Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.S2, 2019, pp. 1-6

A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis

Venkateswarlu Bonta[1], Nandhini Kumaresh[2]and N. Janardhan[3]
[1,2&3]Department of Computer Science, School of Mathematics and Computer Sciences,
Central University of Tamil Nadu, Thiruvarur, India
E-Mail: bonta17@students.cutn.ac.in, nandhinikumaresh@cutn.ac.in,janardhan17@students.cutn.ac.in

Abstract **- In recent years, it is seen that the opinion-based postings in social media are helping to reshape business and public sentiments, and emotions have an impact on our social and political systems. Opinions are central to mostly all human activities as they are the key influencers of our behaviour. Whenever we need to make a decision, we generally want to know others opinion. Every organization and business always wants to find customer or public opinion about their products and services. Thus, it is necessary to grab and study the opinions on the Web. However, finding and monitoring sites on the web and distilling the reviews remains a big task because each site typically contains a huge volume of opinion text and the average human reader will have difficulty in identifying the polarity of each review and summarizing the opinions in them. Hence, it needs the automated sentiment analysis to find the polarity score and classify the reviews as positive or negative. This article uses NLTK, Text blob and VADER Sentiment analysis tool to classify the movie reviews which are downloaded from the website www.rottentomatoes. com that is provided by the Cornell University, and makes a comparison on these tools to find the efficient one for sentiment classification. The experimental results of this work confirm that VADER outperforms the Text blob.**
Keywords: **Sentiment Analysis, Opinion Mining, Sentiwordnet, NLTK, Text blob, VADER**

I. INTRODUCTION

Sentiment analysis is the process of computationally identifying and categorizing the opinions expressed in a piece of text, especially in order to determine the writer's attitude towards a particular topic, product, etc. is positive, negative or neutral[1]. The attitude may be as one of the following scenarios. (a)His or her judgment or evaluation (b) Affective state i.e., the emotional state of the author when writing a review. Sentiment analysis can be highly useful in several cases. The best example is the marketing methodology. Marketing teams can use sentiment analysis to launch a new product or to determine the existing product popularity and preference. Reviews from social media can be gathered and used to assess how good or bad a product or service is doing based on customer response. In computer science literature three main streams of sentiment definitions can be found. The first stream of research definition sentiment is through opinion. The second stream of research is sentiment through focusing on feelings. The third stream is sentiment through focusing on both feeling and opinions[2]. In addition, there are other terms with slight different tasks namely opinion mining, opinion extraction, sentiment mining, review mining etc. Though there are many names, the main aim of them is to know how to extract the hidden polarity of the written text or review.

A. Applications

Sentiment analysis has attracted large number of researchers due to its applications in variety of disciplines. Huge amount of existing work has been focused on online customer reviews including product, hotel, movie, etc. Individual consumers want to know the opinions of existing users of product before purchasing it, and also others opinion about political candidate before making a voting decision in political election. Sentiment analysis has shown its impact on social media such as Twitter, Facebook to understand user tweets and their behaviour. Previously, when an individual wanted to know the opinion about any product or service, he/she used to ask his/her friends or family members.

When an organization or a business needs public or consumer opinion, it is used to conduct surveys, opinion polls, and focus groups. After the explosive growth of social media, there is no need to ask one's friends or family for opinions, and no organization needs to conduct surveys, opinion polls, and focusing groups in order to gather public opinions because there is an abundance of such information available publicly. Due to these applications, industrial activities have flourished in recent years. Sentiment analysis applications have spread to almost every possible domain from consumerproducts to services, healthcare, and financial services to social events and political elections.

On the other hand, shifting to automated sentiment analysis saves time and money. Customers can log in and simply watchthe graphs and tables that show them data about the required brands in a user-friendly environment. Such services also have benefit of being able to capture measure and display data in real-time speed. Sentiment analysis has become the gateway to understand the customer needs, extending customer base and expectations.

II. RELATED WORK

Sentiment analysis is a type of data mining that deals with people's opinion through Natural Language Processing,

Computational Linguistics and text Analysis. There are mainly two approaches to extract the sentiment from given reviews and classify the result as positive or negative.

