Real-time Twitter data analysis using Hadoop ecosystem

Anisha P. Rodrigues & Niranjan N. Chiplunkar |

Published online: 19 Oct 2018.

Submit your article to this journal ↗

Article views: 6477

View related articles ↗

View Crossmark data ↗

*Corresponding author: Anisha P. Rodrigues, Computer Science and Engineering, NMAM Institute of Technology, India
E-mail: anishapr@nitte.edu.in

Reviewing editor:
Marko Robnik-Šikonja, Faculty of Computer and Information Science, University of Ljubljani, Slovenia

Additional information is available at the end of the article

COMPUTER SCIENCE | RESEARCH ARTICLE

Real-time Twitter data analysis using Hadoop ecosystem

Anisha P. Rodrigues[1]* and Niranjan N. Chiplunkar[1]

Abstract: In the era of the Internet, social media has become an integral part of modern society. People use social media to share their opinions and to have an up-to-date knowledge about the current trends on a daily basis. Twitter is one of the renowned social media that gets a huge amount of tweets each day. This information can be used for economic, industrial, social or government approaches by arranging and analyzing the tweets as per our demand. Since Twitter contains a huge volume of data, storing and processing this data is a complex problem. Hadoop is a big data storage and processing tool for analyzing data with 3Vs, i.e. data with huge volume, variety and velocity. Hadoop is a framework which deals with Big data and it has its own family which supports processing of different things which are tied up in one umbrella called the Hadoop Ecosystem. In this paper, we will be analyzing tweets streamed in real time. We have used Apache Flume to capture real-time tweets. As an analysis, we have proposed a method for finding recent trends in tweets and performed sentiment analysis on real-time tweets. The analysis is done using Hadoop ecosystem tools such as Apache Hive and Apache Pig. Performance in terms of execution time is compared for analysis of real-time tweets using Pig and Hive. From the experimental results, conclusion can be drawn that Pig is more efficient than Hive as Pig takes less time for execution than Hive.

Subjects: Computer Engineering; Computer Science; General; Information Technology

Keywords: Apache Flume; Apache Hive; Apache Pig; Hadoop

This article describes a method for finding popular hashtag and classifies the tweets into positive and negative depends on "opinion" expressed by the people. Entire work is carried out using Hadoop framework which is a big data storage and processing tool. Along with Hadoop, Apache Hive and Apache Pig are used for analyzing tweets. This information can be used in economic, industrial, social or government sectors to understand people interest.

ABOUT THE AUTHORS

Anisha P. Rodrigues is working as Assistant Professor in the Department of Computer Science and Engineering at NMAM Institute of Technology, Nitte, India. Her research interests include Natural Language Processing, Machine Learning and Big Data Analytics.

Niranjan N. Chiplunkar is working as Principal and Professor in the Department of Computer Science and Engineering at NMAM Institute of Technology, Nitte, India. He is a Fellow of Institution of Engineers (India) and member of several other Professional bodies like IEEE, CSI and ISTE. His research interests include Data Mining, CAD for VLSI and Big Data Analytics.

Anisha P. Rodrigues

PUBLIC INTEREST STATEMENT

Now a days, people use social media to share their opinions and to have an up-to-date knowledge about the current trends on a daily basis. Considering the popular micro-blogging forum Twitter, where every minute millions of tweets are being posted, which greatly varies with the field of topic, analyzing these tweets, identifying what they express is tedious. This information can be used in economic, industrial, social or government sectors to understand people's interest. This article describes a method for finding popular hashtag and classifies the tweets into positive and negative depends on "opinion" expressed by the people. Entire work is carried out using Hadoop framework which is a big data storage and processing tool. Along with Hadoop, Apache Hive and Apache Pig are used for analyzing tweets.

cogent ··engineering

1. Introduction
With digitization, there has been a drastic increase in the usage of some of the popular social media sites such as Twitter, Facebook, Yahoo, YouTube as well as e-commerce sites as in Flipkart and Amazon, which have resulted in the generation of large sets of data. If the data is of small size, it is very easy to extract useful information, but if the size of data is huge, then it is quite difficult to analyze what that data actually intends. Ultimately, processing and extracting only the useful information is a tedious job. The advancement in technology has made it easy for people around the globe to express, share their views openly and every potential web user depends on this review and opinions to take decisions.

Considering the popular micro-blogging forum Twitter, where every minute millions of tweets are being posted, which greatly varies with the field of topic, analyzing these tweets, identifying what they express is tedious. In a situation where an organization is interested to know its customers' opinion regarding the product that it has manufactured, it is very difficult for organizations to keep track of each and every opinion. In this scenario, sentiment analysis plays a very important role. Sentiment analysis is identifying the polarity of the text, i.e. it identifies the attitude of the user toward a particular topic. Sentiment analysis greatly helps in knowing a person's behavior toward a particular topic. Importantly, all the data from the internet will be in unstructured or in semi-structured format. It turns out that the biggest challenge is in processing this type of data efficiently so that the information out of this data can be extracted for further survey. All the traditional data analysis techniques have gradually failed to perform analysis effectively on larger data sets. Recently, the powerful framework that has proved to be efficient in processing large sets of data is Hadoop, which is considered to be efficient for distributed processing as well as distributed storage of large sets of data. The main core components of Hadoop framework are MapReduce and HDFS. MapReduce is a programming model for processing larger data sets and HDFS is a Hadoop Distributed File System that stores data in the form of memory blocks and distributes them across clusters.

