COMPARATIVE ANALYSIS OF SENTIMENT ANALYSIS USING LEXICON BASED APPROACH

[1]Monali V.Waghmare, [2]Rajkumar S. Jagdale,[3]Vishal S. Shirsat
[1,2,3] Department of CS and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad.

Abstract*—Individuals take help from friends, family member, or of any relatives in their decision making process. They depend on their opinions or suggestions, while business rely on surveys, focus groups, exit polls and counselors. This paper presents comprehensive description about lexicon based approach for the need of sentiment analysis. Sentiment analysis is the way to analyze written or spoken language to determine if the expression is positive negative or neutral and to what degree it is. Sentiment Lexicon is a process which gives rise to sentiments within the large collection of textual data.The lexicon-based approach determines sentiment polarity using semantic orientation of words, sentences or document for review. A comprehensive, good quality lexicon is extremely important for fast, accurate sentiment analysis.The major focus of paper is to provide the information of lexicon approach which can be useful for research in the sentiment analysis.

IndexTerms*—Lexicon Based Approach, Machine learning Approach, Resources for lexicon based SA, Applications of sentiment analysis.

I. INTRODUCTION

In today’s digital world,Digital information is becoming constantly dynamic and the development of digital social media and user-generated content further provoking this experience. When millions of customer reviews and discussions flood the Internet every day, while individual people feel overwhelmed with information, it is as well impossible for business to keep that up manually. It is difficult for a person or an organization to get the latest trends and to figure out the general opinions about products due to the huge diversity and size of social media, and thus, there is obvious need of computational methods for automatically analyzing sentiment using unstructured text from social media to assist people on information discomfort.Andthis builds the need of automated and real time opinionextraction and mining. The different articles presents for sentiment analysis fields. The number of articlesin this has increased variously.

This makes a need to have survey papers that summarize the recent research trends and directions of SA.Sentiment analysis is gaining popularity day by day due to recent advances in online technologies[1] (blog, review site, forums, social media).It refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, analyze, extract, quantify, and study state level and subjective information. Sentiment analysis is widely applied to voice of the customer material such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

These reviews can gather from customers/users. Review given by the users gives intuition into product reception and quality, which can be used to make crucial business related decisions. The reviews sources are mainly review sites.

With the massive growth of user generated messages, Twitter has become a social site where millions of users can share their opinion. Sentiment Analysis on Twitter data has provided a practicaland effective way to disclose public opinion.There have been some research work which focusing on evaluate the relations between online public sentiment and real-life events (e.g., consumer confidence, stock market [2], polls [3]). From this it is proclaimed that events in real life absolutely have a meaningful and actual response on the public sentiment in Twitter.

Facebook also has become a popular social site where large volume of data and opinions can be shared between tremendous users. Researchers have proposing different methods to support sentiment analysis in Facebook. Based on such social platform sentiments and real life events can be balanced in day by day lives, where different methods are implementing to handle sentiment polarity (positive, neutral or negative [4]).

The exploring work of computation of application and claiming challenges in the field of SA was given by Pang and Lee [5] and Liu [6]. They discussed the techniques used to solve every problem in Sentiment analysis. The sentiment analysis process can be illustrated as given below.

II. SENTIMENT ANALYSIS PROCESS

Data Extraction

Data extraction is the way of retrieving data from data sources for further data processing task. It is the identification of given set of textual documents, phrases, clauses, sentences or entire documents that express attitudes, and determine the polarity of these attitudes [7].

