Data Analytics Framework: R and Hadoop – Geo-location based Opinion Mining of Tweets

K. Santhiya
PhD Research Scholar
Dept. of Computer Application,
Bharathiar University, Coimbatore, India
krsanthiya@gmail.com

V. Bhuvaneswari
Assistant professor
Dept. of Computer Applications
Bharathiar University, Coimbatore, India
bhuvanes_v@yahoo.com

***Abstract*– Internet social media services such as Twitter have seen phenomenal growth as millions of users share opinions on different aspects of life every day. This tremendous growth has induced an interest in making use of such data for extracting valuable information, such as their opinions, location of the users and certain other information. In this paper we have analyzed the tweets related to crime attributes against women and children, different sort of crimes that are prevailing , the location in which the users tweets are more frequently occurring related to crimes. The proposed work make use of R language for extracting real time tweets and relies upon Hadoop-based framework for storing the tweets as they are larger in number. The tweets are parsed under Hive environment and we build a sentiment classifier in R that is able to determine positive, negative and neutral sentiments for a given phrase. We observe that the elapse time for processing under Hadoop based framework significantly outperforms the other conventional methods and is more suited for real time streaming tweets.**

Keywords– Big data, Hadoop, Hive, R, Sentiment, Tweets, Location, Crime

I. INTRODUCTION

Data has evolved to greater heights and the digital globe is in data deluge due to massive amount of data generated by various sources through business transactions, social networking, mobile communication, search engines etc., Human and machine generated data is expected to grow exponentially in the era of IoT. Microblogging has become a very popular communication tool among internet users. Few instances of microblogging services like Twitter, Facebook, Tumblr, LinkedIn, Instagram etc., generate more than petabytes of data which stores the views of human in all arenas. In the current data era, large number of internet users tends to migrate from traditional communication tools (such as blogs or mailing lists) to microblogging services because of free format of messages and an easy accessibility of microblogging platforms[1]. Authors of those messages share opinions on variety of topics, express their views on products and services and discuss current issues. Such data can be efficiently used by organizations to improve their business decisions and also believe data will provide with Return of Investment (ROI) when analyzed by providing with deep insights on their own business.

In our paper, we concentrate on the dataset comprised of collected messages from twitter. Twitter contains millions of very short messages (approx. 140 characters) created by the users of microblogging platform. We focus on collecting the messages related to different sorts of abuse against women and children. Table 1 shows typical example of tweets from twitter.As the users of microblogging platforms and services grow every day, millions of tweets (counting nearly to petabytes) started to accumulate which can be effectively used for opinion mining and other sentiment analysis tasks. But the traditional methods, algorithms and frameworks for managing this enormous amount of tweets have become both inadequate for storage and processing.

In this context, big data has emerged as a new paradigm that aims to provide an alternative to traditional solutions in terms of storage and processing. Big data is not just about storage or access to data; its solution aim to analyze data in order to exploit their value[2]. Big data refers to collection of datasets that are terabytes to petabytes in size. The new term coined as "social big data" represents the interaction between social media and big data. Therefore, social big data will be based on the analysis of tremendous amount of data that

ISSN: 2349-6363

could come from multitude of sources but with a strong emphasize on social media. The data extraction, fusion, processing and analysis of the big social media data to extract value is an extremely difficult task which has not been completely solved[3]. Different big data frameworks such as Apache Hadoop and Spark have been emerged which outperforms the existing traditional tools and technologies.

In this paper, we propose a Hadoop-based framework that allows the user to store tweets in a distributed environment. Further we use natural language processing (NLP) techniques like POS tagging, parsing, text mining and sentiment analysis.

