AN APPROACH FOR EFFICIENT ANALYSIS OF SOCIAL MEDIA DATA

Chetan Pandey
Graphic Era Hill University, Dehradun (UK), India

Durgaprasad Gangodkar
Graphic Era Deemed To Be University, Dehradun (UK), India

Varsha Mittal
Graphic Era Deemed To Be University, Dehradun (UK), India

ABSTRACT

Analysis of social media data is now become an important part of almost every organization. Organizations analyse the past and present records to get informative insights of customer feedback about services/products being offered. However, the analysis of data is mostly based on the point of view of emotions expressed in the dataset. A Data Analysis becomes a treasure for an organization when identification of expressed views are accompanied with the hidden information reside within a social media data. This paper will present a framework for Data Analysis which includes Sentiment Analysis and Feedback Analysis. The proposed framework is an approach to perform data analysis which not only discovers the sentiment views but also discovers the hidden knowledge in terms of suggestions and complaints. The proposed work is performed over a Twitter dataset by using N-Gram approach & Naïve Bayes classifier and is implement over Hadoop framework. Performance of proposed framework is examined by calculating Pre-cision, Recall, F1 Score and Accuracy, which are very satisfactory.

Keywords: Naive Bayes, N-Gram, Bag of Words (BOW), Hadoop, Map Reduce

Cite this Article: Chetan Pandey, Durgaprasad Gangodkar and Varsha Mittal, An Approach For Efficient Analysis of Social Media Data, International Journal of Advanced Research in Engineering and Technology (IJARET), 11(6), 2020, pp. 1157-1166

http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=11&IType=6

1. INTRODUCTION

Data Analysis [3] is the process of determining the hidden emotion and discovery of knowledge within a given dataset. Such analysis is used to gain an understanding of the attitudes, opinions and emotions expressed within a given social media dataset which are the feedbacks for the related organization. The analysed data would be either a user’s experience or user’s good
experience or user’s suggestion or user’s bad experience about a product or a service. In simple term Data Analysis [5] includes: collection of data, cleaning of data and finally transformation of resultant data in such a form that it must give some information and some knowledge. It is not wrong to say that a Data Analysis process is also a Knowledge Discovery [18] process.

Although there are sufficient research successfully completed in the field of Sentiment Analysis [2] [3]. It is observed that generally sentiments’ value [16] are either positive or negative or neutral, but also they are on the basis of groups of emotions & feelings, for example the sentiments’ value [16] may be like happy, sad, surprise, angry and more. If one consider such analysis from the organization’s point of view then one will wonder that for an organization it is very little helpful to figure out that N number of people have X type of sentiment among the data sets which they provided for Data Analysis. There must be some additional information after analysis which must be beneficial for an organization. It is also observed that it is good to say that N number of people have positive sentiment regarding some service or product but point to be noted here is that “positive” word also consist more positive and most positive also. This means that if one give rank to all possible sentiment values, like most positive, more positive and positive, then there “positive” has lower rank. An organization may re-implement or re-plan their strategy regarding a particular product or service after getting output of Data Analysis. So if an organization come to know that for a product X they have let 86% positive sentiment then they may terminate their re-implementation plan, which is surely not good for them because “positive” has lower rank. There is also a possibility that may be this 86% positive figure is in actual has only 11% most or more positive sentiment and rest 75% is a positive sentiment, means 75% has lower rank sentiment. Note that these 75% also means then they fell just “OK” type satisfaction with that service or product and there is much probability that these 75% may be shifted to the services or products of other organization, so this need to be addressed in Data Analysis Report. It is also observed that the end result of a Data Analysis must contain such information which make an organization come to know that what their customer or users have problems and what are the suggestions they have for them. Since a person who actually using a product or service will really know how good the product or service is, benefits of it, problems with it, what type of improvement are possible and so more. That means we have to mine the given data set in such a manner that at end we have problems, suggestions and types of sentiment values [5] [10], which is observe that is much beneficial for an organization and after all it is beneficial for us also since we are also a part or user of that service or product.

The authors’ of this paper proposed a framework which is capable of performing complete analysis of any social media data in a batch processing as well as real time processing. This paper covers the discussion of some related works of data analysis, then brief description of proposed approach with details of experiment setup, then analysis of the suggested methodology along with algorithms, then we observe the results we obtained on the basis of performance parameters and finally conclusion & future work will be discussed.

