I J C T A, 9(34) 2016, pp. 745-754
© International Science Press

Extraction of Tweets using Wrappers and Streaming API

Haritha Akkineni [#1], P. V. S. Lakshmi[*2] and B. Vijaya Babu[#3]

ABSTRACT

Ubiquitous online communication is producing massive amounts of data on an un-precedential scale. Twitter stands as a zeal bucket in sharing the meaningful public conversations, experiences and opinions on various topics like discussions on the policies launched by the government. Working on the premise that online social media conversations might represent a new source of information to throw light on the insights of the policies, this investigate figures out the process of extracting such potentially valuable data. Our paper mainly addresses the methodology of extracting the tweets regarding the government policies. In this paper, we have studied the procedure of extracting data related to government policies from Twitter using wrapper development and streaming API. Retrieving structured data from deep web is a main problem due to the essential convoluted structures of web pages. A comparative study has been done between web wrappers and the algorithm developed based on Streaming API. All of them have their innate margins but the algorithm constructed using streaming API has got its own benefits in extracting on the fly policy related tweets launched by the government.

Keyword: DOM Tree, OAuth, Streaming API,Social Networks,User Generated Content,Web Scrapping, XPath.

I. INTRODUCTION

This The big wave in consumer generated media have created an abundance of user generated content where a vast amount of potentially valuable knowledge is buried therein. [1] It provides continuous quest for the analyst to work with new and fresh content. The digital traces created by the social media that, when anonymized, aggregated and analyzed, can reveal significant insights that help governments make faster and more informed decisions. This sheer scale of content has created the burning need for automated methods of extracting relevant information. The social networking sites like Twitter provides a truly authentic experience and it speaks much louder than content solely available from traditional methods like statistics, household surveys and census data. It provides real-time snapshot in order for policymakers to develop timely actions to protect vulnerable populations against crises. In this paper we are analyzing the process of extracting Twitter conversations related to policies launched by the governments and this can be used to infer real-time information regarding how it can be used to predicts the splatter of the policy in the public and its effectiveness than taken from the normal data.

The regular approach to text mining is Information Extraction, extracting specific templates of information from a document collection. In this work we quantify the contribution of extraction methodologies, by comparing two strategies for IE: Wrappers and streaming API followed by the extraction of suitable tweets. We use the two strategies for the extraction of tweets related to different government policies. We show that the algorithm based on Streaming API provides significantly better precision results. Data is the crucial part of any exploration. People might want to collect and analyze data from several websites. The process of extraction have to face some challenges like variations in the format of the data

[# 1,3] Computer Science and Engineering , KL University, Guntur, Andhra Pradesh, India, *E-mail: akkinenih@gmail.com*; *vijay_gemini@kluniversity.in*

[*1,2] PVP Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India, *E-mail: papinenivsl@gmail.com*

being presented, spanning of data across multiple pages under various sections etc. The Web Data Extraction systems are software applications for the purpose of extracting information from Web sources like Web pages [2]. The Web Data Extraction system usually interacts with a Web source and extracts data stored in it and converts the extracted data in the most convenient structured format and stores it for further usage[3].

(A) Web Scrapping

The process of extracting and creating a structured representation of data from a web site is known as scrapping. The look-and-feel and the data will be continuously being updated in HTML. Since current techniques for web scraping are based on the markup, a change may lead to the extraction of incorrect data.

If the owner of the information does not provide an open API, the remedy is to write a program that targets the markup of the web page. A general approach is to parse the web page to a tree representation and evaluate an XPath expression on it. The XPath denotes a path, possibly with wildcards, and when evaluated on a tree, the result is the set of nodes at the end of any occurrence of the path in the tree[4]. This can be the motivation for the wrapper development.

(B) Scrapper Tools Developed using Different Techniques

Some of the tools that were developed using the process of scrapping are: HarvestMan , Scraperwiki, FiveFilters.org, Kimono, Mozenda, 80Legs, Scrape.it, Scrapy, Needlebase , OutwitHub, irobotsoft [5].

(C) Usage of Tree based Techniques

A commonly used measure for tree similarity is the tree edit distance which easily can be extended to be a measure of how well a pattern can be matched in a tree.

To check weather both the trees are similar Tree Matching Algorithms are used. It gives an indication that the HTML documents they represent are also very similar and the Web data will be extracted from that page. This is all concerned with the theoretical background regarding the data extraction process.

The rest of the paper is organized as follows: in Section 2 we consider the related work on theoretical background and Web data extraction detailing some interesting aspects of algorithms and providing some examples. It also presents the methodology used for wrapper development and for streaming API. Section 3 discuss on the results where Experimentation and evaluation are discussed on. Sections 4 covers the discussions part where the comparative analysis between the two methodologies has been detailed. Section 5 finally presents some conclusive considerations. Section 6 provides the References which resulted in this paper.

