International Journal of Engineering Technology and Management Sciences[IJETMS]

Website: ijetms.in Issue:1, Volume No.2, January 2018

SAMPLING TECHNIQUES FOR TWITTER DATA STREAM USING SENTIMENT ANALYSIS

D.Yazhini priyanka[1], K.S.Suganya[2]

[1,2] *Assistant professor,Bannari Amman Institute of Technology*

[1]yazhinipriyanka@bitsathy.ac.in

[2]suganyaks@bitsathy.ac.in

***Abstract*— Streaming data arrive rapidly in a continuous manner and it may vary in a timely manner. Thus, analyzing the streaming data is a critical task because the flow of the data is enormous and it is impractical to preserve the data completely. Hence, Sampling and Sketches techniques can be applied for analyzing the streaming data. In our proposed work, various Sampling techniques were incorporated to extract the sample dataset from the complete dataset. Thereby, the sample dataset mirrors the properties of the original dataset. Four sampling techniques such as Reservoir Sampling (RS), Bernoulli's Acceptance/Rejection Sampling (A/R), Weighted Random Sampling (WRS) and Hybrid Sampling (HS) were used to choose the sample from the whole dataset. Twitter, Facebook, and LinkedIn etc are the most popular networking site which gives rise to streams of data. Experiments were carried out by using the stream of tweets. Sentiment Analysis (SA) of tweets were carried out by using RS, A/R, WRS and Hybrid sampling techniques. It was observed that the sample dataset is providing the imprecise results when associated with the whole dataset results. Comparative analysis was done for the selected sample dataset which was carried out by the sampling techniques and for the whole dataset. It was perceived that the hybrid sampling technique is best suitable and efficient sampling technique for the twitter streaming dataset.**

.

***Keywords*— Twitter API, Data preprocessing, Four Sampling Techniques, Sentiment analysis.**

I. INTRODUCTION

In numerous current applications, data do not take the form of persistent relations, yet rather enters in the continuous, fast, large amount, time-changing data streams. It acts as a data source for many applications. One feasible methodology for managing the data streams is to yield a sample and do an analysis on the yielded sample. Since approximate results are enough for some data mining applications. Sampling is the most widely utilized and best-understood approximation procedure. Sampling can be used as a synopsis structure, which relies on obtaining the crucial attributes of the data set.

Sampling is the process used in statistical analysis in which the predetermined number of observation taken from the large population. The sampling algorithms are classified as uniform sampling and biased sampling (Wenyu Hu et al., 2012). In uniform random sampling, in which every item in the original data set has the same probability of being sampled, in contrast with biased sampling, each and every element has the distinct probability of being choosen as samples from the original dataset. Reservoir Sampling (RS) and Bernoulli's Acceptance and Rejection Sampling (A/R) are under the classification of uniform sampling whereas the Weighted Sampling (WS) is under the classification of biased sampling.

In the proposed method, four sampling techniques such as reservoir sampling, Bernoulli's acceptance, and rejection sampling, weighted sampling and Hybrid sampling (reservoir and weighted sampling) are used to select the sample from the original data.

Among a wide range of social networking service, Twitter, as a micro-blogging service is the second popular social communication website. Twitter is a massive networking site tuned towards quick communication. More than 140 million dynamic clients distribute more than 400 million "Tweets" each day. With its uncommon limitation that just 140 characters can be entered in every tweet, Twitter provides the first-rate position to social community investigation. Many researchers have focused on Twitter for social network analysis. The twitter dataset has been taken as the input for our proposed work.

SA is a standout amongst the most as often as possible utilized system for analyzing twitter. The reason for the SA is to distinguish the feeling and order the extremity of the content. SA is otherwise called opinion mining procedure, to discover the feeling of the tweet, for example, positive, negative and neutral. There are three characterization levels of SA procedures. They are document level, sentence level, and perspective level SA.

The Document-level SA arrange the feeling of the report, however the subject of the archive is uniform until the end. The Sentence level SA characterize the sentiment of the sentence. The Aspect-level SA arrange the notion of the viewpoints and substance. The Twitter tweets are considered as the sentence. Hence, Sentence level SA has been used for analysing the tweets in this work

II. RELATED WORK

A. Collection of dataset

Twitter has an Application Programming Interface (API) for programmatically accessing tweets by query term. Tweets are collected through twitter API (Pulkit Goyal, Sapan Diwakar, 2011)(Alec Go *et al*., 2009). Gathering Twitter information (tweets and metadata) starts with distinguishing the theme of enthusiasm utilizing a keyword(s) or hashtag(s), and requires the utilization of APIs. This API technique permits getting 1% of the openly accessible Twitter information. Twitter information is additionally accessible through data suppliers (e.g., GNIP, DataSift), otherwise called Twitter Firehoses, which can convey 100% of the Twitter information in light of criteria.

