Imperial Journal of Interdisciplinary Research (IJIR)
Vol-3, Issue-2, 2017
ISSN: 2454-1362, http://www.onlinejournal.in

Predication of Sales Using Sentiment Analysis in Twitter Using Big Data

Prof. Asadullah Shaikh, Afjal Shaikh, Imran Siddiqui & Azeem Shaikh
Dept. of Computer Engineering, M.H. Saboo Siddik College of Engineering, Mumbai, India

Abstract*—On social media network, people write many things about the product to share their experience with friends and acquaintances. The data generated from such activity is very large. Company of any product may want to know their customers' reviews. To read each customer's reviews is not possible by the company. So this all data can be process to generate a conclusion which includes the overall customers' reviews which helps company to know all their customer and subsequently it can take further decision based on the conclusion obtained from the system to increase their business. This paper discuss how to take all customer's reviews and process them using algorithm. To store and process the huge data, Hadoop is used.*

1. Introduction

With increasing use of social network like in today's world, data is growing day by day. This data is useful in prediction of trends, marketing, and opinions of anything from users of social network. The huge data will be generated from social media. This data is useful to carry out analysis on it and deduce some conclusion which makes the analyzer aware of certain facts about people. This conclusion also gives the business idea to the product vendors based on the reviews posted by the user about the products.

To maintain huge amount of data some type of storage is needed for storing and processing it. This huge data is called Big Data. Big Data analysis is building system around generated data. Single computer will take much processing time and it doesn't have storage capacity to store this huge data. To solve the problem of Big Data, Hadoop is provided by Apache Software Foundation in 2011. Hadoop was created by Doug Cutting and Mike Cafarella in 2005.

Hadoop is an open source framework written in java language that supports the processing and storage of extremely large data sets in a distributed computing environment. All modules of this framework ensures that hardware failure can occur and it automatically handled by framework. Apache Hadoop is a framework for running applications on large cluster built of commodity hardware[11]. Hadoop services provide for data storage, data processing, data access, data governance, security, and operations. Twitter, Yahoo (One of the biggest user & more than 80% code contributor to Hadoop), Facebook, Netflix, Amazon, Adobe, etc.

Prediction of sale of a product such as mobile phone brand Samsung Galaxy J5 can be made based on analysis of customer reviews about this product. Similarly other products like shoes, clothing, watches, computers and many other products can be analyzed by our system using the tweets from the twitter and will be fed to our system for further analysis to give out the précised results. Data about any product can be collected from Twitter. This Data (i.e. tweets) can be fed to Hadoop for storing and processing. Hadoop uses parallel processing system, so it can provide the result after processing in less time. The processing part of Hadoop is classification of customers' reviews into positive, negative, and neutral reviews. The numbers of reviews are positive, negative and neutral will be shown as output which can be used to predict the sales of the product by the product manager of the product. The product manager can predict if there is more positive reviews about the product indicates that people are liking their product and can provide the more supply of the product. If product manager find negative reviews, then product can be improved by product manager after knowing about customers reviews about the product. The result obtained from Hadoop can be represented using some programming language in some graphical or tabular representation form to get the overall reviews about the product which will help the buyer and users to purchase product of their choice with ease and hassle free.

2. Literature Survey

2.1. Software Requirement Specification

The SRS specifies requirements of the system. It provide how system will perform in real time environment, its interaction with user and what is the outcome from the system. The SRS provides the complete description of the system. The SRS

provides the complete description of the system. It includes the following:

2.2. Introduction

This system includes interaction between the user and the twitter data i.e. data will be extracted and sentiment analysis will be performed on the products' reviews which is in the form of tweets.

2.3. Overall Description

It includes product perspective and description, product details, users' opinions and reviews about the product, Consideration of the product by the company from the reviews, probability of purchasing a product, constraints, updating of tweets i.e. latest tweets

2.4.Specific requirements

System requires database extraction tool i.e. Hadoop, VMware, Oracle, Eclipse, NetBeans, java development kit.

