INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

# BIG DATA AND TWITTER

**Shilpi Taneja[1], ManishTaneja2**

[1]*University of Delhi, shilpi.asija@gmail.com*
[2] *Independent Scholar, mox7@rediffmail.com*

## Abstract

Web 2.0 (which is driven by social media) has ushered the age of data driven analytics as organizations experiment with newer ways to capture and analyse their data. The traditional approach of using relational databases (using SQL – structured query language) has started giving way to NOSQL databases that help us analyse vast amount of structured / unstructured data. In this paper, we discuss about Big Data – the term and its generally accepted definition and approaches that are available to analyse the same. We then take the example of Twitter, which is a social media giant that generates enormous amount of data, as a reference to explain how the concepts related to Big Data analytics can be applied. We provide an overview of commonly used approaches to access and analyse twitter data. There are many online sites / tools available which generate twitter related analytics for their clients / users. Some of the features of these tools are available for free (basic analytics), but for advanced analytics users are expected to subscribe and pay a monthly fee. We compare features of four such online tools and recommend those which provide better features in its free form.

**Keywords***:* Big Data, Twitter, Data Analytics, Twitter Analytic Tools

## 1. What is Big Data?

Big data is a term for collection of data sets so large and complex that are difficult to manipulate or interrogate with standard database management tools (for ex. Relational databases, desktop statistics or visualization packages). In 2001, Doug Laney, an analyst with Meta Group (now Gartner) published a research note titled “3D Data Management: Controlling Data Volume, Velocity and Variety”. These 3Vs have now become the defining characteristics of big data which distinguish it from other types of Business Intelligence concepts. To these 3Vs, analysts/organizations have added a 4th V called Veracity (Veracity is based on IBM research and there are other Vs being discussed like Value, Validity, Volatility which we’ll not be going into much detail).

a) Volume – Refers to the vast amount of data that gets generated every second. Government and Enterprises are now flooded with data coming in from various sources and channels and need to make sense of it. Data is being generated from mobile devices, social media (Facebook posts, twitter, Instagram etc.), traditional business interactions (like Walmart / online purchases), financial transactions, sensors in everyday objects like cars, airplanes, refrigerators, etc.

b) Velocity – Speed at which data is generated / delivered / captured / analysed. Twitter users generate on average ~500 million tweets / day. Or for catching credit card frauds, companies must analyse 5 million trade events daily to identify fraud. Big data technologies allow us to analyse data while it is being generated without ever putting it into a database.

c) Variety – Data can come in any form / type. Streaming / non-streaming. Structured / Unstructured / Semi structured. For example, emails, videos, audio, tweets, clicks, log files, sensor data, transaction data etc..

d) Veracity - Refers to uncertainty in the data. With many forms of data (for example twitter posts, Facebook, websites – a metric like social sentiment – quality and accuracy of which can be questioned. The volumes involved often make up for lack of quality / accuracy.

## 2. Approaches to Analyse Big Data

The traditional BI(Business Intelligence) approaches do not work well when analysing big data because it require some pre-processing of data before it can be put into relational database for analysis. Also, the scalability of some of these databases given volumes of data involved is suspect. Storage connections are also a bottleneck. The following are generally used for managing Big Data

a) Distributed Parallel Processing – Means moving data and processing to many slave nodes at the same time and tracking progressing. Apache Hadoop and MapReduce allow for implementing the same.

b) In Memory Databases – These platforms use memory as a system for data access. The data is loaded into the memory for faster processing. No storage is needed for analytics purpose – almost real time analytics is available, but data availability is suspect and redundancy needs to be built in to avoid loss of data in event of power failure / shutdown.

c) NO SQL Databases – NO SQL Databases are not based on rigid schemas and can easily incorporate new data types. They deliver higher throughput (query response times) when big data is considered compared to relational databases. They allow linear and unlimited scalability. We can use various approaches of data representation to manage big-data on these databases.