The data accumulation and sorts of collecting tweets are quickly examined in (Bongsug Kevin Chae, 2015). There are three sorts of data accumulation, for example, Geo-tagged tweets, Tweets about a subject, Tweets from a group of people. Geo-tagged tweets are the tweets containing Geolocation. Twitter's Tweet with Location highlight permits clients to specifically add area data to their Tweets. Tweet about a topic is the tweet can be extracted based on the keyword. Tweets from a group of the user are the tweet can be extracted from the group of user and their friends and followers.

The study showed that the tweets have been crawled from the twitter. Twitter Authentication (RAuth) was required to extract the twitter. In the proposed work, the twitter dataset was gathered from the twitter by means of interface Twitter API.

B. Sentiment Analysis

The categories involved in SA and its application are discussed in (Walaa Medhat et al., 2014). This survey deliberated the related fields to SA including emotion detection, building resources and transfer learning. Sentiment Analysis techniques are analysed by using various methods in the extracted twitter (Sarlan A *et al*., 2014)(Shahheidari S, Hai Dong, 2013) (Othmura M *et al*., 2014). The keyword search method is used to extract the keyword in the tweet dataset (Chatterjee R, Goyal M, 2015). Different type of analytic techniques are proposed in (Bongsug chae, 2015). They are Content Analytic, Descriptive Analytic, Sentiment Analytic and Network Analytic techniques are used to analyse the tweets.

Sentiment Analysis is the strategy for finding, for example, positive and negative from a content information in (Koto F. Adriani M, 2015). Some component determination procedures, for example, Mutual data, Chi-Square, Information pick up and TF-IDF were proposed in (Shahana P.H, Bini Omman, 2015) to choose highlights from the high dimensionality of the list of capabilities. Considering unigram, bigram, POS (Parts Of Speech) labels of words and capacity words as a list of capabilities.

SentiWordNet (Bruno Ohana, Brendan Tierney, 2009) is a feeling dictionary got from the WordNet database where every term is connected with numerical scores demonstrating positive and negative assumption data. The utility of phonetic components for recognizing the conclusion of Twitter messages and assess the handiness of existing lexical assets are examined in (Bhuta*et al*., 2014). And additionally highlights that catch data about the casual and imaginative dialect utilized as a part of microblogging. An administered way to deal with the issue,however, influences existing hashtags in the Twitter information for building preparing information. Utilizing hashtags to gather, preparing information proved helpful, as did utilize information gathered in view of positive and negative feelings. The strategy proved that investigates the social sentiment from extracted tweets that have no less than 1 of 800 wistful or enthusiastic descriptive words in (Koto F, Adriani M, 2015)

The commitments of the paper (Apoorv Agarwal *et al*.,2011) are (1) Introduce POS-particular earlier extremity highlights. (2) Investigate the utilization of a tree kernel to block the requirement for tedious feature engineering. Compare different algorithms and methods for Sentiment Analysis polarity detection and sentiment summarization for the tweet in (Bahrainia S.A, Dengel A, 2013).

Approach to naturally arranging the supposition of Twitter messages are proposed in (Alec Go *et al*., 2009). These messages are named either positive or negative concerning a questionable term. The preparing information contains Twitter messages with emoticons, which are utilized as uproarious marks. Machine learning calculations (Naive Bayes, Maximum Entropy, and SVM) have exactnesses above 80% when prepared with emoticon information.

In this proposed work, SA was used to analyze the twitter dataset. In the three existing SA techniques, Sentence level SA was used in the proposed work. Each tweet has considered as the sentence. So, Sentence level SA has correctly predicted the sentiment of the tweet and also POS tag has been used in this work to discover the polarity of the tweet.

C. Sampling Techniques for Streaming Data

The examination of the sample and complete twitter dataset are portrayed in (Yazhe Wang *et al*., 2015). Generally, sample sizes have been founded on accessibility and useful contemplations. 1% sample dataset and 10% sample dataset gathered from the original dataset. Contrast the two sample dataset and the original dataset and break down the outcomes by utilizing different strategies.

Sampling is an engaging procedure for data mining, in light of the fact that rough arrangements, by and large, may as of now be of incredible fulfillment to the need of the clients. Using sampling techniques to address the problem of maintaining find out the association rules is described in (Lee S.D *et al*., 1998). A sampling of streaming data is concentrated on in (Walaa Medhat et al., 2014). Consequently, the application and advancement of sampling algorithm are deeply discussed in (Walaa Medhat et al., 2014), particularly those customary examining strategies in the information stream model.

Reservoir sampling of data streams is described in (Jeffrey Scott Vitter, 1985). Algorithms for choosing a sample of size n in a random manner from a document containing N records, in which the estimation of N is not known to the algorithm. The last sample is acquired by choosing n records indiscriminately from this larger sample.

