BURSTY TOPIC DETECTION IN TWITTER

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF

Master of Technology
in
Software Engineering

Under the esteemed guidance of
Mr. Manoj Kumar
(Associate Professor)
Computer Science and Engineering
Delhi Technological University

Submitted By-
Shivani Singh
(Roll No. - 2K16/SWE/14)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

SESSION: 2016-2018

DECLARATION

We hereby declare that the thesis work entitled "**Bursty Topic Detection in Twitter**" which is being submitted to Delhi Technological University, in partial fulfilment of requirements for the award of degree of Master of Technology (Software Engineering) is a bonafide report of thesis carried out by me. The material contained in the report has not been submitted to any university or institution for the award of any degree.

Shivani Singh
2K16/SWE/14

CERTIFICATE

This is to certify that Shivani Singh (2K16/SWE/14) has completed the thesis titled "**Bursty Topic Detection in Twitter**" under my supervision in partial fulfilment of the MASTER OF TECHNOLOGY degree in Software Engineering at DELHI TECHNOLOGICAL UNIVERSITY.

Supervisor
Mr. Manoj Kumar
Associate Professor
Department of Computer Science and Engineering
Delhi Technological University
Delhi -110042

ABSTRACT

Twitter has turned out to be one of the biggest microblogging stages for people around the globe to share anything happening around them with companions and past. A bursty subject in Twitter is one that triggers a surge of important tweets inside a brief time of time, which frequently reflects essential occasions of mass intrigue. The most effective method to use Twitter for early location of bursty subjects has accordingly turned into an essential research issue with huge viable esteem. In spite of the abundance of research chip away at point displaying and investigation in Twitter, it remains a test to distinguish bursty themes progressively. Moreover no work has been done in the direction of analysing those bursty tweets. As existing strategies can scarcely scale to deal with the errand with the tweet stream progressively, the Topic Sketch methodology combined (a draw based point show together with an arrangement of methods to accomplish constant recognition)combined with naïve Bayes is proposed to detect as well as to analyse those bursty topics in terms of their polarity or sentiments.The analysis of this method on the tweets comes about with the result that shows both efficiency and effectiveness of this approach. Bursty-Event discovery calculations, have shown that the proposed approach can: (1) accomplish better model execution concerning the assessment criteria; (2) accomplish more exact bursty events on long/short content information.

ACKNOWLEDGEMENT

I am very thankful to Mr. Manoj Kumar (Associate Professor, Computer Science Eng. Dept.) and all the faculty members of the Computer Science Engineering Dept. of DTU. They all provided immense support and guidance for the completion of the project undertaken by me.

I would also like to express my gratitude to the university for providing the laboratories, infrastructure, testing facilities and environment which allowed me to work without any obstructions.

I would also like to appreciate the support provided by our lab assistants, seniors and peer group who aided me with all the knowledge they had regarding various topics.

Shivani Singh
M. Tech. in Software Engineering
Roll No. 2K16/SWE/14

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ABBREVIATIONS

SVD : Singular Value Decomposition

SA: Sentiment Analysis

NB: Naïve Bayes

DTM: Dynamic Theme Demonstrate

CHAPTER 1

INTRODUCTION

With millions of dynamic users and billions of tweets per month, Twitter has turned out to be one of the biggest data entries that gives a simple, speedy and dependable stage for clients to share anything occurring around them with companions and different adherents. Previous experiences have proved that, in many life-basic catastrophes, Twitter has been the most critical and convenient resource from which individuals discover and trace many big news even before any standard media grabs on those news and broadcast the recording. For case, on March 11.2011 ,Japan experienced tremor and ensuing tidal wave, the quantity of tweets sent shot up to in excess of 5,000 every second when the individuals posted news about the circumstance alongside transfers of portable recordings they did record during that time. Such occasions which trigger an extremely large number of (surge) a substantial number of important tweets are called ‘bursty topics’.

To distinguish bursty occasions, content stream could be viewed as a grouping of sequentially requested archives and bursty occasion is viewed as an arrangement of profoundly associated words show up in a brief timeframe of each particular word with high report recurrence. At that point, the mix of these particular bursty words is normally utilised to speak to the first bursty events. For illustration, the shocking news of Iraq-Iran War amid 1987 was generally detailed by numerous daily papers. Right off the bat, these archives are assembled together as per the timestamp. Also, bursty expressions of the occasion will be extricated in each gathering of records: at time t, "Iraq, Iran, armed force, troops" are extricated, though "assault, war, inlet" are removed at the time t + 1. In this way, extraordinary bursty words will be gathered together in light of their separation into occasions like: (1) armed forces amongst Iraq and Iran, (2) the warship fighting battle in the Gulf of Persia. Clearly, the bursty occasions are accurately distinguished. Be that as it may, such methodologies may be risky once there exist more than one occasion.

