I J C T A, 10(8), 2017, pp. 429-434 ISSN: 0974-5572

Sentiment Analysis of Social Media

Pranoti Kale[1], Ragini Ranjan[2], Tejal Bagade[3], Anjali Sapkal[4] and Kirti Singh[5]

ABSTRACT

Nowadays almost everyone is a part of a social media either for expressing views or for getting the opinions of others about any product, business organizations, industries, educational institutes etc. So to categorized these fluctuating views or opinions sentiments are analyzed and classified respectively with the help of lexicons, for better understanding of whether the person commenting on a subject is in favor of positive side or negative side and can even have no favor at either side (neutral). Basically sentiment analysis is the process in which subjective information is found from an extracted raw data. This paper describes how to classify the real time sentiments from social media along with accuracy and the speed. As to obtain the optimistic results our proposed system uses combination of two architectures explained below.

Keywords: lexicons, sentiment analysis, accuracy, speed, sentiment classification.

1. INTRODUCTION

There has been an increase in the social media (e.g., review sites, forum discussions, Blogs, Twitter, comments, and posts) content over the past few years. This huge amount of data present on the Web is being used by individuals and many organizations for a better decision making process as opinions present on these sites helps them to gauge a wider public opinion. In today's world, for buying a consumer product there is no need to depend on a limited circle of your friend's opinions, there are many review sites and forums which will provide you with all the necessary information about the product and thus helps you to make a decision. Similarly, organizations don't need to conduct polls and surveys as the opinions are present on the Web publicly. There is abundant amount of data present on various kinds of sites. Therefore, the most crucial task is to extract the useful information present on the internet and organize them in such a way that it assists many individuals and organizations.

An opinionated post about any issue is now recognized all over the world and thus has helped in formation of better business, social and political systems. Therefore it is need of the present world to explore and study opinions on the Web.

One approach of performing sentiment analysis is observing the opinion words and its usage. Opinion words are those words which are used to express the sentiments or opinions about any topic of interest. It can be a positive or negative opinion. Examples of positive opinion words are good, happy, wonderful and negative opinion words are bad, poor, awful, etc. A list or a dictionary containing all such opinion words is termed as opinion lexicon. This approach of using opinion words for sentiment analysis is purely lexicon based approach [1]. This approach has many advantages as it is useful and easy to analyze texts at various levels of sentiment analysis whether it is document, sentence or entity level. But this approach also has very issues associated with it like it is unable to analyze some expressions, abbreviations and emoticons used in the text [2]. For example a statement like: "I borrowed his headset and it sucksss!! :-)" cannot be

[1-5] BVCOEW, Department of Computer Engineering, Savitribai Phule Pune University, *Katraj, Pune, pin - 411043, Maharashtra, India, Emails: kpranoti2005@yahoo.com, raginiranjan13@gmail.com, bagadetejal@gmail.com, sapkalanjali17@gmail.com, kirtisingh24june1995@gmail.com*

understood by the lexicon based method, as it contains word like “sucksss” and emoticons like “:-)” which are not common opinion words so it will just consider the entire statement to be of neutral sentiment value. Moreover, meaning of a sentiment word can change depending on the application domains. For example consider two statements: “You have to do well in these exams” and “I am not feeling well”. The circumstantial meaning of the word “well” is different in both the domains. Also, any sentence does not necessarily have sentiments if it has any opinion word. For example consider this statement “Is this hat looking bad on me?” Here bad is not expressing any sentiment in this sentence.

Another approach for sentiment analysis is machine learning based methods [3]. This approach is highly adaptable to particular context and involves a classifier and a training set of data. It has the task of classifying opinions (positive, negative or neutral) in the text after analyzing it. Machine learning based methods can be classified further mainly into two types: supervised learning approach and unsupervised learning approach. Supervised learning includes a classifier which is trained on a manually labeled training set of data. Manual labeling is very efficient but also very time consuming and requires labor work. Whereas unsupervised learning method involves a classifier which works on the association principle which means the classifier compares the characteristics of a given text against the sentiment lexicons whose sentiment value is already known. After comparison the given text is classified to the category of the lexicon with which it matches and respective sentiment value is assigned to it.