The Weighted Random Sampling over data streams is proposed in (Pavlos S. Efraimidis, 2010). Sampling with and without substitution and show adjustments of the algorithm for a several WRS issues and advancing data streams are discussed.

Bernoulli sampling is briefly discussed in (Dhiren Ghosh, Andrew Vogt, 2000). Bernoulli sampling is the process of selecting a sample in which each and every element has the same probability of being sampled element. The main objective is to regularize the sample size. Rejection sampling is also briefly explained in (Dhiren Ghosh et al., 2000). When the sample is rejected unless the desired sample size is achieved. Otherwise, there is an alternate way to reject the sample if the size is smaller than the desired or randomly trim few elements from the obtained sample to attain the required size. A new sampling technique is introduced in (Luke Shrimpton et al., 2015). The first really streaming cross-document coreference determination (CDC) framework. For the sampling to be illustrative it ought to speak to an extensive number of elements whilst considering both worldly recency and far off references is presented in (Luke Shrimpton et al., 2015). And Uniform Reservoir Sampling is briefly discussed. Insert and remove an element from the sample dataset is also explained.

In the proposed work, four Sampling Techniques, such as, RS, Bernoulli A/R sampling, WS and HS were utilized. These Sampling Techniques has been utilized to choose the sample from the original dataset. Hybrid Sampling is the combination of Reservoir Sampling and Weighted Sampling.

III. SAMPLING TECHNIQUES AND SENTIMENT ANALYSIS

A. Problem definition

The Twitter dataset is the fast-growing social networking site. SA for the twitter dataset has been more and more difficult task because the tweet feed has been updated daily and also it must consume more time for processing. On the behalf of analyzing the whole tweet, sampling techniques provided the best solution for select the sample twitter dataset and SA were carried out on the sample twitter dataset.

B. Proposed Work

The twitter dataset consists of noise such as punctuation marks, URL (Uniform Resource Locator), Retweets, and stop words. The noise was removed in the data

preprocessing step. Further, the whole twitter dataset has been analyzed by using Sentiment Analysis (SA) technique. In SA technique, Sentence level SA was used to analyze the sentiment of the tweet. In our proposed work, four sampling techniques were used to choose the sample from the whole dataset. The proposed sampling techniques are the Reservoir Sampling (RS), Bernoulli's Acceptance/Rejection Sampling (A/R), Weighted Sampling (WS), and Hybrid Sampling (HS). Hybrid sampling technique is the combination of Reservoir Sampling and Weighted Sampling technique. Further, each sample dataset was analyzed by using the SA technique. Subsequently, the Comparative analysis was done for perceiving that the proposed hybrid sampling was efficient for the streaming dataset.

C. Dataset Extraction

The twitter dataset extraction is depicted in Procedure 3.1. The twitter dataset extraction has been done by creating the Twitter App. Only one percent of twitter dataset is publicly available. The streaming API has been used to extract the twitter dataset from that publicly available twitter data. And afterward the consumer key, consumer secret, the access key and access secret token were important to approve the Twitter application. RAuth was required to approve the application. After the approval procedure has been finished, the tweet was extricated by utilizing keyword search command.

Procedure 3.1. For extracting Tweet via API

Input : Consumer key, Consumer Secret Key, Access Key, Access Secret
Output: Extract tweet and saved in the rstudio
1 Create the Twitter Application.
2 Building the corpus in the rstudio
3 Perform handshake by using Consumer key, Consumer secret, Access secret token, and Access Key
4 Authenticate for the Twitter App is done via RAuth.
5 The Keyword search method is utilized to gather the Tweets.
6 Convert the downloaded list file into the Comma Separated Values (CSV) format in the rstudio.

The extracted twitter dataset has consisted of a retweet, stop words, punctuation marks, and URL. Data preprocessing has been used to remove the noise in the twitter dataset.

D. Sentiment Analysis

Sentiment Analysis (otherwise called supposition mining) alludes to the utilization of common dialect handling, content examination, and computational linguistics recognize and extract subjective data in the source materials. SA is broadly connected to surveys and online networking for an assortment of uses

Figure I show the convenience of the sample data on analysing the Twitter content, an sentiment classification task on the sample datasets and the total dataset was performed in (Yazhe Wang et al.,), and after that look at the outcomes. Supposition classification is an opinion mining activity concerned with deciding the general sentiment orientation of the conclusions contained inside a given record (e.g., tweet). The sentiment orientation was delegated positive or negative. Terms in SentiWordNet database take after the classification into part-of-speech labels. Given a term and its part-of-speech tag, Senti-WordNet returns three sentiment scores going from - 1 to 1: positivity, negativity, objectivity, each indicating the term's sentiment bias.