Generalizing, the sheer size of Twitter has already made it unimaginable for conventional media, or some other manual sources, to catch the vast majority of such bursty themes progressively despite the fact that their revealing team can get a subset of the drifting ones.

This hole brings up an issue of monstrous handy esteem: Can we use Twitter for mechanized bursty subject location continuously?

Lamentably, this continuous undertaking has still not been tended to by current works on Twitter subject examination. Initially, the own slanting subject of Twitter rundown does not helps that much as it only reports for the most part those record-breaking well known themes, rather than the trending(bursty) ones that are more of our piece of enthusiasm for this work. Secondly, most of the earlier works of research characterize bursty point as a collection which comprises of few of the bursty words .As just bursty words can be caught, the spoken to bursty theme is a long way from enlightening to actually reflect what is the point truly . Thirdly, most works based on theme demonstration contemplate the themes in the Twitter in the form of a review disconnected way, e.g., performing theme displaying, examination and following for all tweets produced in a specific day and age While these discoveries have offered fascinating bits of knowledge into the themes, it is our conviction that the best estimation of Twitter bursty theme discovery presently can't seem to be brought out, which is to distinguish the bursty points without a moment to spare as they are occurring. This constant errand is trying for existing calculations in view of the high computational many-sided quality inalienable in the theme models and also the manners in which the subjects are learnt generally, e.g., Sampling as explained by Gibbs [14] or like variational induction [5].

The key research challenge is the manner by which to unravel the following two issues continuously:

i. How to proficiently keep up appropriate insights to trigger identification.
ii. How we can show bursty points without having the opportunity to inspect the whole arrangement of important tweets as the case in conventional point demonstrating.
iii. How to analyse those detected bursty topics for their polarity.

1.1 MOTIVATION OF STUDY

This study is closely related to the one performed by Wei Xie et al. [1]. They proposed a recognition system called TopicSketch. TopicSketch can distinguish the bursty theme not long after the first tweet about an episode is created, exactly when the tweet begins to become viral, substantially sooner than the primary news report or report by media.

Their contributions can be summarized as under:

i. To begin with, they proposed a two-arrange incorporated arrangement TopicSketch.

1. In the primary stage, they proposed a little information outline which effectively keeps up at a low calculational cost the speeding up of two amounts: the event of matching of each word and the event of matching of each triple-word. These increasing velocities give as right on time as conceivable the pointers of a potential increase of tweet notoriety. They are likewise planned with the end goal that the bursty point deduction would be activated and accomplished in view of them. The way that we can refresh these insights effectively and summon the all the more computationally costly subject surmising part as it were at the point when fundamental at a later stage makes it conceivable to accomplish continuous recognition in an information flow of scale of Twitter.
2. Next, that is in second step, they proposed an outline based theme model to surmise both bursty themes and speeding up of the themes in view of the insights kept up in the information draw.

ii. Second, they proposed dimensional reduction methods in view of hashing to accomplish adaptability and, at the same time, keep up theme quality with strength.

In this research work, I used the TopicSketch approach as proposed by Weu Xie et al[1] to detect the bursty topics.

Further on work has been done in the direction of analysing the detected bursty topic in terms of its sentiment.

Briefly stating there are two main objectives of this research work:

1) Detecting the bursty topics using the running tweets.
2) Analysing the detected bursty topics in terms of their sentiment scores.

1.2 ORGANIZATION OF THESIS

The thesis is organized in various chapters as follows:

- Chapter 1 gives an introduction to the research work carried out and the main motivation of the work.

- Chapter 2 gives an overview of the related work of the study that is what is the various research works have been done in this area and how all those work helped in evolution of our study.
- Chapter 3 summarized research methodologies used in this thesis including overview of the recommended framework and the description of both the various techniques involved.
- Chapter 4 describes the experimental setup required to implement the research work.
- Chapter 5 states result of the study.
- Chapter 6 summarizes the research work under conclusion and suggests some future work.

CHAPTER 2

LITERATURE REVIEW

This chapter gives an overview of the research work done with relation to our thesis.

Event Detection has been contemplated for quite a long time, with developing interests on sites news[4],[27],[34], and as of late social media [28],[25]. Since there are various research works concentrating on the topic, here the ones generally related to this research work, i.e. bursty subject discovery in Twitter, are ordered , along two measurements: the method for characterizing a theme and next the method for handling information (as depicted in Table 2.1).