One more approach can be to utilize both lexicon based method and learning based method [4]. Text will be analyzed at document level which includes opinions about a single entity and gives the final result in terms of sentiment values. Firstly lexicon based approach for sentiment classification is applied and a result is obtained having opinion words and its sentiment values. To extract more opinion words Chi-square test is applied on the already obtained results from the lexicon based method. This step helps in finding out more opinionated texts using the newly added opinion words (obtained from Chi- square test). In the second level, the new extra opinionated texts which could not be analyzed earlier are classified by the classifier and are assigned polarity values.

Lexicon based approach gives high accuracy but has many issues as discussed earlier so to increase the quality of data at the second stage learning based method is applied as it has high adaptability depending on the context’s requirements. The classifier will be able to train itself based on new trends and languages. Therefore classification and analysis of the sentiments can be done more effectively and qualitatively if lexicon based and learning based method both is applied.

For an unsupervised approach classifier can use the result of lexicon based approach as the training set of data automatically, thus requiring no labor and for a supervised approach a classifier can be build by manually training set of examples which will classify the opinions of the text, it is efficient but requires labor.

2. RELATED WORK

For the sentiment analysis, two main approaches are available: lexicon based approach or machine learning based approach.

The lexicon based approach [5] uses the opinion words for classification of sentiments into categories like: positive, negative or neutral. But this approach has many issues associated with it as discussed earlier.

The learning based approach [3] includes usage of classifiers which are trained upon a training set of data and has high flexibility to include new trends.

Another way is to utilize both lexicon based and learning based methods. For example: sentiment analysis of reviews [6] which classified the opinions into two categories only: positive and negative. This makes the problem much easier. A classifier was [7] that classify the tweets into positive, negative and neutral.

3. ARCHITECTURE

In the above figure, extracted data is provided to the preprocessor for removing irrelevant data. After that lexicon based method is applied to the data with the help of lexicon rules and the result is given to the sentiment classifier for further classification of the extra or new data in the document. Finally extracted opinionated data is classified.

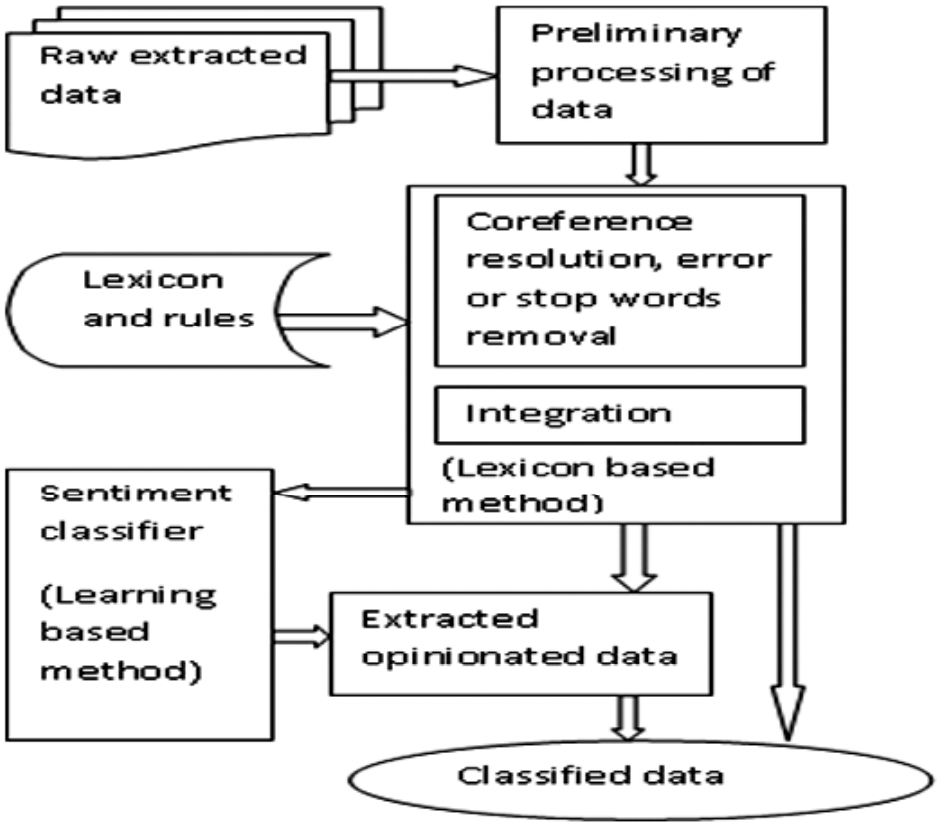

Figure 1: Architectural view

3.1. Lambda Architecture

For data processing, lambda architecture can be used as it has many advantages [8].

