DOI: http://dx.doi.org/10.26483/ijarcs.v9i2.5581

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A HYBRID METHOD PROPOSED FOR BEHAVIOURAL ANALYSIS ON TWITTER OPINION DATA USING DICTIONARY AND SEMANTIC BASED APPROACH

Shelly Gupta
Assistant Professor
CSE dept., IPEC Ghaziabad, India

Lalit Kumar Sharma
Student
CSE dept., IPEC Ghaziabad, India

Kunwar Prashant
Student
CSE dept., IPEC Ghaziabad, India

Mohit Panwar
Student
CSE dept., IPEC Ghaziabad, India

Abstract: The world of technology is growing by leaps and bounds and the arena in technology that is going to be explored is Data Mining. It is estimated that till 2025, most of the world's trade will be based on Data Mining [1]. There is vast availability of people opinion data on twitter for almost every product and service. The challenge is to interpret this data and to extract the information which can lead a decision maker to take better decisions. In dictionary-based approach every word with some positive, negative or neutral value is mapped but opinions are not always direct, hence the sense of the sentence or sub-sentence doesn't agree with its numeral weight. This short coming of this approach lead us to come up with some strategies to increase the accuracy of this method by multiplying the weights together and using some fundamental semantic rule to classify sarcastic tweets. Hence in this paper a hybrid approach is implemented which ensures the sign of total weight of the sentence according to its indirect sense. The positive outcome is that opinions which were earlier treated as neutral are now retaining their sense and add up to our decisions. The hybrid approach is using the concepts of dictionary-based approach and semantic-based approach i.e. matching words from the dictionary and assigning their sentimental value and also using some specific semantic rules used for analyzing sarcastic or neutral tweets for gaining more information about the opinions. The proposed mining of opinions has become easier and more accurate that can be utilized for product's sale forecasting.

Keywords: Twitter Data, Twitter API, Hadoop, Hive, Flume, Sentiment Analysis, Tweets

1. INTRODUCTION

Behavior analysis is the science of studying the comportment of a person to establish a specific profile about it. Hence the process of computationally identifying opinions expressed in a piece of text, in order to determine the writer's attitude towards a particular political issue, breaking news, movie, sports, celebrity or product and categorizing it into Positive, Negative or Neutral sentiment is known as opinion mining or behavioral analysis. It has firstly been used in psychology and since a few years, it has been implemented in information technology programs to understand the needs of the users. There are mainly two approaches to achieve the above defined objective[2]. First is Dictionary based approach which uses a dictionary that contains the sentiment value corresponding to huge number of words and phrases. After the tokenization, the corresponding values to all the words are being assigned to them. After that the total weight age of a tweet is computed by summing up all the sentiment values for a single tweet. If it comes to a negative value then this is a negative tweet, If it comes a positive value then this is a positive tweet and if value comes zero on computation, it's a neutral tweet. Second is the semantic based approach in which a tweet is broken down in the form of tokens using the rules of grammar. Then based on the structure of sentences and meaning of the words in sentences, it is categorized into various types of sentences which further will be classified as positive, negative and neutral tweets.

In dictionary-based approach every word is mapped with some positive, negative or neutral value but the opinions are not always direct, hence the sense of the sentence or sub-sentence doesn't agree with its numeral weight. This short coming of this approach motivated us to come up with some new strategies to increase the accuracy of this method. Hence, in this paper we are going to propose a method which analyzes twitter data. Twitter data is in the form of text which is opinions, feelings of people.

A hybrid approach of dictionary-based approach and semantic-based approach is proposed here. This method uses both the concepts; one is matching words from the customized dictionary and assigning their sentiment value. The customized dictionary is formed using existing dictionary (consist of all the words with their sentiment values) and added different sentiment values for celebrities, political figures etc. Second concept is using some specific semantic rules for analyzing sarcastic or neutral tweets and gaining more information about the opinions.