2.1 CATEGORIZING THE RELATED WORK

The related work is sorted along two measurements: the method for characterizing a theme and the method for handling information.

Table 2.1: Categorisation of related work

	Clustering - Based	Topic- Modelling Based
Retrospective	[34]	[11] [30] [31] [35]
Online or Real Time	[3] [4] [13] [8] [24] [34] [29] [28] [25] [26] [17] [32] [7]	[1]

2.1.1 Clustering-Based vs. Topic-Modelling Based: Diverse approaches are available to characterize the point of an occasion. In the starting work of Detection of First Story [34], [4] and the successors [26], [7], a point is spoken to as a bunch of similar(related) archives. By abusing the worldly vicinity of stories of news talking about a particular given occasion, Yang and others [34] utilize refined progressive & online report bunching calculations to identify occasions from a stream of news. In [4] each report is spoken to as the point present in a space of vectors (e.g. vector TF-IDF), and for each new approaching report, think about it against prior focuses. On the off chance that the new found point is sufficiently close to the closest

neighbour, consumed by the closest neighbour. Something else, this newer record is marked as another occasion. Brants and others [7] expand [4] by utilizing incremental approach in TF-IDF demonstrate, advanced likeness score standardization and so forth. Be that as it may, this method does not expand to the staggering information volume like that in Twitter, as closest neighbour seek is exorbitant on huge informational collection. Petrovic and others[26] utilize area touchy hashing known as LSH [20] to expand this method for streams of Twitter. While in different works, a point is characterized as the lucid set (or group) of watchwords [17], [8], [28], [25], [32], [29], hashtags [13], [3], fragments [24] or phrases in [23]. In these works[9], [15], [16], [33], more often than not an accumulation of the bursty terms are identified using the report stream in view of a few criteria, & perhaps afterwards these bursty words are assembled into a few bunches which speak to the bursty themes. For example, the best in class arrangement SigniTrend [29] to start with identifies the critical slanting terms (co-occurrences of word and the words themselves) in view of the proposed measurement by which the term significances of the terms is measured. With less memory, the measurement can be effectively refreshed in an increasing way. Additionally, by utilizing hashing procedures, it makes conceivable to trace the watchword combines under a settled sum of memory. At long last, a finish of-day investigation is performed, which totals the recognized watchwords into bigger points by utilizing grouping approaches.

Then again, utilizing a likelihood circulation over words to speak to a subject has been very normal in subject displaying [18], [5]. Particularly, the words having high probabilities would portray the theme nicely. It is direct to take in the themes in the archive stream utilizing subject models, & after that discover the bursty ones by thinking about their fleeting data, for example, [30], [11]. Takahashi and others [30] firstly utilize DTM or (dynamic theme demonstrate) [6] where the themes are taken from the stream of news, and then Kleinberg's model is applied [22] to recognize the bursty subjects. So also, Diao and others [11] construct a theme display which at the same time considers both the fleeting data of tweet and client's individual premiums to take in the points from the stream of tweets, and at that point model of Kleinberg model is utilized to discover the bursty themes. More of the related work incorporates [31] wherein a point model is fabricated to find corresponded bursty subjects from composed content streams, what's more, [35] which makes a brought together model to recognize worldly subjects from relatively stable points. In late researches [36], [19] subject models are worked to find land themes from Twitter. In spite of the fact that these works do not directly focus on

bursty themes, utilizing the comparative route as above, land bursty subjects could be well found.

In addition, Ahmed and others [2] proposed a period subordinate point bunch show, which consolidates LDA [5] and grouping to take in the point of every storyline, and in the meantime, to group records into storylines. Also, like [1], in most recent research [12] Du and others group consistent time record streams with the help of Dirichlet-Hawkes Process. In spite of the fact that their models are based on general record streams, they still can be conceivably received to distinguish bursty themes in Twitter.

Another contribution is by Chenliang Li et al., who gave the concept of Twevent. One very new component comprising Twevent is that it utilizes the thought of tweet section rather than unigram for identifying and portraying occasions. A portion of tweet is one or all the more back to back words (expression) in a tweet. We watch that the tweet fragments present in countless are liable to be named substances (e.g., Steve Jobs) or some semantically important unit (e.g., Argentina versus Nigeria). Along these lines, a tweet fragment regularly contains considerably more particular data than any of the unigrams contained in the fragment. The utilization of tweet section rather than unigrams in this way incredibly decreases the clamor in the occasion location process and furthermore makes the occasion distinguished substantially simpler to translate. For instance, Twevent distinguished an occasion with accompanying five portions [Greece, Korea, South Korea versus, Korea won, Greece, Korea] on the 12th of June 2010; the occasion is explanative in itself. Another newer component of the Twevent is use of outside learning base in managing the occasion location process. Also, in the accompanying, the fundamental strides in the Twevent used for occasion location from tweets is briefed.

