Opinion Mining of Tweets on Social Issues – Using R and Hadoop

Barshneya Talukdar

1. Introduction:

Data is being generated at all times. Every digital process and social media exchange produces it. Each day 2.5 exabytes of data is created. It comes from text messages, search engines, posts on social media sites, digital pictures and videos, and purchase transactions to name a few. In 2013, the total amount of data in the world was 4.4 zettabytes. That is set to rise steeply to 44 zettabytes by 2020. To put that in perspective, one zettabyte is equivalent to 44 trillion gigabytes. The majority of this data is created by individual users through social media. Social networking platforms with micro blogging features like Twitter, Facebook, LinkedIn, YouTube, Instagram, Tumblr and Google+ are very popular among Internet users. In fact, its convenience and ease of use has made micro blogging the preferred communication tool today. Millions of posts are created daily on the social networking websites where people write about their lives, share breaking news, express their opinions on a wide range of topics and discuss current issues. These short messages can come in the form of a variety of content formats including text, images, video, audio, and hyperlinks.

Every day, Facebook users post 4.3 billion messages and approximately 5.75 billion posts. There are 6 billion Google searches daily, 3.6 billion Instagram likes each day, and more than 4 million hours of content is uploaded on YouTube every day. Twitter users send out short messages called 'Tweets'. These micro messages are limited to 140characters. While this may seem like barely enough space to convey a message, it forces users to really think about what they want to convey. Users post 500 million tweets daily and share another 40 million tweets every day on Twitter. That is around 350,000 tweets per minute.

Twitter also gives businesses and governments a platform to stay connected with their audience and disseminate information, to get their feedback and respond to their concerns. As such, Twitter has become a repository of valuable data that can be used in opinion mining and sentiment analysis tasks by businesses and governments to make

strategic and tactical decisions in various fields such as advertising, political polls, scientific surveys, market prediction and business intelligence.

Sentiment analysis relates to the problem of mining the sentiments from online available data and categorizing the opinion expressed by an author towards a particular entity into at most three preset categories: positive, negative and neutral. It is one of the most important areas of analysis of Twitter posts that can be very useful for decision-making. However, performing sentiment analysis on Twitter is trickier than doing it for large reviews, as the tweets are very short (approximately 140 characters) and usually contain argot, emoticons, hash tags and other Twitter specific jargon. Added to that, while the amount of Twitter data is huge, the traditional methods, algorithms and frameworks for managing this enormous amount of tweets have become inadequate for storage and processing.

This led to the emergence of Big Data as an alternative to such traditional methods in storage and processing of data. Big Data is a blanket term for the non-traditional strategies and technologies needed to gather, organize and process insights from large datasets. Apache Hadoop and Spark are two of the most widely used Big Data frameworks. In this paper, a proposal is made for pose a Hadoop-based framework that allows the user to store tweets in a distributed environment. I shall concentrate on the dataset that comprised collected messages from Twitter. The focus is on collecting tweets related to physical, sexual and psychological abuse against women and children.

2. Proposed Methodology:

1. Twitter streaming API is used for fetching real time Twitter. Extracting and processing real time tweets are done using R and Hive under the Hadoop Framework.

2. This method allows collecting of negative and positive sentiments in such a way that no human effort is needed for classifying the documents.

3. Regular pattern is used to parse the tweets under the Hadoop framework.

4. Packages are used to estimate the geo-location of Twitter users.

5. Experimental evaluations are conducted on a set of real micro blogging posts to prove that presented technique is efficient and performs better.

6. Tweets are pre-processed for removing noise and meaningless symbols.

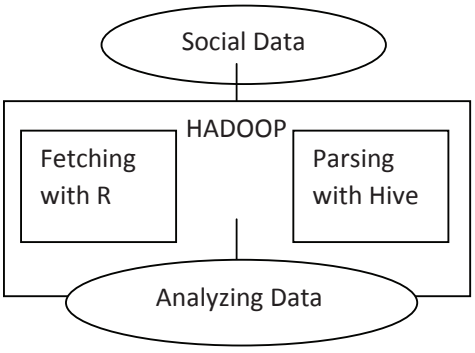


Figure 1: A general workflow diagram of the proposed system

In this paper, the literature survey is done in three stages:

1. At first capturing and preprocessing of the real time tweets is surveyed.
2. Secondly, opinion mining is followed.
3. Finally, the estimation of the geographical location of the twitter is surveyed.

3. Capturing and Preprocessing of Tweets in Massive Amounts:

Twitter users post 500 million tweets daily. Storing and processing of such large datasets is a very complex issue. Most researchers use Tweepy (A Python library for accessing the Twitter API) and Twitter4J (A Java library for accessing the Twitter API). The Twitter API provides a streaming API to allow developers to obtain real time access to tweets. However, Twitter limits the use of their retrieval APIs, due to which one can download only a limited number of tweets in a specified period using these APIs and libraries. This makes collecting a larger number of tweets in real time a challenging task. There is a need for efficient techniques to acquire a large amount of tweets from Twitter. Researchers are evaluating the possibility of using the Hadoop ecosystem for the storage and processing of large amounts of tweets from Twitter, for example,

using Apache Flume with the Hadoop ecosystem or using Topsy with the Hadoop ecosystem for gathering tweets from Twitter.

