Twitter Sentiment Analysis on Big Data in Spark Framework

S Nirmala Sugirtha Rajini¹, K Anuradha², S Umadevi ³ and E Mercy Beulah⁴

¹,²,³,⁴ Department of Computer Applications, Dr. MGR. Educational and Research Institute, Chennai, India.
¹sugirtharaja77@yahoo.com, ²anu.navine@gmail.com, ³umakalyani70@gmail.com, ⁴mercybeulah@yahoo.co.in

Abstract. Earlier for communication between people is something which is delivered by hand from one person to another. In some words, or letter Social Media an important and also an integral part of everyone’s life now days. Twitter is a social media site and its primary purposes are to convert people and allow people to share their thoughts. Twitter an American Microbe Using twitter social media users post their views, opinions and to communicate with messages called as tweets. It is one of the social media that is gaining popularity now days. As of 2014 twitter social media has more than 284 million users are in active in monthly basis and above 500 million messengers sent messages on a daily base. Twitter is created by Jack Dorsey, Noah Class, Biz stone and Evan Williams in March 2006 and it was launched in the month of July in that year. Sentimental analysis means the analysis and classification of the emotions and feelings such as positive thought, negative thought and neutral thought using text analysis techniques within text data. The aim of the proposed analysis is to identify the public opinion using NLP (Natural Language Processing) with n-gram stemming algorithm in Spark framework.

Keywords: Twitter, Sentiment Analysis, Spark, Framework.

1. Introduction

Social media is one of the major platforms to share feelings. Social media contents are the one of the resource issues various services for people to covey opinions about people, services, goods or themselves. From the various social network media twitter is one of the important media where large number of peoples transfer their feeling and ideas. The twitter data is not sufficient to analyze because limited characters are used in the twitter message. By using this twitter data collect the sentiment of the various people. Extraction of proper data is used for business and research purposes. Sentiment of the software peoples mainly controls the quality and productivity of the software.

In this proposed system the tweets which the users tweet can be analyzed using the sentimental analysis the reaction of the users can be known. All the reaction such as positive, negative or neutral using sentimental can be analyzed.

2. Literature survey

Yash Sharma et al., demonstrates a new method of analyzing people opinion available on social network media with the help of Language Processing technique. Here twitter data can be collected from
stakeholders of IIT. The output contains the basic emotional values [1]. Richard O. Sinnott et al., studies different analysis methods and estimate various sentiment analysis methods and evaluate their accurateness of the sentiment study on twitter social media in Australia. Here the authors presented a framework based on cloud [2]. Swati Powar et al., proposes architecture for fragment tweet data by using batch processing framework known as TweetSeg. It consists of limited background data with common data for attain enhanced solution for entity recognition. To find the significant data from the twitter data Hadoop architecture can be used here [3].

Yonas Woldemariam introduces the development and combination of a sentiment data analysis into cross media architecture pipeline. This pipeline consists of cleaner for chat, NLP and analyzer for sentiment. Here hadoop architecture with prediction concept based on lexicon and RNTN form can be used to sentiment data analysis. This proposed concept achieves 9.88% exactness of positive, negative and neutral comments on variable length data. But this method shows the better level for classification of positive type comments [4].

Jin Ding et al., conducted a sentiment analysis based on entity point. Initially they develop a dataset with 3000 comments from GitHub. Then the authors developed SentiSW tool for classification and recognize entity. To assess the performance using ten-fold type validation method, it attains 68.71 precision values, 63.98 recall values and 77.19 exactness values [5]. Abdullah Alfarrarjeh et al., proposes a new architecture for analyzing sentiment data by using disaster data value. This new architecture provides the solution for three challenges. The proposed framework can be tested by using Twitter and Flickr dataset values during the time of Napa earth shaking [6].

3. Proposed System Architecture
This proposed system to identify the public opinion using NLP with n-gram stemming algorithm in Spark architecture. This “Trending Topic Analyzer” is developed using theme based sentiment data classification and multi type tweet summarized data. It aims at creating ordered sub summary data, which understanding the topic development of the recent trends. Theme under learning is assumed to contain several unseen sub topics are exposed with the help proposed modelling. This enables a user to completely investigate a trending matter to a better level.

Target data gathering is performed by removing the trending theme on a area basis. Pre-processing steps are used to arrange the end user information. In pre-processing phase, the system also manages non-English tweet that is necessary to avoid disposal of public views about the current trend topics.

Pre-processed data is later analyzed through theme adaptive sentiment data classification to identify the public opinion. Sentiment labelled data will then be processed for sequential summarization. Sub topic detection is achieved using Stream based approach and Semantic based approach. Finally, sub summary candidate selection is executed which gives the highest scored tweets incorporating certain exclusive features of tweets. Redundancy check is performed to take away duplicates from the selected tweet data followed by a threshold check to make sure fair combination of user’s opinions.