2.1.2 Retrospective vs. Online or Real-time: In earlier research [34] Yang and others proposed strategies for the review & online occasion identification both. In the previous case, it was expected that there will be a review perspective of the information completely. Then again, on account of online occasion identification, the framework forms current record before taking a gander at any resulting archives. It isn't shocking that [34] demonstrates the aftereffects of review location are greatly improved than the one's online, as more data is accessible from a review way. As outlined in Table 2.1, generally theme displaying based techniques [11], [30], [31], [35] fall into this classification. The many-sided quality of their models makes them great at taking in themes from the information reflectively, yet in the meantime lose the adaptability

to react to any new approaching information. Conversely, techniques as [29] just need keeping up a measurement for every term, and report those terms when measurements of those terms surpass the hugeness threshold. Under this structure, online identification is a characteristic decision. Be that as it may, the distinguished point which comprises of couple of catchphrases is far less useful than the point gained from subject demonstrating, which is an appropriation over every one of the words.

Moreover, ongoing discovery is very like on the web recognition. The unobtrusive distinction between them is that in ongoing recognition, time is essential, to such an extent that no settled time-window for identification ought to be expected. The main works known about that accomplish constant discovery are [28] and [29]. While [28] detects occasions progressively, it but needs pre-characterized catchphrases for the point, which makes it inapplicable to bursty subject identification in a general way where no earlier information of the point watchwords is accessible. By incrementally refreshing the measurement in a very effective manner, Signi Trend as in [29] distinguishes bursty catchphrases progressively, yet before it totals watchwords into bigger themes, it is needed to hold up until finish of day or settled era.

In this research work, I expect to accomplish continuous bursty point location from point of view of subject demonstrating, which recognizes us from the most existing researches in scientific classification (as appeared in Table 2.1). Considering learning energy of theme displaying, the technique is expected to give more enlightening bursty points than other prevalent on the web identification arrangements.

CHAPTER 3

Research Methodology

This chapter first describes the methodology used for solving the problem. The steps involved and the algorithms used are explained in this chapter. Also the architecture and the framework involved is depicted.

3.1 PROBLEM FORMULATION

A topic as in this thesis work is spoken to as a dispersion over the words. Especially, in characterizing a bursty theme, we assess the accompanying two criterias: (1) There must be a sudden increase or surge in theme's popularity which we estimate by the aggregate of the pertinent tweets. These record-breaking prominent themes along these lines will not be checked; (2) The point must be sensibly prominent. This in turn would channel away extensive quantity of inconsequential themes which, in spite of the surge (spike) in the fame, are hence considered as noise in light of the immaterial quantity of important tweets.

For point (1), our work is to measure how much bursty a subject is by the speeding up (acceleration) of its fame or popularity. Numerically, speeding up catches the adjustment in the rate of prominence of a point. The more suddenly the change occurs, the bigger the speeding up is. Further, it is discussed how to assess the speeding up of any topic without prior knowledge of which tweets actually are related to it. For rule (2), once we discovered bursty point hopefuls, we essentially tally the pertinent tweets of those topics, and sift through the minor ones.

The main aim in this research work is hence to detect bursty topics from the given tweet stream as early as possible and then to analyse those detected bursty topics in terms of the sentiments they express over time.

3.2 SOLUTION OVERVIEW

The answer, referred as TopicSketch, depends on two principle methods — an outline based point demonstrate and a hashing-based measurement diminishment method. Our portray based subject model gives a coordinated two-advance arrangement. In the initial step, it keeps up as an outline of the information the speeding up of two amounts: (1) each combine of

words, and next (2) each triplet of words, where these triplets are early markers of ubiquity surge & can be refreshed proficiently with ease, making early location conceivable. In the second step, in light of the information outline, it takes in the bursty subjects by tensor disintegration. To play out the location effectively in expansive scale constant setting, we propose a measurement decrease system in view of hashing which gives an adaptable answer for the first issue without trading off much the nature of the subjects.

