IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 53-57
www.iosrjournals.org

An Approach To Sentiment Analysis Using Lexicons With Comparative Analysis of Different Techniques

Tanvi Hardeniya[1], D. A. Borikar[2]
[1]*M. Tech. Student, Shri Ramdeobaba College of Engineering and Management Nagpur, India*
[2]*Assistant Professor, Shri Ramdeobaba College of Engineering and Management Nagpur, India*

Abstract *: The World Wide Web is growing at an astonishing rate. This has resulted in enormous increase in online communication. The online communication data consist of feedback, comments and reviews that are posted on internet by internet users. To analyze such opinionated data sentiment analysis is required. Sentiment analysis is a natural language processing technique which classifies the data into positive, negative and neutral. This paper proposes a framework for sentiment analysis using dictionary-based approach and brings out a comparative study on sentiment analysis techniques including machine learning technique and lexicon based technique. The comparisons are majorly drawn based on features such as preprocessing, technique employed, dictionary, datasets, and soft-computing approaches. An approach to sentiment analysis using dictionary-based approach incorporating fuzzy logic is proposed.*

Keywords : *Fuzzy Logic, Lexicon Based Technique, Machine learning, Natural Language Processing, Sentiment analysis.*

I. Introduction

The rapid growth of World Wide Web has resulted in substantial increase in use of social media. The social media data consist of comments, feedback and reviews. This data is of great consequence for business organization as well as for customers. Customer often refers to reviews of other customer before buying any product. The companies need feedback from their customers to adopt changes in their product or to take future decision and develop business strategy. To facilitate this communication between the customer and business organization, it is required to analyze this social media data. For analyzing social media data sentiment analysis is required. Sentiment analysis is a natural language processing task which is used to obtain customers feeling about several product and services which are posted on internet through various comments and reviews. Sentiment Analysis is used for text classification which classifies the text into positive, negative and neutral.

The major categories under which the sentiment analysis approaches fall include - Machine learning based techniques, Lexicon based techniques and the hybrid of these. In machine learning techniques various classification methods like Support Vector Machine (SVM), Naive Bayes (NB) and maximum entropy (ME) are used for sentiment classification. Machine learning methods maintain two datasets, namely the training dataset and the testing data set. Lexicon based approach can be further divided into Dictionary-based and Corpus-based. In Dictionary-based approach, firstly the opinion word from review text are found, which is followed by finding their synonyms and antonyms from dictionary. The dictionaries like WordNet, SentiWordNet, SenticNet may be incorporated for mapping and scoring. Corpus-based method helps to find opinion word in a context specific orientation. Beginning with a list of opinion word, the corpus-based approach finds other opinion word in a huge corpus. A hybrid approach combining the machine learning and the dictionary-based approaches may be used for sentiment analysis. It employs the lexicon-based approach for sentiment scoring followed by training a classifier assign polarity to the entities in the newly find reviews. Hybrid approach is generally used since it achieves the best of both worlds, high accuracy from a powerful supervised learning algorithm and stability from lexicon based approach [11].

The main issues in sentiment analysis are negation handling and domain dependency. Negation words are the words which reverses the polarity of sentence if occur in a sentence. Domain dependency is there because the word has positive orientation in one domain and the same word has negative orientation in differentdomain. It is most important to handle this issue for correct classification of reviews.

The fundamental process of sentiment analysis is often attributed to two stages - Opinion Extraction and Sentiment Classification. Opinion Extraction aims to extract opinion words from the target text, whereas Sentiment classification categorizes and ranks the opinionated text phrases based on polarity orientation. Different classification techniques are used for classification.

This paper is organized as follows. Section 2 discusses the literature review done in sentiment analysis. Section 3 describes the comparative analysis of sentiment analysis technique. In section 4 a detailed proposed approach to sentiment analysis using dictionary-based approach has been deliberated. Section 5 concludes the discussion in earlier sections.