   a. Key Value Stores – Consist of large number of key and value pairs. Value is accessed using a key which references the value. They are very suitable for internet data (click streams) have to be processed at high speeds.
   b. Document Stores – data is stored as documents. Documents are referenced by unique names (keys). They are used for storing contiguous data (like HTML Page) or serialized object structures (e.g. in JSON format) .
   c. Columnar Stores – Most common representation of data used in areas where there is large amount of structured data that cannot be processed in a relational database.
   d. Graph Databases – In graph databases information is represented by vertices and edges . Use cases include determining relationships between people in social networks, searching (PageRank algorithm).

d) Complex Event Processing – Are used we need to analyse data in real time (or in motion). Depending upon the need we will need to decide on an approach to parallel processing / in memory database / NO SQL Database to implement this set up. Which triggers a particular action whenever an event is captured for which action is needed. For example, information being shared by jet engines and sent to manufacturer / airline and only in certain conditions do we need to trigger event to inform users of a malfunction so that necessary action can be taken.

In the following sections of this paper, we will take Twitter as a reference to explain how some of the concepts related to Big Data and its analytics are applied in analysing data generated on Twitter.

## 3. What is Twitter?

Twitter (www.twitter.com) is an online social networking and micro blogging platform that enables a user to send and read short 140 character text messages, called “tweets”. It is ranked amongst the top 10 most visited websites by Alexa’s web traffic analysis for the last 2 years. It has ~250+ million monthly active users who send about ~500+ million tweets per day (ref. Wikipedia). 78% of the twitter users access it on a mobile device and 77% are living outside US. It supports 35+ languages. Its vast global reach and usage has made Twitter into a remarkable platform for analysing and understanding trends and events shaping the world either on a global or a local scale. It is a great medium to understand ground realities of global events (like natural disasters, terror-attacks) as users tweet their impressions even before news crews are able to reach the areas. Given the reach, businesses, politicians, and celebrities are using it to market their messages / views / opinions to their followers. The volume, velocity and variety of data that gets generated on twitter fits the description of big data. We now explore some approaches on analysing the same.

### *3.1 Twitter Data Analysis*

Twitter shares its data in document store format (JSON – JavaScript Object Notation) and allows developers to access it using APIs. Twitter APIs can be accessed only via authenticated requests. Twitter uses Open Authentication (OAuth) and each request must be signed with valid twitter credentials. OAuth provide a safer alternative because the user generally logs into twitter and approves sharing of his information to each application. The user can anytime de-authenticate certain applications from accessing his tweets. API users can use the REST API (for pull access – where we must specify the user credential whose data is sought) or Streaming API (for push access – once a request is made they provide continuous stream of updates with no further input required from user). Twitter also places something called a Rate Limit (restriction on amount of data calls that a particular application can make) based on which data gets shared. Once we have received data from twitter and placed it in our NOSQL database we can apply Map-Reduce / Graph theory approaches to analyse the same.

***3.1.1 Map Reduce* -** is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster. It allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster or even standalone computers. Originally, it was developed at Google for indexing Web pages but it has now become a generic term. It has two main functions – Map and Reduce.

Map – is a function that parcels out work to different nodes in the distributed cluster. It does filtering and sorting (for example identifies unique words in a message, makes one queue for each word).

Reduce – is a function that collates the work and resolves the results into a single value. It performs what we call summarizing operation (like counting the number of times a word appears in the text and giving frequencies).

The MapReduce framework is fault-tolerant because each node in the cluster is expected to report back periodically with completed work and status updates. If a node remains silent for longer than the expected interval, a master node makes note and re-assigns the work to other nodes.

### *3.1.2 Map Reduce Analysis of Twitter Data –*

a) Trending – By analysing individual tweets and looking for certain words amongst them and using map-reduce we can filter out key words that are trending (or being used by a lot of users on twitters). It gives view of what are the key topics.

b) Sentiment Analysis – looking for keywords about brands and analysing them to compute a score of sentiment for that brand.

