Analysis of Popular Hashtags on Twitter Data: Impact on Marketing

Tanvir Faysal[1], Asif Imran Chowdhury[2], Ahmed Salman Tariq[3], Md. Harun Or Rashid[4]

[12]B.Sc. Students, Dept. of Computer Science and Engineering, Bangladesh Army University of Engineering & Technology (BAUET), Natore-6431, Bangladesh
[3]Senior Lecturer, Dept. of Computer Science and Engineering, Bangladesh Army University of Engineering & Technology (BAUET), Natore-6431, Bangladesh
[4]Lecturer, Dept. of Computer Science and Engineering, Bangladesh Army University of Engineering & Technology (BAUET), Natore-6431, Bangladesh

---***---

Abstract - *Twitter is one of the most popular social networking and real time communication app that allows its user to communicate. It is a micro blogging site where people can express themselves creating status called tweets. Tweets are actually the opinion or idea about different things. Creating status people use hashtags for important topics. For that reason, the main aim of this work is the analysis of popular hashtags on twitter data using Scala and Spark tools to analyze twitter hashtags on real time and making some decision on real time data. This actually opens up an opportunity to read peoples mind and allows you to organize content and track discussion topics based on those keywords. It is a short way to classify and call attention to what you are saying. Hashtags empower peoples saying and makes them a part of global multi-platform dialogue. That means that understanding hashtags and hashtags trends related to brand can help the brand to connect and engage with the audience, quantify the impact of marketing efforts and discover important influencer. This project will help to understand people's opinion based on analysis of twitter hashtags using Scala and Spark. This project can be able to count number of hashtags on a topic in each second interval. A graphical representation has been presented to reflect public demand in business.*

Key Words: Twitter; Hashtags; Scala; Spark tools; Real time data; Analysis; Decision making.

1. INTRODUCTION

The Twitter micro blogging site operates an algorithm to determine which topics are the most discussed via Twitter users at any given time. The most popular topics are known as "trending topics." Significant world events, international sports results, and news about popular celebrities are among the items that commonly "trend" on the Twitter site [1]. Twitter is a widely used social networking site that allows users to post short text "tweets" of up to 280 characters. Twitter works by allowing you to "follow" other users, who may be friends, celebrities or companies. In turn, other Twitter users can choose to follow you. When you log on to Twitter, you see a feed, which shows you the tweets of all the users you have chosen to follow. It is common for many Twitter users to be tweeting about the same topics at the same time. Twitter's trending algorithm identifies these popular topics by detecting when words or phrases are being frequently mentioned in tweets. Twitter then lists the current top 10 topics on your Twitter home page. You can click on any of the topics in the "Trends" list to view a feed of all the recent tweets containing the trending topic. This includes tweets from all Twitter users, not just those you have chosen to follow. Twitter users commonly use "hashtags" to participate in trending topics. To use a hashtag, you include the name of the topic after the hash (#) symbol within your tweet. For example, during an international sports event such as the World Cup, Twitter users will append "#worldcup" to the end of their tweets to show their tweet is related to the topic. The tweet then appears as part of the feed if people click on the "#worldcup" topic on the Trends list. Twitter sorts trending topics by geographical region. By clicking the "Change" link near the Trends list, you can choose to view worldwide trending topics, or to view the topics trending in a specific country or city.