The paper is presented in various sections where section 2 presents the motivation and literature review which is illustrated in the form of table, section 3 presents the proposed hybrid method , section 4 presents the methodology used for analysis of Twitter data using hybrid method, Section 4 illustrates the conclusion of our work and at last Section 5 presents the future scope and ideas in the field of behavioral analysis.

2. MOTIVATION

This section provides the brief description of various research papers studied for this study. The given below table 1 represents the summarization of various methods applied for behavioral analysis. All these research papers lead us to a thought that there is much scope for more work to be done on simplification and improvement of the accuracy of text opinion analysis because human language is always evolving and changing.

TABLE I
METHODS APPLIED FOR BEHAVIORAL ANALYSIS

S.No.	Title	Method	Description
1	Mining comparative opinions from customer reviews for competitive intelligence.	2-level CRF	Xu Kaiquan et al, [3] proposed a graphical model to extract and visualize comparative relations between products from customer reviews, to help enterprises discover potential risks and further design new products and marketing strategies. They verified this method on Amazon customer reviews.
2	Blogger-centric contextual advertising.	SVM, Chi Square	Fan Teng-Kai et al, [4] used text mining techniques to discover 'bloggers' immediate personal interests in order to improve online contextual advertising. They extended the concept of long tail theory.
3.	Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities.	Semantic	Zhou L, et al, [5] used Rhetorical Structure Theory (RST) along with a small set of cue phrase-based patterns to collect instances and convert them to semantic sequential representations (SSRs). Finally, an unsupervised method was adopted to generate, weight and filter new SSRs.
4.	Fine-grained sentiment analysis with structural features	Statistical (MM), semantic	Zirn C et al, [6] present a fully automatic framework for fine-grained sentiment analysis. They used Markov logic to integrate polarity scores from different sentiment lexicons with information about relations between neighbouring segments.
5.	Manipulation of online reviews: an analysis of ratings, readability, and sentiments	Statistical	Hu Nan et al, [7] proposed a simple statistical method to detect online reviews, and assess how consumers respond to products with manipulated reviews. In particular, the writing style of reviewers is examined, and the effectiveness of manipulation through ratings, sentiments, and readability is investigated
6.	A lexicon model for deep sentiment analysis and opinion mining applications	Semantic	Maks Isa, et al, [8] they presented a lexicon model for the description of verbs, nouns and adjectives to be used in applications like sentiment analysis and opinion mining. The model aims to describe the detailed subjectivity relations that exist between the actors in a sentence expressing separate attitudes for each actor.
7.	Sentiment analysis via dependency parsing.	Context Based Method, NLP	Caro Luigi Di et al, [9] proposed a context-based model where the users' sentiments (or opinions) are tuned according to some context of analysis. Finally, they presented the system called SentiVis which implements these ideas.
8.	Identifying the semantic orientation of terms using S-HAL for sentiment analysis.	S-HAL, SO-PMI	Tao Xu et al, [10] 's method takes a classification approach that is based on a novel semantic orientation representation model called S-HAL (Sentiment Hyperspace Analogue to Language.