2. RELATED WORKS
After the year 2005 there is sudden increase in social media communication and there are presently lots of such platform like Facebook and Twitter. On one side it allows people from different place to communicate and share their thoughts with each other and on the other side it gives an opportunity of Data Analytics. Some authors [1] implemented Data Analytics on the Hadoop’s Map Reduce for storing big data across different data centres. They concluded that large amount of data required large number of distributed computing centres for storing and accessing the database and for that Map Reduce is most acceptable and successful approach. Most of the research are carried with the help of either machine learning approach or Natural
Language Processing (NLP) approach for evaluating the sentiment of a social media data set. SentiWordNet, Fuzzy logics, Naïve Bayes, Neural Network, etc. are some Machine Learning approach. One of the author [2] proposed an algorithm named matrix based Fuzzy system to mine the accurate opinion among the twitter dataset. They first construct a full list of tweets from the extracted unstructured twitter dataset. Finally Fuzzy Logic technique is applied on each tweet to identify the emotion word and its sentiment score w.r.t. SentiWordNet. NLP is a technique of understanding human verbal & written language and it is very easy to use & understand and also they are very effective. Collaborative Filtering, Part of Speech, Bag of Words, etc. are some of most commonly used NLP approach. Another author [3] proposed a work in which they implement NLP along with Hadoop framework. Their suggested work is based on Map Reduce approach in which Collaborative filtering is used to generate recommendation on the basis of user data and NLP along with Text analysis techniques is used to predict the sentiment on the basis of polarity of all possible sentiment values and it is set between +3 (positive) to -3 (negative). Pre-processing is an initial and must-to-implement process which was discussed in detail by a author [4] in which there is a suggestion about the importance of pre-processing and need to apply properly before any data analysis algorithm. To understand its importance and effects, they compare four supervised classification method: Support Vector Method (SVM), Naïve Bayes (NB), Logistic Regression (LR) and Random Forest (RF). An effective pre-processed dataset yields better analysis result and also reduce the time complexity of an algorithm. One author [5] used Naïve Bayes classifier, for which stop-word removal (pre-processing) plays a vital role, for their research work in which they uses Hadoops Word Count to perform Sentiment and Feedback analysis. They focuses on one point that Feedback is valuable content along with the Sentiment of that analysed data.

By analysing some more papers it is observed by authors that although there was much research implemented in the area of data analysis but still there are some parameters needs to explore more in order to obtained an efficient analysis of data. Following are the observed parameters:

- Most research works are based on broad division of classes of emotions like good/bad, happy/angry/sad, etc. There is need of sub-division of these classes on the basis of their comparative and superlative degrees.
- There are many words whose sentiments are depend on the word which is next to it, for example "damm cool" and "damm fool". As authors observed there are no clear explanations available to identify the sentiments of such words.
- Some sentences' words individually has no sentiment, but collectively they express some sentiment like "feel like home". Previous research papers does not talk much about it and how to identify the hidden sentiment value.
- Authors of this paper uses Naive Bayes Algorithm for their proposed work. But authors observe the dominance of prior probability while making the prediction over test data. This need to be resolved in order to get more accurate analysis result but there is no discussion about this issue in previous research papers.

3. PROPOSED METHODOLOGY AND ALGORITHM

The proposed framework, given in Fig. 1, for the Analysis of Social Media Data (for example Twitter Data) covers two processes, one is Sentiment Analysis [4] [5] and another is Feedback Analysis [5] [10]. This framework make an effort to provide a detailed sentiment analysis in which an organization have seven type of sentiments. Seven sentiment values according to their rank from positive to negative are: Extreme Positive, Moderate Positive, Positive, Negative, Moderate Negative and Extreme Negative, these are six sentiment values and the seventh is
Neutral which means neither positive nor negative. Note that here authors distributed a sentiment value on the basis of comparative and superlative degree of an adjective. Thus a sentiment value represents either positive or negative, comparative degree represents the “moderate” value of a sentiment and superlative degree represents the “extreme” value of a sentiment.

One more thing is that the proposed framework also works on Feedback Analysis which means it also provide all those sentences among the given data set which have some suggestions & complaints and which are definitely beneficial for an organization. Authors first implemented above discussed framework by using only Naive Bayes, but later authors' experiences some challenges which are discussed in the next section along with their proposed solution.

