ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

A Framework for Real-Time Twitter Data Analysis

A. Charlin Monisha[1], P. Rajkumar[2], N.Mohan Prabhu[3]

M.E. Computer Science and Engineering, Mount Zion College of Engineering and Technology, Lena Vilakku, Thirumayam Tk, Pudukkottai, Tamil Nadu 622507, India.

Assistant Professor, Department of Computer Science and Engineering, Mount Zion College of Engineering and Technology, Lena Vilakku, Thirumayam Tk, Pudukkottai, Tamil Nadu 622507, India.

Assistant Professor, Department of Computer Science and Engineering, Mount Zion College of Engineering and Technology, Lena Vilakku, Thirumayam Tk, Pudukkottai, Tamil Nadu 622507, India.

ABSTRACT: Twitter has turned out to be one of the biggest stages for clients around the globe to impart anything occurring around them to companions and past. A bursty point in Twitter is one that triggers a surge of pertinent tweets inside a brief timeframe, which regularly reflects essential occasions of mass intrigue. The most effective method to use Twitter for early identification of bursty points has along these lines turn into a critical research issue with enormous commonsense esteem. In spite of the abundance of research work on point demonstrating and examination in Twitter, it remains a tremendous test to recognize bursty subjects continuously. As existing strategies can barely scale to deal with the errand with the tweet stream continuously, we propose in this paper TopicSketch, a novel outline based subject model together with an arrangement of methods to accomplish constant location. We assess our answer on a tweet stream with more than 30 million tweets. Our investigation comes about show both proficiency and adequacy of our approach. Particularly it is additionally exhibited that TopicSketch can conceivably deal with many millions tweets every day which is near the aggregate number of day by day tweets in Twitter and present bursty occasion in better granularity.

KEYWORDS: Tweets, Data Stream, Bursty Topic Modelling, Practical Exposure, Twitter Analysis.

I. INTRODUCTION

In world 200 million dynamic clients and more than 400 million tweets for every day as in a current report, Twitter has turned out to be one of the biggest data entrances which give a simple, snappy and solid stage for customary clients to impart anything occurring around them to companions and different devotees. Specifically, it has been watched that, in life-basic calamities of societal scale, Twitter is the most vital and opportune source from which individual find out and track the breaking news before any prevailing press grabs on them and rebroadcast the recording.

For instance, in the March 11, 2011 Japan quake and resulting tidal wave, the volume of tweets sent spiked to more than 5,000 every second when individuals post news about the circumstance alongside transfers of versatile recordings they had recorded 2. We call such occasions which trigger a surge of countless tweets "bursty themes".

Bursty theme on November first, 2011. A 14-year-old young lady from Singapore named Adelyn (not her genuine name) created a huge hubbub online after she was despondent with her mother's unremitting annoying and turned to physical mishandle by slapping her mother twice, and gloated about her activities on Facebook with

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

vulgarities. Inside hours, it soon became a web sensation on the Internet, drifting worldwide on Twitter and was one of the top Twitter inclines in Singapore expansive news media.