2. BACKGROUND STUDY

Analysis of popular hashtags on twitter data using Scala and spark is the project that will help people to understand the current trend on twitter. In this project we used spark and Scala language. Apache Spark is an open-source cluster computing framework. Spark is a general-purpose data processing engine, suitable for use in a wide range of circumstances [2]. Spark Streaming supports the ingest of data from a wide range of data sources, including live streams from Apache Kafka, Apache Flume, Amazon Kinesis, Twitter, or sensors and other devices connected via TCP sockets. Data can also be streamed out of storage services such as HDFS and AWS S3. Data is processed by Spark Streaming, using a range of algorithms and high-level data processing functions like map, reduce, join and window. Processed data can then be passed to a range of external file systems, or used to populate live dashboards. On the other hand, Scala is practically the de facto language for the current Big Data tools like Apache Spark, Finagle, Scalding, etc. Many of the high-performance data science frameworks that are built on top of Hadoop usually are written and use Scala or Java. The reason Scala is used in these environments is because of its amazing concurrency support, which is key in parallelizing a lot of the processing needed for large data sets. It also runs on the JVM, which makes it almost a no-brainer when paired with Hadoop [11].

Although the notion of sentiment analysis, or opinion mining, is relatively new, the research around this domain is quite extensive. The area of an opinion mining also known as sentiment analysis has recently enjoyed a huge burst of research activity [3]. The year 2001 or so seems to mark the beginning of widespread awareness of the research problems and opportunities that sentiment analysis and opinion mining raise due to the following factors: (i) the development of machine learning methods in natural language processing and information retrieval (ii) the availability of training datasets for machine learning algorithms, and (iii) realization of the fascinating intellectual challenges and commercial and intelligence applications that the area offers. Early studies focus on document level sentiment analysis concerning movie or product reviews and posts published on web pages or blogs. Due to the complexity of document level opinion mining, many efforts have been made towards the sentence level sentiment analysis [4]. A less investigated area is the topic-based sentiment analysis due to the difficulty to provide an adequate dentition of topic and how to incorporate the sentiment factor into the opinion mining.

3. METHODOLOGY

Analysis of popular hashtags on twitter data using spark and Scala is a project where a huge amount of data is loaded in and processed in real-time. For this project spark and as a language Scala is used. For running the project first, we need to load data form twitter. Real-time. The real authentication is needed for this most importantly. At the time of creating twitter app we need to provide all the information for running the project successfully. Twitter works with many different types of data. But the project working with a particular type of data called hashtags. So, project pull the hashtag in every two seconds for analysis. Then need to look particular Hashtags related to brand. If the Hashtag is relevant to the topic is will be counted. That means it will be provided for processing.

After processing the data, a graphical representation will be presented to describe the public opinion. It pulls the real scenario what people thinking right now.

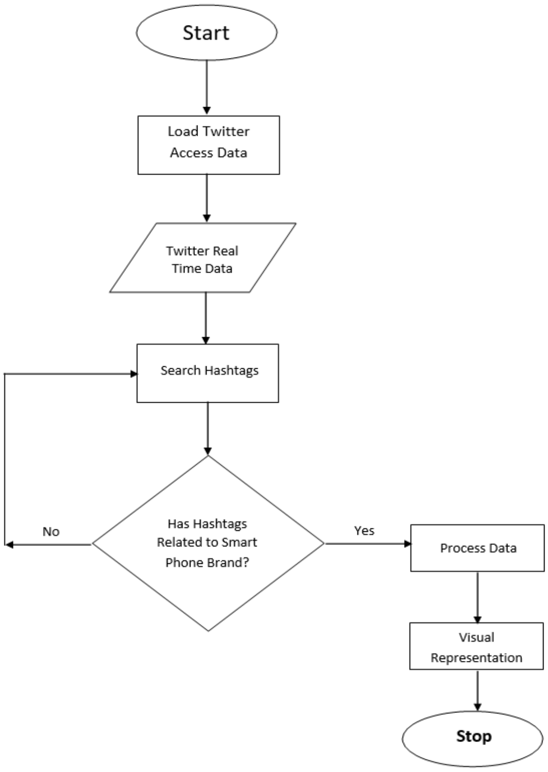


Fig. 1: System flowchart.