In addition to core components, Apache also delivers different tools/components to satisfy developer's needs. These tools along with core components of Hadoop are called the Hadoop Ecosystem. We have considered real-time streaming data on Indian political issues which are stored in Hadoop's file system. The real-time Twitter data is extracted using Apache Flume (https://flume.apache.org/). The data being extracted would be in JSON (JavaScript Object Notation) format. In JSON format, every data are represented in key/value pairs and separated by commas. The data stored in the HDFS are analyzed using data access components of Hadoop ecosystem Apache Pig (https://pig.apache.org) and Apache Hive (https://hive.apache.org).

1.1. Data access components of Hadoop ecosystem
The two key data access components of Hadoop Ecosystem are Apache Pig and Apache Hive. Hadoop's basic programming layer is MapReduce but these components ease the writing of complex Java MapReduce program. Apache Pig is an abstraction over Map Reduce and it provides a high-level procedural data flow language for processing the data known as Pig Latin. Programmers use Pig Latin and write Pig scripts to analyze the data which is internally converted to Map Reduce jobs. Pig reduces the complexity of writing long lines of codes, and using built operators, users can develop their own function.

Apache Hive is a data warehousing software that address how data is structured and queried. Hive has a declarative SQL like language, i.e. HiveQL or HQL for querying data. Traditional SQL queries can easily be implemented using HiveQL. In Hive, queries are implicitly converted to the mapper and reducer job. The advantages of using Hive are the features it provides, i.e. fast, scalable, extensible, etc. People who are not good at programming too can go in for this to analyze data on HDFS (KadharBasha & Balamurugan, 2017).

The rest of this paper is organized as follows. Section 2 describes different methods used for sentiment analysis and tools used for extracting tweets. The suggested methodology is discussed

in Section 3. Section 4 explains and analyses the results obtained from our proposed method. Eventually, the conclusion and future work are drawn in Section 5.

2. Related work

In this section, the literature survey focuses on fetching tweets, Hadoop being efficiently used as a feedback system, recommendation system and Sentiment analysis of Twitter data.

2.1. Fetching tweets

Nowadays, people use social media sites to express their opinions toward some issues, or product. Twitter is one of the social media sites which generates a huge amount of data. Collecting and processing this huge amount of data is challenging. Twitter Application Programming Interface provides a streaming API to stream real-time data. Tare, Gohokar, Sable, Paratwar, and Wajgi, 2014 and Sheela (2016) have used Twitter REST API to gather tweets from Twitter. Ha, Back, and Ahn (2015) have used Topsy to acquire historical data from Twitter. Some researchers (González-Ibánez, Muresan, & Wacholder, 2011; Kumar & Bala, 2016; Riloff et al., 2013) used Tweepy and Twitter4j to stream tweets from Twitter. Because of a few restrictions set by Twitter on streaming API, one can download a limited number of tweets in a given time frame. Therefore, we require an efficient tool to retrieve huge amount of data from Twitter. Apache Flume is an efficient tool to retrieve real-time huge amount of data from Twitter. Authors (Barskar and Phulre, 2017; Nadagoud & Naik, 2015; Sangeeta, 2016; Yadav, Pandey, & Rautaray, 2016) used Apache Flume to acquire a huge set of data from Twitter and they proved that Apache Flume is one of the efficient tool for real-time streaming.

2.2. Big data analytics

Sangeeta (2016) have discussed the usage of Apache Flume and Apache Hive which is built on top Hadoop for analyzing Twitter data. Verma, Patel, and Patel (2015) discussed a recommendation system that provides a summary of users' reviews, comments, feedback about any subject using Hadoop framework. Similarly, Shrote and Deorankar (2016) designed a recommendation system that recommends services. Bhardwaj, Kumar, Narayan, and Kumar (2015) have compared the effect of technologies such as Hive, Apache Pig, Flume, Zookeeper, HBase and Sqoop on Hadoop's performance. They have also analyzed the performance of Hive queries by considering two factors such as total MapReduce CPU time spent for running Hive query and total time taken to execute the job. Ennaji, El Fazziki, Sadgal, and Benslimane (2015) designed Hadoop framework for deriving and analyzing the customers' opinion toward an item from social networks, the designed framework extracts and analyzes the opinions of social customer relationship management. Jain and Bhatnagar (2016) have proposed a method that analyzes the crime dataset to keep track of the crimes happened so far. They used Apache Flume to stream data and processed it using Pig. Khade (2016) proposed a model to predict customer behavior. She used decision tree classifier and implanted this algorithm using the Map Reduce programming model.

2.3. Sentiment analysis of twitter data using Hadoop

Tare et al. (2014) have used a Naïve Bayes algorithm to classify the large number of tweets. Tweets were collected using Twitter4j library which internally uses the Twitter REST API. Sheela (2016) performed sentiment analysis on Twitter data using Twitter streaming API, and for the storage of Twitter data, Hadoop's file system was used. Ha et al. (2015) have proposed a method to perform sentiment analysis on Twitter data. They have written map reduce functions to classify tweets. Kumar and Bala (2016) have used Hadoop framework for performing sentiment analysis on a large set of Twitter data as well as for storage purpose. Barskar and Phulre (2017) have discussed an efficient mechanism for performing opinion mining of Twitter data, with the help of Hadoop components such as Apache Flume and Apache Pig. Yadav et al. (2016) presented a feedback analysis system for efficient mining of data wherein analysis is done using the MapReduce framework and Hadoop is used for storage and text classification can be performed by using one of the popular supervised classification method, Naive Bayes algorithm. Nadagoud and Naik (2015) used Twitter data to identify the popularity of Flipkart using Apache Hive. Apache Flume was used to stream continuous information from Twitter. Shang, Shi, Shang, and Hong (2015) have proposed a system that processes public opinion using Hadoop Mahout