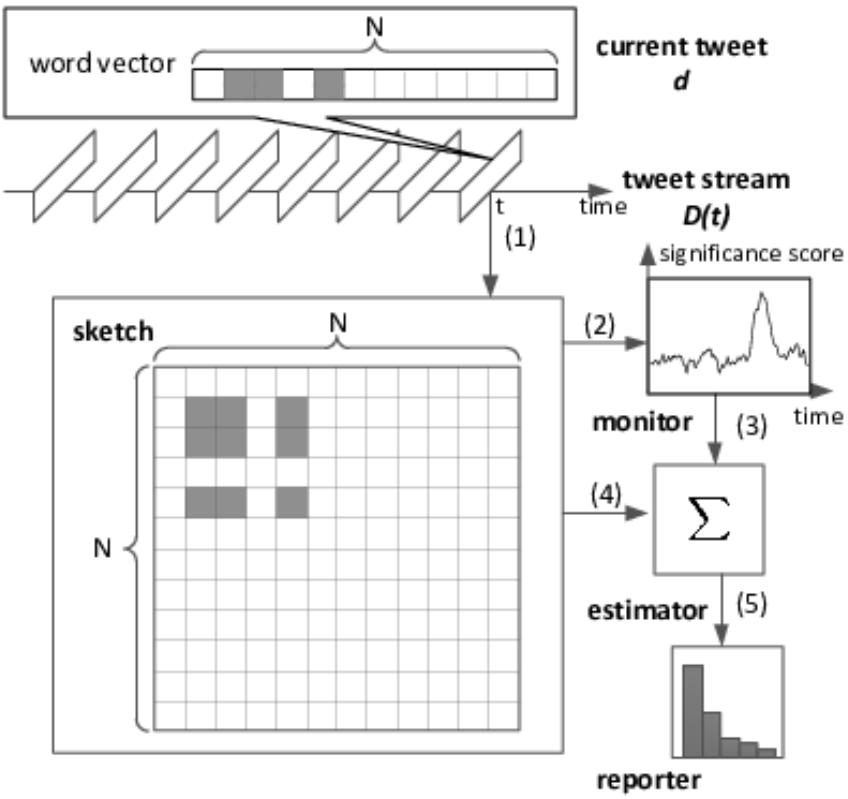

Figure 3.1: System Architecture

Figure 3.1 shows the basic functioning of the TopicSketch model. The flow for detecting real time bursty tweets goes as follows:

1) As soon as a tweet arrives, the tweet stream is refreshed.

2) Changes obtained are sent to the screen, if sketch did get refresh.

3) If obtained difference is more prominent than the pre-decided threshold, the screen ways the information draw and contrasts it and old averages and the estimator is then incited to find the subjects which are trending.

4) Estimator gets a picture(snapshot) of sketch and make sense of the trending subjects, upon supply of the notification from screen.

5) Induced trending subjects are then send to the reporter to assess and then report.

TopicSketch is composed with the end goal that steps (1) to (3) are computationally inexpensive to empower constant reaction and earlier discovery. Step number (4), which proves to be costly if done gullibly, is incredibly assisted with the help of dimension reductioning procedures as depicted further.

The last step is then the analysis of the detected bursty topic or in other words analysing the sentiment of the topic(SA) which is done mainly through Naïve Bayes (NB) classifier.

As a matter of fact, as specified in work [17], the expression "bursty topic" is extremely vague, and can be seen in altogether several ways. Different instincts and relating explanations on the topic lead to assorted arrangements [17], [3], [29], [22]. The instinct behind the work originates from the perception that, the entire stream of tweet is brimming with expansive measure of tweets about general points, for example, auto, music and nourishment. In spite of the fact that they take a substantial extent in the entire stream of tweet, they are most certainly not supportive for the bursty theme identification assignment. In this manner, we endeavour to isolate the bursty points from them. We found that, following day by day schedule, individuals ordinarily tweet about general themes in an unfaltering pace. Interestingly, bursty themes are regularly activated by a few occasions, for example, some big news or a convincing ball game, where these topics get a ton of consideration from individuals, and compel individuals to tweet regarding the topics strongly. In material science, this above "force" which can be communicated by "acceleration", which in this setting portrays the difference in "velocity", i.e.which describes the rate at which tweets arrive. Bursty themes can get critical increasing speed whenever they are blasting, while general themes typically get almost zero speeding up. So the "increasing speed" trap can be utilized to safeguard the data of bursty subjects yet sift through the others. Nonetheless, as the subjects are shrouded, we cannot ascertain their increasing speeds specifically. A conceivable path is to gauge them by figuring the increasing speeds of words. Equation 3.1 indicates how the "acceleration" aˆ (t) and the "velocity" vˆ(t) of words can be figured.