Table 1. Examples of tweets with expressed opinions on abuse against women and children

@Deb_Hitchens 2. Of course kidnapping, child abuse, abuse, sexual slavery are human rights issues but don't blame innocents for it.
@SkyNews 5 year old children do not commit sexual abuse. They explore as in nature.
@GBVnet: The social norm which says children must be seen & not heard enables child sexual abuse to proliferate. We must learn to listen…
@daily_texts: Royal Commission findings: #Watchtower policies place #JW children "at significant risk of sexualabuse".
@PearsonElaine: Deliberate abuse, self-harm, sexual harassment taking place under our noses. New @hrw @amnestyonline report on Nauru htt.

A. Contributions

i. Capturing and processing real time tweets using R and Hive under the Hadoop Framework.

ii. We propose a method to collect a corpus with positive and negative sentiments, and a corpus of objective texts. Our method allows collecting negative and positive sentiments such that no human effort is needed for classifying the documents.

iii. We propose a regular pattern to parse the tweets under the Hadoop framework.

iv. We make use of packages to estimate the geo-location of twitter users.

v. We conduct experimental evaluations on a set of real microblogging posts to prove that our presented technique is efficient and performs better than previously proposed methods.

B. Organizations

The rest of this paper is organized as follows. Section 2 presents related work for capturing and processing data acquired through the Twitter streaming API followed by the location prediction of the twitter user. Section 3 explains the methodology and the framework used in this proposed work. In Section 4 we describe the details of the experimental setup and the also explains the results obtained from the experiments. Finally, the conclusion and recommendations for future work are drawn in Section 5.

II. RELATED WORK

In this paper, the literature survey is done on three folds. At first capturing and preprocessing of the real time tweets are surveyed. Secondly, the literature on opinion mining follows. Finally, the estimation of geographical location of the twitter user is surveyed.

A. Capturing and preprocessing of tweets in massive amount

Social networking platforms facilitate users to generate data at an alarming rate. Twitter is one such platform that generates data regularly. Twitter claims to have more than 500 million users as of 2016, out of which more than 332 million are active[4]. Users post more than 340 million tweets every day. In the existing literature, most of the researches used Tweepy (A Python library for accessing the twitter API) and Twitter4J (A

Java library for accessing the twitter API) [5,6]. Befit and Frank discuss the challenges of capturing Twitter data streams[7].Tufekci and Zeynep examined the methodological and conceptual challenges for social media based big data operations[8]. Due to some restrictions placed by Twitter on the use of their retrieval APIs, one can only download a limited amount of tweets in a specified time frame using these APIs and libraries. Getting a larger amount of tweets in real time is a challenging task. There is a need for efficient techniques to acquire a large amount of tweets from Twitter.Shirahatti et al. make use of Apache Flume with the Hadoop ecosystem to collect tweets from Twitter[9]. Ha et al. used Topsy with the Hadoop ecosystem for collecting tweets from Twitter[10]. Furthermore, they analyzed the sentiment and emotion information for the gathered tweets in their research. Taylor et al. used the Hadoop framework in applications in the bioinformatics domain[11].

B. Opinion Mining

With the population of blogs and social networks, opinion mining and sentiment analysis became a field of interest for many researches. A very broad overview of the existing work was presented in (Pang and Lee,). In their survey, the authors describe existing techniques and approaches for an opinion-oriented information retrieval. However, not many researches in opinion mining considered blogs and even much less addressed microblogging. In (Yang et al.,), the authors use web-blogs to construct corpora for sentiment analysis. The authors applied SVM and CRF learners to classify sentiments at the sentence level. (Alexander et al.,) presented a method for an automatic collection of a corpus that can be used to train a sentiment classifier.

C. Estimating the geo-location of the twitter user

The geographic location estimation problem has been studied deeply by researchers who propose different ways to retrieve user location information from internet social media platforms.These works rely on external resources such as gazetteers and databases, to identify the related geographical information.Smith et al., studied the variation of language usage on Twitter. This can also be used to augment our work to improve the accuracy of predicting user geographic location.There have been works on: Lee et al., surveyed the relations between geotags [12], Backstrom et al., studied geo-location estimation in search engine query logs [13], Friedland et al., studied the user privacy of geotags [14], Backstrom et al., worked on predicting geographic location on proximity.