cogent ·· engineering

Algorithms. Selvan and Moh (2015) used twitter data to provide feedback by performing sentiment analysis using the Hadoop framework with the help of Cloudera setup. Chauhan and Shukla (2017) have discussed the use of different big data technologies in identifying the opinion of tweets. They have used Naïve Bayes, OpenNLP, LM Classifier, Hadoop and Apache Mahout to build a model, and classification of texts. Fernandes and Rio D'Souza (2017) discuss the semantic analysis of text using rule-based and supervised machine learning techniques which do not deal with huge data set.

Summary of Bigdata analytics using Hadoop is described in Table 1.

3. Proposed method

This section describes the overall framework for capturing and analyzing tweets streamed in real time. As a first part, real-time tweets are collected from Twitter. These tweets are retrieved from Twitter using Twitter streaming API as shown in Figure 1. The Flume (https://flume.apache.org/) Fernandes & Rio D'Souza, is responsible for communicating with the Twitter streaming API and retrieving tweets matching certain trends or keywords. The tweets retrieved from Flume are in JSON format which are passed on to HDFS. This semi-structured twitter data is given as input to the PIG module as well as the HIVE module which will convert nested JSON data into a structured form that is suitable for analysis.

The proposed work is divided into two sections: one is to find the recent trends from real time tweets and another is performing sentiment analysis on trending topic tweets.

3.1. Finding recent trends

Trend is a subject of many posts on social media for a short duration of time. Finding recent trends means to process the huge amount of data collected over the needed period of time. Algorithm 1 briefs on finding popular hash tags.

(1) Finding popular hashtags using Apache Pig

To find popular hashtags of given tweets, the tweets are loaded into Apache Pig module, wherein these tweets are passed through a series of Pig scripts for finding popular hashtags. Following are the steps to determine the popular hashtags in tweets:

(a) Loading the Twitter data on Pig

This streamed twitter data is in JSON format and consists of map data types that is data with key and value pair. For analysis, the tweets stored in HDFS are loaded into PIG module. To load the Twitter data, we used elephant bird JsonLoader jar files which supports to load tweets of JSON format.

Algorithm 1: Finding popular hashtag
Data: *dataset*: = Corpus of tweets
Result: *popular hashtag*
   Load tweets from HDFS to Hadoop ecosystem module
   for each tweet in module
     feature = extract(extract id, hashtag text)
   end
   **for each** feature
     count_id = Count(id€hashtag text)
   end
    popular_hashtag = max(count_id)

---

cogent engineering

Table 1. Summary of big data analytics approaches

Study	Analysis approaches				Twitter4j	Data extraction tools		
	MapReduce	Hive	Pig	Mahout		Flume	Topsy	REST API
Bhardwaj et al. (2015).		✓				✓		
Khade (2016)	✓							
Barskar and Phulre (2017)			✓			✓		
Jain and Bhatnagar (2016)			✓			✓		
Ennaji et al. (2015)	✓					✓		
Ha et al. (2015)	✓						✓	
Verma et al. (2015)				✓				
Shrote and Deorankar (2016)	✓							
Yadav et al. (2016)	✓					✓		
Sheela (2016)	✓							✓
Selvan et al. (2015)		✓		✓		✓		
Kumar and Bala (2016)				✓	✓			
Nadagoud and Naiket (2015)		✓				✓		
Sangeeta. (2016)		✓				✓		
Shang et al. (2015)								
Tare et al. (2014)	✓			✓	✓			✓
Chauhan and Shukla (2017)				✓	✓			

Figure 1. System model for capturing and analyzing the tweets.
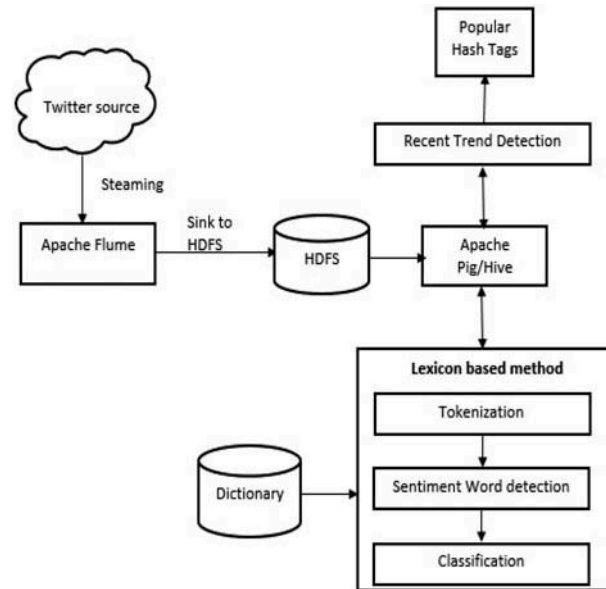
(b) Feature extraction

This step is called preprocessing where Twitter messages containing many fields such as id, text, entities, language, time zone, etc. are looked at. To find famous hashtags, we have extracted tweet id and entities fields where the entity field has a member hashtag. This member is used for further analysis along with tweet id.