II. literature review

Gonçalves and Araujo have explained different methods of sentiment analysis in their work. They are as below

- Emoticons are face-based expression represent happy or sad feelings. To calculate polarity of emoticons a set of three common emoticons is used.
- Linguistic Inquiry and Word Count (LIWC) is software which uses dictionary to calculate the polarity and also find the related word.
- SentiStrength it is based on machine learning technique. It added new features to LIWC like a list of negative and positive words, a list of booster words to strengthen or weaken sentiments, a list of emoticons, and the use of repeated punctuation to strengthen sentiments.
- SentiWordNet is based on WordNet it associate three sentiment score to the word positive, negative, objective. This score is calculated by a semi-supervised method.
- SenticNet it implement artificial intelligence along with semantic web technique. It calculate the polarity of common sense concepts from natural language text at a semantic level not at a syntactic level.
- A machine learning-based tool called the SailAilSentiment Analyzer (SASA) is developed. It is based on SentiStrength.
- Happiness Index uses the popular Affective Norms for English Words (ANEW). The score to the text is ranges from 1 to 9.
- PANAS-t is a psychometric scale proposed for detecting mood fluctuations of users on Twitter. The method consists of an adapted version of the Positive Affect Negative Affect Scale (PANAS), which is a well-known method in psychology. The PANAS-t is based on a large set of words associated with eleven moods: joviality, assurance, serenity, surprise, fear, sadness, guilt, hostility, shyness, fatigue, and attentiveness. The method is designed to track any increase or decrease in sentiments over time [12].

Wang, Chuan-Tong and Chan have performed analysis on non-English data language and integrated with analysis of data in English language to improve the sentiment understanding and sentiment analysis. The approach adopted a Fuzzy Inference Method with Linguistic Processor to minimize semantic ambiguity and multi source lexicon integration and development. This approach used LIWC method, the pleasure, arousal and dominance (PAD) model of emotional states and the affective norms for English words (ANEW) approach. It handled issue like complexity and semantic ambiguity of languages, the requirement domain-dependent adaptive methods to obtain high accuracy readings, dependence of training dataset and to handle the data in other languages besides English [13].

Jose and Chooralil have used WordNet, SentiWordNet lexical resource and Word Sense Disambiguation for finding political sentiment from real time tweets. A lexicon based sentiment analysis method is adopted which exploits the sense definitions, as semantic indicators of sentiment. Negation handling is done at pre-processing stage to increase the accuracy [14].

Mikula and Machova have accomplished Topic identification as a mechanism to increase the accuracy of system. Topic identification is done by increasing the weight of sentences containing the word related to the topic. A topic lexicon is constructed which contain all topic extracted from the datasets. An algorithm is designed that works in three steps. In the first step, algorithm obtains the text that is to be analyzed removing diacritics, changes all letters to lowercase and then removes all stop words. Then it creates list of topic words and removes words with polarity. This is followed by sentiment analysis of text and then the algorithm writes results into the file [15].

III. comparative study on sentiment analysis technique

Sentiment analysis and classification methods can be compared on the basis of different aspects. The major aspects used are as below:

3.1 Technique(s) Used

As there are three methods supervised, unsupervised and semi-supervised method which uses technique like SVM, Naive Bayes (NB) and maximum entropy (ME) and lexicon based technique. Some method also use fuzzy logic and rule based technique to get more accurate result.

3.2 Use of Lexicon/Dictionary

Next aspect is if the approach uses the lexicon based method then which lexicon is used like SentiWordNet or WordNet, SenticNet or other. Some approaches used machine learning method also apply dictionary for more accuracy.

3.3 Training dataset

Supervised techniques require training datasets so it is another aspect. Training datasets is required to train the classifier and this classifier is then applied for classification.

Some other aspects are of preprocessing step as some method requires stop word removal and some does not require stop word removal. Feature extraction is the important step in sentiment analysis which is performed by maximum methods.

Following table show the comparative analysis of some dictionary based approaches.

