Dictionary Based Approach to Sentiment Analysis - A Review

Tanvi Hardeniya [1], Dilipkumar A. Borikar [2]

[1] M.Tech Student, CSE Department , Shri Ramdeobaba College of Engineering and Management Nagpur, India
[2]Assistant Professor, CSE Department, Shri Ramdeobaba College of Engineering and Management Nagpur, India

Abstract— *Due to the fast growth of World Wide Web the online communication has increased. In recent times the communication focus has shifted to social networking. In order to enhance the text methods of communication such as tweets, blogs and chats, it is necessary to examine the emotion of user by studying the input text. Online reviews are posted by customers for the products and services on offer at a website portal. This has provided impetus to substantial growth of online purchasing making opinion analysis a vital factor for business development. To analyze such text and reviews sentiment analysis is used. Sentiment analysis is a sub domain of Natural Language Processing which acquires writer's feelings about several products which are placed on the internet through various comments or posts. It is used to find the opinion or response of the user. Opinion may be positive, negative or neutral. In this paper a review on sentiment analysis is done and the challenges and issues involved in the process are discussed. The approaches to sentiment analysis using dictionaries such as SenticNet, SentiFul, SentiWordNet, and WordNet are studied. Dictionary-based approaches are efficient over a domain of study. Although a generalized dictionary like WordNet may be used, the accuracy of the classifier get affected due to issues like negation, synonyms, sarcasm, etc.*

Keywords— Sentiment Analysis, Natural Language Processing, SenticNet, SentiFul, SentiWordNet, WordNet.

I. INTRODUCTION

Human decision making or thinking is always affected by others thinking, ideas and opinions. The growth of social web gives a huge amount of user generated data such as comments, opinions and reviews about products, services and events. This data will be useful for consumers as well as manufacturer. While buying any product online consumers usually check comments or opinion of others about the product. Manufacturer can understand the response of that product and get insight into its products strength and weaknesses based on the sentiment of the customers. These opinions are helpful for both business organizations and individuals but the huge amount of such opinionated text data becomes burden to users. To analyze and summarize the opinions expressed in these enormous opinionated text data is a very interesting domain for researchers. This new research domain is typically called Sentiment Analysis or Opinion Mining.

Sentiment analysis is used to automatically mine the opinions and emotions from text, speech, and database sources with the help of Natural Language Processing (NLP). Sentiment analysis does the classification of opinions in the text into categories like “positive” or “negative” or “neutral”. It’s often referred to as subjectivity analysis, opinion mining and appraisal extraction.

For buying any product customer wants to see the opinion of other about the product. Customer reviews are very important for business process since to make future decision business organizations should know what customers are saying about their product or service that an organization is providing. It will provide important functionality for voice of customer and brand reputation management. Thus it is helpful in business process and also for the customers

Major areas of research in Sentiment analysis are Subjectivity Detection, Sentiment Prediction, Aspect Based Sentiment Summarization, and Text Summarization for Opinions, Contrastive Viewpoint Summarization, Product Feature Extraction, and Detecting Opinion Spam. Subjectivity Detection is a task of finding whether text is opinionated or not. Sentiment Prediction is about predicting the polarity of text whether it is positive or negative. Aspect Based Sentiment Summarization generates sentiment summary in the form of star ratings or scores of features of the product. Text Summarization generates a few sentences that summarize the reviews of a product. Contrastive Viewpoint Summarization puts an emphasis on contradicting opinions. Product Feature Extraction is a work that extract product feature from its review. Detecting Opinion Spam is concern with identifying fake or bogus opinion from reviews [5].

Sentiment classification is carried out at three levels Document level, Sentence level and Aspect or feature level. In Document level the task is to classify complete document into positive or negative class. Sentence level sentiment classification classifies sentence into positive, negative, neutral class based on each sentence level. First the polarity of each word of a sentence is calculated and then the overall sentiment of the sentence is calculated. Aspect or Feature level sentiment classification identifies and extract product features from the source data and do the classification [1].

Sentiment analysis is generally carried out by two approaches: machine learning based and dictionary based. Machine learning based approach applies classification technique to classify text such as support vector machine or neural network. Dictionary-based method uses sentiment dictionary with opinion words and match them with the data to determine polarity. They assign sentiment scores to the opinion words describing the Positive, Negative and Objective score of the words contained in the dictionary.

Sentiment Analysis is generally done in two phase subjectivity detection and polarity assignment. Firstly the subject towards which the sentiment is directed is found called subjectivity detection then, the polarity is assign using dictionary such as SentiWordNet, WordNet, SenticNet, SentiFul and others.