Lambda architecture provides four important attributes:

- Volume-Data size, Retention granular level
- Variety-Data sources, Data formats (semi-structured, mannered)
- Value-Quality of data, Improve data quality, deliver hidden insights
- Velocity-Speed of change, Speed of reaction.

The designing of architecture for activities of sentiment analysis of social network which are involved with Big Data needs to deviate from the traditional data warehouse or intelligent systems for businesses. Systems handling Big Data work with semi-structured data and hence, should be able to process and analysis data not only in Batch mode but also in Real-time mode.

Multiple features of Big data generated by the social media need to be taken care of. Most important features are the following:

(1) Dimensions

(2) uniqueness

(3) Source

(4) Reliability.

Initially, it was assumed that the data should always be processed and then made available, regardless of time aspect. This process is batch processing. Actually, the batch processing models do not work real time data, due to long time duration of operations. On the other hand, real-time architecture is implemented utilizing real-time data but with lower accuracy rate.

Therefore, a feasible solution for handling Big Data is to combine these two methods into a single architecture. This architecture with Batch and Real-time processing is called Lambda Architecture (Marz and Warren 2015). It includes three different layers [9]:

- Batch Layer
- Speed Layer
- Serving Layer

1) *Batch Layer:* This layer is responsible for storing the input datasets into the master dataset. Map reduce approach is used to periodically process datasets. Pre-computation of results are carried out

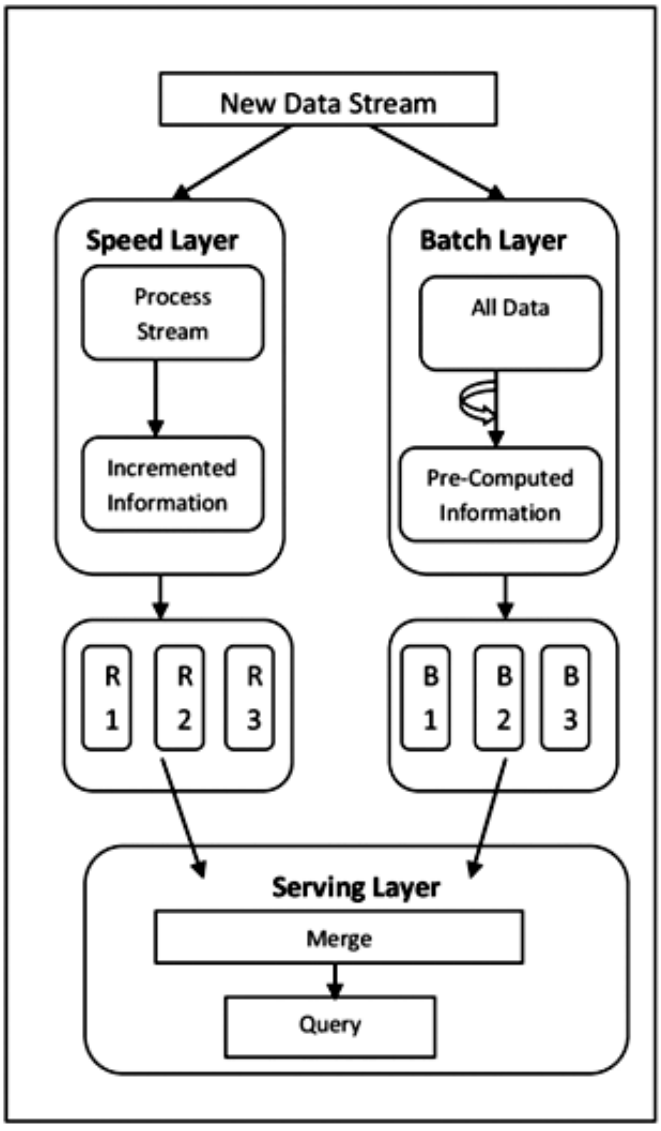

Figure 2: Lambda architecture for data processing

by batch layer using distributed systems. This layer can fix errors by re computing data set, then updating existing views.

2) *Speed Layer:* This layer is responsible for processing data streams in real-time. Main aim of this layer is to provide real-time views with the help of most recent data. The speed layer is used to fill the 'gap' caused by the Batch layer processing operations. This layer does not provide accurate results like the Batch layer, but they are immediately available after the data is received. These results will be replaced as soon as the Batch layer provides results for the same data.