A Sample of the result obtained after this phase is shown as follows:

[symbols#{},urls#{([expanded_url#http://smarturl.it/gst-app,indices#{(15),(38)},display_url#smarturl.it/gst-app,url#https://t.co/7MkIaMHzzy])}**,hashtags#{([text#GST,indices#{(0),(4)}]),([text#tax, indices#{(54),(58)}])}**,user_mentions#{}];910449715870818304.

(c) Extract hashtags

Each hashtag object contains two fields: they are text and indices where text field contains the hashtag. So to find famous hashtags, we have extracted text field. The output of this phase is hashtag followed by the tweet id.

For example, GST; 910449715870818304

(d) Counting hashtags

After performing all the above steps, we get hashtag and tweet ids. To find popular hashtags, we first group the relation with respect to hashtag, next we count the number of times the hashtag appeared. Hashtags, which have appeared highest number of times, are categorized as famous hashtags or recent trends.

(2) Finding recent trends using Apache Hive

Recent trends from real-time tweets can also be found using Hive queries. Since tweets collected from twitter are in JSON format, we have to use JSON input format to load the tweets into Hive. We have used Cloudera Hive JsonSerDe for this purpose.

This jar file has to be present in Hive to process the data. This jar file can be added using following command.

add jar <path to jar file>;

Following steps are performed to find the recent trend:

(a) Loading and Feature extraction

The tweets collected from twitter are stored in HDFS. In order to work with Data stored in HDFSusing HiveQL, first an external table is created which creates the table definition in the Hive metastore. Figure 2 shows the query used to create the Twitter table. This query not only creates a schema to store the tweets, but also extracts required fields like id and entities.

(b) Extracting Hashtags

In order to extract actual hashtags from entities, we created another table which contains id and the list of hashtags. Since multiple hashtags are present in one tweet, we used UDTF (User Defined Table generation function) to extract each hashtag on the new row.The outcome of this phase is id and hashtag.

(c) Counting hashtag

After performing all the above steps, we have id and hashtag text.A hive query is written to count the hashtags.

3.2. Sentiment analysis

Sentiment is defined as an expression or opinion by an author about any object or any aspect. The main focus of sentiment analysis is parsing the text and finding the opinion word. After identifying opinion words, sentiment values are assigned to these words. Finally, we need to detect the polarity of the text. Polarity can be positive, negative or neutral. We have used Lexicon approach for sentiment classification. As a parsing step, the sentence is split into words. This step is also called tokenization. These tokenized words are taken as an input for identifying opinion words. We have performed sentiment analysis of real time tweets using Pig and Hive. Algorithm 2 is used to perform the sentiment analysis.

(1) Sentiment analysis using Apache Pig

Figure 2. Query to create a table in Hive.


```
Logging initialized using configuration in jar:file:/home/archanarao/
apache-hive-1.2.2-bin/lib/hive-common-1.2.2.jar!/hive-log4j.propertie
s
hive> CREATE EXTERNAL TABLE twit1
    > (
    > id BIGINT,
    > entities STRUCT<hashtags:ARRAY<STRUCT<text:STRING>>>
    > )
    > ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'
    > LOCATION '/user/goi';
```

The tweets from trending topics fetched using Apache Flume are stored in Hadoop's file system. In order to perform sentiment analysis, these data have to be loaded on Apache Pig. Following are the steps to detect the sentiment in tweets using Pig:

(a) Extracting the tweets

The twitter data loaded into Apache Pig is highly nested JSON format that consists not only of tweets, but also other details such as image, URL, Twitter user id, tweet id, user profile description, location from where the tweets were posted, tweet posting time, etc. The sentiment analysis is done only on tweets. As a preprocessing step of sentiment analysis, we have extracted Twitter id, and tweet from the JSON Twitter data.

Algorithm 2: Sentiment Classification

Data: *dataset*: = Corpus of tweets, list of dictionary words with rating
Result: *classification:* = positive, negative or neutral
Notations: d_r: dictionary_word_rating, T:Tweets, C:Corpus
While *T in C* **do**
    while *words in tweet* **do**
        if *word = = any phrase in dictionary* **then**
            word_rating = d_r;
            continue;
        end
    end
    avg_rating = avg(word_rating)
    if *avg_rating ≥ = 0.0* **then**
        Given tweet is positive
    end
    else if *avg_rating < 0.0* **then**
        Given tweet is negative
    end
    else
        Given tweet is neutral
  end
end

(a) Tokenizing the tweets

Sentiment analysis means finding polarity of sentiment words. In order to find sentiment words, we have to split the entire sentence into words. The process of splitting up of stream of text into words is said to be tokenization. All the tweets collected from the previous step are tokenized and split into words. This tokenized list of words is fed as input for further processing of sentiment analysis.

(a) Sentiment word detection

In order to find the sentiment words from tokenized tweets, we have created a dictionary of sentiment words. This dictionary consists of a list of sentiment words which are rated from + 5 to

−5 depending on their meaning. The words which are rated from 1 to 5 is considered to be positive and the words rated from −5 to −1 are considered to be negative. With the help of this dictionary, tokenized words are rated. To rate the tokenized words, we performed a map side join by joining the token statement and the contents of the dictionary.

(a) Classification of Tweets

After performing all the above steps, we have tweet id, tweet and its associated rating. We have grouped all the tweets that are rated with respect to the tweet id. We now have to calculate the average of the rating given to the tweets. Average rating, AR, of the tweets are calculated using formula (1), where SWR indicates the sum of word ratings and TWP denotes total number words in a tweet.

$$AR = \frac{SWR}{TWP} \qquad\qquad (1)$$

Based on the calculated average rating, we can classify the tweets into positive tweets and negative tweets. The tweets that are rated above zero are treated as positive tweets and below zero are treated as negative tweets. The tweets which do not contain any sentiment words are considered as neutral.