Each and every tweet has been analyzed by the parts of the speech (POS) tag is described in procedure 3.2. After that, the word in the tweet has been determined as a positive or negative word. In view of that analysis, the score was figured. In the event that the score is positive, then it is as the positive tweet.

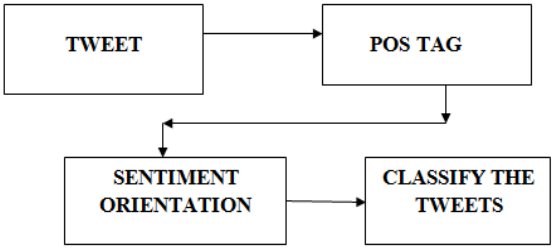


Figure I. Flow Diagram Of Sentiment Analysis On Tweet

Procedure 3.2. Sentiment analysis on tweets

Input: Tweet complete dataset and sample dataset
Output: Group the tweet as positive, negative and Ascertain the score for each tweet
1 For each tweet
2 Data preprocess is done to remove the stop words
3 Analyze the POS for each tweet
4 Analyze the word is either positive or negative
5 By comparing it with two lists such as positive and negative list.
6 Calculate the score for each tweet

7 The score can be Positive or Negative values.
8 Counting the positive and Negative Score to decide the Sentiment Orientation.
Overall Score = Positive score – Negative score
9 Calculate the overall score for the user

For example, a person tweets a tweet "Today is going to be a good day", in this tweet word good indicates the positive word. So the tweet is considered as the positive tweet. If the score has negative value, then it is deliberated as the negative tweet, for example, A person tweet "I am very sad", In this tweet word sad specify a negative word otherwise consider the tweet as neutral.

E. Adaptive Sentiment Analysis

The Adaptive Sentiment Analysis of the tweet dataset is described in procedure 3.3. For example, 1000 tweets were already analyzed by SA technique. Further, 500 tweets were extracted and analyzed by using the SA.

Procedure 3.3. Adaptive Sentiment Analysis

Input : Twitter dataset
Output: Merge the result of the input dataset with the results obtained from the SA phase
1 Let A be the result of the SA
2 Let B be the upcoming Twitter dataset
3 Enter the B data into the SA phase
4 Merge the result of B with A
5 Sum up the overall score for the user.

Then, the results of the 500 tweets have been merged with the results of 1000 tweets. The score of the user has been generated dynamically based on the upcoming tweets.

F. Sampling Techniques

Big data is very large in size. It will take high computation time for processing and analyze the complete dataset and it may consume huge storage space for storing the bigdata. Rather than storing and preparing the original dataset, the sampling technique was given the surmised result when contrasted and the original dataset. In our proposed work, four sampling methodologies were utilized to choose the sample from the entire twitter dataset, for example, Reservoir Sampling (RS), Bernoulli's Acceptance/Rejection (A/R), Weighted Random Sampling (WRS) and Hybrid Sampling (HS). Fundamentally, the sampling procedures were broadly arranged into two types, for example, uniform sampling and biased testing.

In uniform sampling, every element in the dataset has an indistinguishable likelihood of being chosen as a sample. In any case, on account of the biased sampling, each element has the diverse likelihood of being chosen as a sample. The proposed Reservoir Sampling (RS) comes under the classification of uniform sampling. Bernoulli's A/R and weighted Sampling come under the classification of biased sampling technique. Let us discuss the each sampling technique in detail manner.

G. Reservoir Sampling(RS)

In RS procedure, every data in the twitter dataset has the equivalent likelihood of being picked as a sample. In procedure 3.4, the size of the streaming data has not known ahead of time. With a specific end goal to choose the sample of size k from the entire dataset estimate N.

Procedure 3.4. Algorithm for Reservoir Sampling

Input: Tweet dataset whose size is not known
Output: Sample dataset of array size t
1 Initialize array whole[N]
2 For each data i in 1: t do
3 Sample[i]:=whole[i]
4 done
5 for each i in t+1 to N do
6 j:=random(1,i)
7 if(j<=t) then
8 sample[i]=whole[j]
9 continue until i=N

This algorithm depicted the first t data's in the entire dataset N were specifically embedded into the sample array of size t. After that t+1 data was gone into the sample dataset , then it has been created the random number between 1 to t+1. On the off chance that the random number was not as much as t means, it has been embedded into the sample array, otherwise dismiss that data.

H. Bernoulli's Acceptance/Rejection Sampling

The Bernoulli's Acceptance/Rejection Sampling (A/R) is the one of the types of the uniform sampling technique. That means every last element has the same likelihood of being chosen as a sample.