Table:1 Comparative Study Of Sentiment Analysis Technique

SrNo	Author	Technique used	Lexicon used	Training datasets	Stopword removed	Feature extraction
1	AndreeaSalinca (2016) [1]	SVM, Naïve Bayes, Logistic Regression,SGD	SentiWordNet	Yes	Yes	Yes
2	Ghag and Shah (2015) [2]	TSC, ARTFSC, SentiTFIDF, RelativeTFSC	No	Yes	No	No
3	K. Indhuja, Reghu Raj P. C. (2014) [3]	Tree Bank Model, Fuzzy Opinion Mining Model	FOLH	No	No	Yes
4	Vipin Kumar, S. Minz(2013)[4]	NB, KNN, SVM	SentiWordNet	No	Yes	Yes
5	A. Yeole, P. Chavan (2015) [5]	Affective words and sentence context analysis methods	SentiWordNet	No	Yes	Yes
6	Y. Wang, Baoxin Li (2015) [6]	RSAI, USEA	MPQA	Yes	No	Yes
7	A. Cernian, V.Sgarciu, B. Martin (2015) [7]	Lexicon based method	SentiWordNet	No	No	Yes
8	D. Yuan, Yanquan Zhou(2014) [8]	SVM	HowNet	Yes	No	Yes
9	Lizhen Liu, XinhuiNie (2012) [9]	FDSOT	No	No	Yes	Yes
10	AleksanderWawer(2015) [10]	CRF algorithm	Domain Independent	No	No	Yes

Andreea Salinca has proposed a technique using four learning models: Multinomial Naïve Bayes, Support vector machines, Linear Support Vector Classification, Logistic regression and Stochastic Gradient Descent Classifier for sentiment analysis. Classification is performed on Yelp challenge dataset [1]. Ghag and Shah uses four sentiment classifier that are traditional sentiment classifier, average relative term frequency sentiment classifier, senti-term frequency inverse document frequency and relative term frequency sentiment classifier. This classifier is applied on datasets in which stop words are removed and also without removing stop word [2].

K. Indhuja, Reghu Raj P. C. have performed identification of opinion phrases from text, dependency between words and tagging the sentiment polarity of opinionated phrases. Stanford dependency parser is applied to find the dependency between words. The fine features are checked with the Feature Orientation dictionary with Linguistic Hedges to extract its fuzzy value [3]. Vipin Kumar, Sonajharia Minz method used SentiWordNet for feature extraction and various classifiers are used to classify the text. SVM suited the best and give highest accuracy [4]. A. Yeole, P. Chavan done sentence context analysis and affective word is found. SentiWordNet dictionary is used for sentiment calculation [5].

Yilin Wang, Baoxin Li applied sentiment analysis on image based on image feature and contextual social network information. Both visual feature and textual feature are used and find prediction on two scenarios supervised and unsupervised. By using supervised sentiment analysis it proposed an effective method name Robust Sentiment Analysis for Images. For unsupervised sentiment analysis Unsupervised E-Sentiment Analysis is used [6].

A. Cernian, V. Sgarciu, B. Martin presented a semantic approach for a sentiment analysis which is found using the SentiWordNet lexical resource [7]. Ding Yuan, Yanquan Zhou used dictionary based approach. Rule based approach is applied for feature extraction and weight computing and classification is done using SVM [8]. Lizhen Liu, Xinhui Nie described an approach using Fuzzy Domain Sentiment Ontology Tree model which is constructed by a set of seeds based on Synonyms set and domain sensitive words [9]. The last method by Aleksander Wawer used a domain-independent sentiment dictionary this is applied with a machine learning method based on CRF algorithm [10].

IV. proposed approach

The proposed approach mainly deals with classifying the reviews as positive, negative and neutral based on score that is calculated using SentiWordNet and WordNet dictionary and by applying some fuzzy logic to handle negations. The system consists of mainly three modules.

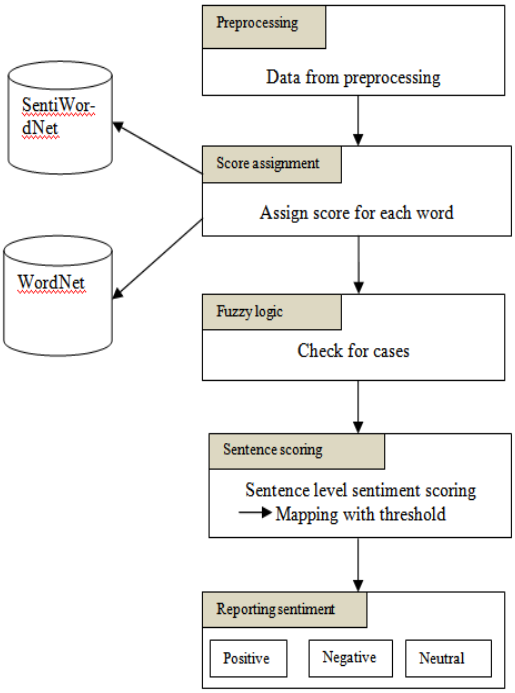

Fig. 1. Framework of sentiment analysis using dictionary based approach.