9.	Ontology-based sentiment analysis of twitter posts.	FCA	Kontopoulos Efstratios et al, [11] 's paper proposed the deployment of original ontology-based techniques towards a more efficient sentiment analysis of Twitter posts. The novelty of the proposed approach is that posts are not simply characterized by a sentiment score, as is the case with machine learning-based classifiers, but instead receive a sentiment grade for each distinct notion in the post.
10.	Deriving market intelligence from microblogs.	SVM	Li Yung-Ming et al, [12] the consideration of user credibility and opinion subjectivity is essential for aggregating microblog opinions. The proposed framework is designed to cope with the following tasks: trendy topics detection, opinion classification, credibility assessment, and numeric summarization
11.	Sentiment analysis algorithms and applications: A survey	Sentiment analysis techniques	W. Medhat et. al.[13] have studied and summarized various sentiment analysis techniques applied for sentiment analysis. In their paper they have concluded that enhancement in sentiment classification and feature selection algorithms is an open research area and also found that SVM and NB classification methods are most promising algorithms for sentiment analysis.
12.	Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis	Lexicon and learning based approach	A. Z. H. Khan et. al. [14] have presented a new entity-level sentiment analysis method based upon the lexicon based approach. This method gave high precision, but low recall. To improve recall, Chi-square test applied on its output, and additional opinionated tweets are identified. A classifier is then trained to assign polarities to the entities in the newly identified tweets.
13.	SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives	Hierarchical clustering	E. Cambria et. al.[15] have used an ensemble of hierarchical clustering and dimensionality reduction for discovering the primitives for both noun and verb concepts in SenticNet to extend the coverage of the commonsense knowledge base and boosted the accuracy of SenticNet for sentence-level polarity detection .
14.	SemEval-2017 Task 4: Sentiment Analysis in Twitter	CrowdFlower annotation method	S. Rosenthal et. al. [16] have described the fifth year of the Sentiment Analysis in Twitter task with two major changes i.e. introduction of Arabic language for all subtasks and availability of information from the profiles of the Twitter users who posted the target tweets.
15.	Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier	Gini Index feature selection , SVM	A. S. Manek et. al. [17] have proposed a Gini Index based feature selection method with Support Vector Machine (SVM) classifier for sentiment classification for large movie review data set. This proposed has improved the accuracy of sentiment polarity prediction in comparison with other feature selection methods on movie reviews.

3. PROPOSED HYBRID METHOD

The Hybrid method is developed by using the two Pre-existing models namely Dictionary-Based and Semantic-Based Model under Lexicon Analysis Methods to improve the accuracy of sentiment analysis for particular tweets which are complicated in the sense of sarcastic tweets. This method uses both the concepts; one is matching words from the customized dictionary and assigning their sentiment value. The customized dictionary is formed using existing dictionary (consist of all the words with their sentiment values) and added different sentiment values for celebrities, political figures etc. Second concept is using some specific semantic rules for analyzing sarcastic or neutral tweets and gaining more information about the opinions. The given below figure 1 represents the conceptual architecture of the proposed hybrid method.

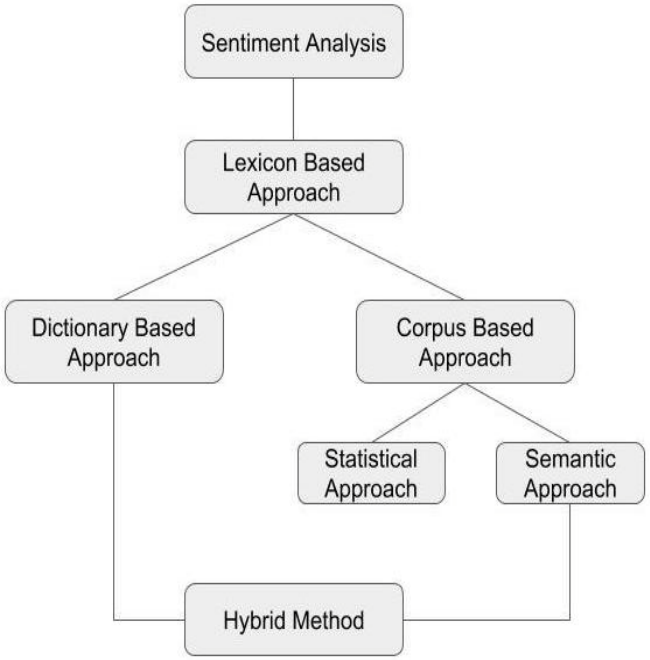

Figure 1: Hybrid method

After data reduction and applying the sentimental methods for behavior analysis we will get the outcome in the form of classified tweets that is Positive, Negative and Neutral tweets.

Where, the output value 1 means a positive tweet which show a good or positive response towards something, -1 means a negative tweet which shows a negative response towards something and at last the 0 value means a neutral tweet which neither shows a support or appreciation nor oppose or depreciation towards anything. It also includes tweets which are facts or theories.