Procedure 3.5. Bernoulli's Acceptance/Rejection Sampling

Input: stream of N elements
Output: sample elements of n size
1 Calculate the count of tweets for each user
2 Set a minimum threshold value α
3 If(count[i]>α)

4 Accept the tweet i as sample
5 Insert it into the sample
6 Else if(count[i]<α)
7 Reject the tweet
8 Continue until i=N

. In procedure 3.5, a person who tweet number of tweets has considered them as a regular user. The first step was to count the tweets for each and every user. After that, threshold value has been fixed. If the count of the tweet for the user was greater than the threshold value, then the tweet has been accepted. Further, the tweet has been inserted into the sample. If the count of the tweet for the user has lesser than the threshold value, then the tweet has been rejected.

I. Weighted Random Sampling

The proposed Weighted Random Sampling (WRS) is one of the types of Biased Sampling technique which means each and every element in the whole dataset has the different probability of being selected as a sample. In this WRS technique, the sample elements have selected based on its weight. The element who has a higher weight that must be included in the sample dataset.

Procedure 3.6. Algorithm for Weighted Random Sampling

Input: Twitter dataset N and its weight w
Output: Sample dataset of size n
1 Assign the weight for each tweet i
2 For each tweet i in 1 to N
3 m=average(w[i])
4 If(w[i] > m)
5 Insert into the sample
6 Else if(w[i] < m)
7 Discard it
8 Until i=N

The algorithm of weighted sampling technique described in procedure 3.6. The input for the algorithm was the streaming dataset whose size was not known in advance and also its weight. The weight has been assigned for the tweet in a random manner. The probability of each data has selected by its weight with respect to the weight of other data in the dataset. The weight w[i] is a positive real number w[i] > 0 and the weights of all items has been considered as unknown. The WRS algorithms have generated a weighted random sample of size m.

J. Hybrid Sampling

The proposed Hybrid Sampling (HS) consisted of the Reservoir Sampling(RS) and Weighted Sampling(WS) technique. In the first part, Reservoir Sampling has been implemented and the key has been generated based on the weight of the tweet.

The algorithm of Hybrid sampling technique described in procedure.3.7. The first n elements were inserted into the reservoir and the key was generated for each tweet. The minimum key value in the reservoir has been set as the

Procedure 3.7. Algorithm for Hybrid Sampling

Input: Streaming dataset of size m
Output: sample of size n
1 First n tweets are inserted into the reservoir R[n]
2 For each tweet t_i € R,
3 Calculate the key $key_i=(Z_i)^{1/w_i}$ where Z_i=random(0,1)
4 Repeat step 1 to 3 for i=1:n
5 For $t_i=t_{n+1},\ldots t_m$
6 Set the threshold T for the tweet whose have minimum key key_i
7 For tweet t_{n+1} :calculate the key key_{n+1}
8 If key key_{n+1} > T then
9 Tweet with minimum key is replaced with t_{n+1}
10 Otherwise, discard it.

threshold value. After that, n+1 th element was entered into the reservoir, then the key for that tweet had generated. Then the key was compared with the threshold value obtained in the above steps. If the key value was greater than the threshold value, then that element has been inserted into the reservoir by replacing the element which has minimum key value. Instead of random replacement,the key with the minimum key has replaced with the new upcoming data in the data stream. This is the main advantage of the Hybrid sampling technique.

K. Performance Metrics

The accuracy of the four proposed sampling techniques has been calculated by using the equation 3.1

$$Accuracy=TP+TN/(P+N) \quad (3.1)$$

Where, TP = True Positive, TN = True Negative, FP = False Negative, FN = False Negative, P = TP+FP, N = TN+FN.

Precision and Recall were the basic performance measure. Precision is the proportion of the number of relevant datasets retrieved with respect to the total number of irrelevant and relevant datasets retrieved. It is usually expressed as a percentage. The precision has been used as one of the metric in the performance evaluation in order to measure the ratio of the number of related tweets with respect to the total number of related and unrelated tweets in the twitter dataset. The precision of the Sampling technique has been calculated by applying the equation 3.2.

Precision of sampling technique (PR)= TP/(TP+FP) (3.2)

The recall has been used as the performance measure in order to measure the proportion of the number of related tweet datasets retrieved with respect to the total number of related tweet datasets. It is usually expressed in percentage. The recall for the sampling techniques has been calculated by applying the equation 3.3

Recall for sampling techniques(R) = TP/(TP+FP) (3.3)

High precision relates to a low False Positive rate, and high recall relates to a low False Negative rate. High scores for both shows that the proposed work has returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

The F-measure for the sampling have been calculated by using the obtained values of precision and recall. F-Score, which has defined as the harmonic mean of precision and recall.