III. METHODOLOGY: OPINION MINING

This section describes the overall framework for capturing and analyzing tweets streamed in real time. In addition, the HDFS architecture followed by POS tagging, parsing, sentiment analysis and location estimation of the given tweet is elaborated.

A. Corpus Collection

The tweets which were collected in the proposed work is related to abuse against women and children. Violence strikes women from all kinds of backgrounds and of all ages. It can happen at work, on the street or at home. The type of violence against women can be classified as Dating Violence , Domestic and Intimate Partner Violence, Emotional Abuse, Human Trafficking , Same-sex Relationship Violence, Sexual assault and abuse, Stalking, Violence against immigrant and refugee women, Violence against women at work, Violence against women with disabilities. In this paper , tweets related to sexual abuse , child abuse and psychological abuse are taken into consideration.

B. Framework for sentiment analysis in real time tweets

The proposed system shown in Figure 1.uses R-programming language to extract the tweets and the Hadoop framework to store the tweets streamed in real time. The twitteR and httr packages available in R enable R to extract tweets in real time. These packages are responsible for communicating with the twitter streaming API and retrieving tweets matching certain criteria or keywords. The retrieved tweets are then stored in HDFS using rhdfs API. The tweets are then passed on to the Hive module, using rhive API, where the tweets are parsed into a suitable format for analyzing.

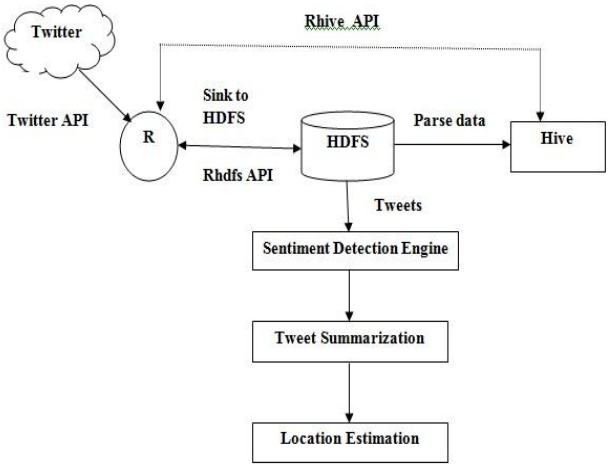

Figure 1: System model for capturing and analyzing tweets

C. Parallel HDFS

To increase the throughput of a system and handle the enormous amount of tweets, the parallel architecture of HDFS that is used is shown in Figure 2.HDFS is a clustered approach to manage files in a big data environment[15]. It breaks larger files into small pieces called blocks and distributes those blocks across different data nodes. HDFS cluster consists of a single Name node, a master server that manages the file system namespace and regulates access to files by clients. It instructs the data node to perform certain operations like create, update, delete and even replication of blocks. The Secondary name node takes a snapshot of metadata available in the name node at intervals specified in the hadoop configuration to facilitate fault tolerant mechanism.

D. rhdfs API

It is an R interface for providing HDFS usability from R interface. The R rhdfs package calls HDFS API in backend to operate data sources stored on HDFS. It facilitates the programmer to perform read and write operations on distributed data files. With the help of this API, we extract tweets from R environment and store it in HDFS as mentioned in the figure 1.

E. rhive API

This API provides integration between R console and hive. It also facilitates distributed computing via hive query. The components of Rhive are shown in table 2.

In this proposed work, with the help of this package, the tweets stored in HDFS are parsed in Hive and again it is fed into R for further analysis.

Table 2. Components of Rhive

Rhive Components	Description
UDF	Allow users to use R functions and R objects in Hive.
rhive	Allow to interact with hive within R
rhive. hdfs	Allow to interact with HDFS from within R

F. Parts-of-Speech tagging

Parts-of-Speech (POS) tagging divides sentences or paragraphs into words and assigning corresponding parts-of-speech information to each word based on their relationship with adjacent words in a sentence.