II. METHODOLOGY/ EXPERIMENTAL

Tweets posted by the users are mostly in an unstructured manner. To change it over to a structured format we should take the help of wrappers. Any procedure that aims at extracting structure data from unstructured data sources is usually referred as wrapper.

It is a procedure, that executes one or many different classes of algorithms, which seeks and finds data required by a human user, extracting them from unstructured Web sources, and transforming them into structured data, merging and unifying this information for further processing, in a semi-automatic or fully automatic way.[6]

(A) Methodology Used in Wrappers

The following methodology as shown in Fig. 1 is followed to extract **“Make in India”** related tweets from web.

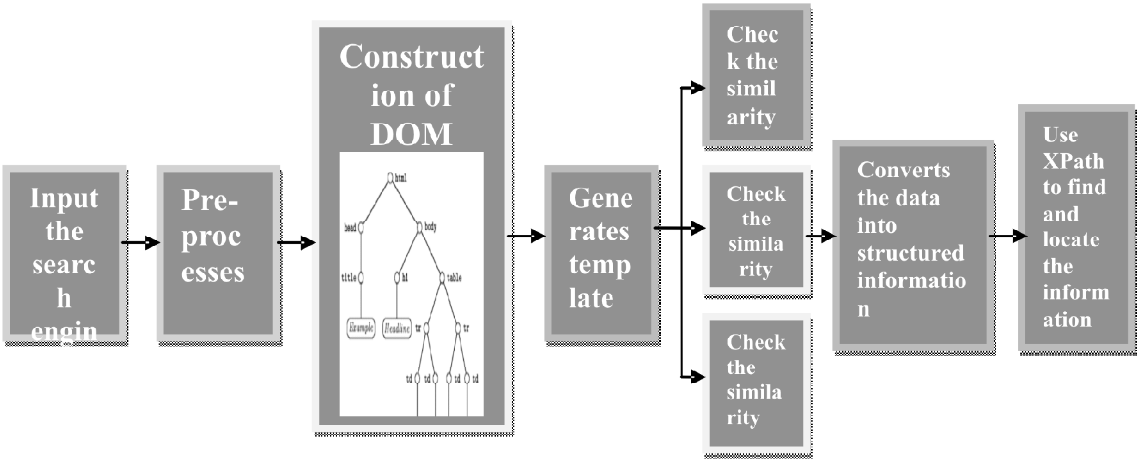


Figure 1: Methodology for Twitter Wrapper

The web page is represented as plain text in the form of labeled ordered rooted trees, where labels represent the tags, and the tree hierarchy represents the different levels of nesting of elements constituting the Web page. This representation of a Web page is referred as DOM (Document Object Model). The DOM is used to build the data as a tree.The tree starts at the root node and branches out to the text nodes at the lowest level of the tree[7]. This is the place where we locate our data.XPath is a query language for selecting nodes from an XML like document, such as HTML [8].

HTML DOM is in a tree structure, usually called an HTML DOM tree. Figure 2 illustrates a simple HTML document and its corresponding DOM tree. We are interested only in the node and its offspring. In this example, body node has three children: element nodes <B> and <I>, and text node #and. Element node <B>has a text node child #Wrapper, and element node<I> has a text node #Streaming API. Following the DOM convention, we use <> to indicate element node, and use # to indicate text node[9].

```
<html>
    <head>
            <title>Extraction</title>
    </head>
    <body>
            <b>Wrappers </b>and<i>Streaming API</i>
    </body>
</html>
```

To reach out to the exact content the procedure followed is:

Input: HTML Page with DOM Structure

Process:

Step 1: Navigate through the parsed hierarchy tree.