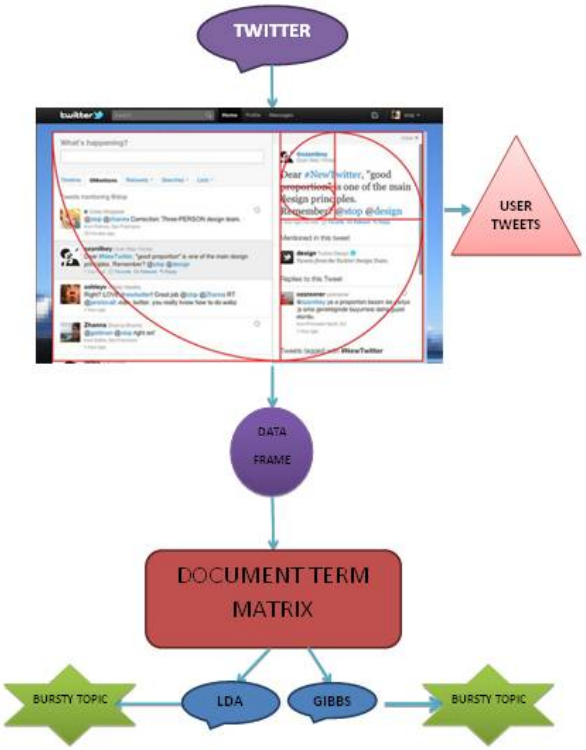

Fig.1 System Architecture Design

II. SYSTEM OVERVIEW

The three primary research challenges here are How to identify the bursty topics, i.e., what are the keywords of the topics, How to detect a bursty topic as early as possible, and How to perform the task efficiently in large-scale real-time setting as Twitter. Our solution, called TopicSketch, is based on two main techniques a sketch-based topic model and a hashingbased dimension reduction technique. Our sketch-based topic model provides an integrated two-step solution to both challenge and above. In the first step, it maintains as a sketch of the data the acceleration of three quantities: the whole tweet stream, every word, and every pairs of words, which are early indicators of popularity surge and can be updated efficiently at a low cost, making early detection possible. In the second step, based on the data sketch, it learns the bursty topics by solving an optimisationframework with five parts: the tweet stream, the data sketch, the monitor which tracks changes in the data sketch and triggers the detection when certain criteria are satisfied, the estimator which infers the topics, and the reporter which evaluates the bursty topics provided by estimator and reports the detection result. The real-time detection flow of TopicSketch is the following stepsUpon the arrival of each tweet, the sketch is updated, which is an efficient step as detailed in Section IV-A and V-B, Once the sketch is updated, the change is sent to the monitor. The monitor tracks the data sketch, compares it with historical average, and triggers the estimator for potential bursty topic detection if the difference is larger than pre-determined threshold. Upon notification, the estimator takes a snapshot of the sketch and infers the bursty topics as described. The inferred bursty topics are sent to the reporter to evaluate and report. TopicSketch is designed such that steps to)are computationally cheap to enable real-time response and early detection. Step which is expensive if done naively is greatly expedited by dimension reduction techniques as used.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

III. DATA EXTRACTION

Data Extraction will be used to pull the tweets from R code using Oauth facility by Submitting Twitter credentials. Similarly, Facebook Posts will be pulled from R code using FB oauth facility by Submitting Facebook credentials.Corpus Cleaning will be used to clean present the extracted data to R engine.

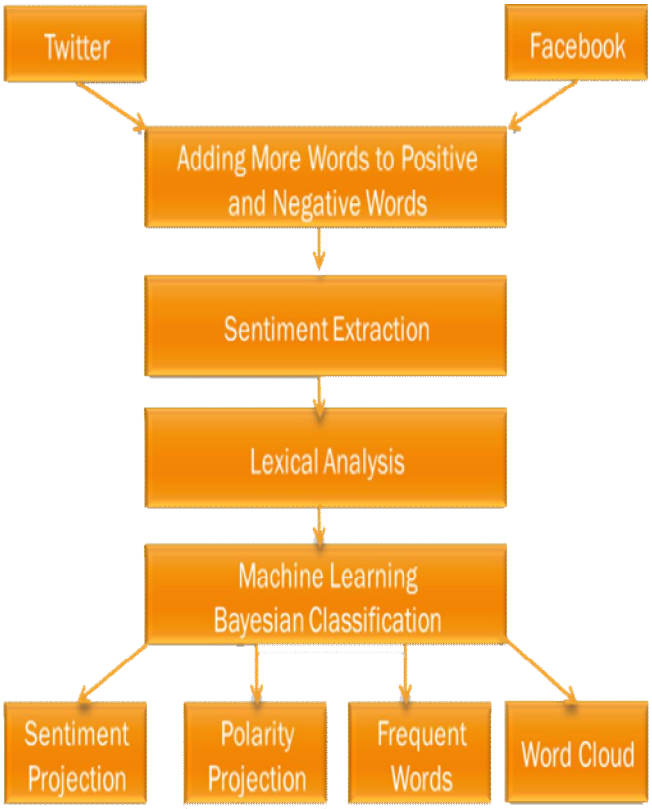

Fig.2. Data Extraction Model

Lexical Analyzer analyzes the Extracted data as positive or negative tweets. Machine learning will take place to project the result. Bayesian Learning will happen. The end result will be projected Sentiment or Opinion projection, polarity analysis (trend towards positive or negative trait), word cloud most frequent words spoken.

