Introducing an Automated System for Social Media Data Analytics

Ashish Deshpande[1] Anuradha Thakare[2] Anish Katkamwar[3] Mahima Khandelwal[4] Srushti Kharat[5]
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]Pimpri Chinchwad College of Engineering, Pune University, India

Abstract— Social media is not only source of communication between people but also a source of communication between brands and their potential and existing customers. The term influential marketing though coined in recent past, has immense psychological effect over people. Decision to buy a specific product or choose a service is highly influenced by "Feeds" in social media. Feedback in the form of reviews play an important role in decision making of brands to improve their products and services according to expectations and needs of customers. Hence there is need to design and develop a system for social media data analytics which helps the business to build their brand image from users perspective. This article proposed an automated system for social media data analytics. The system keeps an eye on competitor's strategy giving us an advantage in the global market. The system keeps an eye on competitor's strategy giving us an advantage in the global market. Age volume of social media data will be pre-processed using modern engineering tools. A system will be implemented with hadoop cluster. For the conversion of unstructured data to structured form, the JSON files which are in the form of key-value form will be stored in relational database, preferably hive. An important goal is to analyze this social media data from user's perspective. In order to achieve this, various classification algorithms are analyzed for various business strategies and it is observed that predictive Naive Bayes model supersedes the existing algorithms. It is expected that the proposed listener will help to improve the brand image for any business and engage with audience in a direct manner.

Key words: Influential Marketing, Feeds, Social Media Data, Data Analysis, Relational Database

I. INTRODUCTION

With the introduction of internet, people now are connected to each other more closely than ever. Opinions can be expressed almost instantly irrespective of the location and on any topics. Social media, serves as a platform not only for personal use but also for commercial use. The huge data that it generates every second is of immense value to the brands. The advertisements and marketing strategies are designed to target particular society or age group. Brands are turning towards the digital platform than the print media. Statistics have shown a huge rise in the revenue earned through advertisements. The response to a particular digital campaign or to the launch of product can be obtained in the terms of numerical statistics. This helps to plan the future campaigns, right time to launch the product, and the customer's expectation from the product. Also once a product is launched, the immediate response and reaction that customer's express through their opinions, feeds or posts is the deciding factor whether the product launch is successful or not. Responding to the customer's grievances and complaint which is a part of customer service has increasingly changed from traditional phone calls to tweets and comments on the brand's social handle.

However the concern over analyzing this social media data and its accuracy is increasing due to introduction of bot's and fake user's. It is very easy for the competitors to malign the image of brand's through trolls and fake reviews. Targeted advertisements also cause nuisance to the customers who are not interested in the product. The manner in which personalized advertisements are formulated poses a huge risk to safety and privacy of the user's data.

II. RELATED RESEARCH

There are various papers which describe the strategy to classify feeds in the form of reviews using various approach.

A strategy to classify tweets sentiment using Naive Bayes techniques based on trainers perception into three categories; positive, negative or neutral was proposed[1].Specific keywords are chosen for demonstrating ability if Naive Bayes for sentiment classification in political and business environment. This method is suitable to train and classify sentiment from twitter and high degree of accuracy can be achieved using Naive Bayes technique.

Twitter reviews have been using machine learning algorithm like Naive Bayes and logistic regression for classification[2].Hadoop along with Mahout have been used to implement both the classifier. The performance has been evaluated on the basis of different parameter like accuracy, precision and throughput.

Attempt to examine the impact of different number training data set on the accuracy of sentiment classification using Naive Bayes techniques[3]. The accuracy level of the analysis increases as the number of data set of training data set increases upto a certain threshold value.

The analysis of the contents on the web which are growing exponentially in numbers as well as in volumes as sites are dedicated to specific types of products and they specialize in collecting users reviews from various sites. A set of techniques of machine learning with semantic analysis for classifying the sentence ad product reviews based on twitter data is proposed[4].The Naive bayes technique gives better result than maximum entropy and SVM.