G. Parsing

Parsing is a process of analyzing grammatical structure, identifying its parts of speech and syntactic relations of words in sentences. When a sentence is passed through a parser, the parser divides the sentence into

words and identifies the POS tag information. In this paper, an R based package called koRpus which uses Tree Tagger for POS tagging has been used. An example of parsing of text

H. Sentiment Detection Engine

To identify the sentiment of a given tweet, it passes through the score sentiment function written in R for sentiment classification. The tweet is classified into either a negative, positive or neutral based on the detection engine. Figure 2.depicts an automated SDE which takes tweets as an input and produces the actual sentiment of the tweet as an output.

I. Location Estimation

This deals with extracting geographic location specific information of the twitter user using longitude and latitude values. Weassume that each user belongs to a particular city, and thus his/her tweets also belong to that city. That is, the terms occurring in the user's tweet can be assigned as terms relatedto the user's city.

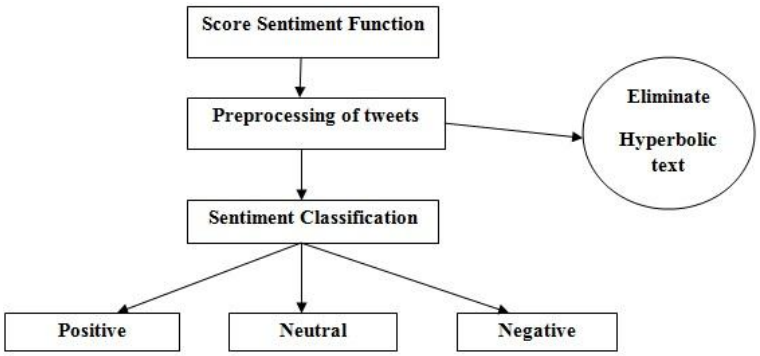

Figure 2: Sentiment Detection Engine

IV. RESULTS AND DISCUSSION

This section describes the experimental results of the proposed scheme. Three datasets are crawled using R and Twitter Streaming API and it is stored under Hadoop Framework. We also discuss about the preprocessing of tweets, sentiment analysis of the tweets with the corresponding sentiment score values. Location estimation results are also predicted and it is depicted.

A. Datasets collection for experiment and analysis

The dataset for the experimental analysis are shown in table 3. There are four sets of tweets crawled from the Twitter using the Twitter Streaming API and processed through R before being stored in HDFS. In total, 5,00,0000 tweets were collected using keywords sexual abuse, women abuse, child abuse, psychological abuse. After preprocessing, approximately 3,56,000 tweets were found related to abuse. The remaining tweets approximately 1,44,000 tweets were retweeted. Every set contained a different number of tweets. Depending on the number of tweets in each set, the crawling time (in hours) is given in table 3.

B. POS Tagging for the datasets

In this paper, POS tagging is an essential phase for all the proposed approaches.

Table 3. Datasets captured for experiment and analysis

Datasets	No. Of Tweets (approx)	Extraction Period (h)
Set 1	1,39,000	11
Set 2	1,20,000	10
Set 3	1,50,000	11.5
Set 4	83,000	9

Therefore in R, the koRpus library uses the TreeTagger for POS tagging. In other words, the TreeTagger has to be installed prior torunning the script as it accesses the TreeTagger via R. The results after implementing TreeTagger is shown in table 4.