Term Document Matrix

A term-document matrix is also known as document-term matrix which is also shortly called as TDM. Document-term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms.

Frequent Words

Frequent Words can be identified from the Term Document Matrix. Frequent Word and all its associations can be find through this data sets. These frequent words and all of its associations are projected as a graph with two axes like the count and words. These are visualized in a graph with the words on the mentioned frequency. Screenshot of projecting the Frequent Words is as shown in Frequent Words for the Facebook data retrieved by using the hash tags

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

mentioned in the section 3.3.2.1.7 with a minimum frequency as 48 is shown in . Frequent Words for the Twitter data retrieved by using the hash tags mentioned in the section 3.3.2.1.7 with a minimum frequency.

Word Cloud

A Word Cloud is also called as Tag Cloud. Word Cloud is a visual representation of text data, typically used to depict keyword metadata or tags on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color. This format is useful for quickly perceiving the most prominent terms. Screenshot of projecting the Word Cloud is as shown in. Word Cloud for the Facebook data retrieved by using the hash tags mentioned in the section 3.3.2.1.7 with a minimum frequency of 28 is shown in. Word Cloud for the Twitter data retrieved by using the hash tags mentioned in the section 3.3.2.1.7 with a minimum frequency of 28.

Data Cleaning

A Web Application will be developed to process all a new Land registration, new electricity (EB) connection, approval of plan for construction of building and water supply connection. It will work based on the last five year data. If it is an agricultural land or dried water body. Then, it won't be allowed to register a Land. Similarly, it won't allow registering a land when it is a dried water bodies.

Filters:

Kalman Flter assumes that the posterior density at every time step is Gaussian and that it is therefore parameterized by a mean and covariance.

Resampling:

(a) Re-sample N particles from a particle set St using weights of each particles and allocate them on the map. (We allow to re-sample same particles more than one.)
(b) Generate a new particle set St+1 and weight them based on weight distribution Dw(x,y).

Partial Filters:

Sequential Importance Sampling (SIS) algorithm is a Monte Carlo method that forms the basis for particle filters. The SIS algorithm consists of recursive propagation of the weights and support points as each measurement is received sequentially.

Sentiment and Polarity Analysis

Polarity analysis is classifying the polarity of a, a sentence or an entity feature/aspect is positive, negative, or neutral. Screenshot of projecting the Polarity Analysis is as shown in. Polarity analysis for the Facebook data retrieved by using the hash tags mentioned in the section 3.3.2.1.7 is writer. Sentiment Analysis works based on the identified emotions such as "angry", "sad", and "happy". A basic task in sentiment analysis is classifying the polarity of a given text at the document and sentence. Screenshot of projecting the Sentiment Analysis.

Report and Generation

A Word Cloud is also called as Tag Cloud. Word Cloud is a visual representation of text data, typically used to depict keyword metadata or tags on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color. This format is useful for quickly perceiving the most prominent terms.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

IV. EXISTING SYSTEM

In existing system, online portal of media department of and twitter follow up and face book, micro blogs, E-mail and more social media networks.