Figure 1 Proposed Framework by using N-Gram and Naive Bayes Classifier

To achieve better result, the dataset must be concise and specific. For that authors selected Twitter's dataset according to its hashtag value [14]. The collected raw dataset will be then pre-processed with the help of Bag of Words (BOW) which include correct spelling of wrong ones (e.g. acheives→achieves), stop words (e.g. to, another, for, etc.), full form of short form word (abt→about) and more. After pre-processing of dataset, firstly N-Gram algorithm will be applied over each tweet of dataset. This will include only N-Gram of 2 & 3 words only and identifies that whether sentiment is present or not. If sentiment is present then it proceeds further and the algorithm will discover the exact type of that sentiment for the given tweet. However if N-Gram fails to find a sentiment present in a tweet, Proposed Framework will then only use Naive Bayes and definitely discover the sentiment for the given tweet. Once all tweet of dataset is processed, then a Graph is drawn for the discovered sentiments along with their percentage. After that, same raw dataset is used for Feedback extraction. There is a separate Pre-processing procedure for Feedback extraction after which all tweets containing feedback will be extracted.

3.1. Challenges with Naive Bayes framework and Proposed Solution

Our proposed approach is basically based on Naïve Bayes classification algorithm and N-Gram is used to enhance its reliability & accuracy. Before discussing about the algorithm used, let first discuss about the challenges authors’ faced while implementing only Naive Bayes:

- Naive Bayes usually calculate the posterior probability of each word of a given tweet, either those words exists in the BOW or not. Since Naïve Bayes process words of a tweet one by one, it ignores those comparative or superlative degree of a word (which has more than one word like "most valuable", "more expensive", etc.) and mislead the sentiment analysis process. This can resolve by incorporating N-Gram approach in which we make use of 2-Gram model.

- Another challenge exists that some words may behave positively or negatively. For example “features of this phone are damn cool” shows positive sentiment but ‘damn’ in sentence “you are damn fool dude” shows negative sentiment. Note that “damn” is a negative word but such type of words are completely depend on the
word (may be adjective) next to it, like here “fool” and “cool”. This can also resolve by incorporating N-Gram approach by including 2-Gram model so that the proposed algorithm will analyse each couple of words and will predict "damm cool".

- It is also observe that many tweets contains such words which does not express any sentiment separately but still there is some hidden sentiment within it together. For example, a tweet about a hotel say that “In this hotel I feel like home”. If we observe each word separately then it is clear that none of them express any sentiment. However there is a hidden “positive sentiment” which this tweet express and it is because “feel like home” together express a positive sense of sentiment. This can be resolve by incorporating N-Gram approach in which we make use of 2-Gram and 3-Gram model which results into an effective analysis of data.

- Apart from above three challenges one more challenges author faced is that the dataset used as test dataset will misbehaves if all classes or group of train dataset does not have equal number of elements. This happens specially when one class of train dataset have very high number of elements which means it also has high prior probability value. During calculation of posterior probability, if likelihood value of two predicted class has very little difference then instead of predicting correct value it becomes dependent over that class which have highest value of prior probability. This can resolve by supervising train dataset and ensuring that each class must contains equal number of elements. This will make value of prior probability of each class equal and remove the dependency of it for predicting the correct class value.

3.2. Proposed Methodology

Following are the suggested version of tools and system requirements require to implement the proposed framework: CPU – Intel Core i3, RAM – 8 GB, HDD – 500 GB, Operating System – Windows 10 (64 bit), Java – jdk1.8.0_144, Eclipse – Oxygen Release (4.7.0) and Hadoop – Hadoop-2.7.4. Below are the steps of Proposed Methodology which will be performed with the help of the proposed algorithm (will discuss it in next section):

1. Pre-Processing of a twitter dataset obtained from different sources. Here prior to this process, it is must to make three different collection of some words which is technically known as BOW. Then apply Removal of noisy data, Replace all short form of words with their original form according to the BOW, Replace all wrong spelling with correct spelling according to the BOW and Remove all words contained by ‘Stop Words’ BOW.

2. Make a BOW, which is a testing dataset for Naïve Bayes classifier, which contain sentiment type and its corresponding word. Make sure each sentiment type must contains equal number of words.

3. Apply proposed classification algorithms (given in next section), which will include N-Gram and Naïve Bayes.

4. Collect the final sentiment value of each tweet and find the percentage value for each sentiment type and construct a graph according to the percentage obtained.

5. Make a BOW which will contains the words, phrases and clauses which are related to some suggestions, complaints, emotions and more.