***3.1.3 Graph Theory*** is the study of graphs, which are mathematical structures used to model pairwise relations between objects. A "graph" is made up of "vertices" or "nodes" and lines called edges that connect them. A graph may be undirected, meaning that there is no distinction between the two vertices associated with each edge, or its edges may be directed from one vertex to another (ref. Wikipedia). Relationships in a social network are treated like graphs. For example, User (A) is a friend of User (B),User (B) is a friend of User (A) and a follower of User (C). We can depict this relationship like a graph such as the one shown in figure1.

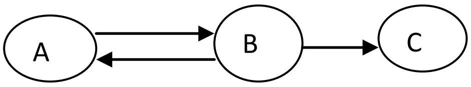


Figure 1 - Twitter Counter Screenshot

Unlike Facebook, where a connection is bi-directional (if A is a friend of B then B is a friend of A – both are friends and follow each other), Twitter's relationships are asymmetric. On Twitter, if A is following B then A is a friend of B, but B is not friend of A. A will get tweets of B reporting on his home page but B will not see tweet of A until he/she explicitly starts to follow A. We can use concepts of Graph Theory to analyze twitter data.

### *3.1.4 Graph Theory Analysis of Twitter Data*

c) Influencer – Re tweeting (Degree Centrality) – Has two approaches in-degree and out-degree. How many times a user re-tweets others messages (out-degree) and how many times messages of a user are re-tweeted by others (in-degree) give a measure of influence of the user.

d) Shortest Path / Most Network Connections – Allow us to figure out who are most influences. And we can use it for other analysis as well – depth of network etc.

### *3.2 Twitter Analytic Tools.*

Now, we explore few tools available on the internet that help analyse twitter experience for their clients and even make recommendations on when to tweet, whom to follow, who are the big influencers etc. The tools use approaches highlighted above (on analysis) and many of their own interpretations to arrive at specific analysis for their users.

a) Twitter Counter – started as a self-funded start up in Amsterdam in 2008. It is a third party application that provides analytics of twitter usage. It allows some basic features as free use and other features tagged as premium which comes with subscription.