(2) Sentiment Analysis using Apache Hive

The tweets of trending topics stored in HDFS are used for sentiment analysis and processed using HiveQL. The steps are discussed in the following section.

(a) Loading the tweets and feature extraction

In order to perform sentiment analysis, we first need to load the tweets on Hive warehouse. To do that, we create an external hive table in the same directory where the tweets are stored in HDFS. To perform sentiment analysis, we only need tweet id and text.Therefore, we create a table with these two fields. Figure 3 shows the table created to store the Twitter data.

(b) Tokenizing the tweets

In order to find the sentiment words, the tweet is split into words using one of the Hive UDF functions. A Hive table is created to store the tweet id and the array of words present in each tweet. As multiple words are present in an array, we used some built in UDTF function to extract each word from an array and created a new row for each word. Another table is created to store id and word.

(c) Sentiment word detection

Sentiment analysis is done using dictionary-based method. A table is created to store the contents present in the dictionary.

Figure 3. Hive table to store twitter data.

```
Logging initialized using configuration in jar:file:/home/archanarao/
apache-hive-1.2.2-bin/lib/hive-common-1.2.2.jar!/hive-log4j.propertie
s
hive> create external table load_tweets
    > (
    > id BIGINT,
    > text STRING
    > )
    > ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'
    > LOCATION '/user/goi';
```

In order to rate the tokenized words, the tokenized words have to be mapped with the loaded dictionary. We performed left outer join operation on a table that contains id, word and dictionary table if the word matches with the sentiment word in the dictionary, then a rating is given to the matched word or else NULL value is assigned. A hive table is created to store id, word and then rating.

(d) Classification of tweets

After performing all the above steps, we have id, word and rating. Then "group by" operation is performed on id to group all the words belonging to one tweet after which average operation is performed on the ratings given to each word in a tweet. Based on the average ratings, tweets are classified into positive and negative.

4. Results and discussion
In this section, the experimental results of the proposed scheme are discussed.

4.1. Experiment environment
The experimental setup consists of Hadoop implemented in pseudo distributed mode. Hadoop was installed on an Ubuntu 16.04 operating system with 8 GB of RAM. Other components such asApache Flume 1.6.0, Apache Pig 0.16.0 and Apache Hive 1.2.2 were installed on top of Hadoop.
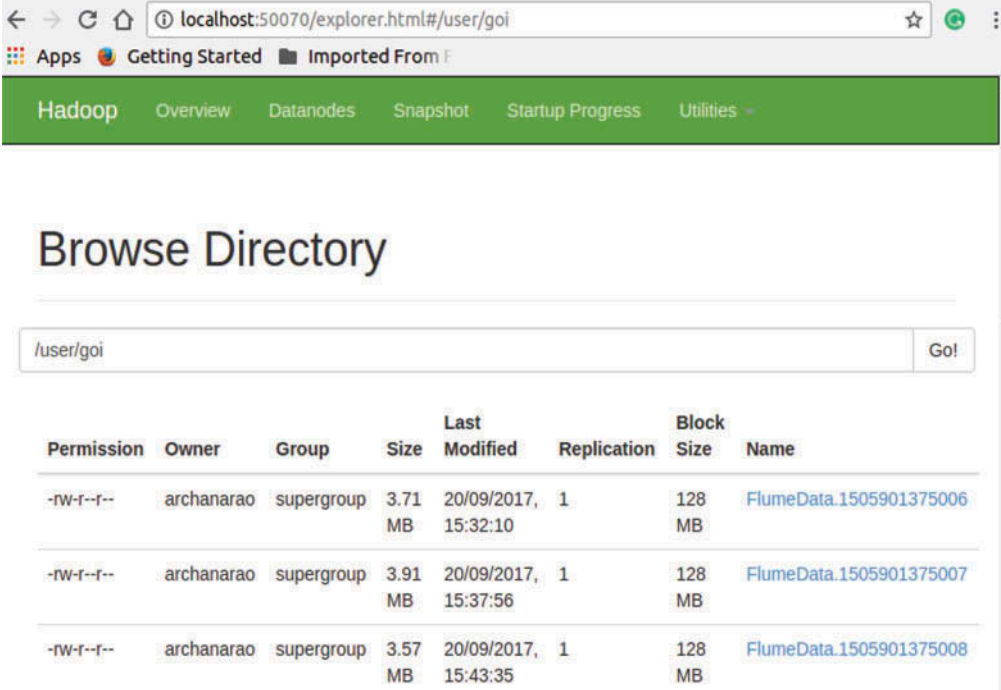
4.2. Datasets fetched for analysis process
The sets of data related to political issues were fetched from Twitter using the Twitter Streaming API and passed through Apache Flume. These fetched tweets are stored in the HDFS. We collected tweets in the month of September 2017. Figure 4 shows the tweets collected from Twitter.

Table 2 shows the datasets for the analysis process. There are six sets of tweets crawled from Twitter. Depending on the number of tweets in each set, the crawling time (in minutes) is given in Table 2.

Figure 4. The real-time tweets collected from Twitter.