4. Opinion Mining:

Opinion mining and sentiment analysis became a field of interest for many researchers with the increase of users in social networking sites. Existing work was presented by Pang and Leein a broad way. In their survey, the authors describe existing techniques and approaches for an opinion-oriented information retrieval. Alexander et al. presented a method for automatic collection of a corpus that can be used to train a sentiment classifier.

5. Estimating the Geo-location of the Twitter User:

Twitter has over 328 million active accounts spread across the globe. It allows its users to specify their geographical location as user information. This location information is manually entered by the user or updated with a GPS enabled device. However, this location data for most users may be missing or incorrect as a significant number of users do not turn on the location services feature due to privacy concerns or to conserve battery. In addition, users may have multiple locations or may not provide the correct location information while manually entering their details.

Researchers have been trying several techniques to determine Twitter users' location even where they do not purposefully give it away. Various studies have deeply analyzed the geo location estimation problem with the help of which one can retrieve user location information from internet social media platforms. Also, Twitter users may use different languages to communicate. This greatly complicates the estimation of location. Smith et al. studies the variation of language usage on Twitter. This can also be used to improve the accuracy of predicting user geographic location. Lee et al. surveyed the relations between geo tags, Backstrom et al. studied geo-location estimation in search engine query logs, Friedland et al. studied the user privacy of geotags, Backstrom et al. worked on predicting geographic information on social and spatial proximity.

6. Methodology:

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.

7. Corpus Collection:

In this paper, tweets related to child abuse, sexual and physical abuse, and psychological abuse on women and children are taken into consideration. Victims of such abuse can be from all kinds of backgrounds and of all ages. It can happen anytime and anywhere, on the street, at school, at work or even at home.

8. Framework for Sentiment Analysis in Real Time Tweets:

The proposed system 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.

To handle the enormous amount of tweets, the parallel architecture of HDFS is used. HDFS is a clustered approach to manage files in a Big Data environment. 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 tolerance. The R rhdfs package which is an interface for providing HDFS usability from R interface, 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, tweets are extracted from R environment and are stored.

rhive API provides integration between R console and Hive. It also facilitates distributed computing via Hive query. The tweets stored in HDFS are parsed in Hive and again it is fed into R for further analysis.

Parts-of-Speech Tagging:

Parts-of-Speech (POS) tagging divides sentences or paragraphs into words and assigning corresponding parts-of-speech in formation to each word based on their relations. When a sentence is passed through

a parser, the parser divides the sentence into words and identifies the POStag information. In this paper, an R based package called koRpus which uses Tree Tagger for POS tagging has been used.

Sentiment Analysis:

Tweets are classified into negative, positive or neutral based on the detection engine. The score sentiment function written in R is used to identify the sentiment of a particular tweet. Finally, on the detection engine the tweet is classified in to positive, negative and neutral tweet.

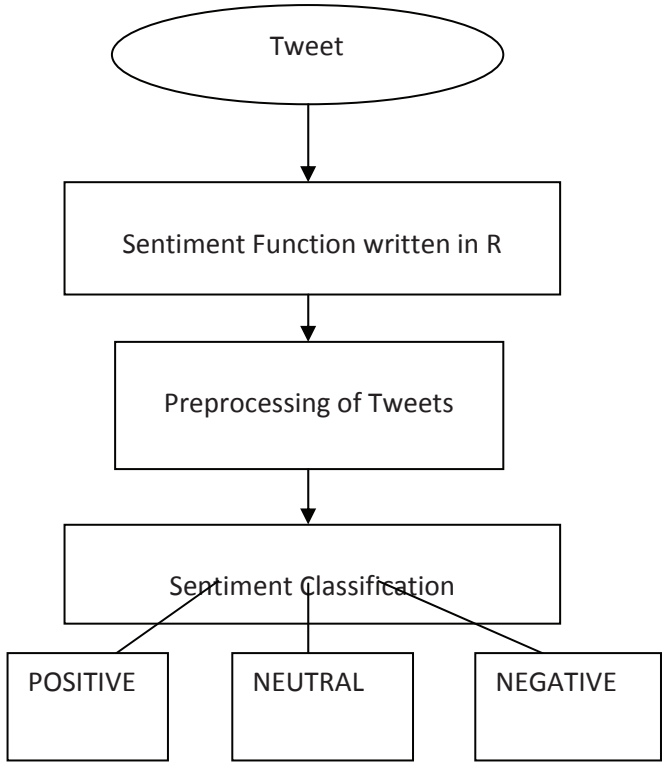


Figure 2: Sentiment Analysis

***Location Analysis*:**

In order to know the location from where the user tweeted, the geographic location of the user is extracted using longitude and latitude values.

9. Experimental and Result Analysis:

The experimental results of the proposed scheme is discussed below:

A dataset is crawled using R and Twitter streaming API and it is stored under Hadoop Framework. The preprocessing of tweets, sentiment analysis of the tweets and location estimation is done.