Table 4 : POS tagging for tweets

Token	Tag	Lemma	Lttr	Wcl
Man	NN	man	3	noun
Held	VBN	held	4	Verb
For	IN	for	3	preposition
sexually	RB	sexually	8	adverb
abusing	VBG	abusing	7	Verb
daughter	NN	daughter	8	noun

C. Sentiment analysis

To identify sentiment in a given phrase, we use a pre-defined list of positive and negative words such as Sentiwordnet. It is a standard list of positive and negative English words. Using the Sentiwordnet lists along with equations (1) – (3), we find the sentiment score for a given phrase or sentence:

$$PR = \frac{NPW}{TW} \quad (1)$$

$$NR = \frac{NNW}{TW} \quad (2)$$

$$Sentiment\ Score = PR - NR \quad (3)$$

where PR is the positive ratio, NR is the negative ratio, NPW is the number of positive words in a given phrase , NNW is the number of negative words in a given phrase, and TW is the total words in a given phrase. The sentiment score for the first set of data is shown in Figure3 . It is very clear that the users of the twitter have predominantly given negative comments regarding the topic sexual abuse.

	score	topic	very.pos	very.neg
1	-4	sexualabuse	0	1
2	0	sexualabuse	0	0
3	0	sexualabuse	0	0
4	0	sexualabuse	0	0
5	-1	sexualabuse	0	0
6	0	sexualabuse	0	0
7	0	sexualabuse	0	0
8	0	sexualabuse	0	0
9	0	sexualabuse	0	0
10	0	sexualabuse	0	0
11	0	sexualabuse	0	0
12	0	sexualabuse	0	0
13	0	sexualabuse	0	0
14	3	sexualabuse	1	0
15	3	sexualabuse	1	0

Figure 3: Sentiment score for sexual abuse tweets.

The same score is visualized in the form of histogram as shown in Figure4 which depicts clearly that most of the tweets are neutral and negative and some of them has given positive comments about the topic.

Word cloud Generation

In this experiment, a word cloud is generated consolidating the opinions of different users around the world regarding the types of abuses against women and children.

Analysis of Location Estimation

We assume that each user belongs to a particular city, and thus his/her tweets also belong to that city. That is, from the screen name (user name) available in the tweets, we are extracting thelocation of the user's tweet. Thisforms the basic distribution of terms for the set of cities considered in the complete data set. The probability distribution of term *t* over the entire data set, for each city *c*, is given as

$$p(t|c) = |\{t|t \in terms\ \wedge t\ occurs\ in\ city\ c\}|/\ |t| \quad (4)$$

That is, the number of occurrences of term *t* for a city *c* divided by the total occurrences of the term *t* in the entire dataset. In this proposed work, we have estimated that most of the people who belong to United States have recently tweeted more regarding to "gymnastic sexual abuse of US Olympic committee". The other people who also commented regarding the same issue belongs to the location shown in the table 5.

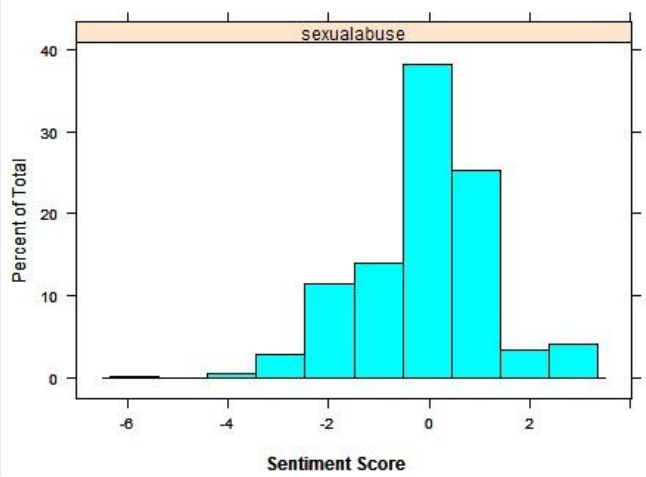

Figure 4: Sentiment Score visualization in histogram

Table 5. Location of user

Location	No.of People tweeted
USA	67,000
China	5600
Sweden	375
England	178
Ireland	57
Dubai	28
New York	13

V. CONCLUSION

Crime detection and analysis done with the help of microblogging platform such as twitter provides invaluable insight into public opinion on this particular topic. It also provides good insight into the problem of parsing unstructured data(tweets) and estimating the users geographical location.Further, the quality of this work can be refined by considering a larger dataset deployed under hadoop multi-node cluster setup. We would also like to see further improvements by combining the information from multitude of sources like articles from newspapers , message from blogs and even the complaints registered in corps office.