$$\hat{v}_{\Delta T}(t) = \sum_{t_i} \leq T \, Xi. \left(\exp\left(\frac{t_i - t}{\Delta T}\right)\right) / \Delta T)$$

$$\hat{a}(t) = (\hat{v}_{\Delta T_2}(t) - \hat{v}_{\Delta T_1}(t))/(\Delta T_1 - \Delta T_2) \qquad (3.1)$$

In Equation 3.1, X_i is the frequency or recurrence of a word (maybe a couple of words, or might be a triple of words) present in i-th tweet, t_i is the timestamp. Exponential part in the $v\Delta T$ (t) actually works as a moving window which is delicate, and gives high weight to the ongoing terms, however gives lower weightage to the terms which are far away, & the smoothing parameter ΔT is actually the measure of the window. To catch the difference in speed, quickening a(t) is characterized as the contrast of speeds with various window estimate $\Delta T1$ and $\Delta T2$. (Like the uniqueness of five day normal and ten day normal in securities exchange, which is utilized to gauge stock drift.).

Simply speaking, the methodology for detecting and analysing the bursty topic can be stated in the following steps:

1) Extracting the running tweets. Running tweets imply the live tweets, which are currently being tweeted by users all over the world.
2) The second step involves pre-processing which mainly involves data cleaning. It involves removal of stop words, URL's , number and other unnecessary words not required for further processing.
3) Then we find out the data acceleration per unit of time (which may be hour), that is to say how the trend of the tweets change with respect to time.
4) Next, we find out those similarity level of the tweets representing the same topic.
5) The next step is to calculate the significance level of those similar tweets, and the tweets which exceed the predefined threshold then qualify as those concerning the bursty topic.
6) Lastly, the bursty topic is analysed by the sentiment analysis of the tweets comprising the bursty topic which is done by using Naïve Bayes classifier.

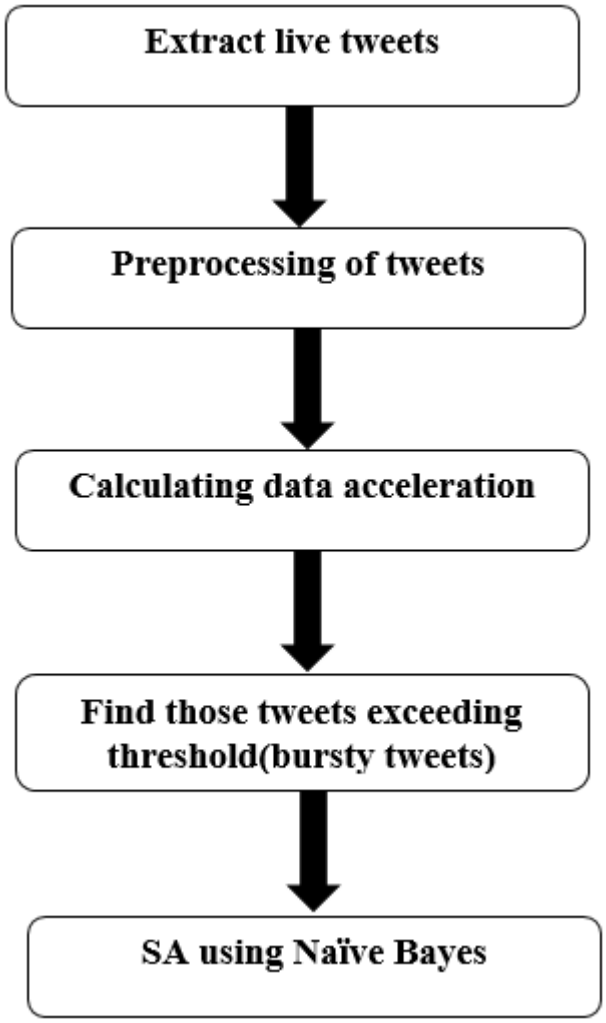

Figure 3.2 Flowchart of the proposed work

3.2.1 Sketch based Topic Model

Signify D = {di} as arrangement of all the tweets created in the stream of tweets, where di is said to be a tweet in the given stream D, and ti is the timestamp. Additionally signify $_{Ci}$ the quantity of appearance of the word w in the tweet di , and Ci as aggregate num of words in the tweet di . w ∈ [N], where the symbol N is quantity of particular words present in vocabulary.