F_Measure=2*PR*R/(PR+R) (3.4)

L. Comparative Analysis

The comparative analysis of the sample and whole dataset was described. If a user was considered as a positive minded person in the whole dataset, then the sample dataset also showed that the user was a positive minded person. This analysis has been done in this section. This analysis has been used to find out the optimal sampling selection technique for streaming dataset.

M. Optimal Sample Selection

The optimal sampling technique for twitter dataset was deeply described. This has been found out by which sampling technique has given approximate results when compared with the complete dataset.

IV. EXPERIMENTS AND RESULTS

A. Dataset Extraction

The first step to perform Twitter Analysis has to create a twitter application. This application has allowed you to perform analysis by connecting your R console to the twitter using the Twitter API. The handshake has been performed by using the Consumer Key and Consumer Secret number of your own application. Authorize the twitter app and then collect the tweet by using the search function. Using these steps, an emotion dataset from twitter was extracted via twitter API.The dataset contains 5000 tweets and the size of the dataset is 1362 KB.

B. Sentiment Analysis for the Tweets

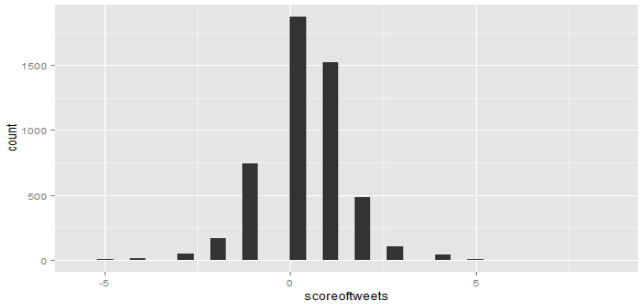
In SA phase, Sentence level Sentiment Analysis is used to analyze the tweets. The score is calculated for each and every tweet. The score for the tweet is depicted in Table I.

TABLE I

SNIPPET FOR SCORE OF THE TWEETS

TWEET	SCORE
A hug from your best friend can make everything feel so much better	3
youre turning 20 soon how do you feel	0
It's hard to feel pretty, don't make it impossible for others	-1
i feel you though no one can ever get mine right	1
I feel sorry for these highline schools who gotta play seattle schools	-1
Just wanna feel appreciated	1
I feel like when girls post photos of basically all cleavage and the guy says, "you have such beautiful eyes!" He means the bewbz	2

Sentiment Analysis is performed in collecting twitter dataset. Results are analyzed and count the number of tweets having the positive, negative and neutral is calculated. The Figure 4.2. represent the number of scores of the tweets with respect to the count of the tweets.



V.

FIGURE II. SENTIMENT ANALYSIS OF TWEETS

The number of neutral,positive and negative tweets is described in Figure II. The extracted tweets can be analyzed by using sentiment analysis.Approximately, out of 5000 tweets extracted from the Twitter, 1800 tweets are neutral tweets, 2510 tweets are positive tweets, and 690 tweets are negative tweets.

C. Proposed Sampling Techniques

The proposed four sampling techniques are implemented. The performance evaluation of the four sampling techniques is briefly discussed in this section. The Input twitter dataset consists of 5000 tweets.

The SA for the sample dataset is depicted in Figure 4.8. The four sample dataset are selected by using four sampling techniques such as Reservoir Sampling (RS), Bernoulli's Acceptance/Rejection Sampling (A/R), Weighted Random Sampling (WRS) and Hybrid Sampling (HS).

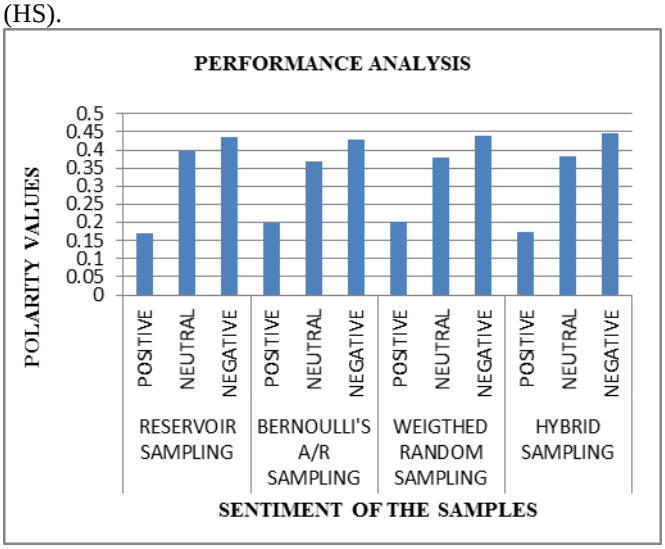


FIGURE III SENTIMENT ANALYSIS FOR SAMPLE DATASET

The polarity values are identified by using the SA technique. Each and every sample dataset consists of positive, negative and neutral tweets. The X-axis represents the sentiment of the sample dataset and Y-axis represents the polarity value in Figure III.