There are some special cases i.e. in some sentences or phrases there may be equal number of positive and negative words leading the whole sense to neutral but in actual it may carry a positive or negative sense. Such as when any negative word attached with positive word it conveys a negative sense. In existing method when it get added gives a neutral sense but in this proposed method when we apply multiplication, it retains its negative sense. When two entities (people, product or service) are compared, sentence may contain both positive and negative views about both the entities whereas the tweet can be in favour of anyone entity. So, interpreting whole tweet may lead to false sense. In these cases, sense can be concluded in the last sentence of tweet. For the same we can write rules by which we could directly interpret the last part of tweets comparing two entities.

There are some tweets in which some view is expressed by using the name of some famous celebrity, political party, organization, political leader etc. Then the whole sense of the tweet depends on the public impression of the particular entity (mentioned earlier). In this proposed method we have used an improved version of dictionary which also contains the sentiment value for this type of famous personalities and organization.

4. METHODOLOGY

The various steps involved for Behavior analysis are shown in the figure 2 below:

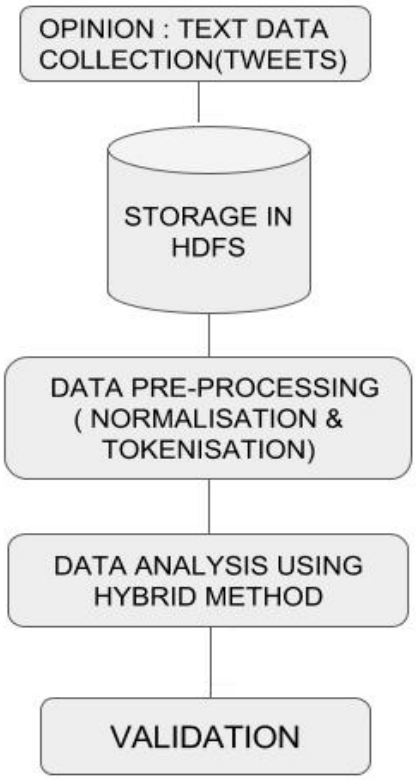

Figure 2: Methodology steps

Step 1) Opinion collection: In this step the large amount of data is collected using Twitter streaming API and Apache Flume 1.8.0

Step 2) Storage: The collected data is stored in a certain format Apache Hadoop (HDFS: Hadoop Distributed Filesystem) so as to form a key-value pair which is needed to feed to mapper in map-reduce programming approach.

Step 3) Data Pre-Processing: Data collected over a period of time is processed by using distributed processing software framework using Map-reduce programming model and queried using Apache HIVE 0.12.0.

Following are the steps for pre-processing the twitter data

1) Tokenization: All the words in a tweet are broken down into tokens. This is the tokenization process. For example, '@Jack That is an awesome car!' is broken down into individual tokens such as '@Jack', 'is', 'that', 'an', 'awesome', 'car'. Emoticons, abbreviations, hashtags, and URLs are recognized as individual tokens. Each word in a tweet is separated by a space. Therefore, on encountering a space, a token is identified.

2) Normalization: The normalization process verifies each token and performs some computing based on what kind of token it is.

Step4) Data Analysis: In this step the processed data is analyzed using the proposed hybrid method for classifying the tweets as positive, negative or neutral tweets.

Step5) Validation: in this step the results are validated and represented for decision making.

5. CONCLUSION

It has been observed that by implementing our proposed method most of the indirect opinions which were earlier treated as neutral are now giving a positive or negative sense, It is also adding up to the accuracy.

However, natural languages are not easy to interpret for machines, people have shown an inclination of giving opinions in an indirect way but we can constantly work to enhance the accuracy of it and it will lead us to an easy way of interpreting the opinions, resulting in providing vital information about launch and sale of any product in just 2-3 days and this will help in setting better strategy for marketing of any product.