Table 2. Dataset captured for analysis		
Datasets	No. of tweets (approx.)	Extraction period (min)
Set 1	5,000	4
Set 2	51,000	40
Set 3	100,000	80
Set 4	200,000	160
Set 5	300,000	240
Set 6	500,000	400

4.3. Discussion on finding recent trends process

We have used Indian political issues for finding recent trends. To find the recent issues in Indian politics, tweets are collected using general keywords such as Government of India, Indian Parliament, Indian Government, and so on. As a first step, real-time tweets are fetched using Apache Flume and stored in Hadoop's file system. Figure 5 shows the sample of tweets loaded into HDFS.

In order to process the tweets, it has to be loaded into Pig and Hive. Next step is to extract features required for further analysis, i.e. id and entities, where entities field has an object called hashtags. Next, we count, how many times the hashtag has appeared in the dataset. Then hashtags having highest count are classified as popular hashtags. Figure 6 shows some hashtags and the count that it appeared in the dataset. According to the result, GST and demonetization are recent trending issues.

4.4. Discussion on sentiment analysis process using Apache Pig and Apache Hive

We have used algorithm 2 to perform sentiment analysis. Samples of the tweets classified as positive and negative are shown in Figures 7 and 8.

We tried extracting the tweets without Hadoop framework, i.e. using Twitter4j API, and found that it is not suitable for large number of data. Hence, traditional method, i.e. without using Hadoop framework, is not flexible and efficient to load real-time streaming. In the proposed

Figure 5. Sample of the collected twitter data.

```
([in_reply_to_status_id_str#,in_reply_to_status_id#,created_at#Wed Sep 2
0 11:33:11 +0000 2017,in_reply_to_user_id_str#,source#<a href="http://tw
itter.com/download/android" rel="nofollow">Twitter for Android</a>,retwe
eted_status#{in_reply_to_status_id_str=null, in_reply_to_status_id=null,
 created_at=Wed Sep 20 06:06:42 +0000 2017, in_reply_to_user_id_str=null
, source=<a href="http://twitter.com" rel="nofollow">Twitter Web Client<
/a>, retweet_count=686, retweeted=false, geo=null, filter_level=low, in_
reply_to_screen_name=null, is_quote_status=false, id_str=910384703558971
392, in_reply_to_user_id=null, favorite_count=953, id=910384703558971392
, text=Things that are good for nation:
- Expensive petrol
- Complicated GST
- Statues costing 6K Crore
- Fake news
Things that are bad:
- Dissent, place=null, lang=en, quote_count=21, favorited=false, coordin
ates=null, truncated=false, reply_count=22, entities={urls={}, hashtags=
{}, user_mentions={}, symbols={}}, contributors=null, user={utc_offset=-
25200, friends_count=34, profile_image_url_https=https://pbs.twimg.com/p
rofile_images/644074186814566400/KGIJWzZc_normal.jpg, listed_count=149,
profile_background_image_url=http://abs.twimg.com/images/themes/theme1/b
g.png, default_profile_image=false, favourites_count=46, description=Sta
ndup comedy, music, and stories of our complicated & very funny country.
 (Contact: aisitaisidemocracy at gmail), created_at=Tue Aug 04 12:25:52
```
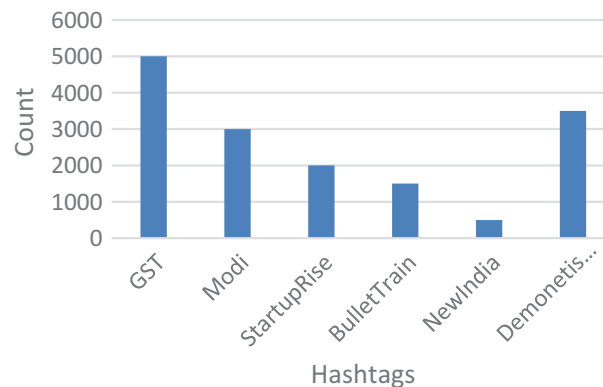
Figure 6. Sample of popular hashtags.

Figure 7. Tweets classified as positive.

work, we used Hadoop framework with Apache Flume to stream the data and were able to load huge tweets with less extraction time.

A total of approximately 500,000 tweets were loaded into Hadoop components, i.e. Apache Pig and Apache Hive. Of them, approximately 300,000 were related to GST, and remaining tweets were related to the issue of demonetization. Under GST, approx. 150,000 tweets were classified as positive and 90,000 tweets were classified as negative.

Similarly, with demonetization, 100,000 tweets were classified as positive and 75,000 tweets were classified as negative. Remaining tweets which were not related to the domain were classified as neutral. Figure 9 shows that the count of positive tweets is greater than the count of negative tweets. Hence, conclusion can be drawn from the above classification that users have a positive attitude toward the issue of demonetization and GST being discussed.

4.5. Performance evaluation
We have used Pig Latin and HiveQL languages to process real-time tweets. Depending on type of data and purpose, developers can choose either Pig or Hive. Some differences are, Pig is mainly used by researchers and programmers whereas Hive is ETL (Extract-Transform-Load) tool used by