1. Lexicon Based Approach
2. Machine Learning Approach

Lexicon based approach requires predefined lexicon while Machine Learning approach automatically classifies the review which requires training data. A lexicon is a stock of terms that belongs to a particular subject or language. Hybrid approaches are also used to overcome the drawbacks of the individual techniques.

A. Lexicon-Based Approach

Lexicon-Based Approach uses sentiment lexicon with information about which words and phrases are positive and which are negative [3]. A sentiment lexicon is a list of lexical features which are generally labelled according to their semantic orientation as either positive or negative. Researchers first create a sentiment lexicon through compiling sentiment word lists such as manual approaches, lexical approaches, and corpus-based approaches, then determine the polarity score of the given review based on the positive and negative indicators which are identified in the lexicon.

There are some lexicons like LIWC (Linguistic Inquiry and Word Count), GI (General Inquirer) that categorizes the words into positive and negative according to their context free semantic orientation. LIWC consists of almost 4,500 words organized into one of 76 categories, including 905 words in two categories especially related to sentiment analysis.

LIWC was well-established and validated in a process spanning tree more than a decade of work by sociologists, psychologists, linguists [4]. Though its extensive use to find sentiment analysis in social media text, LIWC does not include acronyms, initialisms, emoticons, and slang which are important factors for sentiment analysis of social media text.

However, other lexicons like ANEW (Affective Norms for English Words), SentiWordNet, and SenticNet are associated with valence scores for sentiment intensity. SentiWordNet consists of 1,47,306 synsets are annotated with three sentiment scores such as positive negative and objective [5].

Though, it is not a gold standard resource likeWord Net, but is useful for wide range of tasks. One of the major advantages of Lexicon-Based approach is, its domain independence, and also it can be easily extended and improved.

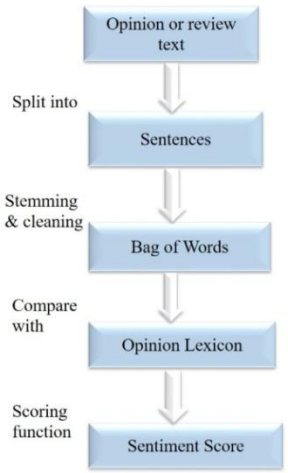

Fig. 1 Fundamental approach for Sentiment classification using Lexicon

B. Machine Learning Approach

Machine Learning Approaches are used to construct an algorithm and build a model by feature selection and by learning from labelled training datasets [6]. Naïve Bayes Classifier, Support Vector Machine (SVM), and Random Forest are the well-known methods for sentiment classification through Machine Learning.

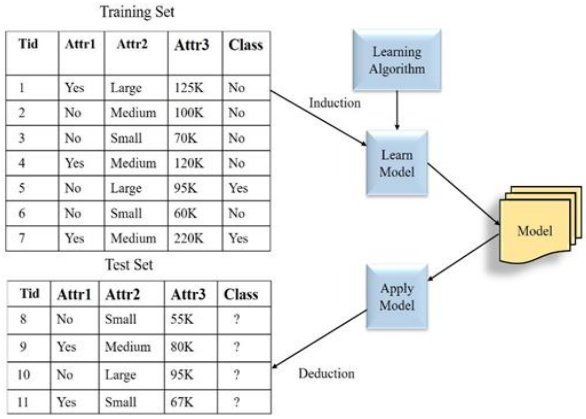

Fig. 2 Fundamental approach for Sentiment classification using Machine Learning

These algorithms can automatically learn all kinds of features forclassification through optimization. Since the sentiment classifier is trained on the labelled data from one domain, oftenit does not work with another domain. To overcome problem, Lexicon based approaches are recommended.