The issues in sentiment analysis are:

1. In different domain a positive or negative sentiment word may have opposite orientations. For example, "frozen" generally indicates negative sentiment for software engineering but it can also imply positive sentiment in air conditioning and refrigeration.
2. A sentence which contains sentiment words may not imply any sentiment. This phenomenon occurs frequently in several types of sentences. Question (interrogative) sentences and conditional sentences fall into this category. For example, "Can you tell me which smart phone is good?"This sentence contain the sentiment word "good" but does not state a positive or negative opinion about the smart phone.
3. Many sentences which do not contain any sentiment words can also express opinions. These sentences are called objective sentences which are used to express some factual information.
4. For example "This washer uses a lot of water", implies a negative sentiment about the washer as it uses a lot of resource (water).

The natural language processing issues are:

1. Negation words are the words which reverse the polarity of sentence. These are dealt with under negation handling. For example, in the text "this smart phone is not good", the negation word "not" reverses the polarity of sentence.
2. Word sense disambiguation is the lexicon ambiguity that may be syntactic or semantic. It refers to the words with more than one meaning that completely different. For example, "I love this movie" and "This is the love movie". In this the word "love" has different meanings.

This paper is organized as follows. Section 2 discusses the current perspectives in sentiment analysis. Section 3 describes the main approaches to sentiment analysis. In section 4 a detailed view of sentiment analysis using dictionary-based approach has been deliberated. Section 5 covers the literature review on sentiment analysis. Section 6 concludes the discussion in earlier sections.

II. SENTIMENT ANALYSIS– CURRENT PERSPECTIVE

Since form past few years sentiment analysis was done in English language only. As many social networking sites allow different languages for communication web content are increased in a faster rate for other language also and hence it is necessary to do sentiment analysis for other languages. Recently Hindi opinion mining has been carried out using two methods Machine learning method uses Naïve-Bayes classifier. Part-Of-Speech tagging finds adjective and consider it as opinion word and based on their count document is classified. It uses TnT() POS tagger and extract the adjective. The overall classification accuracy is 87.1% [16]. Different language dictionaries were also created for sentiment analysis. In recent times sentiment analysis has been carried out in many languages other than English. Sentiment analysis is not only done on text data but also on visual images. For extracting textual information embedded on images text mining techniques incorporating on sentiment analysis will be useful [17].Personality-based sentiment analysis is performed as personality affect the ways people write and talk and it is important for government and public agencies to analyze the information propagating in social networks. It provides higher accuracy value for both positive and negative tweets than the baseline and ensemble learning method [18].

An automated construction of the terminological thesaurus for a specific domain is done. It uses explanatory dictionary as the initial text corpus and a controlled vocabulary related to the target lexicon to begin extraction of the terms for the thesaurus. Sub-division of the terms into semantic clusters is built on the CLOPE clustering algorithm. It diminishes the cost of the thesaurus creation by involving the expert only once during the whole construction process, and only for analysis of a little subset of the initial dictionary [19].

The values of precision and coverage depend on the technique to build a lexicon. SentiWords a prior polarity lexicon that is of approximately 155,000 words is constructed that has both high precision and a high coverage. It uses the experience of automatic derivation of prior polarities from the SentiWordNet resource and a collection of learning framework that take advantage of manually built lexica [20].

III. SENTIMENT ANALYSIS– APPROACHES

Sentiment analysis approaches may be classified into – machine learning approaches, lexicon-based approaches and the hybrid approaches.

3.1 Machine Learning Approach

Machine learning approaches include Support Vector Machine (SVM) and Naïve Bayesian classification. SVM is a supervised learning method used to analyze the data and recognize data patterns that can be used for classification and regression analysis. Naïve Bayesian Classification is based on Naïve-Bayes theorem and uses the concepts of maximum likelihood and Bayesian probability. The limitation of this method is that the model needs to be trained with a large data volume before testing. It is time consuming and low on accuracy when training data is not sufficient.

3.2 Lexicon Based Approach

These approaches can be divided into two methods– Dictionary based method, first finds the opinion word from review text then finds their synonyms and antonyms from dictionary. The dictionary used may be WordNet or SentiWordNet or other. Corpus-based method helps to find opinion word in a context specific orientation start with a list of opinion word and then find other opinion word in a huge corpus.

SentiWordNet 3.0 is most useful dictionary used. It is a lexical resources publically available made up of "synsets" each is associated with a positive, negative numerical score range from 0 to 1. This score is automatically allotted from the WordNet. It uses a semi-supervised learning method and an iterative random walk algorithm [6].