The family of naive bayes classifiers is used for detecting the polarity of English tweets. The experiments have shown that the best performance is achieved by using a binary classifier trained to detect just two categories: positive and negative[5].In addition, in order to detect tweets with and without polarity, the system makes use of a very basic rule that searches for polarity words within the analysed tweets/texts.

Issues on Naive Bayes are discussed along with its advantages and disadvantages [6]. It also presents an overview of Naive Bayes variants and provide a categorization of those methods based on four dimensions. These include manipulating the set of attributes, allowing interdependencies, employing local learning and adjusting the probabilities by numeric weights. Here several methods are reviewed for improving Naive Bayes algorithm Text

Document Classification is a task of classifying a document into predefined categories based on the contents of the document. From the survey the inference made is that the Naïve Bayes technique performs better and yields higher classification accuracy when combined with the other techniques[7].

Sr. No.	Title	Author/Publication	Strength
1.	Twitter Sentiment Classification Using Naïve Bayes Based on Trainer Perception	Mohd Naim Mohd Ibrahim,Mohd Zaliman Mohd Yusoff"Twitter Sentiment Classification Using Naïve Bayes Based on Trainer Perception"2015 IEEE Conference on e-Learning, e-Management and e-Services.	Specific keywords are chosen for demonstrating ability of Naïve Bayes for sentiment classification in political and business environment.
2.	Sentiment classification on Big Data using Naïve Bayes and Logistic Regression	Anjuman Prabhat, Vikas Khullar 2017 International Conference on Computer Communication and Informatics (ICCCI - 2017), Jan. 05 – 07, 2017, Coimbatore, INDIA.	Machine learning algorithm like Naïve Bayes and logistics regression for classification task. Hadoop and Mahout have been used
3.	The Impact of Different Training Data Set on the Accuracy of Sentiment Classification of Naïve Bayes Technique	Mohd Naim Mohd Ibrahim, Mohd Zaliman Mohd Yusoff2017 IEEE Conference on Open Systems (ICOS), November 13-14, 2017, Miri, Sarawak, Malaysia	The accuracy level of the analysis increases as the number of data set of training data set increase up to a certain point.
4.	Sentiment Analysis of Twitter Data Using Machine Learning Approaches and semantic analysis	G.Gautam, D.Yadav Seventh International Conference on contemporary computing(IEEE), pp. 437-442,2014.	A set of techniques of machine learning with semantic analysis for classifying the sentence and product reviews based on twitter data is proposed.
5	Citius:A Naive-Bayes Strategy for Sentiment Analysis on English Tweets	Pablo Gamallo,Marcos Garcia, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 171–175, Dublin, Ireland, August 23-24, 2014.	Naive bayes classifiers is used for detecting the polarity of English tweets. The best performance is achieved by using a binary classifier to detect just two categories: positive and negative.
6	Naïve Bayes Variants in Classification Learning	Khadija Mohammad Al-Aidaroos1, Azuraliza Abu Bakar2 and Zalinda Othman3	It provides categorization of methods based on four dimensions. These include manipulating the set of attributes, allowing interdependencies, employing local learning and adjusting the probabilities by numeric weights.
7	A Survey of Naïve Bayes Machine Learning approach in Text Document Classification	Vidhya.K.A, G.Aghila IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 2, 2010	Text Document Classification is a task of classifying a document into predefined categories based on the contents of the document.