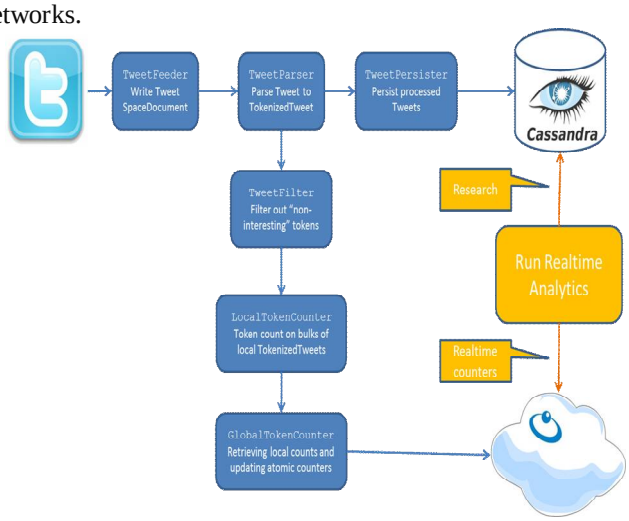

Fig.3 Twitter Real Time Analysis

Twitter works as a huge news media. Especially bursty topics, news first appears in Twitter, rather than traditional news media. It is interesting and also useful to detect bursty topics from Twitter. Rather than keep the big historical data, maybe we can take a snapshot of the current data stream. Atleast, it takes much smaller space and hopefully we can efficiently infer topics from it.

Disadvantages

- In the world are using twitter analysis and the document are sharing all.
- To find out the positive and negative information showing the twitter analysis.
- Tweet parser are used tokenized process so, time extension process.

V. PROPOSED METHODOLOGY

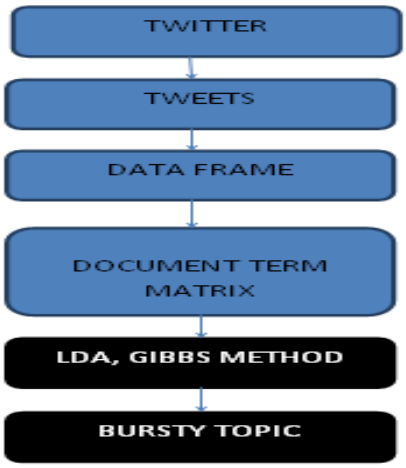

Fig.4 Proposed System

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

Twitter is an online news and social networking service where users post and read short messages called "tweets". Users can post and read tweets, but those who are unregistered can only read them. Users access Twitter through the website interface mobile device. Twitter is an online news and social networking service where users post and read short messages called "tweets". Users can post and read tweets, but those who are unregistered can only read them. Users access Twitter through the website interface mobile device Twitter is based in San Francisco and has more than 25 offices around the world.

Algorithm 1: Topic Interference Algorithm

```
getwd()
library(tm)
docs <- Corpus(DirSource("D:/TopicModelling/TextMining"))
docs
writeLines(as.character(docs[[30]]))
toSpace <- (function(x, pattern) {return (gsub(pattern, " ", x))})
docs <- tm_map(docs, toSpace, "-")
docs <- tm_map(docs, toSpace, ":")
docs <- tm_map(docs, removePunctuation)
docs <- tm_map(docs, toSpace, "’")
docs <- tm_map(docs, toSpace, "‘")
docs <- tm_map(docs, toSpace, "`")
docs <- tm_map(docs, toSpace, "~")
docs <- tm_map(docs, toSpace, "-")
docs <- tm_map(docs, tolower)
docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, removeWords, stopwords("english"))
docs <- tm_map(docs, stripWhitespace)
writeLines(as.character(docs[[30]]))
library(SnowballC)
docs <- tm_map(docs,stemDocument)
writeLines(as.character(docs[[30]]))
dtm <- DocumentTermMatrix(docs)
dtm
inspect(dtm[1:2,1000:1005])
freq <- colSums(as.matrix(dtm))
length(freq)
ord <- order(freq,decreasing=TRUE)
freq[head(ord)]
freq[tail(ord)]
dtmr <-DocumentTermMatrix(docs, control=list(wordLengths=c(4, 20),
bounds = list(global = c(3,27))))
findFreqTerms(dtmr,lowfreq=80)

findAssocs(dtmr,"project",0.6)
wf=data.frame(term=names(freq),occurrences=freq)
 p <- ggplot(subset(wf, freq>100), aes(term, occurrences))
 p <- p + geom_bar(stat="identity")
 p <- p + theme(axis.text.x=element_text(angle=45, hjust=1))
 p
 library(wordcloud)
set.seed(42)
wordcloud(names(freqr),freqr, min.freq=70)
```