3.3 Hybrid Approach

It uses both the machine learning and the dictionary-based approaches. 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 [5].

IV. LEXICON–BASED APPROACHES IN DETAIL

There are many lexicons available for sentiment analysis such as SentiWordNet, WordNet, SentiTFIDF, SentiFul, SenticNet, etc.

SentiTFIDF is based on proportional frequency count distribution and proportional presence count distribution across positively tagged document and negatively tagged document and classify the term as positive or negative. The term with equal proportion in positively tagged document and negatively tagged document were classified as a SentiStopWord and discarded. The process is completed in three parts–calculating positivity of a term, calculating negativity of the term, and classification of the term based on its proportion of positivity and negativity. SentiTFIDF has achieved accuracy of 92% [9].

SenticNet is a publicly available resource for opinion mining construct based on Artificial Intelligence and Semantic Web techniques. Dimensionality reduction is used to deduce the polarity of common sense concepts and hence provide sentiment analysis at a semantic level rather than merely at syntactic level. It uses techniques such as blending and spectral activation with emotion categorization model and ontology for describing human emotions. SenticNet is much more accurate than SentiWordNet [10].

SentiFul database comprises of a reliable lexicon of sentiment conveying terms, modifiers, functional word, and modal operators. It gives a strong analysis on orientation and strength of sentiment text. It differentiates four types of affixes based on the role they play with regards to sentiment features: propagating, reversing, identifying and weakening. The sentiment conveying words are found out through synonymy antonym and hyponymy relations, derivation and compounding. It helps to expand sentiment lexicon and improve coverage of sentiment analysis [21].

Sentiment analysis using dictionary based approach is described in Fig. 01 and may be performed as:

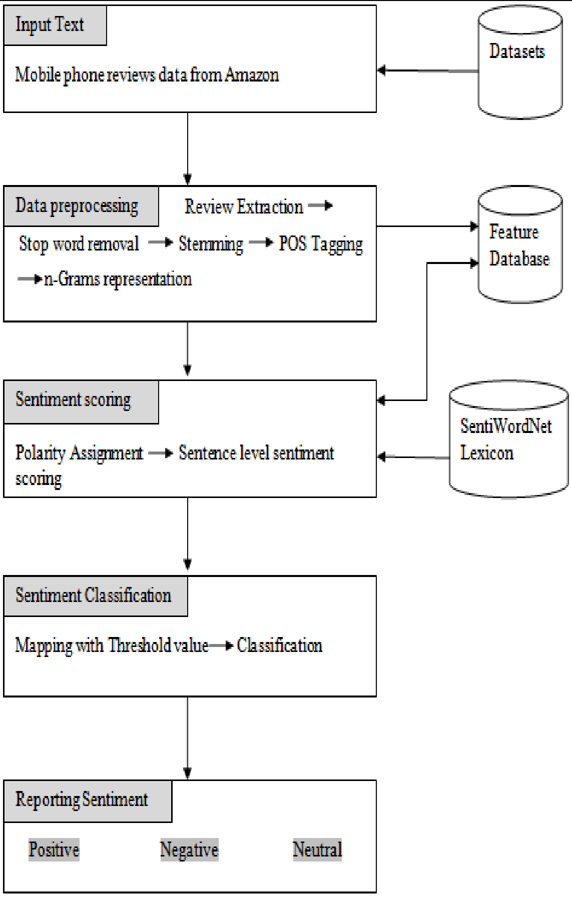

Fig. 1: A Framework for Sentiment Classification.

Initially input data is taken from the datasets then data preprocessing is done. In this preprocessing extracts the reviews and stop words are remove after that stemming is done to convert the derived form of the word into root form. Part-of-Speech tagging is done to assign the speech to each word of the review so as to concentrate on the adjectives, verb and adverbs. These words of review are represented using n-grams. This representation is stored in a database for sentiment polarity calculation. The features from the database are retrieved and the sentiment polarity is calculated using sentiment analysis technique i.e. dictionary based technique. SentiWordNet dictionary may be used for assigning the polarity to each word and then the polarity of whole sentence is calculated. A pre-defined threshold is used to compare the score value and if it exceeds the threshold value it is classified as positive otherwise negative. In situation when the difference is zero, the text is classified to convey a neutral sentiment.