Table 1: Depicts Review of Related Research

III. PROPOSED SYSTEM

Data acquisition is the first step towards building the listener. This can be done in mainly two ways either Web scraping or through API's. Web scraping or crawling is not supported by all websites and is considered as grey area. API's of social site's provide a good tool to mine the data and embed the code as required. The data that is acquired can be converted into JSON format where it is stored in the form of key-value pair. This file is then stored in hadoop cluster. Since large amount of data is to be mined it would be convenient to store data into a database where fast processing can be done. The data that would be acquired is in raw format. There are many unwanted and stop words which won't be contributing towards the analysis of data. The preprocessing module removes all the unnecessary words and sentence. After the preprocessing is done using NLTK library sentiment analysis is performed on the processed data set. In predictive analysis we use machine learning algorithms to train the data set. The knowledge base will contain the exact data on which analysis is to be done. This analysis is the final stage which will be customized according to the requirement. The output of this analysis is then represented in visual form using charts, graphs and summaries to the end user.

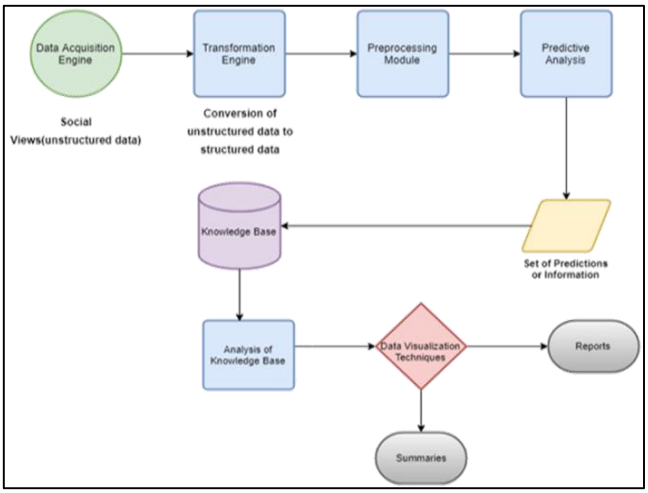

Fig. 1: Represents Workflow of Proposed System

Data acquisition would be done using python language as the libraries like beautiful soup and requests handle web scraping very well. The data is stored in hive. Data transformation is done using NLTK. Classification of data set is done primarily using Naive Bayes algorithm. Visual representation of analysis in form of reports would be done using C3.js.The web application is developed using HTML, Node JS and AngularJS.

IV. CONCLUSION

Thus, we have studied and analyzed existing approaches for data mining in social media. We have also designed computational model for proposed work. Both social media and person-to-person information-gathering have value, but social media listening is quickly becoming an important customer intelligence tool. There are several ways to use social media to gain insight, including monitoring online customer support forums, using software tools to gather comments from social outlets such as Facebook and Twitter and encouraging customers to suggest new product features and vote on their favorites.

REFERENCES

[1] Mohd Naim Mohd Ibrahim, Mohd Zaliman Mohd tYusoff "Twitter Sentiment Classification Using Naïve Bayes Based on Trainer Perception"2015 IEEE Conference on e-Learning, e-Management and e-Services.
[2] Anjuman Prabhat,Vikas Khullar,"Sentiment classification on Big Data using Naive Bayes and Logstic Regression"2017 International Conference on Computer Communication and Informatics (ICCCI-2017),Jan. 05-07,2017,Coimbatore,INDIA.
[3] Mohd Naim Mohd Ibrahim,Mohd Zaliman Mohd Yusoff, "The Impact of Different Training Data Set on the Accuracy of Sentiment Classification of Naive Bayes Technique"2017 IEEE Conference on Open Systems(ICOS),November3-14,2017,Miri,Sarawak,Malaysia
[4] G. Gautam, D.Yadav, "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and semantic analysis.
[5] Pablo Gamallo,Marcos Garcia,"Citius:A Naive-Bayes Strategy for Sentiment Analysis on English Tweets", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 171–175, Dublin, Ireland, August 23-24, 2014.
[6] Khadija Mohammad Al-Aidaroos1, Azuraliza Abu Bakar2 and Zalinda Othman3,"Naïve Bayes Variants in Classification Learning".
[7] Vidhya.K.A, G.Aghila ,"A Survey of Naïve Bayes Machine Learning approach in Text Document Classification ",(IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 2, 2010