C. Levels of Sentiment Analysis

Based on the level of analysis which they involve, the levels of sentiment analysis are categorized into 3 types, namely

1. Document level analysis
2. Sentence level analysis
3. Entity and Aspect level Analysis.

The task of Document level analysis is to classify whether the whole opinion document expresses a positive or negative sentiment [7]. This level of analysis assumes that each document expresses opinions on a single entity. The Sentence level analysis deals with sentences and determines whether each sentence expressed is positive, negative, or neutral. This level of analysis is mostly related to subjectivity classification. Entity and Aspect based analysis. Goal of the third level of analysis is to discover sentiments on entities and their aspects. Instead of looking at the language constructs such as reviews, sentences, etc., the aspect level directly looks at opinion itself. It is based on idea that an opinion consists of a sentiment and a target entity.This article presents our work on Lexicon based approaches to identify the sentiment of the given movie reviews.

D. Lexicon-Based Sentiment Analysis Tools

*1. NLTK:*NLTK (Natural Language ToolKit) is an open source Natural Language Processing platform for python, developed in conjunction with computational linguistics at university of Pennsylvania in 2001. It provides easy-to-use interface over 50 corpora, lexicon resources such as SentiWordNet with a suit of text processing libraries for classification, tokenization, and semantic reasoning. In NLTK sentiment score is calculated from SentiWordNet which consists of polarity score of each synset of WordNet with three sentiment numerical scores positivity, negativity values ranging from 0.0 to 1.0 and their sum is 1.0 for each synset. However, you can also determine the objectivity score of the synset of Word Net using the formula given below.

$$1 - (PosScore + NegScore) \quad (1)$$

The scores are calculated using a complex mix of semi-supervised algorithms. It is not a gold standard resourcelike WordNet and LIWC. However, it is useful for a wide range of tasks.

The WordNet synsets are uniquely identified by POS, ID pairs [8]. Synset Terms column shows the included terms with the sense numbers. The SentiwordNet lexicon is very noisy; a large majority of synsets have no positive or negative scores. It also fails to account for sentiment bearing lexical features relevant to text in micro blogs.

TABLE I SAMPLE SENTIWORDNET VALUES

POS	ID	PosScore	NegScore	SynsetTerms
A	00071142	0.5	0.5	Impressed#1
A	00070111	0	0	Enhansive#2
A	00065064	0.625	0	Good#5
A	00061664	0.625	0	Neat#1
A	00035868	0.5	0	Blusting#1

2. TextBlob: Textblob is a python library for processing textual data. It provides a consistent API for common natural language processing (NLP) tasks [9]. Textblob is just like a python string.

Features of Textblob

a. Tokenization
b. Noun phrase extraction
c. POS tagging
d. Sentiment analysis
e. Language Translation and detection
f. n-grams
g. spelling correction
h. WordNet integration

*3. VADER:*VADER (Valence Aware Dictionary and sEntimentReasoner) is a lexicon and rule-based sentiment analysis tool. It is an open source under the MIT license developed by George Berry, Ewan Klein, and Pier Paolo. Vader lexicon performs exceptionally well in the social media domain. VADER retains the benefits of traditional sentiment lexicons like LIWC (Linguistic Inquiry and Word Count). It is bigger, simply inspected, understood, quickly applied and easily extended. The VADER sentiment lexicon is gold-standard quality and has been validated by humans. VADER distinguishes itself from LIWC as sensitive to sentiment expressions in social media context and also generalizing more favourable to other domains.

III. METHODOLOGY

Movie reviews are analysed for their sentiment during Movie release to predict the movie response as positive or negative. The overall methodology follows four steps

1. Data collection
2. Pre-processing
3. Sentiment extraction
4. Classify sentiment as positive or negative.

A. NLTK

Fig. 3 shows how NLTK classifies a review into either positive or negative. It first tokenizes the reviews into words and then remove the stop words such as a, an, the, for, is, etc. After removing stop words, the words are stemmed to get their root words. For example, “disappointed” is reduced to “disappoint”. This helps in reducing the time while searching the word in the SentiWordNet. All special symbols and numbers are also removed from the reviews. Now it performs the POS (Parts of Speech) tagging on the purified reviews. It involves stringent grammar rules while performing the tagging. Thus, the data is ready for classification by extracting the positive and negative words from the given review and match them with respected sentiment score given in the SentiWordNet. Finally, by counting the positive and negative terms which are found in the review, and using sentiment polarity, the class receives the highest score.