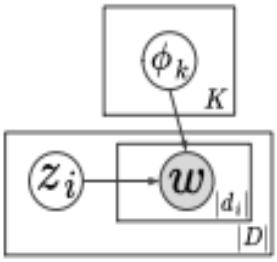

Figure 3.3: Single Topic Model

Considering the accompanying single subject model (appeared in Figure 3.2). There are K points $\{\varphi_k\}$ [k=1 to k], where every point φ_k is a dissemination over words. Here the key supposition is each tweet di is just related to one inactive point z_i . (This supposition depends on the way that measurement of each of the tweet is less, restricted by usage of 140 - characters) & each of the word in di is then taken from Multinomial (φz_i).

Accept points $\{\varphi_k\}$ [k=1 to k] and point pointer z_i are obscure yet settled, for the recurrence of a word w in di

It implies the single word recurrence f1, i mirrors the theme φz_i. Notwithstanding, as the point pointer z_i is obscure, we cannot straightforwardly induce theme from it. We know that, word sets are valuable to aggregate diverse words into subjects, for example, the "goads"-"amusement" obstruct in the warmth maps. Here, we consider computing the recurrence of each of the word combine. Especially, with the interchangeability of the words in the tweet di (which is based on single theme display above), the recurrence of a word match (w_1, w_2) is then characterized in a tweet di as takes after,

$$f_{2,i}[w_1, w_2] = \left\{ \frac{P(C_{i,w_1}, 2)}{P(C_i, 2)} \quad , \quad w_1 = w_2 \right\}$$

$$f_{2,i}[w_1, w_2] = \left\{ \frac{C_{i,w_1}\, C_{i,w_2}}{P(C_i, 2)} \quad , \quad w_1 \neq w_2 \right\}$$

(3.2)

where P(C, 2) tallies the 2-changes (permutation) of C. Here, the denominator is quantity of every single conceivable case picking 2 out of C_i words in tweet di, whereas the numerator is number of the cases picking P 2 particular words. As the words in di are drawn from a similar

circulation Multinomial (φzi), it can be demonstrated that E[f2,i[w1, w2]] = φzi [w1] · φzi [w2].

For the purpose of simplicity, the notation of tensor product ⊗ is adopted, and f2,i is denoted in a matrix form where the [w1, w2] element is f2,i[w1, w2]. So we get an equation which is equivalent to the previous one:

$$E [f2,i] = \varphi z_i \otimes \varphi z_i.$$

As discussed previously, we ascertain the acceleration of word sets to protect the vast majority of the data for bursty points, and in the meantime channel other general points out. Set Xi = f2, i[w1, w2] and apply condition 3.1, we can compute the acceleration of any word match (w1, w2). Notice that the speeding up a(t) in Equation 3.1 is to be sure a weighted aggregate of grouping {Xi}, and the weight just relies upon $\{t_i\}$. At the end of the day, a(t) can be characterized as a direct capacity At(·) on {Xi}. Expressly we mean At({Xi}) as the acceleration a(t) on {Xi}.

For each of the tweet di, the word pair frequency matrix f2,i, is calculated ,so a sequence of matrices is obtained, i.e.{f2,i}.Hence , based on {f2,i},acceleration for every word pair frequency, i.e. At({f2,i[w1, w2]}) is calculated. So that we have a matrix At({f2,i}) in which the [w1, w2] element is At({f2,i[w1, w2]}).

In this manner we get an approach to connect the noticeable increasing speed of word combine recurrence to the obscure increasing speed of subject k without the knowledge of the tweets which are related to it. All the more critically, also it suggests that by keeping up the speeding up of word combine recurrence, i.e. At({f2,i}), we can protect the data of the themes with higher increasing velocities (bursty subjects), what's more, in the meantime sift through the subjects with almost zero increasing speed (stable points which are general in nature). This is precisely what is needed.

In the event that for any two subjects k_1 and k_2, φk_1 furthermore, φk_2 are symmetrical, i.e. $h\varphi k_1$, φk_2 I = 0, every one of the themes $\{\varphi_k\}$ will be the eigenvectors of At({f2,i}). Assuming this is the case, simply a single SVD (Singular Value Decomposition) can be performed on At({f2,i}) to construe these themes. Be that as it may, $h\varphi k_1$, φk_2 I = 0 implies subjects k1 and k2 have no regular words, which is a long way from reality. In this manner considering much higher request data, the recurrence of each word triple. Like f2,i, recurrence of word triplet (w1, w2, w3) in a tweet di can also be characterized.

Additionally, it can be demonstrated that

$$E[f3,i] = \varphi z_i \otimes \varphi z_i \otimes \varphi z_i$$

So at this point, we have two representations: M2 and M3. Notice that every one of these increasing velocities are anything but difficult to figure and refresh upon the landing of each tweet (as will be discussed in further sections), which is basic for adaptability progressively setting.