Step 2: Identify relevant nodes

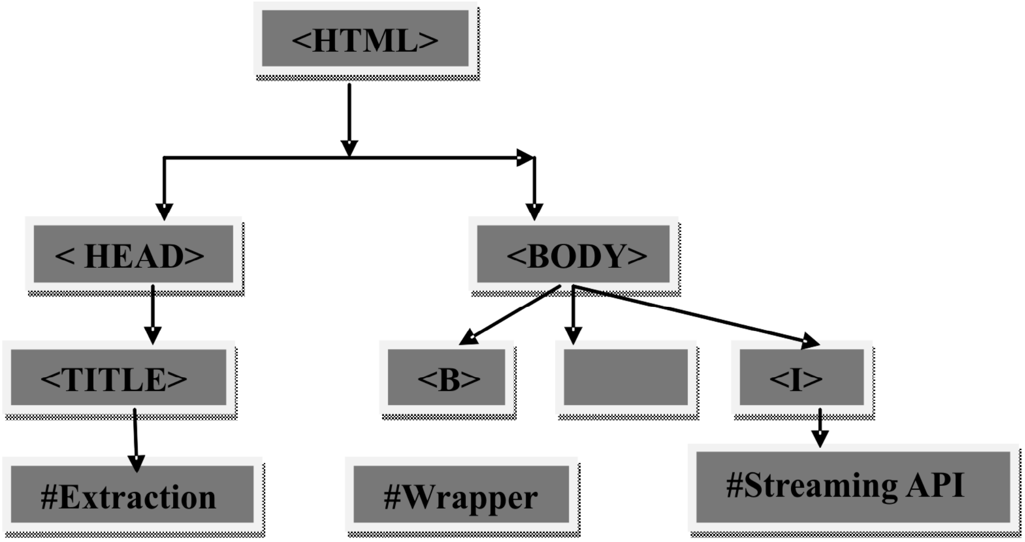


Figure 2: HTML document modelled as a tree

Step 3:Extract relevant values of the attributes [10]

Output: The text data

(B) Methodology to Extract Tweets using Streaming API

1) The Search API: Using this API we can mine for tweets posted in the past .The Search API is a Representational State Transfer API which allows users to request specific queries of recent tweets by using HTTP methods to execute different operations. A delay while fetching the tweets is incorporated as there is a rate limit associated with the query. To actually fetch tweets, we continuously send queries to the Search API.

   Both of these APIs require the user to have an API key for authentication. Once authenticated, we were able to easily access the API through a Java library called Twitter4j a simple application for the Twitter API[11].

2) Limitations of Search API:With the Search API you can only sent 180 Requests every 15 min timeframe. With a maximum number of 100 tweets per request this means you can mine for 4 x 180 x 100 = 72000 tweets per hour.The REST APIs support short-lived connections and are rate-limited [12].

3) The Streaming API: The Streaming API allows users to obtain real-time access to tweets from the users input query. The user requests a connection to a stream of tweets from the server. Then, the server opens a streaming connection and tweets are streamed in as they occur, to the user. Streaming goes forward in time and captures tweets as they are posted.

   With the Streaming API we can collect all tweets containing the keyword, up to 1 % of the total tweets currently being posted on twitter. At present the amount of tweets being posted per day account to 500+ million, so 1 % of all tweets still gives us 1+ million tweets a day.

4) Limitations of Streaming API:However, there are a few limitations of the Streaming API. First, language is not specifiable, resulting in a stream that contains Tweets of all languages, including a few non-Latin based alphabets.

   OAuth is an open protocol that Twitter implemented in March 2009, to tackle the downfalls of basic authentication. Using OAuth, users give your application permission to interact with their Twitter account, Twitter gives you a token to authenticate with, and you never have to ask for or handle the users passwords. Twitter provides four methods for working with OAuth.

authenticate() , authorize(), request_token(), access_token()

Authenticate and authorize are used as links for your users to login.

From the developer's point of view, OAuth takes six steps[13].

(C) Algorithm to Access Tweets Using Streaming API

Input

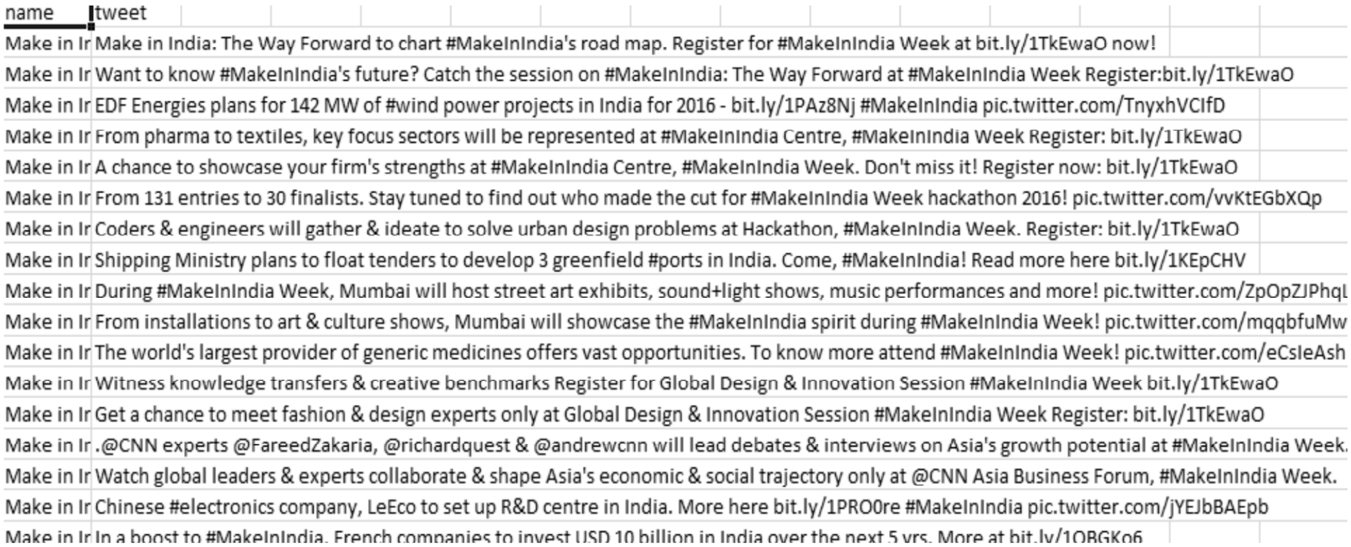
Register with Twitter for consumer token and secret and input it to request_token () method[14].

```
Process:
Do
{
Step 1:Present the user with a link to either authenticate() or authorize() method
    Step 2:Include the request token as a query string value named oauth token.
    Step 3: The user logs on to twitter to get the application approved.
Step 4: Twitter issues the original request token included in the URL query string labeled oauth token.
    Step 5: Once the user is back pass the request token to the access_token() method.
    Step 6: Use the access token to make your API calls to Twitter on behalf of the user.
    Step 7: Request for content given a particular keyword.
}
Output: Requested Tweets.
```