4.5 COMPARATIVE ANALYSIS

The performance analysis of the sampling techniques is depicted in Figure IV.

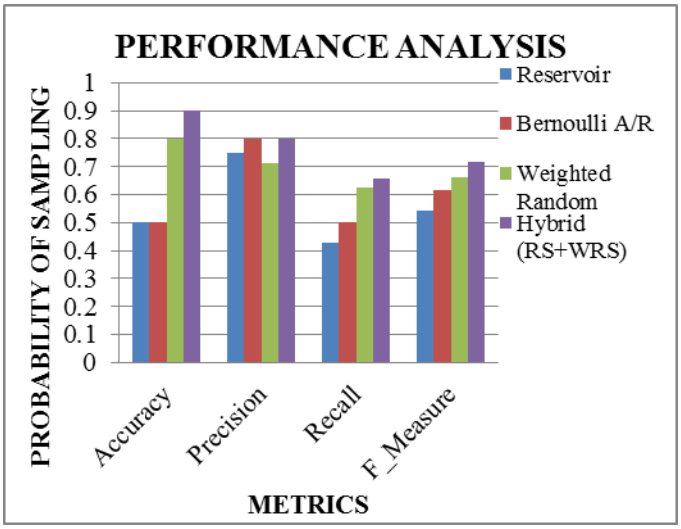


FIGURE IV. PERFORMANCE ANALYSIS

The X-axis represents the performance metrics and the Y-axis represents the probability of the sampling techniques. Accuracy, Precision, Recall, and F_Measure are used for evaluating the sampling techniques. This can be calculated by applying the equations 4.1, 4.2, 4.3, 4.4. The proposed Hybrid Sampling technique yielded 90% Accuracy, 80% Precision, 65% Recall and F_Measure value is 72%. Figure 4.8, proved that proposed Hybrid Sampling (HS) technique is best suitable and efficient sampling technique for the Twitter dataset when compared with other sampling techniques.

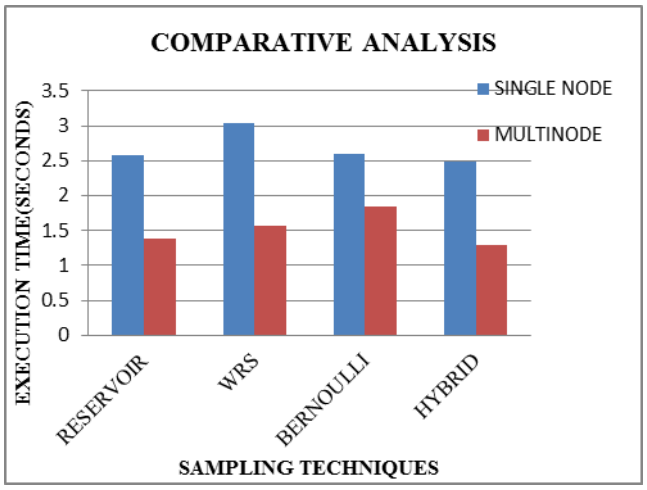


FIGURE V. COMPARATIVE ANALYSIS OF THE SINGLE AND MULTINODE SYSTEMS

The comparative analysis between the single node Hadoop with R and the two distributed system Hadoop with R is depicted in Figure V. The X-axis represents the proposed four sampling techniques and the Y-axis represents the execution time in seconds. Hadoop is the parallel distributed platform. It is mainly used for processing semistructured, unstructured and structured data across a cluster of commodity server. The main advantage of Hadoop is to reduce the execution time for processing bigdata. This graph clearly says that the execution time is reduced when deployed the sampling algorithm in a distributed manner.

VI. CONCLUSION AND FUTURE WORK

The Twitter dataset has been extracted through twitter API . sentiment Analysis has been performed on the twitter dataset to analyze the opinion of the tweets. The score for tweets has been ascertained. Results obtained from the dataset was analyzed and polarity has been identified. Assist, the Sample twitter datasets has been gotten from the entire twitter dataset and sampling procedures, such as, Reservoir Sampling (RS), Bernoulli's Acceptance/Rejection Sampling (A/R), Weighted Random Sampling (WRS) and Hybrid Sampling (HS) strategies were connected to the total twitter dataset to extricate the sample dataset and execution analysis has accomplished for each example dataset.Further, the SA technique has been carried out for the sample datasets. Comparative analysis of the whole dataset and sample dataset has done. The Hybrid Sampling techniques provided a more approximate result when compared with the complete dataset. The proposed work concluded that the Hybrid Sampling (HS) technique is the best suitable sampling technique for the streaming dataset and yielded 90% accuracy when compared with other sampling techniques.In future work, Machine Learning techniques will implement and may produce the optimal result for streaming dataset.