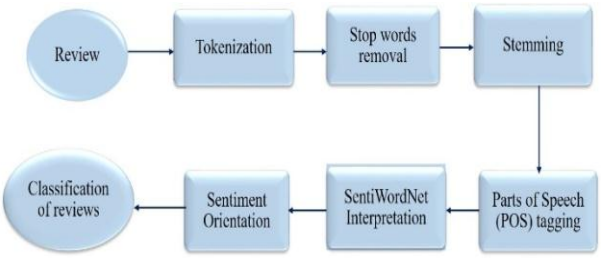

Fig. 3 Overall process of NLTK to classify a review

For each Synset the score in SentiWordNet lexicon, can be calculated by

$$SynsetScore = PosScore - NegScore \quad (2)$$

For a term with specific POS tag, if k synsets contain it, then the sentiment score of the term can be calculated by following expression

$$TermScore = \frac{\sum_{n=1}^{k} SynsetScore(r)/r}{\sum_{n=1}^{k} 1/r} \quad (3)$$

Where n is the sense number.If a term not in the SentiWordNet, we assume that its sentiment score is 0. If a negation word appears in front of a term, we simply reverse the sentiment value of the respected term. The sentiment score of the target review can be calculated by adding up all the term sentiment scores as shown in below:

$$PosScore(p) = \sum_{i=1}^{m} TermScore(T_i) \quad (4)$$
$$PosScore(p) = \sum_{i=1}^{m} TermScore(T_i) \quad (5)$$
$$SentiScore(p) = PosScore(p) + NegScore(p) \quad (6)$$

Where p is a review which contains m positive terms and n negative terms.PosScore(p) and NegScore(p) represents the positivity and negativity of the corresponding review p. SentiScore of p represents the final sentiment score of the review p.

B. TextBlob

Textblob is a python library that provides text mining, text analysis and text processing modules for python developers. Textblob reuses NLTK corpora, and if NLTK has been installed beforeTextblob, then the Textblob will be installed with a great ease. Textblob supports the python versions 2.6 and the latest.

Installation:
$ pip install -U textblob
$ python -m textblob.download_corpora

The above commands install Textblob and download necessary NLTK corpora, and if NLTK is installed before Textblob, there is no need to download corpora.
Textblob is a sentence level analysis. First, it takes a dataset as the input then it splits the review into sentences. A common way of determining polarity for an entire dataset is to count the number of positive and negative sentences/reviews and decide whether the response is positive and negative based on total number of positive and negative reviews. Polarity and subjectivity of a given review can be known using *sentiment()* function. It returns a named tuple with two parameters called polarity and subjectivity. The polarity score is ranging from -1 to 1 and subjectivity ranges are from 0 to 1 where 0 is most objective and 1 is most subjective.

Example:
review=Textblob ("the movie was interesting.")
review.sentiment
Sentiment(polarity=0.5, subjectivity=0.5)

C. VADER

As mentioned earlier, VADER is a lexicon and rule-based sentiment analysis tool. It uses a combination of a sentiment lexicon, a list of lexical features which are generally labelled according to their semantic orientation as either positive or negative.VADER has been quite successful when dealing with social media texts, movie reviews, and product reviews. This is because VADER not only tells about the positivity and negativity score but also tells about how positive or negative a sentiment is. The developers of VADER have used Amazon's Mechanical Turk to get most of their ratings.

1. Advantages of VADER

a. Works perfectly on social media type text.
b. It does not require any training data but constructed from generalizable, valence-based, human-curated gold standard lexicon.
c. VADER supports emoji for sentiment classification.
d. It is fast enough to be used online.
e. It does not severely suffer from a speed-performance tradeoff.

Installation:

C:\Users\Admin>pip installvaderSentiment

VADER analyses a piece of text to see if any of the words from the text are present in the VADER lexicon. It can find the polarity indices using polarity_scores() function. This will return the metric values of the negative, neutral, positive, and compound for a given sentence. The compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between *-1* and *+1* where *-1* indicates most extreme negative and *+1* indicates most extreme positive. It is useful to set the standardized thresholds for classifying sentences as positive, neutral or negative. The typical threshold values are given below

Positive Sentiment: compound score >= 0.05
Neutral Sentiment: compound score > -0.05 and < 0.05
Negative Sentiment: compound score <= -0.05

These are the most useful metrics for multidimensional measures of sentiment for a given textual review.The below figure shows the VADER lexiconcontaining words along with their sentiment ratings.