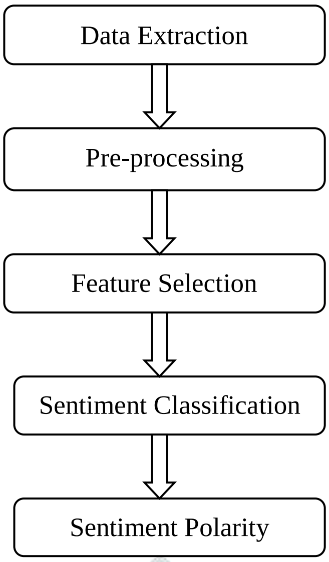

Figure 1: Sentiment Analysis Process

Pre-Processing

Basically pre-processing is used to clean the data later on which is fit for text classification .In the work of [8], author have shown that appropriate text pre-processing include data transformations and filtering can significantly improve the performance. In [9], Rajput et al. have proposed preprocessing strategy which gives SA of twitter feeds for English tweets. To retrieve pre-processed data certain steps need to be followed.

- Removing Non English words- The focus is only on English tweets so non-English tweets must be removed.
- Removing Special symbol, Uniform Resource Locators-The unwanted data can be deleted and then replace it with the blank spaces.
- Slang word translation-For slang word translation the internet slang dictionary can be used.
- Stemming-Stemming gives the stem word. For Example such as ‘happiest’, ’happier’ are replaced with ‘happy’.

Feature Selection

Feature extraction is used for sentiment analysis. It is the process of selecting a subset of relevant features.There are various feature selection methods like TF-IDF (Term Frequency–Inverse Document Frequency), Unigram, Bigram and N- gram. Count Vector and TF-IDF are two feature selection techniques which are described in [10].

Sentiment Classification

Sentiment classification is a special task of text classification whose purpose is to classify a text according to the sentimental polarities of opinions. The main purpose is to classify polarity of given text document, sentence, feature have to be check whether the expressed opinion is in which category such as positive, negative or neutral.

Polarity Result

Sentiment polarity gets resulted by distinguishing words into two class such as positive words and negative class. In some cases third class may also available which a neutral class.

III. Literature Review

The general literature on Lexicon based sentiment analysis is vast, in which discussed work range from sentiment analysis to different types of sentiment lexicons. Here, we focus on reviewing related work on affective sentiment analysis, the use of lexicons for sentiment polarity.

In this paper Valentin et al. has done work for automatically [11] generating focused and accurate topic specific subjectivity lexicons. This work based on general purpose polarity lexicon that allow user to pin-point subjective on-topic in-formation in a set of relevant documents.Theyhave used WordNet for the construction of lexicon. Based on word sentiments, a decision is made at the sentence level. In evaluation section they assess the quality of generated topic- specific lexicons numerically and extrinsically.

Author Veronica [12] has present a generalized framework to derive sentiment lexicons in target language by using automatically annotated data available. To demonstrate work methods, author focus on Spanish as thelanguage in which we seek to develop sentiment lexicons and employthe Spanish WordNet6 for experiments and evaluations.

For this purpose Multilingual WordNet structure allows to generate a high accuracy.Lexicons have been widely used for sentiment and subjectivity analysis, as they represent a simple, yet effective wayto build rule-based opinion classifiers.

The research paper [13] proposes lexicon based approach for sentiment classification of microblog posts. The author’s approach is based on the exploitationof widespread lexical resources such as SentiWordNet, WordNet-Affect, MPQA and SenticNet. In the experimental session they describes effectiveness of the approach was evaluated against two state of the art datasets.clearly, the effectiveness of the whole approachstrongly depends on the goodness of the lexical resource it relies on. In this study they defined four different implementations of such approach: Basic,normalized,emphasized and emphasized-normalized.Specifically, author evaluated the accuracy of their lexicon-based approach on varying boththe four lexical resources as well as the four versions of the algorithm. Statistical significance assess through two test-sets.

Article [14] introduces sentiment analysis based on dictionary based approach. Here they have present word level Sentiment analysis & sentence level sentiment analysis for text mining. Thus, a simple technique in this approach is to use a few seed sentiment words to bootstrap based on the synonym and antonym structure of a dictionary. A notable approach is uses a sentence level sentiment analysis. The word level feature abstraction is done using Naïve Bayesian Classifier. For classifying and analyzing of the sentiment from the reviews, machine learning and lexical contextual information are used.The naive bayesapproach is used to depict each sentence as positive and negative on the bases of useful word level feature. SVM classifier trained on the depicted sentences for the positive and negative classification. Contextual data is used to calculate the polarity of sentence and categorize it as either negative or positive.