2.5. Other Requirements

It includes pre-processing module which has keywords, stop words, abbreviations, hash tags, symbols, and emoticons meaning which will be useful in filtering the data using NLP.

2.6. Functional requirement

System should take input file contain tweets extracted from twitter website using twitter API.
System should be able to break the huge collection of tweets in form of files of small collection of tweets into different data nodes (i.e. different system) for storing using HDFS. This set of data nodes is the cluster which make the parallel processing of data to do processing in less time.

Different data nodes should be able to map the algorithm with the data for data processing. Algorithm is to classify the data into positive, negative, or neutral with the help of predefined keywords for positive and negative in algorithm. After processing done by each data node, Reduce module of Hadoop gather the result from each data node and give the result of huge data which is fed as input. For this processing, MapReduce module of Hadoop is used. System should provide the percentage or number of positive, negative tweets.

2.7. Natural Language Processing (NLP)

The effective way of communication is "language" for human. In the case of computer, the area of research and application which shows how computer is used to process and understand the natural language text and voice speeches. The main aim of NLP is to analyzing and understanding and base on gathered information it generates a language in both written and spoken form. NLP is nothing but the making the computer to understand how humans learn and uses language. In this proposed system, NLP is used in algorithm to make the system understand the sentiment of each sentence. Positive, negative, and neutral sentiment score is assigned to each sentence in a file, then average of the positive sentiment scores give the positive sentiment score of the file. Similarly, sentiment score for negative and neutral of a file can be calculated.

3. Existing System

The existing systems which provide sentiment analysis on live tweets, have the below system as a generalized model, where they give input in the form of keyword, as a result of which the tweets consisting of those keywords are fetched or retrieved in a file or dialog box. Those tweets then undergo pre-processing to remove unwanted material and transform those raw tweets into an understandable format, so the classification algorithms can work on clean data. The pre-processed data is then subjected to a classification mechanism which differentiates the words and performs operations on them to achieve the result of the intended operation. After the scores to different words of the documents is applied and they have underwent the operations of algorithm, sentences are differentiated based on the polarity. And overall sentiment score is then evaluated over the entire document.

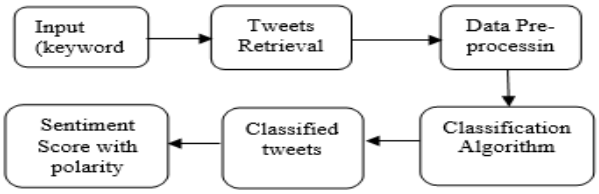

Figure 1. Existing System

4. Proposed System

The proposed system gives to analyze live Tweets. The system makes use of Hadoop's Map Reduce features to make fast the execution speed of the system. The files are stored in HDFS. HDFS stands for Hadoop Distributed File System.

4.1. Working of the system

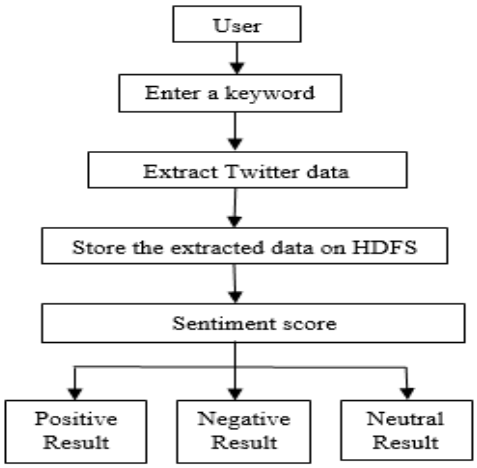

Figure 2. Proposed system

User of the system need to enter the keyword on twitter website whose sentiment user wants to know. This data will be extracted from twitter API and with the help of Hadoop, Sentiment Analysis of the data can be done. NLP (Natural Language Processing) is used to make system understand the language based on some predefined keywords to calculate sentiment score.