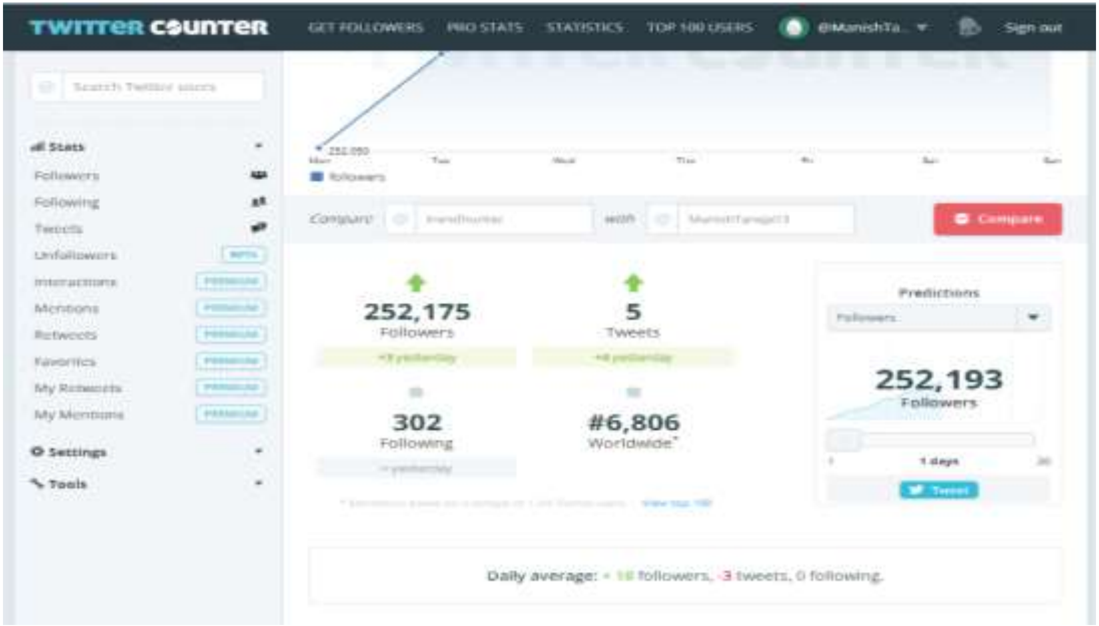


Figure 2 - Twitter Counter Screenshot

b) Twitonomy - provides a good basic statistics free and for other features user needs to sign up for a pro account

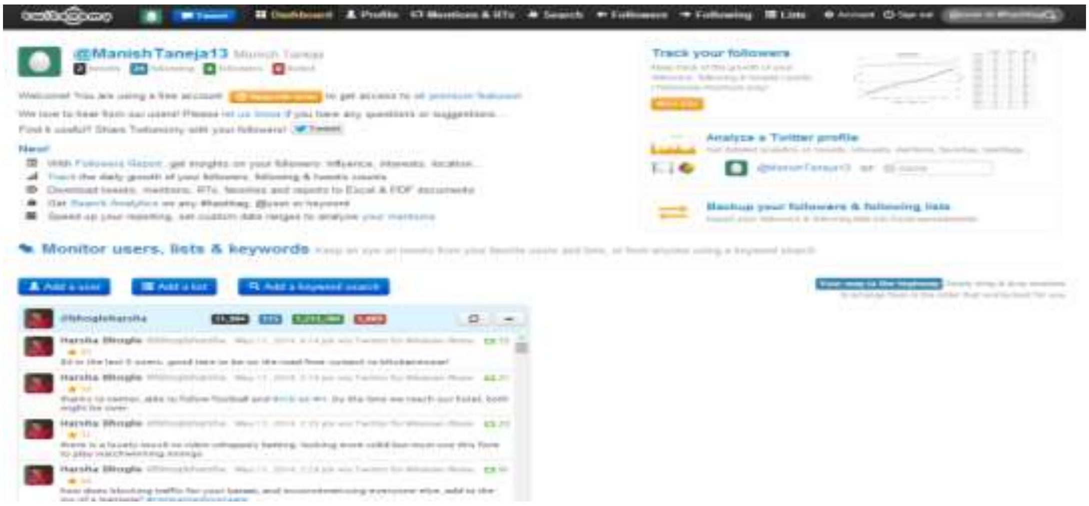
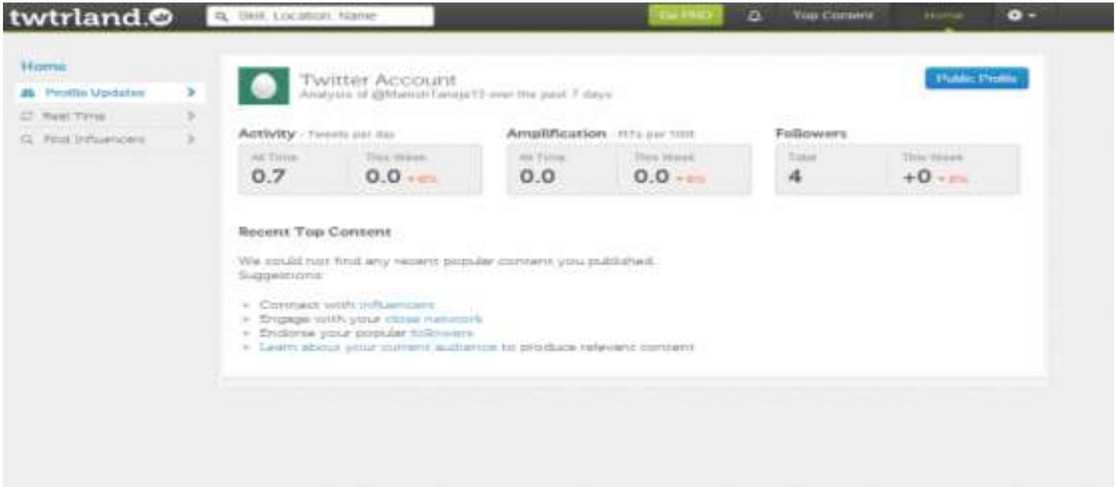


Figure 3 - Twitonomy Screenshot

c) Twtrland – provides some niche features like best time to tweet free along with additional statistics, but keeps things like export, prediction, comparison in its paid version.