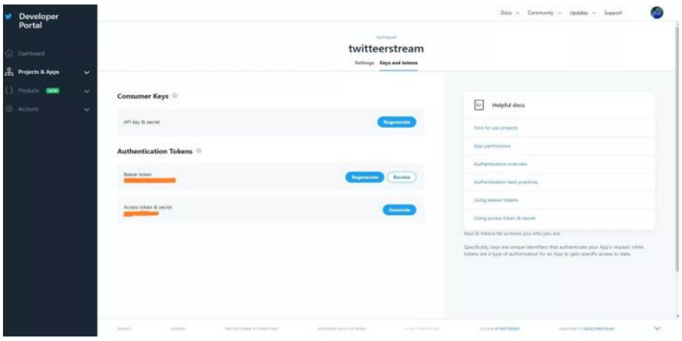
4. PROCEDURE

4.1 Implementation

Implementation is the realization of an application, or execution of a plan, idea, model, design, specification, standard, algorithm, or policy. We have worked hard to put things together and make a better and efficient approach to fetch twitter hashtags. Then we have analyzed our result for comparison and making decision. We have also tested our approach if it is suitable or not for predicting peoples' mind.

4.2 Creating Twitter App

For creating twitter app first, we had to create a twitter developer account then after that we needed to visit **developer.twitter.com** to create a twitter app.



[redacted]

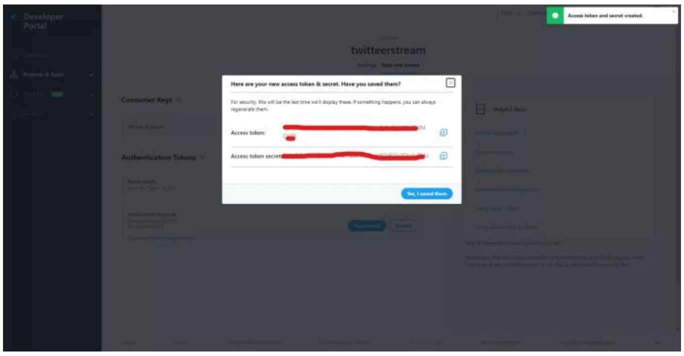
[redacted]

Keys & tokens let us know you who you are.

Specifically, keys are unique identifiers that authenticate your App's request, while tokens are a type of authorization for an App to gain specific access to data.

Fig. 2: Creating twitter app.

We needed to fill up the application. Application Name should be unique. Otherwise, it will show an error. After successfully completing the application form, one has to click on manage keys and access token for creating Consumer Secret Key.


twitteerstream

Here are your new access token & secret. Have you saved them?

For security, this will be the last time we'll display these. If something happens, you can always regenerate them.

Access token: [redacted]

Access token secret: [redacted]

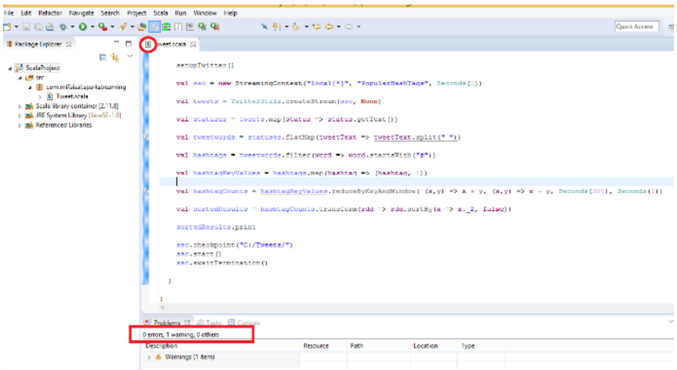
Yes, I saved them

Fig. 3: Creating consumer secret key.

After creating consumer secret key, we will have our API Key and Secret, API private key also accesses token and accesses secret token.

