International Journal of Engineering Research and General Science Volume 3, Issue 6, November-December, 2015
ISSN 2091-2730

Sentiment Analysis of Twitter Data Using Hadoop

Ajinkya Ingle, Anjali Kante, Shriya Samak, Anita Kumari

PES's, MCOE Department Of Computer Engineering, Pune-05, ajinkyaingle05@gmail.com, Mob.:8793646961

Abstract—In today's highly developed world, every minute, people around the globe express themselves via various platforms on the Web. And in each minute, a huge amount of unstructured data is generated. This data is in the form of text which is gathered from forums and social media websites. Such data is termed as big data. User opinions are related to a wide range of topics like politics, latest gadgets and products. These opinions can be mined using various technologies and are of utmost importance to make predictions or for one-to-one consumer marketing since they directly convey the viewpoint of the masses. Here we propose to analyse the sentiments of Twitter users through their tweets in order to extract what they think. Hence we are using hadoop for sentiment analysis which will process the huge amount of data on a hadoop cluster faster.

Keywords— Opinion Mining, Sentiment analysis, Hadoop Cluster, Twitter, Unstructured data, Movie review analysis, Tokenisation.

INTRODUCTION

Sentiment Analysis:

Sentiment analysis also known as opinion mining. The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine the writer's attitude towards a particular topic or product. Sentiment Analysis is the process of detecting the contextual polarity of text. In other words, it determines whether a piece of writing is positive, negative or neutral.

Twitter Data:

Twitter, one of the largest social media site receives tweets in millions every day in the range of Zettabyte per year. This huge amount of raw data can be used for industrial or business purpose by organizing according to our requirement and processing.

About this Project:

In this project, we are going to implement a system in Hadoop which analyses twitter data where cluster of nodes will be formed. Twitter data is in the form of comments which are nothing but sentiments that is opinions, feelings of people. This data will be collected by using Twitter API. By analysing this data, our system will give output in the form of positive, negative and neutral tweets. In this case, it makes the use of data dictionary for classifying the data. This data can be used further according to particular application. And this analysed data can be represented in the form of pie-charts.

Motivation:

Today we are living in the world which is surrounded by 99% of data. There are different microblogging sites where users express their views about different products these views are nothing but opinions of people and it will go waste if it is not used in proper way so there is a need to use opinions of people in improving productivity, usefulness, functionality of particular product or application or technique or any entertainment resource. Hence, there is a need to develop a product which can analyse opinions of people. This product will be useful in increasing market value of industries as well as satisfy needs of customers.

BRIEF DESCRIPTION

Need of Sentiment Analysis:

Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. Social media monitoring tools like Brand-watch Analytics make that process quicker and easier than ever before. The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organizations across the world. Shifts in sentiment on social media have been shown to correlate with shifts in the stock market. The Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential election.

The ability to quickly understand consumer attitudes and react accordingly is something that Expedia Canada took advantage of when they noticed that there was a steady increase in negative feedback to the music used in one of their television adverts.

The core objective of the project is:-
1) Content Retrieval: The large amount of data is collected using java Twitter streaming API.
2) Storage: This data is stored in a certain format (HDFS: Hadoop Distributed File system) so as to form key value pair which is needed to feed to mapper in map-reduce programming approach. The data is stored in Hadoop Distributed File System.
3) Data Processing: Data collected over a period of time is processed by using java and distributed processing software framework developed by Apache Hadoop and using map reduce programming model and Apache hive frame work.
4) Data Analysis: The output obtained from reducer phase is analysed.
5) Data Representation: Representation of classified data in the form of pie charts.
6) At the end we will get the outcome in the form of classified tweets that is Positive, Negative and Neutral tweets.

This project will mainly analyse the predefined stored twitter data and classify it based on polarity.
Analysis of data consist following steps:

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', 'That', 'is', '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.

- If the token is an emoticon, its corresponding polarity is taken into account by searching the emoticon dictionary.
- If the token is an acronym, it is checked in the acronym dictionary and the full form is stored as individual tokens.
- Intensifiers such as 'AWESOME' are converted into lowercase and the token is stored as 'awesome'.
- Spelling of character repetitions such as 'veryyyyy' are first corrected into 'very' and then stored as 'very'.
- The normalization process also discards all those tokens which, in no way, contribute to the sentiment of a tweet such tokens are called stop word. It also discards URL's.