In this paper [15]author has trying to find out approaches that generate output with good accuracy for text data. In this they present recent updates on papers related to classification of sentiment analysis. Bag of features frameworkimplemented in this paper. Further they have compared lexicon based and ML based approach. The performance of sentiment analysis is calculated by using help of confusion matrix which is generated when algorithm is implemented on dataset.

In this paper [24] different tweets of event from twitter has taken using hashtag and did sentiment analysis. It compares the polarity of different events.

IV. Sentiment classification approaches

Lexicon-based approach

The lexicon based approach involves calculating orientation for a document from the semantic orientation of words or phrases in the document [16].The main motive of lexicon based approach is to extracts the sentiments from the text.Dictionaries for lexicon-based approaches can be created manually, as authors has shown in this article [17] or automatically, using seed words to expand the list of words. Much of the lexicon-based research has focused on using adjectives as indicators of the semantic orientation of text. First a list of adjectives and corresponding

Sentiment Orientation (SO) values is compiled into a dictionary. After compilation for any given text, all adjectives are extracted and annotated with their SO value, using the dictionary scores. The SO scores are in turn aggregated into a single score for the text.However, although an isolated adjective may indicate subjectivity, there may be insufficient context to determine semantic orientation.Therefore the algorithm extracts two consecutive words.The first member is an adverb or an adjective while the second word provides the context.

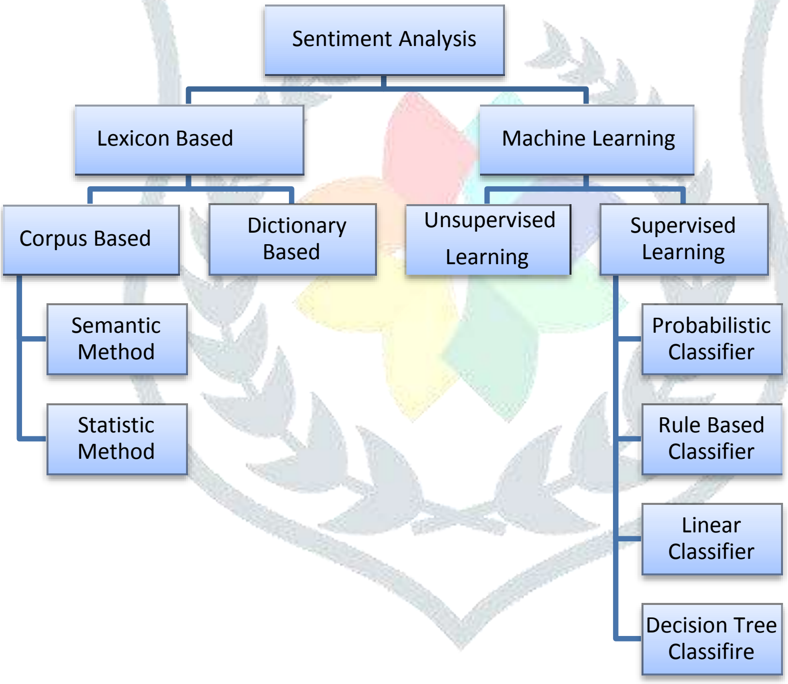

Figure 2: Sentiment Analysis Techniques.

Dictionary-based approach

Dictionary-based methods are the most straightforward approach to obtaining a sentiment lexicon. The dictionary -based approach involves using a dictionary which contains synonyms and antonyms of a word.The dictionary is domain specific. Thus, a simple technique in this approach is to use a few seed sentiment words to bootstrap based on the synonym and antonym structure of a dictionary. Specifically, this method works as follows: A small set of opinion words (seeds) with known positive or

negative orientations is first collected manually, whichis very easy.The algorithm then expands this set by searching in any online available dictionaryfor their synonyms and antonyms.The seed list will be added with the new found words. The process repeatedly keeps on adding the words until no more new words are found. Manual inspection can be used to clean up the list at last by removing or correcting errors.