4.3 Scala Project

Now we had to open Scala IDE. Here we have selected that folder where we kept the twitter.txt file. We have gone to File and created a new Scala Project. File => New => Scala Project, right click on src and create a new Package. Here com.mifaisal is the Domain Name and after that we have used a name as sparkstreaming. One can use any name that he likes. Src => Package creating package for Scala. After creating a package we have to right click on package name and create a new Scala Object. Object name first character must be Capital letter. Here we have used Tweet, one can use any name that he likes most. com.mifaisal.sparkstreaming => Scala Object. Here ssc.checkpoint() set the checkpoint directory. So we have selected that folder directory where we have kept the twitter.txt file.


File Edit Refactor Navigate Search Project Scala Run Window Help
Quick Access
Package Explorer
ScalaProject
src
com.mifaisal.sparkstreaming
Tweet.scala
Scala library container [2.11.8]
JRE System Library [JavaSE-1.8]
Referenced Libraries
Tweet.scala
Problems
Tasks
Console

```
    setupTwitter()

    val ssc = new StreamingContext("local[*]", "PopularHashTags", Seconds(1))

    val tweets = TwitterUtils.createStream(ssc, None)

    val statuses = tweets.map(status => status.getText())

    val tweetwords = statuses.flatMap(tweetText => tweetText.split(" "))

    val hashtags = tweetwords.filter(word => word.startsWith("#"))

    val hashtagKeyValues = hashtags.map(hashtag => (hashtag, 1))

    val hashtagCounts = hashtagKeyValues.reduceByKeyAndWindow( (x,y) => x + y, (x,y) => x - y, Seconds(300), Seconds(1))

    val sortedResults = hashtagCounts.transform(rdd => rdd.sortBy(x => x._2, false))

    sortedResults.print

    ssc.checkpoint("C:/Tweets/")
    ssc.start()
    ssc.awaitTermination()

  }

}
```

0 errors, 1 warning, 0 others

Description	Resource	Path	Location	Type
Warnings (1 item)				

Fig. 4: Scala code for streaming.

Spark Streaming is an extension of the core Spark API that enables scalable, high- throughput, fault-tolerant stream processing of live data streams. In this project we are using spark and Twitter as data source to do a comparative analysis.

5. RESULT ANALYSIS

5.1 Spark UI Analysis

Spark Streaming provides a high-level abstraction called discretized stream or DStream, which represents a continuous stream of data. DStreams can be created either from input data from twitter. Internally, a DStream is represented as a sequence of RDDs. Here, the Fig. 5 shows the URL path of spark UI.

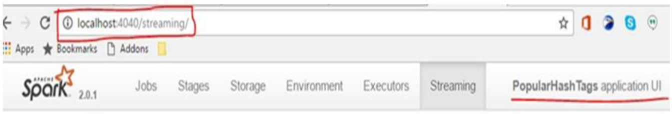


Fig. 5: Spark UI URL.

Streaming context takes two parameters; your application configuration and the streaming time. As Spark streams data in micro batches, we need to set some time so that for every set time (time set), be it seconds or milliseconds, it will stream the data. Here, we have set 5 seconds, so for every 5 seconds, it will stream the data from Twitter and save it in a new file.

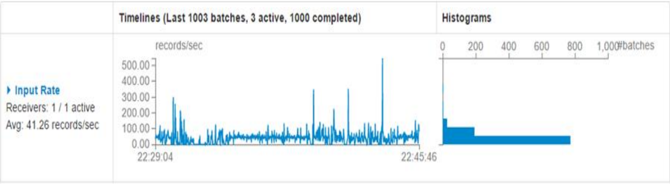


Fig. 6: Input rate of twitter data.

5.2 Product Comparison

Here is a comparative analysis of twitter hashtag between two famous phone company Samsung and Apple. People are discussing now a day's lot of these two companies because they released their new series of phone. Here is a comparative hashtag analysis between Samsung and iPhone. For this analysis we run program for 10 days around 4 hours in each day. Below shows the number of hashtags pulls by using the keyword iPhone and Samsung.