Twitter Account
Analysis of @ManishTaneja13 over the past 7 days

| Activity - Tweets per day | | Amplification - RTs per tweet | | Followers | |
|---|---|---|---|---|---|
| All Time | This Week | All Time | This Week | Total | This Week |
| 0.7 | 0.0 | 0.0 | 0.0 | 4 | +0 |

Recent Top Content

We could not find any recent popular content you published.
Suggestions:

- Connect with influencers
- Engage with your close network
- Endorse your popular followers
- Learn about your current audience to produce relevant content

Figure 4 – Twtrland Screenshot

d) SocialBro – Provides free access to all its features but only for 2 weeks, after which user needs to pay subscription to continue using the services.

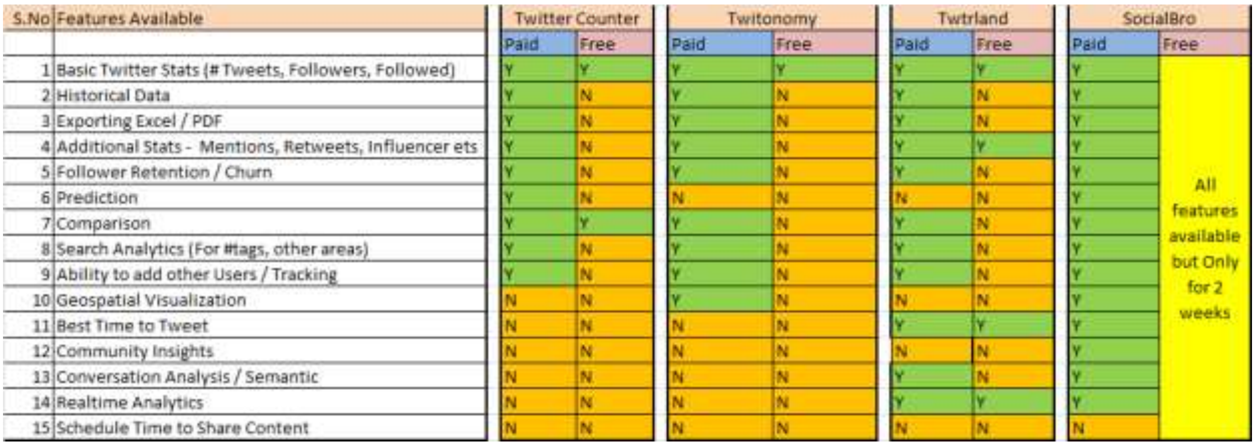
Figure 5 - SocialBro Screenshot

We have categorized features being provided by various players into following buckets and then mapped the tools for their availability / non availability in both free / paid versions.

Key Features of Twitter Analytics Tools

1) Basic Stats – Number of Tweets, Followers etc.
2) Historical Data – Ability to provide historical information about various statistics.
3) Exporting Excel / PDF – Ability to export information on search results / user statistics to excel or PDF for further analysis.
4) Additional Stats – Number of Mentions, Re tweets, Influence categories – which tell how much impact a user's tweet having on his network.
5) Follower Retention / Churn – Tool allows one to become aware of the followers who are no longer following a user. It lists new followers added / left during the week.
6) Prediction – Certain tools allow prediction – what will the number of followers for a particular user based on his tweeting history / some other metric.
7) Comparison – Allow comparison of twitter statics like tweets / followers / retweets between two different users ids.
8) Ability to add other users / lists tracking – Allows users to add other users / lists of users / communities for further tracking.
9) Geospatial visualization – Whether the tool provides map based information. Tweets shown by location on a map to see a visual appeal of influence / other metrics.
10) Best Time to Tweet – Certain tools predict what will be the best time to tweet based on history of user's tweets and responses received on them.
11) Conversation Analysis – Semantic analysis on tweets.
12) Real-time Analytics – providing statics based on real-time data – tweets posted by followers / followed and their impact.
13) Schedule time to share Content – certain tools provide an option to send content at predefined scheduled time or based on analytics release it to the followers for maximum impact.