Corpus-based approach

The Corpus-based approach assists to solve the problem of finding opinion words or thoughts with context specific orientations.Its method depends on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus. Corpus-based semantic orientation approach takes large dataset to detect the polarity of the terms and therefore the sentiment of the text.This approach originally mines sentiment-bearing terms from the unstructured text and further computes the polarity of the terms.There are two methods in the corpus based approach.

Statistical Approach

If the word appears periodically among positive texts, then its polarity is positive. If the word appears frequently among negative texts, then its polarity can be considered as negative. If it has equivalent incidence, then it can be considered as neutral word.Seed of opinion words can be obtained using statistical techniques.Thus, if two words appear together most often within the same context, then there is high possibility that they have same polarity. Therefore, the polarity of an unknown word can be regulated by calculating the moderate frequency of co-occurrence with another word.

This could be done using Pointwise Mutual Information(PMI)as in example shown in [16]. Using part-of-speechpatterns, this technique then classifies the text by extracting the bigrams. PMI is then calculated by using the polarity score for every bigram.

Semantic approach

This approach assigns matching sentiment values to semantically close words. These Semantically close words can be accessed by getting the list of sentiment words and iteratively expanding the initial set with synonyms and antonyms and then deciding the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word in article [17].

V. RESOURCES FOR LEXICON BASED SENTIMENT ANALYSIS

SentiWordNet

SentiWordNet [18] is a lexical resource for opinion mining. It assigns to each synset of WordNet has its own sentiment scores: positivity, negativity, objectivity. Using SentiWordNet for sentiment classification involves scanning a document for relevant terms and matching available information from the lexicon according to part of speech.

WordNet-Affect

WordNet-Affect [19] is a linguistic resource for a lexical description of intuitive knowledge. It is an extension of WordNet which labelsaffective-related synset with affective concepts defined asA-Labels(e.g. the termeuphoriais labeled with the conceptpositive-emotion, the noun Illnessis labeled withphysical state, and so on).

MPQA

MPQA (Multi-Perspective Question Answering) Subjectivity Lexicon [20] is maintained by TheresaWilson, Janyce Wiebe, and Paul Hoffmann.They provide lexicon of 8,222 terms (labeled as subjective expressions), collected from several sources. This lexicon consists a list of words, along with their POS-tagging, labeled with polarity (positive, negative, neutral) and intensity (strong, weak).

Bing Liu's Opinion Lexicon

This lexicon [21] maintains and freely distributes a sentiment lexicon consisting of lists of strings. Liu’s lexicon has a set of around 6800 English words in which 2006 words are positive and 4783 are negative words. The set was initiated using a small set of seed adjective words meaning either good or bad. The seed set was augmented using knowledge discovery methods based on semantic synonym and antonym relations.

Harvard General Inquirer

The General Inquirer lexicons [22] are exists from early work in the cognition psychology of word meaning and also work for content analysis.This General Inquirer is a freely available web resource which contains lexicons of 1915 positive words and 2291 negative words. The Harvard General Inquirer is a lexicon attaching syntactic, semantic, and pragmatic information to part of speech tagged words.

LIWC

Linguistic Inquiry and Word Counts (LIWC) [23] is an appropriate database which maintains protocol for regular expressions. LIWC is a propriety collection of text analysis program available for purchase. It helps to calculate the degree to which various categories of words are used in a text, and can process texts ranging from e-mails to speeches, poems and transcribed natural language in either plain text or Word formats.