TABLE II SAMPLE VADER LEXICON VALUES

Word	Sentiment Rating
Great	3.1
Disaster	-3.1
Good	1.9
Horrible	-2.5
Rejoiced	2.0

VADER analyses sentiments primarily based on certain key points such as Punctuation, Capitalization, Degree modifiers, Conjunctions, Preceding Tri-gram [10].

There are more than 7,500 lexical features with validated valence scores that indicate both the sentiment polarity, and sentiment intensity ranging from -4 to +4.

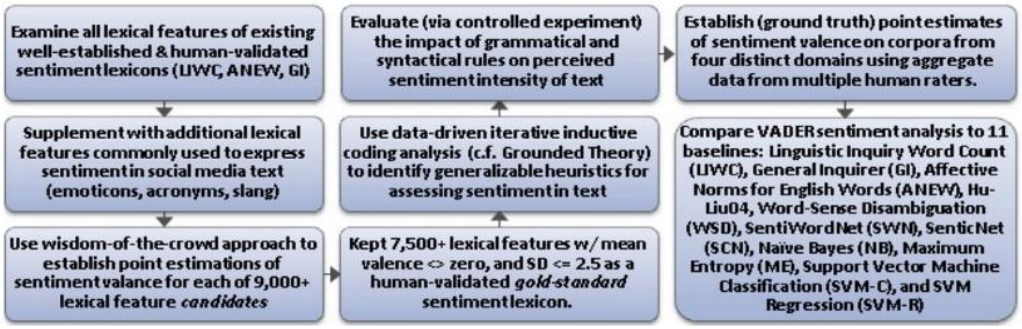

Fig. 4 Methods and process approach overview of VADER

2. Dataset

Dataset includes 11861 sentence-level snippets from www.rotten.tomatoes.com provide by the Cornell University [11]. The snippets were derived from an original set of 10662 movie reviews (5331 positive and 5331 negative) in Pang & Lee (2005 July).

IV. RESULTS AND DISCUSSION

Our experimental approach yields the following results

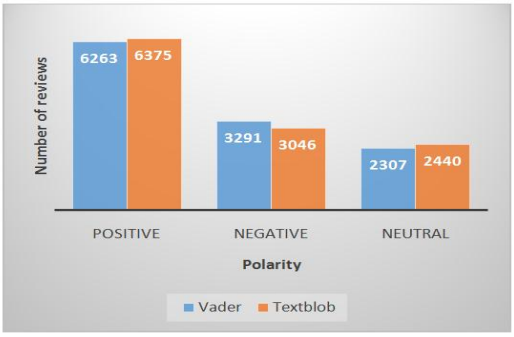

Fig. 5 VADER versus Text blob. VADER shows the close proximity of reviews than Textblob with respect to the actual reviews in the Dataset

A. Lexicon-Classification Performance

TABLE III PERFORMANCE OF LEXICON SENTIMENT ANALYSIS TOOLS

Lexicon	Classification Accuracy metrics			
	Precision%	Recall%	F1 score%	Accuracy%
VADER	78.46	85.0	81.60	77.0
Textblob	76.92	81.96	79.37	74.0
NLTK	60	55.0	57	62.0

From the above table III it is clear that accuracy of VADER is 77% where as Text blob and NLTK has 74% and 62% respectively. It is also identifiedthat how NLTK and Text blob show a concentration of reviews incorrectly that are classified as neutral. Presumably, this is due to lack of coverage for sentiment-oriented language of social media text which consists of emoticons, slangs, acronyms.

The lexicon of machine learning algorithms is constructed by training their modules on half of the data,and remaining half is used for testing.The lexicon of machine learning algorithms is constructed by training their modules on half of the data,and remaining half is used for testing. Most of the algorithms work only in specific domains. Unlike machine learning algorithms VADER performs better across various kinds of domains. As compared to machine learning techniques, VADER has several advantages. Firstly, it is both quick and computationally economic. VADER runs directly from standard modern laptop or computer; a corpus takes a fraction of a second to analyse with VADER, butit approximately takes hours when using more complex models like Support Vector Machine. Second advantage is that the lexicon and the rules used by the VADER are directly accessibleand not hidden. Therefore, VADER is easily understood, extended and modified.