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

```
 library(topicmodels)
 write.csv(freq[ord],"word_freq.csv")
 burnin <- 4000
 iter <- 2000
 thin <- 500
 seed <-list(2003,5,63,100001,765)
 nstart <- 5
 best <- TRUE
 k <- 5
>#Run LDA using Gibbs sampling
ldaOut <-LDA(dtm,k, method="Gibbs", control=list(nstart=nstart, seed = seed,
best=best, burnin = burnin, iter = iter, thin=thin))

ldaOut <-LDA(dtm,k, method="Gibbs",control = list(nstart=nstart, seed =
seed,best=best,burnin=burnin,iter=iter,thin=thin))
ldaOut.topics <- as.matrix(topics(ldaOut))
write.csv(ldaOut.topics,file=paste("LDAGibbs",k,"DocsToTopics.csv"))
ldaOut.terms <- as.matrix(terms(ldaOut,6))
write.csv(ldaOut.terms,file=paste("LDAGibbs",k,"TopicsToTerms.csv"))
topicProbabilities <- as.data.frame(ldaOut@gamma)
write.csv(topicProbabilities,file=paste("LDAGibbs",k,"TopicProbabilities.csv"))

topic1ToTopic2 <- lapply(1:nrow(dtm),function(x)
sort(topicProbabilities[x,])[k]/sort(topicProbabilities[x,])[k-1])

topic2ToTopic3 <- lapply(1:nrow(dtm),function(x)
sort(topicProbabilities[x,])[k-1]/sort(topicProbabilities[x,])[k-2])
write.csv(topic1ToTopic2,file=paste("LDAGibbs",k,"Topic1ToTopic2.csv"))
write.csv(topic2ToTopic3,file=paste("LDAGibbs",k,"Topic2ToTopic3.csv"))
```

V. LITERATURE SURVEY

In the Machine Learning journal titled "Latent Dirichlet Allocation", describes the latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities.

In the paper titled "An improved data stream the count-minimum sketch and its applications", the authors described such as introduce a new sublinear space data structure—the Count-Min Sketch— for summarizing data streams. Our sketch allows fundamental queries in data stream summarization such as point, range, and inner product queries to be approximately answered very quickly; in addition, it can be applied to solve several important problems in data streams such as finding quantiles, frequent items, etc. The time and space bounds we show for using the CM sketch to solve these problems significantly improve those previously known - typically from 1/ε2 to 1/ε in factor.

In the paper titled "Approximate Frequency Counts over Data Streams", the authors described such as Present algorithms for computing frequency counts exceeding a user-specified threshold over data streams. Our algorithms are simple and have provably small memory footprints. Although the output is approximate, the error is guaranteed not to exceed a user-specified parameter. The input to the algorithm is a stream of transactions where each transaction is a set of items drawn front denote the current length of this stream by $. The user specifies two parameters support and use.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