Figure 8. Tweets classified as negative.

```
(892257769046163456,RT @paulkidd: @Asher_Wolf Remember when John H
oward told us the GST would prevent tax evasion? Now consumers are
 being asked to police it.);-1
(892271602204803072,RT @MoeedNj: If opp &amp; media charges agains
t Shahid Khaqqan true then why all ignored till he is nominated as
 PM? https://t.co/RcfGe8Jd8d vi…);-2
(892257769046163456,RT @paulkidd: @Asher_Wolf Remember when John H
oward told us the GST would prevent tax evasion? Now consumers are
 being asked to police it.);-1
(892274152597372928:RT @MarkMcGowanMP: This just won't cut it.
https://t.co/0vp9a6JqJk);-1
(892272909451788288,RT @MarkMcGowanMP: This just won't cut it.
https://t.co/0vp9a6JqJk);-1
(892257769046163456,RT @paulkidd: @Asher_Wolf Remember when John H
oward told us the GST would prevent tax evasion? Now consumers are
 being asked to police it.);-1
(892266500589072384,RT @PTIPunjabPK: Shahid Khaqan Abbasi &amp; Sh
ahbaz Sharif will disqualified under corruption charges, @EjazChau
dhary
#PTI https://t.co/Uut9q…);-2
(892267063925891072,RT @t_d_h_nair: 3/ Modi and Jaitley put LPG un
der 5% GST. Earlier few states charged VAT on LPG ranging between
2%-4%.
https://t.co/4CF1vv…);-3
```

Figure 9. Graph for sentiment analysis of Twitter data.
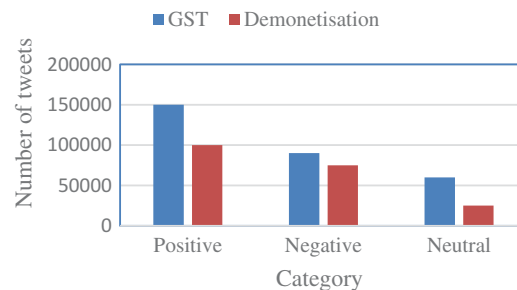
Data Analysts to generate reports. Hive operates on the server side of cluster but Pig operates on client side. In Hive, we need to define the table beforehand and store schema details whereas in Pig there is no dedicated metadata database and schemas. Pig supports complex data types like Map, Tuple and Bags and provides high-level operators, namely ordering and filtering which help in structuring semi-structured data whereas Hive is suitable for structured data. Tweets are stored in semi-structured format and we process them using Pig and Hive. Both gave same results, but Pig execution was faster than Hive. The execution of Pig is faster because its architectural design supports nested data structures and provides high-level operators for processing semi-structured data.

4.6. Execution time for finding recent trends

Finding recent trends is the first stage in the proposed work. The steps followed to find the recent trends have been discussed in the proposed work. We implemented this method on both Pig and Hive and estimated the execution time. Figure 10 shows the time taken for execution by Pig and Hive. In Figure 10, the solid line indicates the time taken for counting hashtags by Hive framework and dashed line indicates the time taken by Pig framework. From Figure 10, we can conclude that Hive module takes more time for execution than Pig.

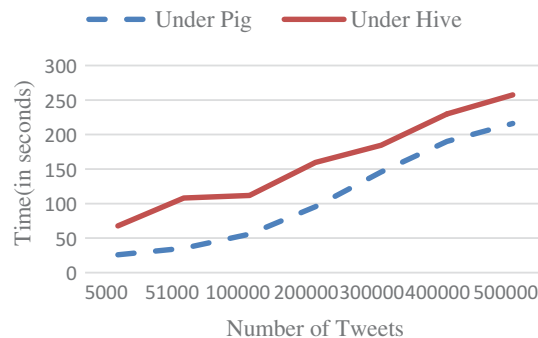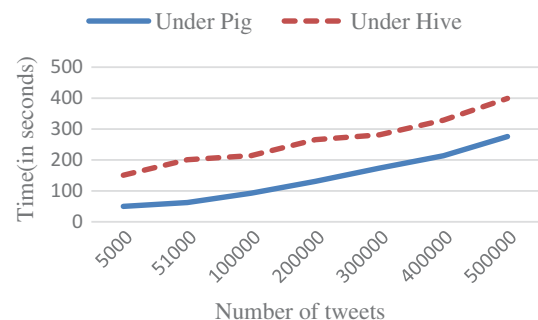**Figure 10. Execution time for finding recent trends under Pig vs. Hive.**

Figure 11. Processing time to analyze sentiment in tweets under Pig vs. Hive.

4.6.1. Execution time for sentiment detection

Sentiment detection is an important phase in the proposed approach. Therefore, we used Algorithm 2 to detect the opinion for all the datasets. We implemented Algorithm 2 on Pig as well as Hive and estimated the execution time as shown in Figure 11. The solid line indicates the time taken for sentiment detection under the Pig module, whereas the dotted line shows time for sentiment detection under Hive framework. Hive module takes approx. 150 s to analyze 5000 tweets whereas Pig module takes around 50 s. Likewise, for 500,000 tweets, Pig module takes approx. 275 s whereas Hive module takes 399 s. From Figure 11 we can come to a conclusion that Pig module works faster than Hive.

5. Conclusion

Social media dependency is inevitable, which has resulted in the generation of an abundant amount of data sets, making the method of processing and analyzing of data a challenge. Extensive dependence on social media data such as Twitter data, e-commerce data, etc. have gained much attention in the area of sentiment analysis. Hadoop proves to be an efficient framework for huge data analysis since Hadoop operates in a fault-tolerant manner. In addition to this, Hadoop can be integrated with Apache Pig, Hive, Oozie, Zookeeper, Sqoop, etc., which promises improved efficiency and performance of Hadoop. Pig Latin and HiveQL languages ease the complexity of writing complex MapReduce programs. In the proposed work, Hadoop framework has been used which is integrated with Apache Flume to fetch data from Twitter and Apache Pig and Hive are used to perform analysis on extracted Twitter data. First, recent trends in the extracted tweets were determined and then sentiment analysis was performed on the retrieved data. The experiment was performed on two Hadoop Ecosystem components, i.e. Pig and Hive and execution time were recorded. From the experimental results, conclusion can be drawn that Pig is more efficient than Hive as Pig takes less time for execution than Hive.

cogent ··engineering

Author details
Anisha P. Rodrigues[1]
E-mail: anishapr@nitte.edu.in
Niranjan N. Chiplunkar[1]
E-mail: nchiplunkar@nitte.edu.in
[1] Computer Science and Engineering Department, NMAM Institute of Technology, Nitte, India.