VADER Sentiment Analysis works better for texts from social media and other Web sources as well, than Text blob because when it comes to analysing comments or reviews from social media, the sentiment of the sentence changes based on the emoticons. VADER takes this into account along with slang, capitalization, and the way the words are written along with their context. For example: “The movie is good” gives a compound score of 0.4404 where as “The movie is GOOD” gives a compound score of 0.5622. Another factor that increases the intensity of the sentiment

in a sentence is the inclusion of the exclamation marks. It considers up to three exclamation marks that addsthe additional positive or negative intensity. For instance, "The movie was GOOD!" will give the result of 0.6027. VADER also takes into account of modifying words in front of a sentiment term, for example"extremely good" would increase the positive intensity. VADER also supports emojisentiments. Hence VADER is better option for tweets analysis and their sentiments.

V. CONCLUSION

VADER is a gold standard list of lexical features which is specially attuned to find semantics in micro blog text. If sentiment was absolutely the only thing planned to do for micro blog text, and if it needs to be processed fast, then VADER is a better choice by considering the threshold as 0.05.VADER also follows grammatical and syntactical conventions for expressing and emphasizing sentiment intensity. VADER performs well than Text blob and NLTK sentiment analysis tools.

REFERENCES

[1] Vikas Malik and Amit Kumar. "Sentiment Analysis of Twitter Data Using Naive Bayes Algorithm", *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol. 6, No. 4, 2018.

[2] J.Ge, M.Alonso Vazquez, and U.Gretzel,"Sentiment analysis: a review", In Sigala, M. &Gretzel, U. (Eds.), *Advances in Social Media for Travel, Tourism and Hospitality: New Perspectives*, Practice and Cases,pp. 243-261. New York: Routledge, 2018.

[3] Z. Nanli, Z. Ping, L. Weiguo, and C. Meng, "Sentiment analysis: A literature review", *Proceedings of the International Symposium on Management of Technology (ISMOT), Hangzhou, IEEE*, pp. 572-576, 2012.

[4] J.W. Pennebaker, R.L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015", *Austin, TX: University of Texas at Austin*, 2015

[5] S. Baccianella, A. Esuli, and F. Sebastiani, "SENTIWORDNET 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", pp. 2200–2204, 2008

[6] M. Hu and B. Liu, "Opinion Extraction and Summarization on the Web", pp. 1621–1624.

[7] B. Pang, L. Lee, H. Rd, and S. Jose, "Thumbs up ? Sentiment Classification using Machine Learning Techniques", pp. 79–86, 2002

[8] Wordnet.com, "WordNet, a Lexical database for English", [online] Available: http://wordnet.princeton.edu/

[9] Textbolb.com, "Textblob Tutorial, Quickstart", [online] Available at: https://textblob.readthedocs.io/en/latest/quickstart.html#quickstart

[10] C. J. Hutto and E. Gilbert, "VADER : A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text", *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, ,* pp. 216–225, 2014

[11] Cornell.edu, "Movie Review data", [online] Available: http://www.cs.cornell.edu/people/pabo/movie-review-data/

[12] Steven Bird and Edward Loper. "NLTK: The Natural Language Toolkit", 2006

[13] Bing Liu,"Sentiment Analysis and Opinion Mining", *Morgan &Claypool Publishers, May 2012*.

[14] Adamo and David. "A Text Similarity Approach to Sentiment Classification (of Movie Reviews) using SentiWordNet" .10.13140/ RG.2.1.3271.1120, 2015

[15] H. Han, Y. Zhang, J. Zhang, J. Yang, and X. Zou, "Improving the performance of lexicon-based review sentiment analysis method by reducing additional introduced sentiment bias",, pp. 1–11, 2018

[16] Steven Loria. "Textblob Documentation",Release 0.15.2,2018.