REFERENCES

[1] S.K. Bharti , B. Vachha, R.K. Pradhan, K.S. Babu, S.K. Jena , "Sarcastic sentiment detection in tweets streamed in real time: A big data approach", Digital Communications and Networks, June 2016.

[2] N.Khan, I,Yaqoob, I.A.T.Hashen, Z.Inayat. W.K.M. Ali, M.Alam, M.Shiraz, A.Gani, "Big data:survey, technologies, opportunities and challenges", The Sci.worldpp 1-18, June 2014.

[3] Z.N. Gastelum, K.M. Whattam, "State-of-the-Art of Social Media Analytics Research", Pacific Northwest National Laboratory, PP. 1-9, 2013.

[4] Chaffey, Global Social Media Research Summary 2016. URL⟨http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/⟩

[5] Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation", in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.704–714, 2013.

[6] Lunando, A. Purwarianti, Indonesian "social media sentiment analysis with sarcasm detection", in: International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, PP. 195–198 , 2013.

[7] Bifet, E. Frank, "Sentiment knowledge discovery in twitter streaming data", in: 13th International Conference on Discovery Science, Springer, PP. 1–15,2010.

[8] Z. Tufekci, "Big questions for social media big data: representativeness, validity and other methodological pitfalls, arXivpreprintarXiv:1403.7400.

[9] A.P. Shirahatti, N. Patil, D. Kubasad, A. Mujawar, "Sentiment Analysis on Twitter Data Using Hadoop".

[10] Ha, B. Back, B. Ahn, "Mapreduce functions to analyze sentiment information from social big data", Int. J. Distrib.Sens. Netw.PP.1–11, 2011.

[11] R.C. Taylor, An overview of the Hadoop/mapreduce/hbase framework and its current applications in bioinformatics, BMC Bioinform. Pp. 1–6, 2010.

[12] S. Lee, D. Won and D. McLeod. "Tag-geotag correlation in social network".In SSM `08 Proceeding of the 2008 ACM workshop on Search in social media.

[13] L. Backstrom, E. Sun, and C. Marlow. "Find me if you can: improving geographical prediction with social and spatial proximity". In WWW, 2010.

[14] Friedland and R. Sommer.Cybercasing the Joint:" On the Privacy Implications of Geo-Tagging". Proceedings of the Fifth USENIX Workshop on Hot Topics in Security (HotSec 10), Washington, D.C.

[15] Shvachko, H. Kuang, S. Radia, R. Chansler, "The Hadoop distributed file system", in: Proceedings of 26th Symposium on Mass Storage Systems and Technologies (MSST), IEEE, pp. 1–10,2010.

BIOGRAPHY

K. Santhiya received her Bachelor's Degree (B.Sc.) in Computer Science from Bharathiar University, India 2006, Master's Degree (MCA) in Computer Applications from Anna University, India 2009. She is currently pursuing her doctoral research in School of Computer Science and Engineering in the area of Big Data mining.

Dr. V. Bhuvaneswari received her Bachelor's Degree (B.Sc.) in Computer technology from Bharathiar University, India 1997, Master's Degree (MCA) in Computer Applications from IGNOU, India , M.Phil in Computer Science in 2003 from Bharathiar University, India and Ph.D in Computer Science in 2013 from Bharathiar University, India She has qualified JRF, UGC -NET, for Lectureship in the year 2003. Her research interests include Big Data, Data mining, Bioinformatics, Soft computing and Databases. She is currently working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, India. She has credit for her publications in journals, International/ National Conferences.