4.1 Dataset

The datasets used is an Amazon dataset taken from web. The reviews are about mobile phone and accessories. It is a text dataset which is labeled. It contain product/product Id,product/title, product/price, review/user Id, review/profile Name, review/time, review/helpfulness, review/score, review/summary, review/text. There are total 7150 entries.

The main focus of the method is to extract the text field from the different field of the datasets.

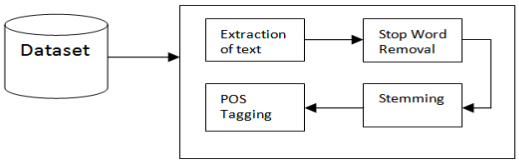

Fig. 2. Preprocessing of datasets

4.2 Preprocessing

Preprocessing is done before sentiment polarity calculation. Preprocessing is required to remove the words which are not useful in polarity calculation and process the words of sentences to match with the dictionary. Following are the steps of preprocessing.

i. The text field from the reviews of the datasets is extracted by matching the index with the product name and then the text in front of text filed is extracted.
ii. The stop words are removed. Stop word are the most common words used in English language. This words are filter out in preprocessing since it is of no use in sentiment polarity calculation. We are using a dictionary of stop words to remove stop words from reviews.
iii. Stemming is a process for reducing derived form of word into base or root form. We are using a dictionary for stemming. It contains 200 root words.
iv. Part of speech tagging is done in the next step. It assigns parts of speech such as noun, verb, adjective, adverb to each word of the text. The information obtained from POS tagging is then used as features to find the emotion information from sentence. The standard Penn Treebank POS tags is used.

4.3 Polarity calculation

Review contain many sentences as a result the score for each sentence is calculated first then the total score for the review is calculated. SentiWordNet dictionary is used for assigning the polarity to each word and then the polarity of whole sentence is calculated by adding the polarity of each word. SentiWordNet is a lexical resource publicly available for research purposes. SentiWordNet is an opinion lexicon derived from the WordNet database where each term is associated with numerical scores indicating positive, negative and

objective sentiment information. SentiWordNet is built via a semi supervised method along with a Random Walk algorithm for refining the score. In this the position of word in the sentence is also refer.

If the word does not found in SentiWordNet dictionary then the word is search in the WordNet dictionary. WordNet is a dictionary for English language it contain synonyms word into a set called synset. The corresponding words associated with the word in WordNet are bring and search in the SentiWordNet and their sentiment score is taken for polarity calculation. This process helps to increase the accuracy of the proposed approach.

Negation words are the words which when present in the sentence reverses the polarity of the sentence. For example, in the text "this smart phone is not good", the negation word "not" reverses the polarity of sentence. To handle this fuzzy logic is used which calculate the polarity based on some rules. The fuzzy logic rules are described as follow.

Case 1. There are few adverbs like very, really, extremely, simply, always, never, not, absolutely, highly, overall, truly, too, etc. which may be used positively or negatively like very good, very bad.

$$Weight = (Value\ of\ (Adj))^{0.5} \quad if\ value\ of\ (Adj) >= 0.5$$
$$= (Value\ of\ (Adj))^{2} \quad if\ value\ of\ (Adj) < 0.5$$

Case 2. Never, not etc. changes the orientation of the opinion. The phrase like "not good" may signify "bad" (although it does not always qualify to be "bad").

Weight = 1 – Value_Of_(Adj or Verb)

Case 3. When Case 1 and Case 2 may appear together like "not very good"

Weight = $(A*B)^{0.5}$

Where A = very/extremely/highly etc (Adj)

And B = (not/never) (Adj)

This fuzzy logic score is calculated and added to the score of sentence and the score for a review is calculated. The threshold value for review classification is set as 0.2. For all the reviews of a product the analysis is done that is a product is positive or negative or neutral based on a threshold value which is set to 0.5.

