Zhang, Yijun Li, “Sentiment classification of Internet restaurant reviews written in Cantonese”, Expert Systems with Applications 38 (2011) ,pp.7674-7682,2011.

[14] Xue Zhang , Hauke Fuehres , Peter A. Gloor ,” Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear”, Procedia - Social and Behavioral Sciences Volume 26, 2011, pp. 55-62,2011.

[15] Kaiquan Xu , Stephen Shaoyi Liao , Jiexun Li, Yuxia Song, “Mining comparative opinions from customer reviews forCompetitive Intelligence”, Decision Support Systems 50 (2011),pp. 743–754,2011.

[16] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan,” Thumbs up? Sentiment classification using machine learning techniques”,Proceedings of EMNLP, pp. 79—86,2002.

[17] Kia Dashtipour,Soujanya Poria,Amir Hussain, Erik Cambria,Ahmad Y. A. Hawalah, Alexander Gelbukh, and Qiang Zhou6,” Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques”, Cognitive Computation, 8, pp.757–771,2016.

[18] V. Vapnik. “The Nature of Statistical Learning Theory.” , NY: Springer-Verlag,1995.

[19] S.Chandrakala, C.Sindhu, “Opinion Mining and Sentiment Classification: A Survey”, ICTACT Journal on Soft Computing, Oct 2012 Vol 3 Issue 1,pp.420-425,2012.

[20] M.Govindarajan, “Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm”, International Journal of Advanced Computer Research, Volume-3 Number-4 Issue-13 December-2013.

[21] Nikolaos Malandrakis, Abe Kazemzadeh, Alexandros Potamianos, Shrikanth Narayanan,"SAIL: A hybrid approach to sentiment analysis",Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 438-442,2013.