For analyzing the tweets, we have to take polarity into consideration using various types of dictionaries.

1) Lexical Dictionary:

It mainly consists of most of the English words which will help us to analyze the tweets by matching the word in the tweet with the words in the lexical dictionary. It also consists of idioms, phrases, headwords and multiwords.

2) Acronym Dictionary:

It is used to expand all the abbreviations and acronyms which will further generate words which can be analyzed using lexical dictionary.

3) Emoticon Dictionary:

A tweet containing emoticons can be analyzed by using this dictionary. Emoticons are basically the textual portrayal of the tweeter's mode which conveys some meaning.

4) Stop Words Dictionary:

These are the words in the tweet which do not have any polarity and they need not be analyzed. So they are eliminated and tagged as stop words. We maintain a dictionary with the list of all stop words for example able, are, both, etc.

Sentiment Classifier:

The tweets are broken down into tokens where each token is assigned polarity which is a floating point number ranging from 1 to -1.

A. Positive Tweets:

Positive tweets are the tweets which show a good or positive response towards something. For example tweets such as "It was an inspiring movie!!!" or "Best movie ever".

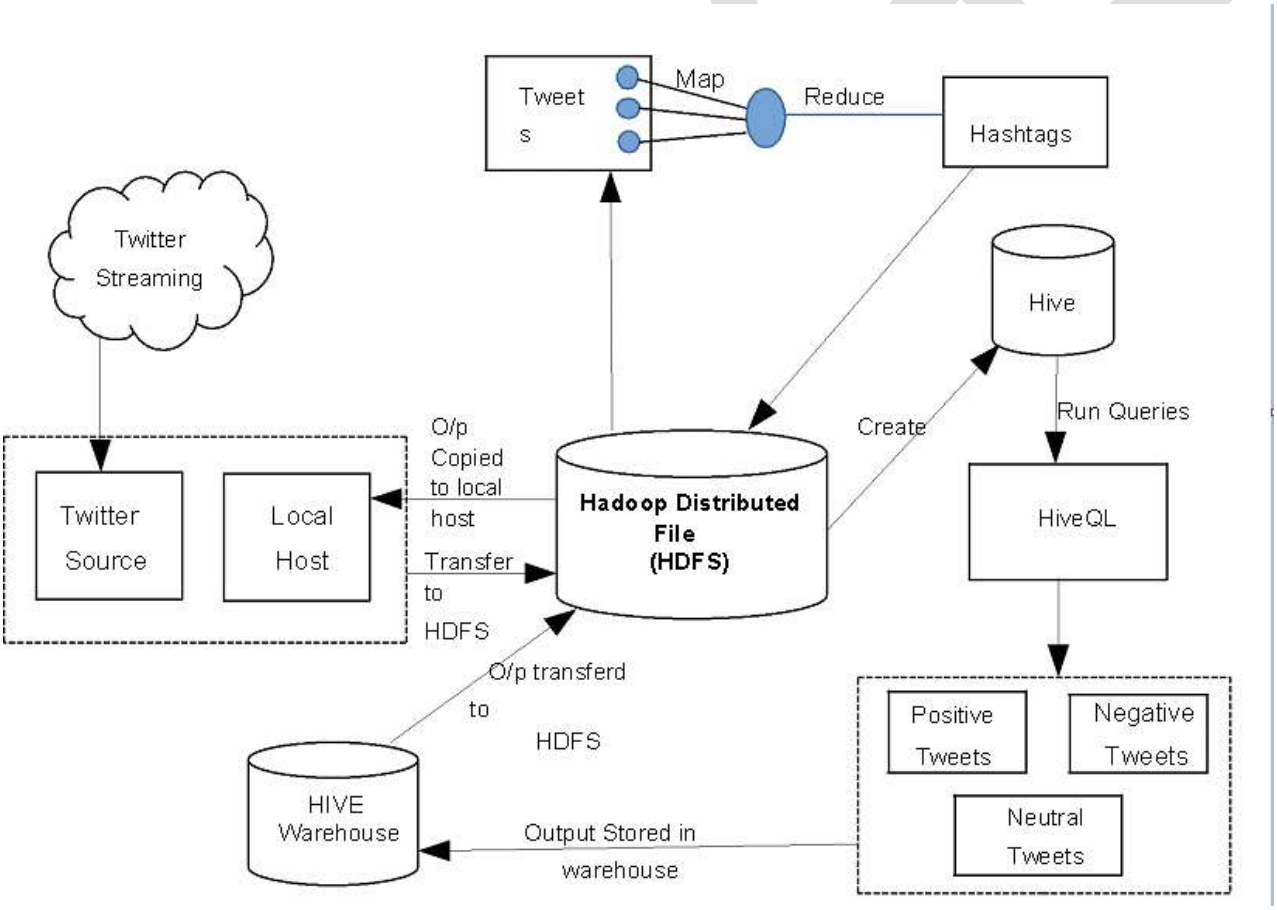
B. Negative Tweets:
Negative tweets can be classified as the tweets which show a negative response or oppose towards something. For example tweets such as "Waste of time" or "Worst movie ever".