Table 1: Distribution of words with sentiment

Positive	admire, amazing, assure, celebration, charm, eager, enthusiastic, excellent, fancy, fantastic, frolic, graceful, happy, joy, luck, majesty, mercy, nice, patience, perfect, proud, rejoice, relief, respect, satisfactorily, sensational, super, terrific, thank, vivid, wise, wonderful, zest
Negative	abominable, anger, anxious, bad, catastrophe, cheap, complaint, condescending, deceit, defective, disappointment, embarrass, fake, fear, filthy, fool, guilt, hate, idiot, inflict, lazy, miserable, mourn, nervous, objection, pest, plot, reject, scream, silly, terrible, unfriendly, vile, wicked

Table1: Some samples of words with consistent sentiment across three sentiment lexicons: the General Inquirer (Stone et al., 1966), the MPQA Subjectivity lexicon (Wilson et al., 2005), and the polarity lexicon of Hu and Liu (2004)

VI. COMPARATIVE ANALYSIS OF LEXICON BASED APPROACH

S. No.	*Studies*	*Methods used*	*Data-Sources*
1	Heerschop B, Goossen F.	Lexicon-based, semantic	IMDB
2	PrabuPalanisamy , Vinit	Lexicon Based	Twitter
3	CaciliaZirn, Mathias Niepert	Statistical (MM), semantic	Amzaon.com
4	Peter Turney and Michael Littman	Semantic	Amazon.com
5	FazelKeshtkar, Diana Inkpen	Corpus Based	Journal Blogs
6	MoonisShakeel	Lexicon based, Dictionary based	Indiabudget.nic.in

Table 2: Comparative Analysis based on lexicon based approaches

VII. APPLICATIONS OF SENTIMENT ANALYSIS

Capital Expenditures

With the Sentiment Analysis report, hotelier could look at a recent time period and identify topics that were mentioned most frequently with a negative sentiment attached. In the topic cloud below, the topics that came up most frequently are “room odor,” “noise,” and “bathroom condition.” With this information, a hotel would know to evaluate these areas when it comes time to make decisions about property improvements.

Online Commerce

The most common use of SA is in e-commerce enterprise. Web portal permit their users to submit their experience about shopping and reviewing their thoughts, opinions and their take on product qualities. This means simply websites allow users to make review about products. They provide summary for the product and different features of the product by assigning ratings. For example, htttp://www.amazon.com is an online shopping website where users rate the products they have bought critically.

Sentiment Analysis for Surveys

Surveys team customers can view sentiment feedback on both their online reviews and their surveys feedback for a true 360 degree picture of their guests’ likes and dislikes. Surveys capture organic guest feedback with an open text format, which is then evaluated with the same technology as an online review. With the added volume of feedback, your decisions will be informed and strengthened by an even wider audience sample.

Brand Reputation Management

BRM helps in finding how public perception of a certain brand changes positively or negatively. The variation after an event can be verified using Sentiments of reviews or feedback.

CONCLUSION

This study shows that the field of sentiment analysis has been well studied by researchers in the past few years. Many different methods have been developed and tested. However, a lot of work is yet to be done. Here we have surveyed various papers which describes about the field of sentiment analysis and sentiment analysis techniques.

Literature survey shows the ideas which method has used by author for their work and the topic of their work, finally we says that sentiment analysis and opinion mining is very vast research area for research and lot of things are there for work. Sentiment analysis can be applied for new applications. Lexicon based approach does not need any prior training in order to mine the data. It uses a predefined list of words, where each word is associated with a specific sentiment.

A future challenge in applying sentiment classification approaches for sentiment analysis of posts in social media is to overcome the ambiguity that actually represents particular problem since it is not easily make use of reflexive information. Typically the analyzed data contain irony and sarcasm, which are particularly difficult to detect. So an evolution of approaches is needed to resolve this limitation.

ACKNOWLEDGEMENT

The author would like to thank the Dept. of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad for providing infrastructure and also experimental facility to carry out the above work.

REFERENCES

[1] Kaushik, Chetan, and Atul Mishra. "A scalable, lexicon based technique for sentiment analysis." arXiv preprint arXiv:1410.2265 (2014).