6. FUTURE SCOPE

The future scope includes getting the weights of nouns such as celebrities, leaders or famous personalities automatically by using some API instead manually assigning it as in our hybrid approach. More semantic rules can be applied in dictionary-based approach for improving its accuracy.

We know that people use a lot of short words in their posts and tweets which cannot be looked up directly in the dictionary for assignment of weight so one can create a dictionary of short trending short words and use it for improvement of accuracy.

REFERENCES

[1]. F. T.Kai, C.C.Hui, "Blogger-centric contextual advertising", Expert Systems with Applications: An International Journal , ACM, pages 1777-1788, 2011.

[2]. W. Medhat, A. Hassan, H. Korashy., "Sentiment analysis algorithms and applications:A survey", Ain Shams Engineering Journal,Volume 5, Issue 4, pages 1093-1113, El-Abaseya, Egypt, 2014

[3]. X. Kaiquan, L. S. Shaoyi, L. Jiexun, S. Yuxia., " Mining comparative opinions from customer reviews for competitive intelligence", Decision Support Systems journal , Volume 50, Issue 4, pages 743-754, Hong-Kong, China, 2011.

[4]. A. Ingle, A. Kante, S. Samak, A. Kumari., "Sentiment Analysis of Twitter Data Using Hadoop", International Journal of Engineering Research and General Science, Volume 3, Issue 6, ISSN 2091-2730, Pune, India, 2015.

[5]. L. Zhou, B. Li, W. Gao, Z. Wei, K. Wong., " Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities", Conference on Empirical Methods in Natural Language Processing, pages 162–171, Edinburgh, Scotland, UK, 2011.

[6]. C. Zirn, M. Niepert, H. Stuckenschmidt, M. Strube., "Fine-grained sentiment analysis with structural features", International Joint Conference on Natural Language Processing, pages 336–344, Chiang Mai, Thailand, 2011.

[7]. N. Hu, I. Bose, N. S. Koh, L. Liu., "Manipulation of online reviews: an analysis of ratings, readability, and sentiments', Decision Support Systems journal , Volume 52, pages 674-684, Hong Kong, 2012.

[8]. I. Maks, P. Vossen., "A lexicon model for deep sentiment analysis and opinion mining application', Decision Support Systems journal , Volume 53, Issue 4, pages 680-688, The Netherlands, 2012.

[9]. L.D. Caro, M. Grella., "Sentiment analysis via dependency parsing", Computer Standards & Interfaces journal , Volume 35, pages 442-453, Turin, Italy, 2016.

[10]. T. Xu, P. Qinke, C. Yinzhao., "Identifying the semantic orientation of terms using S-HAL for sentiment analysis", Knowledge-Based Systems journals , Volume 35, pages 279-289, The Netherland, 2012.

[11]. K. Efstratios, B. Christos, D. Theologos, Z. Nick., "Ontology-based sentiment analysis of twitter posts", International Journal of Expert Systems with applications, volume 40, issue 10, pages 4065-4074, 2013.

[12] L.Y. Ming, L. T. Ying., "Deriving market intelligence from microblogs", Decision Supprot systems journal, volume 55, pages 206-217, 2013.

[13] W. Medhat, H. Korashy, A. Hassan, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal,Volume 5, Issue 4, Pages 1093-1113, 2014.

[14] A. Z. H. Khan, M. Atique, V. M. Thakare, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477, 2015.

[15] E. Cambria, S. Poria, R. Bajpai, "SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives", the 26th International Conference on Computational Linguistics: Technical Papers, pages 2666–2677, 2016.

[16] S. Rosenthal, N. Farra, P. Nakov, "SemEval-2017 Task 4: Sentiment Analysis in Twitter",11th International Workshop on Semantic Evaluations (SemEval-2017), pages 502–518,2017.

[17] A. S. Manek, P. D. Shenoy, M. C. Mohan, Venugopal K R," Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier", Springer link journal, Volume 20, Issue 2, pp 135–154, 2017.