V. Conclusion

An approach to sentiment analysis using dictionaries is proposed. The proposed work uses reviews from Amazon data about mobile phones and accessories for analysis and classification. The approach incorporates SentiWordNet and WordNet to find the proper word from the dictionary and assign sentiment polarity. Negation handling has been a challenging task in sentiment analysis. We have used fuzzy logic to address negation. The paper has also attempted to bring out comparisons and salient features between various techniques and approaches realized for sentiment analysis and classification.

Sentiment analysis support business organizations and customer for analyzing their reviews. Consequently the proposed system also helps as it classifies the reviews of customer into positive, negative and neutral. As a part of future work, it is planned to incorporate additional fuzzy logic cases for sentences which are compound and complicated. This addition would benefit in more accurate classification of reviews.

References

[1] Salinca A., Business Reviews Classification Using Sentiment Analysis, 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) IEEE, Sep 201, pp. 247-250.

[2] Ghag, Kranti Vithal, and Ketan Shah., Comparative analysis of effect of stop words removal on sentiment classification, Proc. IEEE Conf. In Computer, Communication and Control (IC4), 2015, pp. 1-6.

[3] Indhuja, K. and Raj PC Reghu., Fuzzy logic based sentiment analysis of product review documents, In Proc. Computational Systems and Communications (ICCSC), 2014 First International Conference on IEEE, 2014, pp. 18-22.

[4] Kumar, Vipin, and Sonajharia Minz, Mood classification of lyrics using SentiWordNet, Proc. IEEE Conf. in Computer Communication and Informatics (ICCCI), 2013, pp. 1-5.

[5] Yeole, Ashwini V., P. V. Chavan, and M. C. Nikose, Opinion mining for emotions determination, Proc. IEEE Conf. on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015, pp. 1-5.

[6] Wang, Yilin, and Baoxin Li, Sentiment Analysis for Social Media Images, Proc. IEEE Conf. on Data Mining Workshop (ICDMW), 2015, pp. 1584-1591.

[7] Cernian, Alexandra, Valentin Sgarciu, and Bogdan Martin, Sentiment analysis from product reviews using SentiWordNet as lexical resource, Proc. 7th IEEE Conf. on Electronics, Computers and Artificial Intelligence (ECAI), 2015, pp. WE-15.

[8] Yuan, Ding, Yanquan Zhou, Ruifan Li, and Peng Lu, Sentiment analysis of microblog combining dictionary and rules, Proc. IEEE Conf. in Advances in Social Networks Analysis and Mining (ASONAM), 2014, pp. 785-789.

[9] Liu, Lizhen, Xinhui Nie, and Hanshi Wang, toward a fuzzy domain sentiment ontology tree for sentiment analysis, Proc. 5th IEEE Conf. in Image and Signal Processing (CISP), 2012, pp. 1620-1624.

[10] Aeksander Wawer, Towards Domain-Independent Opinion Target Extraction, Proc. IEEE Conf. 15th International Conference on Data Mining Workshops, 2015.

[11] T. Hardeniya and D. Borikar, Dictionary Based Approach to Sentiment Analysis – A Review, Proc. Conf. on national conference on resent trends in computer science and information technology , 2016.

[12] Gonçalves, Pollyanna, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha, Comparing and combining sentiment analysis methods, In Proceedings of the first ACM conference on Online social networks, 2013, pp. 27-38.

[13] Wang, Zhaoxia, Victor Joo, Chuan Tong, and David Chan, Issues of social data analytics with a new method for sentiment analysis of social media data, Proc. 6th IEEE Conf. in Cloud Computing Technology and Science (CloudCom), 2014, pp. 899-904.

[14] Rincy Jose, Varghese S Chooralil, Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation, Proc. IEEE Conf. on International Conference on Control, Communication & Computing India (ICCC), November 2015, 19-21.

[15] Mikula, Martin, and K. Machov, The use of topic identification in opinion classification, In 2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI), pp. 275-278, 2016.