6. Find those tweets which have the elements as given in the above BOW.

7. Print the graph, sentiment analysis result in percentage form and above extracted tweets along under feedback analysis in a separate document file.
3.3. Data Analysis by using proposed framework on Hadoop

By incorporating N-Gram with Naïve Bayes, the above discussed issues are successfully resolved and the reliability & accuracy of the proposed framework is also enhanced, which definitely helps an organization to better understand the actual sentiment of their users. The proposed framework firstly implement the N-Gram [13] and then Naïve Bayes [13] because here authors’ suggest that if a sentence have sentiments in two or three words, for example “more comfortable”, “not good”, “lower than expected”, etc., then its’ better to analyze them with the help of N-Gram first, since Naïve Bayes cause issues with such words as already discussed in above points. In N-Gram, authors included only 2-Gram and 3-Gram, however one can take up to N-Gram where N < M where M is the total number of words present in a tweet. By considering N = 2 and 3, one can easily include all superlative degree of a word like “most valuable”, comparative degree of a word like “more valuable”, those words which may behave as positive or negative like “damn cool”/”damn fool” and those words in which single word have no sentiment but together they express some sentiments like “feel like home” expresses a positive sentiment. If N-Gram does not identify any sentiment value, then that sentence is further processed by Naïve Bayes which definitely identify some sentiment value. Note that all BOW used while the proposed methodology must be manually prepared according to the content of twitter dataset.

**Table 1** Algorithm for finding Sentiment and Feedback of given Twitter Dataset by implementing our proposed algorithm on Hadoop framework

| Step 1. Begin |
| Step 2. Start Hadoop |
| Step 3. Define input file name with extension alongwith output folder name as an arguments |
| Step 4. Set Mapper Class, Combiner Class and Reducer Class and check for J1.waitForCompletion(true) |
| Step 5. Loop for each instance str1 of map() |
| String tweet ← Pre-processing of str1 |
| String Sentiment1, Sentiment2 ← Call 3 Gram, Call 2 Gram |
| If Sentiment1 and Sentiment2 has no value (i.e. no sentiment value) |
| String Sentiment3 ← Call Naïve Bayes Classifier |
| String FinalSentiment ← Sentiment3 |
| Else |
| Compare values of Sentiment1 and Sentiment2 and select highest sentiment type according to their priority |
| String FinalSentiment ← Max(Sentiment1, Sentiment2) |
| End If |
| End Loop |
| Step 6. Find percentage of each sentiment type and draw Graph |
| Step 7. Make BOW which contain feedback word(s) and select tweets which have at least one word from it |
| Step 8. Stop Hadoop |
| Step 9. End |

The detailed steps of proposed algorithm are discussed in Table 4.1 and note that alias are used in place of sentiment type, which are given in the bracket as following: Extreme Positive (EP), Moderate Positive (MP), Positive (P), Negative (N), Moderate Negative (MN), Extreme Negative (EN) and Neutral (Ne). According to the algorithm first we have to set up a Hadoop Environment and synchronize Mapper and Reducer code alongwith Combiner code. In Step 5, first N-Gram (3-Gram and 2-Gram) will be executed which will find sentiments of a group of 2 and 3 words only. If some sentiment values (one is for 3-Gram and another for 2-Gram) are extracted then both values will be compared and the one with highest rank is the "FinalSentiment" value for that processed tweet. However if N-Gram is fails to extract any sentiment value then Naïve Bayes will be executed which will definitely extract some "FinalSentiment" value from the processed tweet. Step 5 will repeat till all the tweets after Pre-
An Approach For Efficient Analysis of Social Media Data

processing will be processed. After that, percentage of each sentiment value along with with pie chart will be calculated. To extract feedbacks from the raw tweets, authors prefer a separate pre-processing so that the meaning of a tweet will never change. Authors also use a BOW which includes complaints and suggestions showing word/words. At the end, by using later bag, the algorithm will find all such tweets which contains some complaints or some suggestions. After its successful execution, the algorithm will show a pie chart which represent all possible sentiment types with percentage and all those tweets which have some feedback.

4 RESULT ANALYSIS

4.1. Dataset Detail

Authors implemented the proposed framework by using Twitter data. The suggested dataset is on the basis of hash tag value “#OneNationOneTax” [14] which contains more than 15K tweets. All steps for extracting Twitter data like first visiting Twitter's developer site, creating access tokens and rest are followed as given in [14]. The proposed framework is based on a supervised approach and BOW [16] concept is used for the training purpose. Figure 5.2 (a) & (b) shows a screenshot of BOW for Naïve Bayes & N-Grams respectively in which all words are based on sentiment values i.e. Extreme Positive, Moderate Positive, Positive, Negative, Moderate Negative and Extreme Negative. Figure 5.3 (c) showing BOW which is used for analysing those tweets which have some suggestions or complaints or both.