VI. EXPERIMENTAL RESULTS

Fig.5 Emotion categories

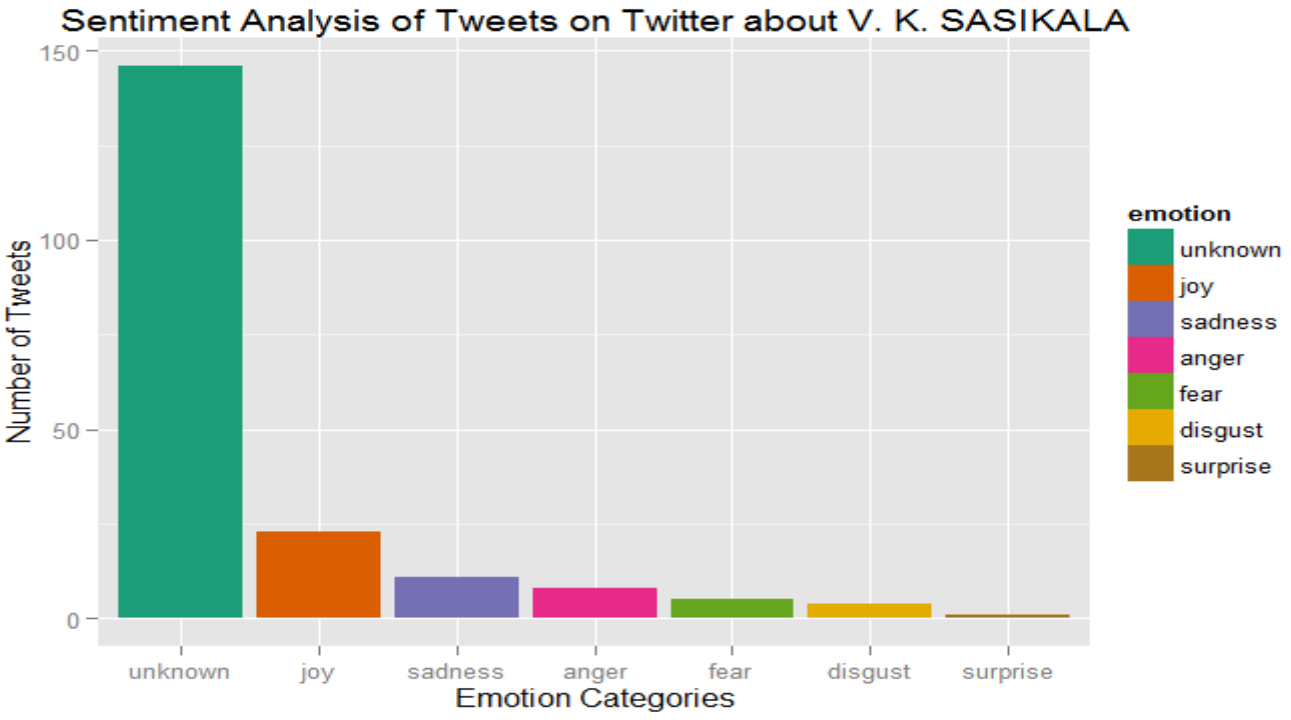

Fig.6 Reviews in .CSV Page

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

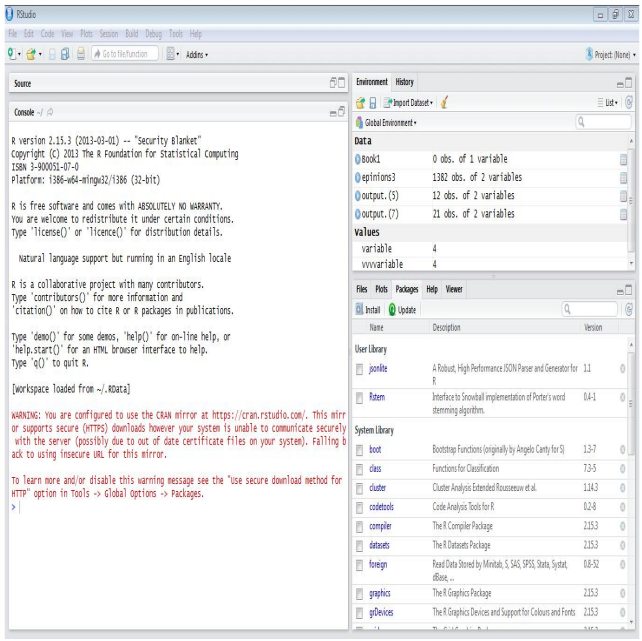

Fig.7 R-Studio

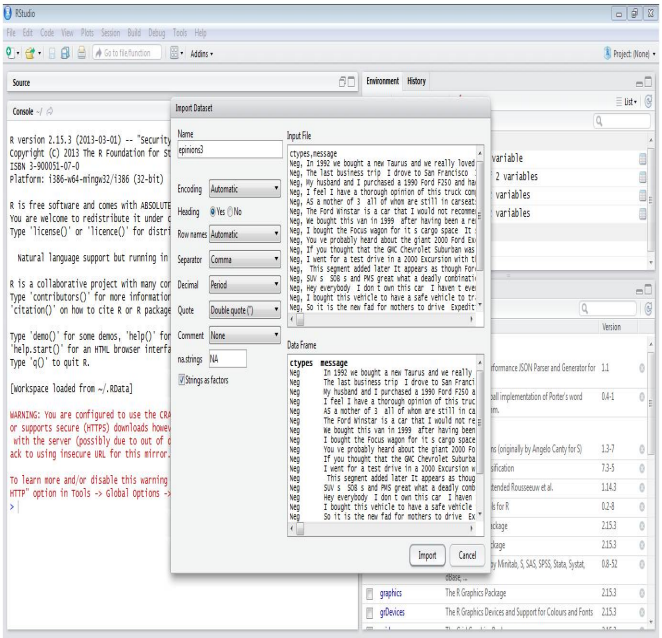

Fig.8 Import Data in R.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

VI. CONCLUSION AND FUTURE SCOPE

Data about tweets from twitter can be getting only for the days. Based on the Sentiment Analysis, Polarity Analysis, Frequent Words and WordCloud generated by using the data pulled from Social Medias like Facebook and Twitter the people felt very sad and fear about the flood caused by rain due to the existing systems proposed. TopicSketch a framework for real-time detection of bursty topics from Twitter. Due to the huge volume of tweet stream, existing topic models can hardly scale to data of such sizes for real-time topic modeling tasks. We developed a novel concept of "Sketch", which provides a "snapshot" of the current tweet stream and can be updated efficiently. Once burst detection is triggered, bursty topics can be inferred from the sketch. Compared with existing event detection system, our experiments have demonstrated the superiority of TopicSketch in detecting bursty topics in real-time. Experimental results indicate that the English reviews based Sentiment analysis and polarity analysis based opinion mining is necessary because nowadays everyone is busy and they don't have a time to read the entire positive or negative reviews if someone just wants to know about some feature of the product.

REFERENCES

[1] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In SIGIR, pages 37–45, 1998.
[2] L. AlSumait, D. Barbar´a, and C. Domeniconi. On-line lda: adaptive topic models for mining text streams with applications to topic detection and tracking. In ICDM, 2008.
[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.
[4] D. M. Blei and J. D. Lafferty. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning, pages 113–120, 2006.
[5] T. Brants, F. Chen, and A. Farahat. A system for new event detection. In SIGIR, pages 330–337, 2003.
[6] K. R. Canini, L. Shi, and T. L. Griffiths. Online inference of topics with latent dirichlet allocation. In Proceedings of the International Conference on Artificial Intelligence and Statistics, volume 5, pages 65–72, 2009.