Table I - Comparison of Twitter Analytics Tools



| S.No | Features Available | Twitter Counter | | Twitonomy | | Twtrland | | SocialBro | |
|---|---|---|---|---|---|---|---|---|---|
| | | Paid | Free | Paid | Free | Paid | Free | Paid | Free |
| 1 | Basic Twitter Stats (# Tweets, Followers, Followed) | Y | Y | Y | Y | Y | Y | Y | All features available but Only for 2 weeks |
| 2 | Historical Data | Y | N | Y | N | Y | N | Y | |
| 3 | Exporting Excel / PDF | Y | N | Y | N | Y | N | Y | |
| 4 | Additional Stats - Mentions, Retweets, Influencer ets | Y | N | Y | N | Y | Y | Y | |
| 5 | Follower Retention / Churn | Y | N | Y | N | Y | N | Y | |
| 6 | Prediction | Y | N | N | N | N | N | Y | |
| 7 | Comparison | Y | Y | Y | N | Y | N | Y | |
| 8 | Search Analytics (For #tags, other areas) | Y | N | Y | N | Y | N | Y | |
| 9 | Ability to add other Users / Tracking | Y | N | Y | N | Y | N | Y | |
| 10 | Geospatial Visualization | N | N | Y | N | N | N | Y | |
| 11 | Best Time to Tweet | N | N | N | N | Y | Y | Y | |
| 12 | Community Insights | N | N | N | N | N | N | Y | |
| 13 | Conversation Analysis / Semantic | N | N | N | N | Y | N | Y | |
| 14 | Realtime Analytics | N | N | N | N | Y | Y | Y | |
| 15 | Schedule Time to Share Content | N | N | N | N | N | N | N | |

Based on data given in table 1), we can see that online tools provide a lot of analytics to their users but most of them are available in the paid versions. In free versions, most tools only provide basic data like number of tweets, followers, followed, what is trending etc.

## 4. Conclusion

Twitter has now become a global platform for mass communication. The data it generates is finding many uses – in disaster management, brand management (understanding brand strengths, sentiments, visibility), politics (campaign management), personal (giving updates to family/friends) etc. Different users of Twitter have different needs of analytics of their twitter usage, hence before deciding which tool to use, they should search the internet to find a tool that suits their particular need best. As there are many tools available and we have only compared 4 of them. Based on our data set, we feel that if any user is looking for a tool that provides some very good features in its free version than Twtrland / TwiterCounter are our recommendations. TwitterCounter for its simple metrics and ability to provide comparisons and Twtrland for some of the advanced analytics it offers free.

## REFERENCES

1. http://www.ibmbigdatahub.com/infographic/four-vs-big-data (retrieved on 11-May)
2. http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/2/ (retrieved on 11-May)
3. http://whatsthebigdata.com/page/5/ (retried on 11-May)
4. http://faculty.smu.edu/tfomby/eco5385/The%20Economist-data-data-everywhere.pdf
5. http://globalsp.ts.fujitsu.com/dmsp/Publications/public/wp-bigdata-solution-approaches.pdf
6. www.twtrland.com
7. www.twitonomy.com
8. www.twittercounter.com
9. www.socialbro.com
10. http://www.kdd.org/sites/default/files/issues/14-2-2012-12/V14-02-02-Lin.pdf
11. http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/