Table-1: Number of Hashtags for iPhone

#iPhone	
Day	Number of Hashtags
1	31
2	47
3	27
4	36
5	40
6	43

7	50
8	29
9	35
10	42

Table-2: Number of Hashtags for Samsung

#Samsung	
Day	Number of Hashtags
1	16
2	13
3	22
4	14
5	17
6	09
7	11
8	12
9	27
10	18

Now for calculation average for the each. We know,

average = $\frac{\sum_{i=1}^{n} x_i}{n}$

For #iPhone, average = 380/10 = 38

For #Samsung, average = 159/10 = 15.9

Sum of both averages = 38+15.9 = 53.9

So, Percentile of iPhone= (38/53.9) *100=70.50%

And Percentile of Samsung= (15.9/53.9) *100=29.50%

So, as we can see from the average value iPhone is much popular and trending then Samsung considering the number of hashtags.

Graphical representation of the comparison satisfies our previous decision that people are more talking about iPhone compare to Samsung. So, iPhone is more trending.

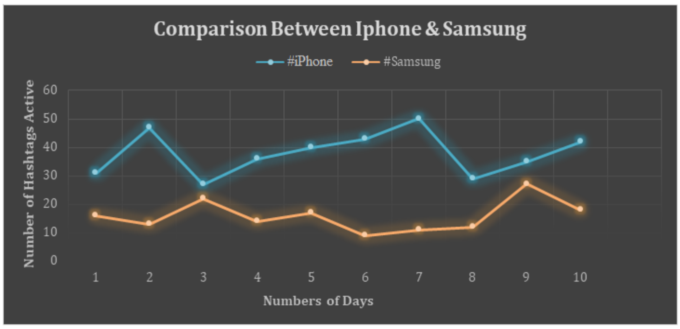


Fig. 7: Comparison graph.

6. CONCLUSIONS

A proper use of our work would significantly help to generate idea about current topic discussing in the world. Apache Spark is the shiny new toy on the Big Data playground, but there are still use cases for using Hadoop MapReduce. Spark has excellent performance and is highly cost-effective thanks to in-memory data processing. It's compatible with all of Hadoop's data sources and file formats, and thanks to friendly APIs that are available in several languages, it also has a faster learning curve. Spark even includes graph processing and machine-learning capabilities. We are using the power of spark to analyze twitter data for exploring current trend. Dealers would be able to make decisions easily about dealing in popular and trendy products on area or time basis. Thus it might have a great impact on marketing.

REFERENCES

[1] Twitter trending, https://www.hashtags.org/featured/what-do-twitter-trends-mean, (November 10, 2020).

[2] Spark,http://searchbusinessanalytics.techtarget.com/definition/Apache-Spark, (November 10, 2020).

[3] Sentiment Mining, http://searchbusinessanalytics.techtarget.com/definition/opinion-mining-sentiment-mining, (January 5, 2021).

[4] Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis." Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005.

[5] Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: a survey." Mobile Networks and Applications 19.2 (2014): 171-209.

[6] Fan, Jianqing, Fang Han, and Han Liu. "Challenges of big data analysis." National science review 1.2 (2014): 293-314.

[7] Lee, Seong-Hoon, and Dong-Woo Lee. "Big Data Processing and Utilization." Journal of Digital Convergence 11.4 (2013): 267-271. http://www.iproject.com.ng/computer-science/final-year-project- topics/design-and implementation-of-student-project-allocation-and-verification-system/project-topics

[8] Watanabe, Takahiro. "Batch processing system." U.S. Patent No. 9,244,719. 26 Jan, 2016.

[9] McKenna, Aaron, et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next- generation DNA sequencing data." Genome research 20.9 (2010): 1297-1303.

[10] Marz, Nathan, and James Warren. Big Data: Principles and best practices of scalable realtime data systems. Manning Publications Co., 2015.

[11] Lee, Kyong-Ha, et al. "Parallel data processing with MapReduce: a survey." AcM sIGMoD Record 40.4 (2012): 11-20.