III. RESULTS

(A) The Data Source

For each experiment, The following tweets are extracted using Wrappers. Fig.3 shows the extracted tweets on "Make in India". Table 1 gives the sample words extracted along with their frequencies. The most frequent words extracted are illustrated in Fig. 4. Fig. 5 gives the word cloud formed on "Make in India" policy using Wrappers.



name	tweet
Make in Ir	Make in India: The Way Forward to chart #MakeInIndia's road map. Register for #MakeInIndia Week at bit.ly/1TkEwaO now!
Make in Ir	Want to know #MakeInIndia's future? Catch the session on #MakeInIndia: The Way Forward at #MakeInIndia Week Register:bit.ly/1TkEwaO
Make in Ir	EDF Energies plans for 142 MW of #wind power projects in India for 2016 - bit.ly/1PAz8Nj #MakeInIndia pic.twitter.com/TnyxhVCIfD
Make in Ir	From pharma to textiles, key focus sectors will be represented at #MakeInIndia Centre, #MakeInIndia Week Register: bit.ly/1TkEwaO
Make in Ir	A chance to showcase your firm's strengths at #MakeInIndia Centre, #MakeInIndia Week. Don't miss it! Register now: bit.ly/1TkEwaO
Make in Ir	From 131 entries to 30 finalists. Stay tuned to find out who made the cut for #MakeInIndia Week hackathon 2016! pic.twitter.com/vvKtEGbXQp
Make in Ir	Coders & engineers will gather & ideate to solve urban design problems at Hackathon, #MakeInIndia Week. Register: bit.ly/1TkEwaO
Make in Ir	Shipping Ministry plans to float tenders to develop 3 greenfield #ports in India. Come, #MakeInIndia! Read more here bit.ly/1KEpCHV
Make in Ir	During #MakeInIndia Week, Mumbai will host street art exhibits, sound+light shows, music performances and more! pic.twitter.com/ZpOpZJPhqL
Make in Ir	From installations to art & culture shows, Mumbai will showcase the #MakeInIndia spirit during #MakeInIndia Week! pic.twitter.com/mqqbfuMw
Make in Ir	The world's largest provider of generic medicines offers vast opportunities. To know more attend #MakeInIndia Week! pic.twitter.com/eCsIeAsh
Make in Ir	Witness knowledge transfers & creative benchmarks Register for Global Design & Innovation Session #MakeInIndia Week bit.ly/1TkEwaO
Make in Ir	Get a chance to meet fashion & design experts only at Global Design & Innovation Session #MakeInIndia Week Register: bit.ly/1TkEwaO
Make in Ir	.@CNN experts @FareedZakaria, @richardquest & @andrewcnn will lead debates & interviews on Asia's growth potential at #MakeInIndia Week.
Make in Ir	Watch global leaders & experts collaborate & shape Asia's economic & social trajectory only at @CNN Asia Business Forum, #MakeInIndia Week.
Make in Ir	Chinese #electronics company, LeEco to set up R&D centre in India. More here bit.ly/1PRO0re #MakeInIndia pic.twitter.com/jYEJbBAEpb
Make in Ir	In a boost to #MakeInIndia, French companies to invest USD 10 billion in India over the next 5 yrs. More at bit.ly/1OBGKo6

Figure 3: Tweets extracted on "Make in India" using Wrapper

The non sparse entries in the document :196

Maximum term length:23.