There are four sets of tweets crawled from the Twitter using the Twitter streaming API and processed through R before being stored in HDFS.

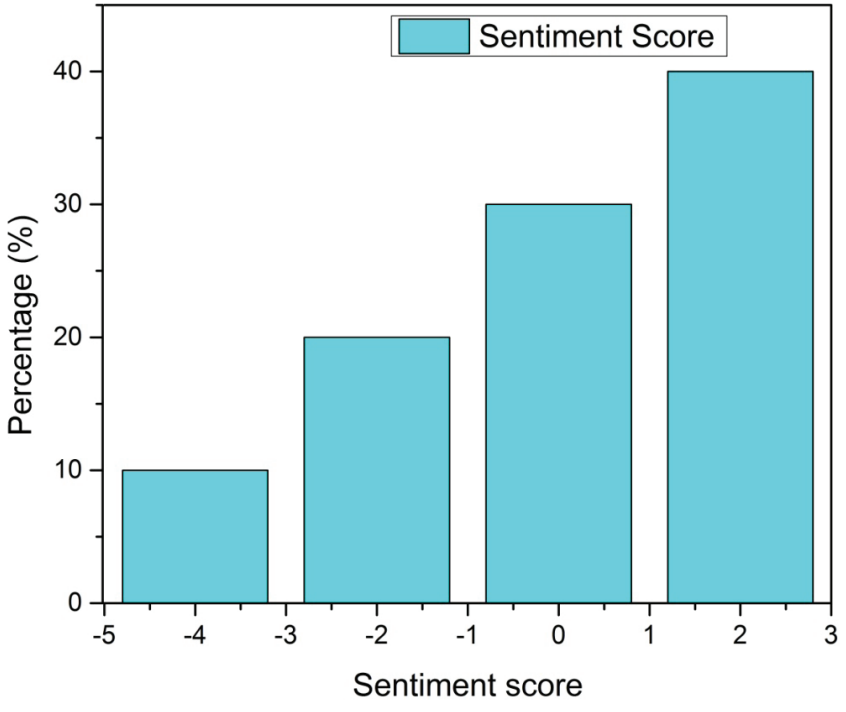
Sentiment Analysis:

To identify sentiment in a given phrase, we use a pre-defined list of positive and negative words. We find the sentiment score for a given phrase or sentence with the help of the following equation.

The equation to find sentiment score is: PR-NR

PR=PW/TW, NR=NW/TW (where PR stands for positive ratio, NR for negative ratio, PW for positive words, NW for negative words and TW for total words in a given sentece).

Figure 3: Histogram reflecting crime against women



Analysis of *Location Estimation:*

From the username available in the tweets, the location of the user's tweet have been extracted. It is assumed that each user belongs to a particular city, and thus his/her tweets also belong to that city. This forms the basic distribution of terms for the set of cities considered in the complete data set. Location of the users is tracked with the information of the user. In the above case, tweets have been extracted from Indian users, and there were about 5000 such tweets.

***Benefits of Sentiment Analysis**:*

Companies mostly benefit from sentiment analysis today. Some refer to it as social media analysis as well, since it also typically analyzes the ongoing activities on major social networking sites. The key points show that businesses can solely track positive and negative reviews of their brands.

With that being said, it also helps them measure their overall performance, especially on their online presence. Companies see sentiment analysis as a major aid in measuring sales and improving their marketing strategies as well. To accomplish this, some companies develop their own tools and others rely on outsourcing companies that specialize in sentiment analysis.

On the other hand, certain individuals can also be benefited from sentiment analysis, whether they are making a brand for themselves or just having that drive to know anything that regards to them. Artists, celebrities, famous authors and all other popular individuals can definitely benefit from the idea of sentiment analysis. They can simply know how they inspire the common public or how (negatively and positively).

An ordinary person, say a fanatic or blogger, can also benefit from sentiment analysis. There are many free sentiment tools available on the internet such as the Sentiment140 , which helps a user, find and learn Twitter sentiments.

Sentiment analysis can help open learners. For example if analysis is made regarding issues faced by the developmental service providers across social sectors, they can get benefits by knowing about people's reaction to any move taken and which of it triggers people's attitude towards them.

10. Conclusion:

After analyzing the complete procedure of fetching the social data and using the same for a social issue, it can be concluded that such a huge and complex data could not have been analysed using traditional analytical tool, hence Hadoop Framework was used for analysis R and Hive were used for analyzing and fetching data. The detection engine helps to detect crime against women. The tweets were extracted from micro blogging site Twitter. Opinion Mining can benefit big companies, actors, open learners, artists, bloggers and also an ordinary person through deep analysis of tweets.

References:

Nadagoud, S., & Naik, K.D. (2015). Market Sentiment Analysis for Popularity of Flipkart. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 4*(5), 2117-2123.

Patel, A.B., Birla, M., &Nair, U. (2012). Addressing Big Data Problem Using Hadoop and Map Reduce. doi: 10.1109/NUICONE.2012.6493198.

Shirahatti, A.P., Patil, N., Kubasad, D., & Mujawar, A. (2015). Sentiment Analysis on Twitter Data Using Hadoop. *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE), 14*(2), 831-837.