C. Neutral Tweets:
Neutral tweets can be classified as the tweets which neither show a support or appreciate anything nor oppose or depreciate it. It also includes tweets which are facts or theories. For example tweets such as "Earth is round".

3. Part-of-speech Tagging:
The valid tokens are then passed to the part-of-speech tagger which attaches a tag to each token, specifying whether it's a noun, verb, adverb, adjective etc. Part-of-speech tagging helps determine the sentiment of the overall tweet because words have different meanings when represented as different parts of speech.

4. Classification**:**
At the end system will classify the twitter data into Positive, Negative, Neutral reviews with the help of data dictionary.

System Architecture:



CONCLUSION

This project will give us hands on experience of handling and parallel processing of huge amount of data. Data collection process will introduce us to Java twitter streaming API. We will get exposure to work with prominent parallel data processing tool: Hadoop.

Apache Hadoop framework is gaining significant momentum from both industry and academia as the volume of data to analyse growth rapidly. This project will help us not only to gain knowledge about installation and configuration of hadoop distributed file system but also map reduce programming model. Amongst the many fields of analysis, there is one field where humans have dominated the machines more than any – the ability to analyse sentiment, or sentiment analysis.

The future of this data analysis field is vast. This project not only analyses the sentiments of the user but also computes other results like the user with maximum friends/followers, top tweets etc. hence hadoop can also be effectively used to compute such results in order to determine the current trends with respect to particular topics. This can be very useful in the marketing sector.

REFERENCES:

[1] Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang, "SentiView: Sentiment Analysis and Visualization for Internet Popular Topics", IEEE Transactions On Human-Machine Systems, Vol. 43, No. 6, November 2013

[2] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, "Sentiment Analysis of Twitter Data", Department of Computer Science, Columbia University

[3] Jianshu Weng, Ee-Peng Lim, Jing Jiang, Qi He, "TwitterRank: Finding Topic-sensitive Influential Twitterers", WSDM'10, February 4–6, 2010, New York City, New York, USA Copyright 2010 ACM

[4] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media Rushabh Mehta, Dhaval Mehta, Disha Chheda, Charmi Shah and Pramila M. Chawan, "Sentiment Analysis and Influence Tracking using Twitter" in International Journal of Advanced Research in Computer Science and Electronics Engineering, Vol 1, Isuue 2, May 2012

[5] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) Aditya Pal & Scott Counts, "Identifying Topical Authorities in Microblogs", WSDM'11, February 9–12, 2011, Hong Kong, China, Copyright 2011 ACM

[6] Hive kiwi at http://www.apache.org/hadoop/hive.

[7] Hadoop Map-Reduce Tutorial at http://hadoop.apache.org/common/docs/current/mapred_tutorial.html.

[8] Hadoop HDFS User Guide at http://hadoop.apache.org/common/docs/current/hdfs_user_guide.html.

[9] Hive Performance Benchmark. Available at http://issues.apache.org/jira/browse/HIVE-396

[10] Running TPC-H queries on Hive. Available at http://issues.apache.org/jira/browse/HIVE-600

[11] Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.

[12] Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., Varma, V.: Mining sentiments from tweets . Proceedings of the WASSA 12 (2012)

[13] A. Bifet, E. Frank, Sentiment Knowledge Discovery in Twitter Streaming Data, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 1–15.

[14] A. Cui, M. Zhang, Y. Liu, S. Ma, Emotion Tokens: Bridging the Gap among Multilingual Twitter Sentiment Analysis, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 238–249.