4.2. Algorithm

Input:
D: Dataset of product reviews
Output:
Dataset sentiment score
Processing:

1. Read Dataset D.
   a. Enter the keyword for extracting live tweets.
   b. Store the extracted tweets on HDFS.
2. Perform Pre-processing
   a. Removal of raw data such as @, #, http removal.
   b. Remove all slang words.
   c. Remove stop words.
3. Classification:
   a. Classify words based on positive and negative dictionary.
   b. Calculate the frequency of each word in the document.
   c. Classify the document and label them as positive, negative, and neutral based on the frequency of word present in each line of the document (i.e. more positive or more negative).
   d. Calculate the sentiment score for each document.
   e. Calculate the sentiment score for the entire dataset.

5. Display the output.

4.3. Expected Results

Get the sentiment score of overall dataset (i.e. tweets from Twitter by searching of product on Twitter using the product name or keyword). List of classified tweets will be represented as a following table of Samsung Galaxy J5. This table will be generated by the system.

Sentiments	Percentage
Negative	30%
Positive	50%
Neutral	20%

Figure 3. Sentiment Analysis on twitter data of Samsung Galaxy J5

5.Future Scope

Our system has capability to analyze live tweets and process tweets to give output to the user. User only needs to enter the keyword on Twitter and extract that data to feed as input to the system. We can extend the capability of the system by providing an option to do sentiment analysis on the Election prediction and movie review prediction. Other things like social media monitoring and survey analysis will also be done as and when required.So the system should be like that,that it can perform sentiment analysis in any field

6. Conclusion

The company and user of any product have to go through all the reviews posted by their customer about their product on social media platform Twitter. Sometime there are so many reviews including positive, negative and neutral reviews. It takes much time and effort for company to deduce any conclusions from so many reviews manually. Company cannot consider only some of the reviews; it has to consider all customers' reviews about the product to get the accurate conclusion. Our system automate this process to save time, effort of the company and give them accurate sentiments of their customers by considering all customer's reviews and also help the customer to buy products from the results obtained.

7. Acknowledgements

Authors would like to express heartfelt and sincere gratitude to the principal Dr. Mohiuddin Ahmed, M. H. Saboo Siddik College of Engineering, Mumbai, for providing the facilities to carry out the work. Authors also wish to express grateful thanks to beloved Head of the Department of Computer Engineering Dr. Z. A. Usmani, who gave us full support and constant encouragement, valuable suggestions and helping tendency which has made us

to carry out and finish the work successfully. Finally, this work might not have been possible had it not been for the efforts of our internal project guide Prof. Asadullah Shaikh, for valuable suggestions and constant encouragement for successful completion of the work

8. References

[1] RemcoDijkman, Panagiotis Ipeirotis, FreekAertsen, Roy van Helden "Using Twitter to Predict Sales: A Case Study" on March 2015

[2] Manoj Kumar Danthala (Author) Dept. Computer Science Engineering Keshav Memorial Institute of Technology (KMIT) "Tweet Analysis: Twitter Data processing Using Apache hadoop**"** on 11 feb 2015

[3] Mr. Swapnil A. Kale, Prof. SangramS.Dandge, Understanding the Big Data problems and their solutions using Hadoop MapReduce, ISSN 2319 – 4847,Volume 3.

[4] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in the 26th IEEE Symposium on Mass Storage Systems and Technologies, pp. 1-10, May 2010.

[5] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters,"
Communications of the ACM, Vol. 51, Iss. 1, pp. 107-113, January 2008.

[6]Bo Pang and Lillian Lee "Opinion Mining and Sentiment Analysis" on oct 2013

[7] Dipak Gaikar,BijithMarakarkandy "Product Sales Prediction Based on SentimentAnalysis Using Twitter Data" on nov 2015

[8] Precise tweet classification and sentiment analysis
RabiaBatool; Asad Masood Khattak; Jahanzeb Maqbool; Sungyoung Lee
2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)

[9]http://hortonworks.com/apache/hadoop/

[10]http://hadoop.apache.org/

[11]http://docslide.us/engineering/big-data-hadoop-55b0d92e4b8d5.html