[7] G. Cormode and S. Muthukrishnan. What's hot and what's not: tracking most frequent items dynamically. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pages 296–306, 2003.
[8] G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. Journal of Algorithms, 55(1):58–75, 2005.
[9] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim. Finding bursty topics from microblogs. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pages 536–544, 2012.
[10] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. Parameter free bursty events detection in text streams. In Proceedings of the 31st international conference on Very large data bases, pages 181–192, 2005.
[11] T. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl 1):5228–5235, 2004.
[12] D. He and D. Parker. Topic dynamics: an alternative model of bursts in streams of topics. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 443–452, 2010.
[13] M. D. Hoffman, D. M. Blei, and F. Bach. Online learning for latent dirichlet allocation. Advances in Neural Information Processing Systems, 23:856–864, 2010.
[14] T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57, 1999.
[15] L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis. A time-dependent topic model for multiple text streams. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 832– 840, 2011.
[16] A. Ihler, J. Hutchins, and P. Smyth. Adaptive event detection with time-varying poisson processes. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 207–216, 2006.
[17] C. Jin, W. Qian, C. Sha, J. Yu, and A. Zhou. Dynamically maintaining frequent items over a data stream. In Proceedings of the twelfth international conference on Information and knowledge management, pages 287–294, 2003.
[18] J. Kleinberg. Bursty and hierarchical structure in streams. Data Mining and Knowledge Discovery, 7(4):373–397, 2003.
[19] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 497–506, 2009.
[20] C. Li, A. Sun, and A. Datta. Twevent: segment-based event detection from tweets. In Proceedings of the 21st ACM international conference on Information and knowledge management, pages 155–164, 2012.

BIOGRAPHY

Ms. Charlin Monisha. A. Studying M.E. Computer Science and Engineering in Mount Zion College of Engineering and Technology, which is located in Lena Vilakku, Pudukkottai, Tamil Nadu, India.