References

[1] Abdelghani Bellaachia and Mohammed Al-Dhelaan "Learning from Twitter Hashtags: Leveraging Proximate Tags to Enhance Graph-Based Keyphrase Extraction",IEEE Conference on Green computing and communication, pp. 348 – 357, Nov 2012.

[2] Alec Go, Richa Bhayani, Lei Huang," Twitter Sentiment Classification Using Distant Supervision",Processing, pp.1-6, 2009.

[3] Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau,"Sentiment Analysis Of Twitter Data",ACM Digital Library,LSM'11 Proceedings Of The Workshop On Languages In Social Media, pp. 30-38, 2011.

[4] Bahrainia S.A. ,Dengel A.," Sentiment Analysis and Summarization of Twitter Data",IEEE on Computational Science and Engineering, pp. 227-234, 2013.

[5] Bongsug (Kevin) Chae, "Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research",International Journal Of Production Economics, Volume 165, pp.247-259, July 2015.

[6] Bhuta, S.; Doshi, A.; Doshi, U.; Narvekar, M.,"A review of techniques for sentiment analysis Of Twitter data", Issues and Challenges in Intelligent Computing Techniques, pp. 583-591, 2014.

[7] Bruno Ohana , Brendan Tierney," Sentiment Classification Of Reviews Using Sentiwordnet" 9th. IT&T Conference, Dublin Institute Of Technology, Dublin, Ireland, pp. 22-23, October 2009.

[8] Chatterjee, R.; Goyal, M. ,"Tactics of twitter data extraction for opinion mining", International Conference on computing for Sustainable Global Development, pp. 761-766, 2015.

[9] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore," Twitter Sentiment Analysis: The Good The Bad And The OMG!", Fifth International AAAI Conference On Weblogs And Social Media, 2011.

[10] Hemalatha I, Saradhi Varma, Govardhan,"Sentiment Analysis Tool using Machine Learning Algorithms", International Journal of Emerging Trends & Technology in Computer Science, Volume 2, Issue 2, pp. 105-109, April 2013.

[11] Jeffrey Scott Vitter,"Random Sampling With Reservoir",ACM Transactions On Mathematical Software, volume 11, Issue 1, pp. 37-57, March 1985.

[12] Koto, F.; Adriani, M.,"The Use of POS Sequence for Analyzing Sentence Pattern in TwitterSentiment Analysis", IEEE 29th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 547-551, 2015.

[13] Othmura M, Kakusho K, Okadome T, "Social Mood Extraction from Twitter Posts with Document Topic Model", International Conference on Information Science and Applications (ICISA), pp. 1-4, 2014.

[14] Pavlos S. Efraimidis, "Weighted Random Sampling over Data Streams",December 2010.

[15] Pulkit Goyal, Sapan Diwakar, "Data Mining And Analysis On Twitter", January 14, 2011.

[16] Sarlan A, Nadam C, Basri S, "Twitter Sentiment Analysis", International Conference on Information Technology and Multimedia", pp. 212-216,2014.

[17] S. D. Lee, David W. Cheung, Ben Kao, "Is Sampling Useful in Data Mining? A Case in the Maintenance of Discovered Association Rules", ACM transaction on Data Mining and Knowledge Discovery, Volume 1, Issue 3, pp. 233-262, Sep 1998.

[18] Septianto, G.R.; Mukti, F.F.; Nasrun, M.; Gozali, A.A.,"Jakarta congestion mapping and classification from twitter data extraction using tokenization and naive bayes classifier",Asia Pacific Conference on Multimedia and Broadcasting (APMediaCast), pp. 1-6, 2015.

[19] Shahheidari S, Hai Dong, "Twitter Sentiment Mining:Domain Analysis", Seventh International Conference on Complex, Intelligent, and Software Intensive Systems(CISIS), pp.144-149, 2013.

[20] Shahana P.H,Bini Omman,"Evaluation of Features on Sentimental Analysis", International Conference on Information and Communication Technologies,Volume 46, pp. 1585-1592, 2015.

[21] Walaa Medhat, Ahmed Hassan, Hoda Korashy," Sentiment Analysis Algorithms And Applications:A Survey", Ain Shams Engineering Journal ,Volume 5,Issue 4,pp. 1093-1113, December 2014.

[22] Wenyu Hu, Baili Zhang, "Study Of Sampling Techniques And Algorithms In Data Stream Environments", 9th International Conference On Fuzzy Systems And Knowledge Discovery , pp. 1028 – 1034, May 2012.

[23] Yazhe Wang, Jamie Callan, Baihua Zheng, "Should We Use the Sample? Analyzing Datasets from Twitter's Stream API", ACM Transactions on the Web(TWEB), volume 9 Issue3,June 2015