V. LITERATURE REVIEW

Emotion estimation is done using affective words and sentence context analysis methods. It generates images according to emotions in the text. For sentences that do not contain the emotion words the emotion feature estimation is done using sentence context analysis. This is done by extracting the hidden knowledge of the sentence by using Natural Language Processing Technique. Emotion lexicon dictionary is used for extracting the emotion feature that is made up of WordNet-Affect database, WordNet 1.6 and SentiWordNet. The approach is applied to two types of sentences – 1) Sentence with direct emotional words for e.g. amazing, annoying.2) Sentences with no emotion words but those contributing for the emotion of writer [2].

SentiWordNet dictionary and smiley dictionary is utilized to score the sentiment into positive, negative and objective. Rule based and fuzzy logic approach is used to handle negation words. Fuzzy Intensity Finder Algorithm is used to find intensity of each word [3].

Sentiment analysis is carried out on Big Data as social media available in internet is very vast. Hadoop is used to perform sentiment analysis on large datasets and performance of the system is measured. Lexicon based technique is more appropriate then machine learning algorithms is concluded for Big Data. Stemming is not done in pre-processing as the dictionary is used which contains all form of words. Negation and blind negation words are found using dictionary and its polarity is reversed. High speed for analysis of big data is present but the accuracy is not too much [4].

The word which reverses the sentiment polarity of other word is handled by a dependency tree based method for sentiment classification using conditional random field with hidden variables and analyzing interaction between words [7].

A linguistic tree transform algorithm is used to eliminate the word sense disambiguation and non-local dependency. An objective sentence removal algorithm is used to account the objective sentence. It performs better than n-gram method [8].

Proportional frequency count distribution and proportional presence count distribution are used for sentiment analysis. SentiTFIDF used logarithmic proportion of TFIDF of a term for positively tagged documents and negatively tagged documents. If the TFIDF of a term in positively tagged documents is larger than TFIDF of same term in negatively tagged documents the term is assigned positive polarity and vice-versa. SentiTFIDF was more accurate than Delta TFIDF [9].

A common sense learning tool, SenticNet is a publicly available semantic recourse for opinion mining built using common sense reasoning technique together with an emotion categorization model and an anthology for describing human emotions. SenticNet is constructed using ConceptNet and AffectiveSpace. SentiWordNet is also incorporated. It is more superior to currently available lexicon recourses [10].

Classification can be done on syntactic patterns and part-of-speech tagging focusing on aspect level analysis using SentiWordNet. This unsupervised method is a domain independent method. SentiWordNet is applied using in two phases. First phase SWN (AAC), considers “Adverb+Adjective” combination and the other phase SWN (AAAVC) considers “Adverb+Adjective”, “Adverb+Verb” combinations. The method is applied for finding out of sentiment polarity of all aspects in one review [11].

A domain-independent lexicon based on Latent Dirichlet Allocation for sentiment analysis is constructed. LDA is a probabilistic model to construct a lexicon. The lexicon constructed is highly related to the dataset. Precision of this lexicon is more than the Liu’s lexicon, MPQA and GI. This method is better than trivial methods in all aspects as trivial approach builds the lexicon based on calculating the words appearing number of times in positive reviews and in negative reviews. [12].

Sentiments of product or services are different from social issues sentiment. Verb plays an important role in analysis of sentiments of social issue. The two main approaches is used for sentiment analysis i.e. bag-of-words and feature-based sentiment. A dictionary is constructed which contain opinion verb, it has 440 opinion verbs. Some Algorithms is given that extract opinions, construct corresponding opinion structures, and at last calculate the sentiment of opinion structures regarding the social issue [13].

The different machine learning techniques and lexicon based techniques, their limitations and the current problem that researcher has studied in their work are domain dependency; sentiment classification based on insufficient labeled data; the lack of SA research in languages other than English; and to deal with complex sentences that requires more than sentiment words and simple parsing [14].

A hybrid approach is mainly used for sentiment analysis. It uses sentiment lexicon for polarity detection and the results from the sentiment lexicon method are then use by machine learning algorithms to train the data. It is accomplishes through three phases, namely - crawler, lexicon and machine learning module. It uses the AFFIN-111 word list developed by FinnArup Nielsen. Feature weighting method is applied since some of the features had less impact on classification. Classification is done by two algorithms Support Vector Machine (SVM) and K-Nearest Neighbor (k-NN). It is observed than SVM outperforms k-NN method [15].