Citation information
Cite this article as: Real-time Twitter data analysis using Hadoop ecosystem, Anisha P. Rodrigues & Niranjan N. Chiplunkar, *Cogent Engineering* (2018), 5: 1534519.

References
Barskar, A., & Phulre, A. (2017). Opinion mining of twitter data using Hadoop and Apache Pig. *International Journal of Computer Applications*, 158, 9. doi:10.5120/ijca2017912854

Bhardwaj, A., Kumar, A., Narayan, Y., & Kumar, P. (2015, December). Big data emerging technologies: A case study with analyzing twitter data using apache hive. In *Recent Advances in Engineering & Computational Sciences (RAECS), 2015 2nd International Conference on* (pp. 1–6). Chandigarh: IEEE.

Chauhan, V., & Shukla, A. (2017, April). Sentimental analysis of social networks using MapReduce and big data technologies. *International Journal of Computer Science and Network*, 6(2), 120–13.

Ennaji, F. Z., El Fazziki, A., Sadgal, M., & Benslimane, D. (2015, November). Social intelligence framework: Extracting and analyzing opinions for social CRM. In *Computer Systems and Applications (AICCSA), 2015 IEEE/ACS 12th International Conference of* (pp. 1–7). Marrakech, Morocco: IEEE.

Fernandes, R., & Rio D'Souza, G. L. (2017). Semantic analysis of reviews provided by mobile web services using rule based and supervised machine learning techniques. *International Journal of Applied Engineering Research*, 12(22), 12637–12644.

González-Ibánez, R., Muresan, S., & Wacholder, N. (2011, June). Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2* (pp. 581–586), Portland, Oregon.

Ha, I., Back, B., & Ahn, B. (2015). MapReduce functions to analyze sentiment information from social big data. *International Journal of Distributed Sensor Networks*, 11, 417502. doi:10.1155/2015/417502

Jain, A., & Bhatnagar, V. (2016). Crime data analysis using pig with Hadoop. *Procedia Computer Science*, 78, 571–578. doi:10.1016/j.procs.2016.02.104

KadharBasha, J., & Balamurugan, M. (2017, May). A review on Hive and Pig. *International Journal of Advanced Research in Basic Engineering Sciences and Technology*, 3(39), 53–58.

Khade, A. A. (2016). Performing customer behavior analysis using big data analytics. *Procedia Computer Science*, 79, 986–992. doi:10.1016/j.procs.2016.03.125

Kumar, M., & Bala, A. (2016, March). Analyzing twitter sentiments through big data. In *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on* (pp. 2628–2631). New Delhi, India: IEEE.

Nadagoud, M. S., & Naik, M. K. D. (2015, May). Market sentiment analysis for popularity of Flipkart. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 4(5), 2117–2123.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013, October). Sarcasm as contrast between a positive sentiment and negative situation. *EMNLP*, 13, 704–714.

Sangeeta. (2016, February). Twitter data analysis using Flume & Hive on Hadoop frame work. *International Journal of Recent Advances in Engineering & Technology*, V4, I–2.

Selvan, L. G. S., & Moh, T. S. (2015, June). A framework for fast-feedback opinion mining on Twitter data streams. In *Collaboration Technologies and Systems (CTS), 2015 International Conference on* (pp. 314–318). Atlanta, GA, USA: IEEE.

Shang, S., Shi, M., Shang, W., & Hong, Z. (2015, June). Research on public opinion based on big data. In *Computer and Information Science (ICIS), 2015 IEEE/ACIS 14th International Conference on* (pp. 559–562). Las Vegas, NV, USA: IEEE.

Sheela, L. J. (2016). A review of sentiment analysis in twitter data using Hadoop. *International Journal of Database Theory and Application*, 9(1), 77–86. doi:10.14257/ijdta

Shrote, K. R., & Deorankar, A. V. (2016, February). Review based service recommendation for big data. In *Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2016 2nd International Conference on* (pp. 470–474). Chennai, India: IEEE.

Tare, M., Gohokar, I., Sable, J., Paratwar, D., & Wajgi, R. (2014). Multi-class tweet categorization using map reduce paradigm. *International Journal of Computer Trends and Technology (IJCTT)*, 9(2), 78–81. doi:10.14445/22312803/IJCTT-V9P117

Verma, J. P., Patel, B., & Patel, A. (2015, February). Big data analysis: Recommendation system with Hadoop framework. In *Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on* (pp. 92–97). Ghaziabad, India: IEEE. doi:10.1002/mus.24271

Yadav, K., Pandey, M., & Rautaray, S. S. (2016, November). Feedback analysis using big data tools. In *ICT in Business Industry & Government (ICTBIG), international conference on* (pp. 1–5). Indore, India: IEEE.

cogent •• engineering

Cogent Engineering (ISSN: 2331-1916) is published by Cogent OA, part of Taylor & Francis Group.

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com