Table I
Frequency of Words Generated using Wrappers

Word	*Frequency*
Make in India	35
Week	22
Regist	11
Design	9
Will	9
India	8
Global	6
Asia	5

Total word :121

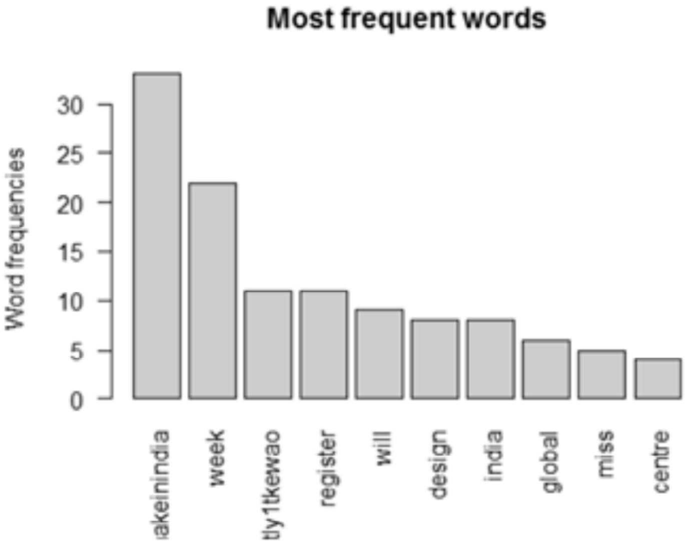


Figure 4: The most frequent words

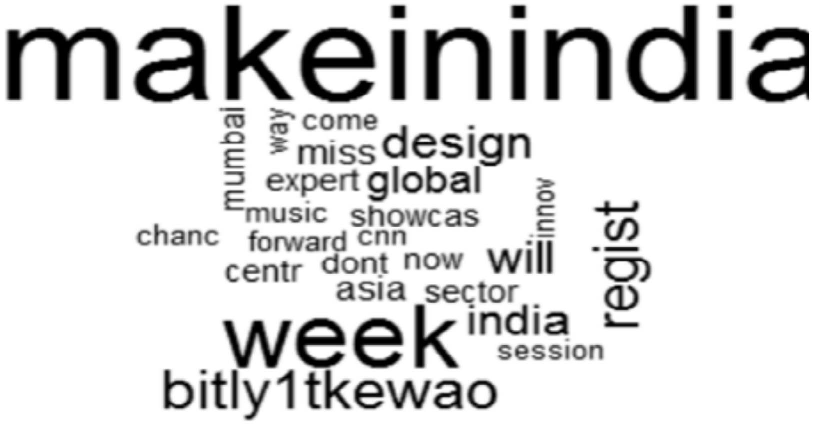
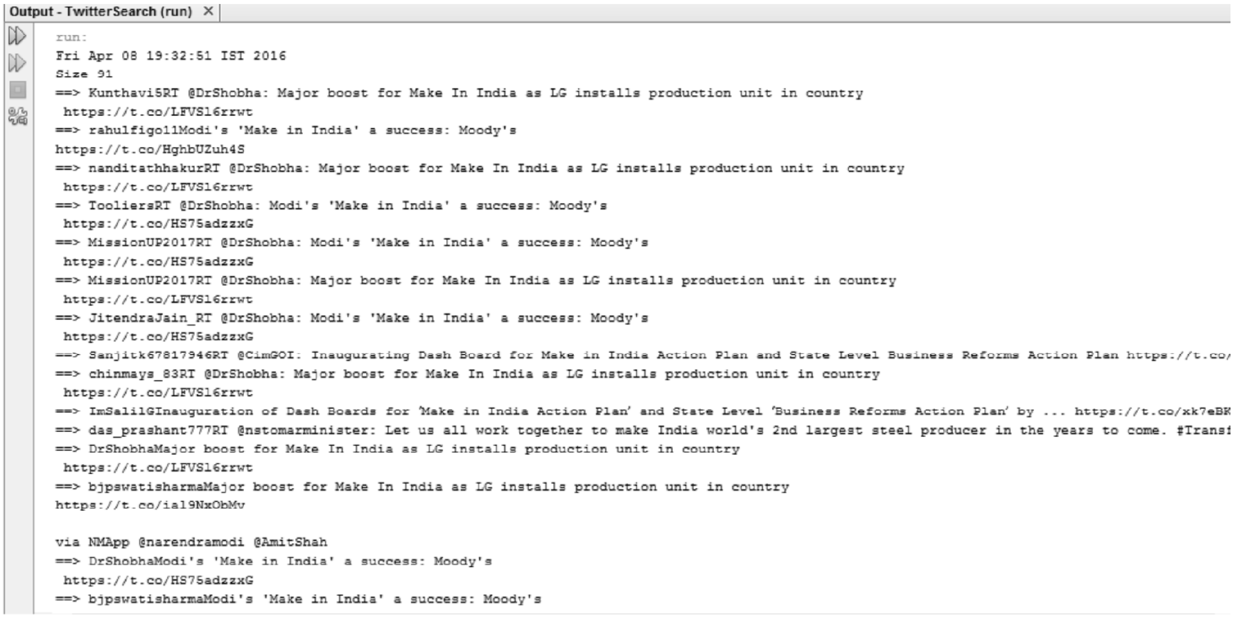