VI. CONCLUSION

From the pioneering contributions to the domain of NLP, IR and in specific to sentiment analysis it is observed that sentiment analysis can play a very important role in the development of successful business. Sentiment analysis can be carried out in different languages and may extend to other areas such as image processing, data aggregation, etc. There are several dictionaries available for sentiment analysis, of which SentiWordNet is used more often. This paper reviews approaches, issues and challenges involved in sentiment analysis and classification. Three types of approaches are described with their relative merits and limitations. It is found that sentiment analysis using dictionary-based approach is swift compared to machine learning-based approach since it require no prior training. A framework of sentiment classification is explained describing the main steps in sentiment analysis using dictionary-based approach.

REFERENCES

[1] Varghese,R., Jayasree, M., “A Survey on Sentiment Analysis and Opinion Mining,” IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11, October 2013.

[2] Patil, S. and T. Patil, “Automatic Generation of Emotions for Social Networking Websites using Text Mining,” 4th International Conference on Computing, Communication And Networking Technologies (ICCCNT) IEEE, 4-6, July 2013.

[3] Pimpalkar, A., Wandhe, T., Swati Rao M., Kene M., “Review of Online Product using Rule Based and Fuzzy Logic with Smileys,” IJCAT – International Journal of Computing and Technology, Volume 1, Issue 1, February 2014.

[4] Kaushik, C.and Mishra A., “A Scalable, Lexicon Based Technique for Sentiment Analysis,” International Journal in Foundations of Computer Science and Technology (IJFCST), Vol.4, No.5, September 2014.

[5] Vohra, S. M. and Teraiya, J. B., “A Comparative Study of Sentiment Analysis Techniques,” Journal of Information, Knowledge and research in Computer Engineering Volume – 02, Issue – 02, October 2013.

[6] Baccianella, S., Esuli, A. and F. Sebastiani, “SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining,” The International Conference on Language Resources and Evaluation (LREC). Vol. 10, 2010.

[7] Nakagawa, T., S. Kurohashi, “Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables,” Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

[8] B Eriksson, "Sentiment Classification of Movie Reviews Using Linguistic Parsing. Natural Language Processing," CS, Volume – 838, 2006.

[9] Ghag, K., K. Shah, "SentiTFIDF – Sentiment Classification using Relative Term Frequency Inverse Document Frequency," International Journal of Advance Computer Science and Application, Volume 5, No. 2, 2014.

[10] Cambria, E., Speer, R.Havasi, C. and A. Hussain, "SenticNet: A Publicly Available Semantic Resource for Opinion Mining," In AAAI fall symposiums: commonsense knowledge, vol. 10, p. 02, March 2010.

[11] Soni, V., Patel M. R., "Unsupervised Opinion Mining from Text Reviews Using SentiWordNet," International Journal of Computer Trends and Technology, volume 11 number 5, May 2014.

[12] Li, C. and R. Li, "Lexicon Construction: A Topic Model Approach," In International Conference on Systems and Informatics (ICSAI), pp. 2299-2303. IEEE, 2012.

[13] Karamibekr, M. and Ghorbani, A., "Sentiment Analysis of Social Issues," International Conference on Social Informatics (Social Informatics), IEEE 2012, December 2012.

[14] Madhoushi, Z.,Hamdan, A. R. and Zainudin, "Sentiment Analysis Techniques in Recent Works," Science and Information Conference (SAI), 2015. IEEE.

[15] Mukwazvure, A., K.P. Supreethi, "A Hybrid Approach to Sentiment Analysis of News Comments. Reliability," In 4th International Conference on Infocom Technologies and Optimization (ICRITO)-Trends and Future Directions, IEEE, 2015.

[16] Jha, V.,Manjunath, N., Shenoy, V.,Venugopal, K. R. andPatnaik, L., "HOMS: Hindi Opinion Mining System," IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), M. 2015.

[17] Giannakopoulos, Theodoros, et al. "Visual Sentiment Analysis For Brand Monitoring Enhancement," 9th International Symposium on Image and Signal Processing and Analysis (ISPA), IEEE, 2015.

[18] Lin, Junjie, and M. Wenji, "Personality Based Public Sentiment Classification In Micro-blog," IEEE International Conference on Intelligence and Security Informatics (ISI), IEEE, 2015, pp. 151-153, M. 2015.

[19] Lagutina, Nadezhda, "An Approach to Automated Thesaurus Construction Using Clusterization Based-Dictionary Analysis," 17TH Conference of Open Innovations Association (FRUCT), IEEE, 2015.

[20] Gatti, L.,Guerini, M. and M. Turchi, "SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis," Affective Computing, IEEE Transactions, 2015.

[21] Neviarouskaya, A., Prendinger, H. and Ishizuka, M. "SentiFul: A Lexicon for Sentiment Analysis," IEEE transactions on affective computing, vol. 2, no. 1, January-March 2011.