<table>
<thead>
<tr>
<th>Positive</th>
<th>Moderate Positive</th>
<th>Negative</th>
<th>Moderate Negative</th>
<th>Extreme Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>@more approvals</td>
<td>@better</td>
<td>@meh</td>
<td>@thems down</td>
</tr>
<tr>
<td>@resolution</td>
<td>@better margins</td>
<td>@more than expected</td>
<td>@@cashing in</td>
<td>@complaints</td>
</tr>
<tr>
<td>@successful</td>
<td>@annoying</td>
<td>@annoying</td>
<td>@more handsome</td>
<td>@suggest</td>
</tr>
<tr>
<td>@jalousy</td>
<td>@more helpful</td>
<td>@most handsome</td>
<td>@most helpful</td>
<td>@advise</td>
</tr>
<tr>
<td>@joyed</td>
<td>@feeling like home</td>
<td>@place</td>
<td>@convenient to use</td>
<td>@please</td>
</tr>
<tr>
<td>@overjoyed</td>
<td>@cloud nine</td>
<td>@sorry to say</td>
<td>@think about it</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2 Screenshot showing (a) BOW for Naïve Bayes, (b) BOW for N-Gram and (c) BOW for Feedback Extraction

4.2. Result

http://www.iaeume.com/IJARET/index.asp 1163  editor@iaeme.com
The proposed algorithm providing satisfactory results and is compatible with any size of dataset. The proposed methodology is successfully implemented on Hadoop [6] which make the data analysis more powerful in terms of parallel and distributed processing. Authors has performed the proposed framework on Hadoop with Single Node Single Cluster in Windows Environment [8]. To analyse the compatibility and performance of the proposed algorithm, authors has implemented it with the help of one Mapper and one Reducer class. Mapper class split the input data one by one and pass the result to Reducer class which finally calculate the total number of each sentiment value. These values are finally access by a control program which manage all activities of Mapper & Reducer class, and with the help of Key/Values the sentiment values are calculated. Finally control program will also identify all possible feedbacks from the given twitter dataset. The sentiment values along with a pie-chart and feedback result will be as shown in Fig. 3 (a) and (b).

The proposed framework is reliable & accurate and in order to validate the result authors opted the four performance parameter i.e. Precision, Recall, F1 Score and Accuracy.

After implementation, the values for Precision, Recall, F1 Score & Accuracy for the Naïve Bayes are 83%, 72%, 77% & 85% respectively and for the proposed framework the values are 83%, 96%, 89% & 93% respectively. So all performance parameters have greater value for the proposed framework, as shown in the graph of Figure 5.3, which validate that the proposed framework provide an efficient analysis of social media data. The algorithm for the proposed framework is proposed in such a way that it maintain the scalability feature of Hadoop and can analyze small to big dataset in a static time as well as in a real time scenario. The proposed
framework is a conjunction of Naive Bayes & N-Gram and here author conclude that N-Gram not only successfully strengthen the Naive Bayes but also demonstrate an effective use of a simple technique along with a more strong technique which perform in depth data analysis which include Sentiment analysis and Feedback analysis together in a single framework by a single algorithm.

5 CONCLUSION AND FUTURE WORK
Data Analysis work must ends with such an output that must have some information which help an organization to enhance its business, improve its services and make better future strategy. The proposed framework is capable of identifying the detailed sentiment results and also capable of extracting the hidden sentiment message in terms of user’s problem, their feedbacks and suggestions. The proposed methodology is reliable & accurate as well as scalable and is successfully implemented for small dataset as well as for large dataset with Hadoop. However this framework is strictly depends on BOWs and as much as BOWs will be strong this framework will also become strong. So it is suggested that instead of using others constructed BOWs, make your own BOWs and also take care that BOWs must be constructed according to the related product or services for which this suggested methodology will be used.

Although the proposed methodology is effectively performing the data analysis process but still there are some areas which need some improvements. It is observed that the results are varies and it depends very much on the strength of BOWs. Future work will include: auto-updating BOW, extract the particular part of a tweet which has some information and further implementing the proposed framework over a multi-node Hadoop cluster which will boost its performance &processing time while analyzing real time big tweeter datasets.

REFERENCES