Figure 5: Word cloud on "Make in India" using wrapper

For each experiment, The following tweets are extracted using Streaming API. Fig.6 shows the extracted tweets on "Make in India". Table II gives the sample words extracted along with their frequencies. The most frequent words extracted are illustrated in Fig. 7. Fig.8 gives the word cloud formed on "Make in India" policy using Streaming API.



```
run:
Fri Apr 08 19:32:51 IST 2016
Size 91
==> Kunthavi5RT @DrShobha: Major boost for Make In India as LG installs production unit in country
 https://t.co/LFVSl6rrwt
==> rahulfigo11Modi's 'Make in India' a success: Moody's
https://t.co/HghbUZuh4S
==> nanditathhakurRT @DrShobha: Major boost for Make In India as LG installs production unit in country
 https://t.co/LFVSl6rrwt
==> TooliersRT @DrShobha: Modi's 'Make in India' a success: Moody's
 https://t.co/HS75adzzxG
==> MissionUP2017RT @DrShobha: Modi's 'Make in India' a success: Moody's
 https://t.co/HS75adzzxG
==> MissionUP2017RT @DrShobha: Major boost for Make In India as LG installs production unit in country
 https://t.co/LFVSl6rrwt
==> JitendraJain_RT @DrShobha: Modi's 'Make in India' a success: Moody's
 https://t.co/HS75adzzxG
==> Sanjitk67817946RT @CimGOI: Inaugurating Dash Board for Make in India Action Plan and State Level Business Reforms Action Plan https://t.co/
==> chinmays_83RT @DrShobha: Major boost for Make In India as LG installs production unit in country
 https://t.co/LFVSl6rrwt
==> ImSalilGInauguration of Dash Boards for 'Make in India Action Plan' and State Level 'Business Reforms Action Plan' by ... https://t.co/xk7eBK
==> das_prashant777RT @nstomarminister: Let us all work together to make India world's 2nd largest steel producer in the years to come. #Transf
==> DrShobhaMajor boost for Make In India as LG installs production unit in country
 https://t.co/LFVSl6rrwt
==> bjpswatisharmaMajor boost for Make In India as LG installs production unit in country
https://t.co/ial9NxObMv

via NMApp @narendramodi @AmitShah
==> DrShobhaModi's 'Make in India' a success: Moody's
 https://t.co/HS75adzzxG
==> bjpswatisharmaModi's 'Make in India' a success: Moody's
```

Figure 6: Tweets extracted on "Make in India" using Streaming API

The non sparse entries in the document :2522

Maximal term Matrix: 26

Table II
Frequency of Words Using Streaming API

Word	*Frequency*
India	1130
make	901
Namorocks2015	213
MakeinIndia	194
Manufacture	140
Leadership	128
Offers	118
Fighter	111