Sentimental analysis allows business to find the user sentiment through their tweets. The term sentimental analysis defined as the use of natural language text analysis processing, computational notations and biometric to thoroughly identify, take out quantify a study the subjective data. The following figure 1 demonstrates the proposed framework used to analyze the twitter data.
Apache Spark is one of the accepted tools used for real-time data processing. Technology is changing from ancient Map-reduce to Apache Spark because of its high speed. Apache Spark is up to 100 times quicker than Map-reduce which become the major cause for various users to move towards Spark. Spark architecture from Apache Foundation is another open source for big data processing which is built specially to overcome the restrictions from the traditional map reduces the jobs. This framework is one of the moving technologies in recent years for the big data growth. Memory abstraction is the most important feature and ability of Spark which enables the allocation of data, it is otherwise called as in-memory data sharing, across the various stages of a map-reduce task.

4. System Implementation

The proposed system can be divided into various phases. The first phase of this system is data extraction. Here the twitter data can be extracted by using flume. It is one of the software used to collect, integrate and transfer large volume of data.

Distinguished characteristics of Flume:

Flume gathers log data from several web servers and stored on centralized area efficiently. With the help of Flume, collect the data from several servers immediately and transfer into Hadoop. Flume is used to extract large amount of event data generated by social media. Flume helps a huge set of sources and destinations types.

Normally tweet information contains noisy data because in twitter the user enters the content with limited letters. So, cleaning and produce meaningful data is important for any type of data analysis. In the pre-processing phase URL link is removed from the tweet data and short terms are also exchanged with proper words.

Detailed pre-processing steps are shown in Figure 2 along with the appropriate algorithm.
In feature extraction phase create the major impact of the performance of classify sentiment data. Because tweet data contains @ symbol (refer the user) and emoticons. Feature extraction process improves the feature removal process. By using text feature extraction common sentiment words and adoptive sentiment words are extracted. Emoticons and network based features are extracted by using non text feature extraction method. This proposed system uses semi supervised classification model is used to classify the data with common values and mixed labelled date from different topics. The final phase is sentiment analysis. In this phase twitter text data is processing and tokenizing. Finding accuracy and predicting scores by using NGRAM classification stemming algorithm.

Algorithm

Establish connection with spark and start the session.
Extract data from twitter through flume and load data
Pre-process data
Do tokenize for each partitioning
Feature extraction using train dataset
Classification of data using binary classification algorithm
Nature Language processing using NGRAM analysis and stemming algorithm
Find each token or word counting
Combine all the word count of all the partitions
Predict negative and positive sentiment words
Find prediction accuracy from is count
Show the prediction and analysis report

5. Result and Discussion
Data science world is full of attractive methods and algorithms to take out unseen insights from the data sources. Twitter is one of the famous social network media in this current digital world. It can be used from normal people to famous personalities also. But it contains large volume of data. Analysis twitter data is very difficult because it contains various symbols and nonverbal words. The analysis can be very useful to various marketing people and researchers to predict the result. Here twitter data can be analyzed by using Spark framework with python programming. The following figure 3 shows the sample screen shot of the proposed classification system result. The proposed system is tested by using kaggle dataset.
The following figure 4 a) and 4 b) shows the classification result of positive and negative sentiment data.

**Figure 3.** Sentiment classification results

**Figure 4.** a) Classification of Positive and Negative Sentiment in SPARK environment

**Figure 4.** b) Classification of Positive and Negative Sentiment
6. Conclusion
Social network is playing a major role in human life. Traditionally peoples convey their opinions and emotion through letter for or telephone. But now most of the people share their feelings and ideas through social network. Social network media data analysis is used to predict the future value. Here positive and negative sentiment data are extracted from tweet data. This system is implemented by using spark environment. Sentiment analysis of Twitter social media using big data helped us to analyze large amount of datasets. Sentimental analysis to identify the public opinion using NLP with n-gram stemming algorithm in Spark framework and the report is generated.

References
[1] Sharma Y, Malik I and Jagwani P 2019 Dual Aspect Sentiment Analysis for Opinion Mining. *IEEE Int. Conf. on Electrical, Computer and Communication Technologies (ICECCT)* 1-4
[2] Sinnott RO and Cui S 2016 Benchmarking sentiment analysis approaches on the cloud *IEEE 22nd Int. Conf. on Parallel and Distributed Systems (ICPADS)* 695-704
[3] Powar S and Shinde S 2017 Named entity recognition and tweet sentiment derived from tweet segmentation using hadoop *1st Int. Conf. on Intelligent Systems and Information Management (ICISIM)* 194-198
[4] Woldemariam Y 2016 Sentiment analysis in a cross-media analysis framework *IEEE Int. Conf. on Big Data Analysis (ICBDA)* 1-5
[5] Ding J, Sun H, Wang X and Liu X 2018 Entity-level sentiment analysis of issue comments *Proc. of the 3rd Int. Workshop on Emotion Awareness in Software Engineering* 7-13.
[6] Alfarrarjeh A, Agrawal S, Kim SH and Shahabi C 2017 Geo-spatial multimedia sentiment analysis in disasters *IEEE Int. Conf. on Data Science and Advanced Analytics (DSAA)* 193-202