3) *Serving Layer:* In this layer, the merging process of Batch layer and Speed Layer is executed. The merge process is necessary in order to obtain a single view of the results. Data synchronization is a critical task handled by this layer. There is a need to integrate a storage engine for all the random reads and bulk writes.

In the given lambda architecture given below includes R1, R2, R3 (Real-time views) and B1, B2, B3 (Batch views) respectively.

4. PROPOSED TECHNIQUES

4.1. Support vector machine

Support Vector Machines (SVM) have recently gained prominence in the field of machine learning and pattern classification. The technique works by maintaining a candidate Support Vector set. It utilizes a greedy method to pick points for inclusion in the candidate set. When a point added to the candidate set which is blocked due to the presence of other points in the set, backtracking method is used for pruning such points.

The technique begins with the nearest pair of points from opposite classes to speed up convergence. Hence optimization based method is used to prune those points in the candidate support vector set.

1) *Direct SVM:* The Direct SVM is an algorithm, which builds the Support Vector set incrementally. Though it has been proved that closest pair of points are the support vectors. Direct SVM starts off with this pair of points in the candidate Support Vector set.
The advantage of the Direct SVM algorithm is that it is geometrically motivated and simple to understand.

2) *Geometric SVM:* The Geometric SVM proposed by us improves the scaling behavior of the Direct SVM by using a method to optimize the solution to add points to the candidate Support Vector set.

For using Direct SVM the steps in processing are:

- Finding the Closest Pair of Points.
  First of all, the closest pair of points requires n^2 computations in the kernel space, where n represents the total number of data points.
- Adding a Point to the Support Vector Set.
  Let us consider a set S which contains only support vectors, add another Support Vector k to S.
- Pruning.
  Keep pruning points from S till k can actually become a Support Vector.
- Scaling.
  The memory requirements of the algorithm scale up as O ($|S|^2$) in the average case.

5. CHALLENGES

1) An opinion word that may contain positivity in one moment and negativity in other.

2) Situations govern people's opinions which lead to inaccuracy.
3) However, in the twitter or blogs, people are likely to combine different opinions in the same moment that is easy for a users to comprehend, but more difficult for a computer to parse.
4) Sometime people have difficulty in understanding what someone thought based on a short piece of text because it lacks context.
5) The main challenging aspects exist in use of other languages.
6) Dealing with negation expressions.
7) Produce a summary of opinions based on product features/attributes.
8) Dealing with complexity of sentence/ document.
9) Handling of implicit product features.

6. CONCLUSION

So the extraction of the data from web can be done successfully with the combined methods (lexicon and learning based methods), which gives a fair decision or favoring side for the person about a particular topic, event or organization etc. Even though with the proposed architecture classified sentiments can be obtained in less time, still the entire process is long, so take most recent 500 comments to increase the efficiency of the system. As to get more fair results the total sentiments must be classified based on percentages for distributed positive, negative and neutral side respectively from future work perspective. So the more informed decision can be obtained.

REFERENCES

[1] Ding, et al, "A Holistic Lexicon-based Approach to Opinion Mining", WSDM, 2008.
[2] Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, May 2012.
[3] Pang B and Lee L, "Opinion mining and sentiment analysis", Foundations and Trends in IR, 2008
[4] Lei Zhang, et al, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", HP Laboratories, 2011
[5] Hu, et al, "Mining and summarizing customer reviews", KDD '04, 2004. Hovy, et al, "Determining the Sentiment of Opinions", COLING'04, 2004. Taboada, et al, "Lexicon-based Methods for Sentiment Analysis", Journal of Computational Linguists, 2010
[6] Tan, et al. "Combing Learn-based and Lexicon-based Techniques for Sentiment Detection without Using Labeled Examples", SIGIR 2008.
[7] Park, A. and Paroubek, P. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", LREC 2010.
[8] Michael Hausenblas, "Implementing the Lambda architecture efficiently with Apache Spark", Hadoop Summit, Brussels, 2015.
[9] Michelle Di Capua et al., "An architecture for sentiment analysis in twitter", International Conference on e-learning, 2015.
[10] V. N. Vapnik. "The Nature of Statistical Learning Theory", Springer, New York, 2nd edition, 2000.
[11] S.V.N. Vishwanathan, M. Narasimha Murty. SSVM, "A Simple SVM Algorithm", Indian Institute of Science, Bangalore 560 012, INDIA.
[12] G. Vinodhini, RM. Chandrasekaran., "Sentiment analysis and opinion mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.