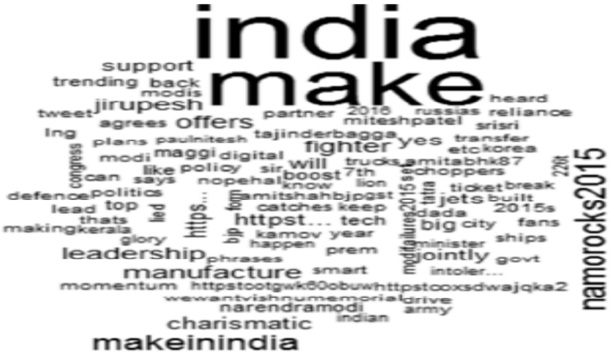


Figure 7: Word cloud using Streaming API

Total Words: 3173

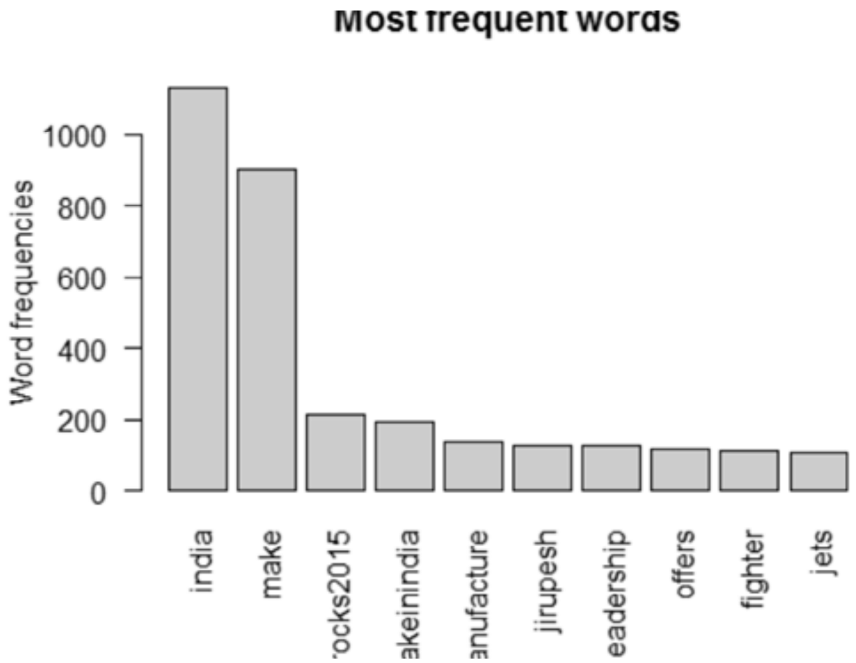


Figure 8: Most frequent words generated

(B) Evaluating Different Strategies

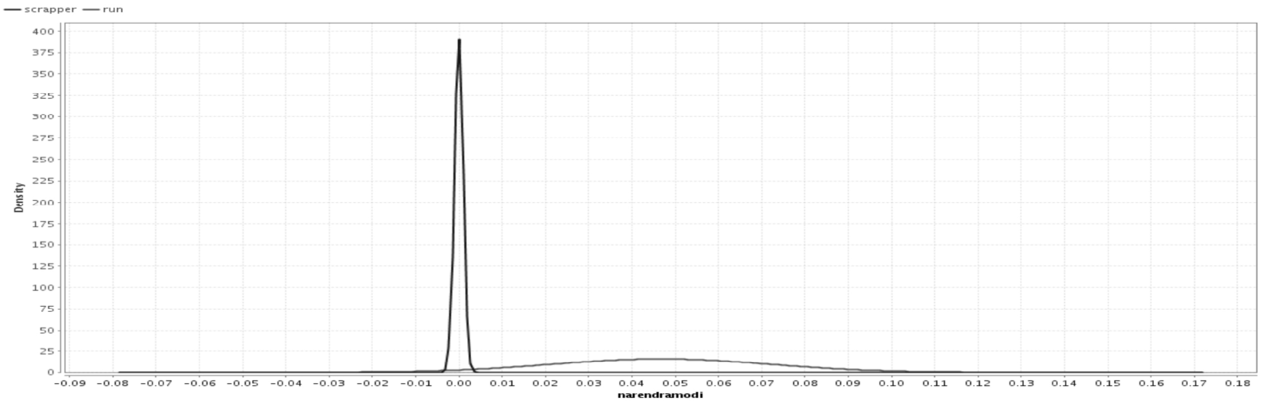


Figure 9: Simple Distribution Result on the word “NarendraModi” using Wrapper and Streaming API

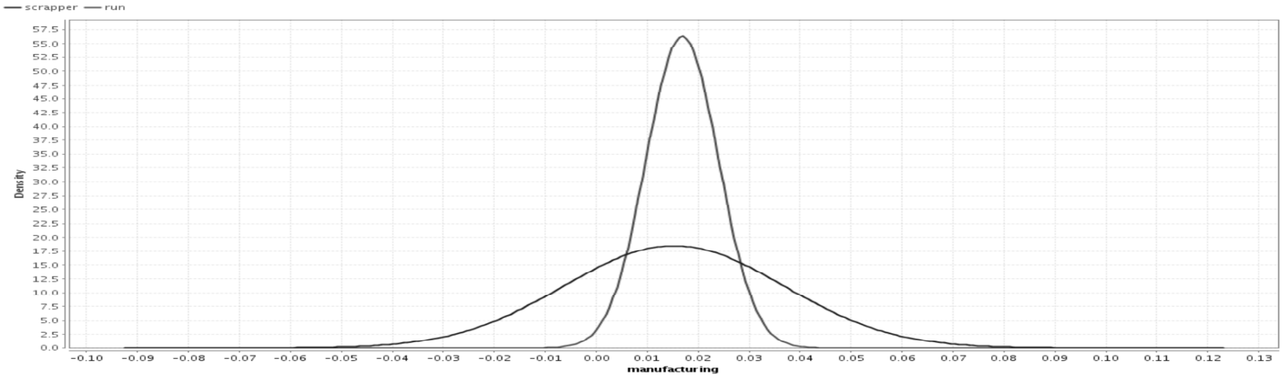


Figure 10: Simple Distribution Result on the word “manufacturing” using Wrapper and Streaming API

IV. DISCUSSION

(A) Experimental Evaluation

We have conducted separate experiments using the two strategies. We used different policies as the keywords and tried for 10 different government policies. In each experiment we extracted tweets on the policy " Make in India" and compared the results. The depicted results are for the policy "Make in India". When comparing the two methodologies Wrapper lead to 35 occurrences and Streaming API lead to 194 occurrences. The distribution result for "Narendra Modi" was shown which the data extracted through scrapper resulted in density of 400 and 20 for Streaming API. The distribution for the word "manufacturing" showed that the streaming API stood highest with density of 57.5 and 17.5 for wrapper.