[2] O'Connor, Brendan, et al. "From tweets to polls: Linking text sentiment to public opinion time series." ICWSM 11. 122-129 (2010):1-2.

[3] Mishne, Gilad, and Natalie S. Glance. "Predicting Movie Sales from Blogger Sentiment." AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006.

[4] Ortigosa, Alvaro, José M. Martín, and Rosa M. Carro. "Sentiment analysis in Facebook and its application to e-learning." Computers in Human Behavior 31 (2014): 527-541.

[5] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis" Foundations and Trends® in Information Retrieval 2.1–2 (2008): 1-135.

[6] Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1 (2012): 1-167.

[7] Kim, Soo-Min, and Eduard Hovy. "Determining the sentiment of opinions."Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

[8] Haddi, Emma, Xiaohui Liu, and Yong Shi. "The role of text pre- processing in sentiment analysis." Procedia Computer Science 17 (2013): 26-32.

[9] Yu, Hongliang, Zhi-Hong Deng, and Shiyingxue Li. "Identifying Sentiment Words Using an Optimization-based Model without Seed Words." ACL (2). 2013.

[10] Tripathy, Abinash, Ankit Agrawal, and Santanu Kumar Rath. "Classification of Sentimental Reviews Using Machine Learning Techniques." Procedia Computer Science 57 (2015): 821-829.

[11] Jijkoun, Valentin, Maarten de Rijke, and WouterWeerkamp. "Generating focused topic-specific sentiment lexicons." Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.Association for Computational Linguistics, 2010. Perez-Rosas, Veronica, Carmen Banea, and Rada Mihalcea."Learning sentiment lexicons in spanish."LREC.Vol. 12. 2012.

[12] Musto, Cataldo, Giovanni Semeraro, and Marco Polignano. "A comparison of lexicon-based approaches for sentiment analysis of microblog posts." Information Filtering and Retrieval 59(2014).

[13] Bhonde, Reshma, et al. "Sentiment Analysis Based on Dictionary Approach." International Journal of Emerging Engineering Research and Technology 3.1 (2015).

[14] Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." Proceedings of the40th annual meeting on association for computational linguistics.Association for Computational Linguistics, 2002.

[15] Goyal, Amit, and Hal DauméIII."Generating semantic orientation lexicon using large data and thesaurus."Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis.Association for Computational Linguistics, 2011.

[16] Neviarouskaya, Alena, Helmut Prendinger, and Mitsuru Ishizuka. "Recognition of affect, judgment, and appreciation in text."Proceedings of the 23rd international conference on computational linguistics. Association for Computational Linguistics, 2010.

[17] Baccianella, Stefano, Andrea Esuli, and FabrizioSebastiani."SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." LREC.Vol. 10.2010.

[18] Wiebe, Janyce, Theresa Wilson, and Claire Cardie. "Annotating expressions of opinions and emotions in language."Language resources and evaluation expressions of opinions and emotions in language." Language resources and evaluation 39.2 (2005): 165-210.

[19] Gupte, Amit, et al. "Comparative study of classification algorithms used in sentiment analysis." International Journal of Computer Science and Information Technologies 5.5 (2014): 6261-6264.

[20] Janyce Wiebe, Theresa Wilson , and Claire Cardie (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210.

[21] Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

[22] Stone, Philip J., Dexter C. Dunphy, and Marshall S. Smith. "The general inquirer: A computer approach to content analysis." (1966).

[23] Gilbert, CJ Hutto Eric. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Eighth International Conference on Weblogs and Social Media (ICWSM-14).

[24] Rajkumar S. Jagdale, Vishal S. Shirsat, Sachin N. Deshmukh, “Sentiment Analysis of Events from Twitter Using Open Source Tool”, International Journal of Computer Science and Mobile Computing, ISSN 2320–088X, Vol.5 Issue.4, April- 2016, pp. 475-485