(B) Comparative Analysis

Thus we have looked into different methods for structured data extraction. Extraction of data using wrappers encompasses usage of DOM tree and XPath construction. This approach is very labor intensive and time consuming. The experimental results shows that it extracted comparatively less number of tweets than the streaming API. The alternative method based on Search API limits the users to request specific queries of recent tweets. With the Search API we can only mine approximately about 72000 tweets per hour. The second approach based on the Streaming API allows users to obtain real-time access to tweets from an input query. It supports long-lived connection and provides data in almost real-time as it is being Twittered. With the Streaming API we can collect all tweets containing the keyword and it accounts up to 1 % of the total tweets currently being posted on twitter. Though it is advantageous in the data extraction it has got its own limitations like the language is not specifiable. The stream may contain Tweets of all languages.

V. CONCLUSION

In this work, we have provided with two methodologies of data extraction from Twitter and explored them. We have extracted the tweets using two different methodologies on the policy "MakeinIndia". We discussed the evolution of scrapping. On the fly tweets could be extracted using Streaming API and we can collect all tweets containing the keyword, up to 1% of the total tweets currently being posted on twitter could be extracted. This approach will be solid enough to be implemented in real systems, ensuring great reliability regarding the opinions on particular policy launched by the government within no time[15]. This enables the policy makers to frame out instantaneous decisions about the policies in aspect based terms.

REFERENCES

[1] LIM, E.,Peng,C,H.,CHEN,G.,(2013).Business Intelligence and Analytics: Research Directions.ACM Transactions on Management Information Systems, 3(4),Research Collection School Of Information Systems. Available at: http://ink.library.smu.edu.sg/sis_research/1966.

[2] SCM,d.,Sirisuriya, S.,(2015), A Comparative Study on Web Scraping, Proceedings of 8th International Research Conference, KDU.p.135-140.

[3] Seema,K.,Jayamalini, K (2013),Web Data Extraction Using Tree Structure Algorithms – A Comparison.International Journal of Recent Technology and Engineering (IJRTE)2(3).p.35-39.

[4] Patrick,H,C.,(2011), Algorithms for Web Scraping. A Thesis Submitted in partial fulfillment of the requirements The University of Denmark DTU Informatics for the master's theses at DTU.

[5] Garethde,(2014). A Guide to Web Scraping Tools.[Dec5,2015] Available from: http://www.garethjames.net/a-guide-to-web-scrapping-tools/.

[6] Emilio,F., Pasquale, D,M., Giacomo,F., Robert,B(2013).Web Data Extraction, Applications and Techniques: A Survey.arXiv:1207.0246v4 [cs.IR].

[7] YesuRaju, P.,KiranSree, P.(2013) A Language Independent Web Data Extraction using Vision Based Page Segmentation algorithm. International Journal of Research in Engineering and Technology. 2(4).p. 635 - 639.

[8] ScreamingFrog(2016). Web Scraping & Data Extraction Using The SEO Spider Tool.[16th Apr 2016] Available from: http://www.screamingfrog.co.uk/web-scraping/.

[9] Jie, Z., Daniel, L.,George, R., Thoma (2006).Combining DOM Tree and Geometric Layout Analysis for Online Medical Journal Article Segmentation. Association for Computing Machinery 6th ACM/IEEE-CS Joint Conference on Digital Libraries. Chapel Hill, NC, USA June 11 - 15, 2006.

[10] Janko,T. (2014) An introduction to analytical Web Scrapping with R at the Munich userR group's.YouTube video. 10 Mar Available from: https://www.youtube.com/watch?v=m6zMBFCfgto.[Accessed:19th Mar 2016].

[11] Linhao,Z.(2013) Sentiment Analysis on Twitter with Stock Price and Significant Keyword correlation. A Thesis Submitted in partial fulfillment of the requirements of The University of Texas for honors theses at Austin. (Dated: April 16, 2013) Report# HR-13-01.

[12] Ahmet, T (2015) Collecting Data from Twitter.[Online] Available from: https://ataspinar.wordpress.com/2015/11/09/collecting-data-from-twitter/.[Accessed:19th Apr 2016].

[13] Dusty, R.(2010).Twitter Application Development For Dummies.Wiley Publishing Inc. Indianapolis, Indiana.

[14] Haritha, A., Lakshmi, P. V. S., Vijay Babu, B., Lakshmi, G.,(2016).Modeling and Visualizing the Extraction of Opinions from Twitter.International Journal of Innovations & Advancement in Computer Science IJIACS. 5(2).p.90-94.

[15] Haritha, A., Lakshmi,P.V.S., Vijay Babu, B.,(2015). Online Crowds Opinion-Mining it to Analyze Current Trend: A Review. International Journal of Electrical and Computer Engineering (IJECE).5(5), p. 1180-1187.