As M3 could also be enormous (a N * N * N framework), we do not by any stretch of the imagination store the M3, yet venture it to a N × N grid M3(η) = P w M3[:, :, w] · η[w], where η ∈ R N , is an irregular vector. Anyway a N * N lattice is as yet immense, in further sections, we talk about additional about lessening the space multifaceted nature.

To distinguish bursty subjects from the information outline, which comprises of two lattices M2 and M3(η), we utilize a tensor disintegration calculation .This calculation first plays out a SVD on M2 to discover a brightening lattice W. Thereafter, brighten M3(η), at that point play out another SVD on brightened M3(η) to discover eigen vectors of it, from where the subject vectors could be recuperated. Here the methodology is given in further section. It mainly has three sections: (1) Whitening, here the main thing is to change M3(η) from a N * N network to a K*K lattice T3; (2) SVD, it gets the summed up vectors $\{v_k\}$ of T3; (3) Reconstruction, it recuperates the subjects $\{\varphi_k\}$ and their comparing increasing speeds $\{a_k\}$. As known , K << N, the tedious task here is part (1), it requires some investment in the request of $O(K \cdot N^2)$.

3.3 METHODS USED

There are basically 3 methods involved in the whole process:

- Tensor Decomposition
- Dimesion Reduction(through hashing)
- Analysing the detected bursty topic.

3.3.1 Tensor Decomposition

In multilinear polynomial math, a tensor decay is any plan for communicating a tensor as an arrangement of basic activities following up on other, regularly more straightforward tensors. Numerous tensor deteriorations sum up some matrix disintegrations.

3.3.2 Dimension Reduction

The principal challenge is mainly the high or large dimension issue, therefore of the tremendous number of particular words N in the stream of tweets, which could without much of a stretch achieve the request of millions or even bigger. In addition, client created newer words or also hashtags dependably show up in Twitter. It outcomes not just in a colossal information outline (review M2 and M3 (η) in draw are N * N grids) yet additionally high measurement contribution to the method.

Since the issue is predominantly in light of the fact that N is too huge, one normal arrangement is to keep just an arrangement of dynamic words experienced as of late, e.g. over the most recent few minutes. Whenever a burst is activated, consider just only the words in this ongoing set. Be that as it may, things being what they are the extent of this lessened dynamic word set used for the stream of tweets is still too expansive to gather the themes proficiently.

To deal with huge number of words, also another normal way is used known as hashing as explained in [29]. The particular words are hashed into B pails (buckets), here B is actually a number considerably littler than N, and all the words in a pail are treated as one single "word". Subsequently, the size of draw moves toward becoming $O(B^2)$, which are fundamentally littler than $O(N^2)$ as in the first issue. Be that as it may, in the wake of hashing, what we acquire is the circulation over containers, as opposed to the conveyance with the words. It implies we need to recoup the probabilities of the words from probabilities of the containers. To tackle this issue, we adjust the Count-Min calculation [21], [10] to the setting useful for us, by utilizing H hashing capacities rather than one. Given, we have, H hash capacities (H_1, H_2, ..., H_H) which outline to basins [1...B] consistently and freely. For any subject k with a word dispersion φ_k, we firstly gauge its circulation over the buckets for all the hash capacities. At that point we recuperate the likelihood of each word i.

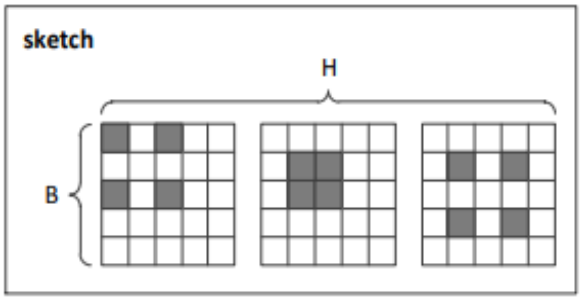

Fig 3.4 Sketch after hashing (dimension reduction)

The sketch depicted in Figure 3.4 is that after applying hashing. As appeared in Figure 3.4, for each arriving tweet, H extraordinary basin combine frequencies are computed, and H networks are refreshed correspondingly. After the measurement diminishment, the cost of memory for the portray is $O(H \cdot B^2)$, & the time many-sided quality for the tensor decay is $O(H \cdot B^2 \cdot K)$, hence which are sufficiently little to be essentially attainable.

Likewise, pool of dynamic words is kept, with the goal that we appraise the likelihood of words just in this particular set rather than every one of the words in entire vocabulary. This calculation will gauge the likelihood of every word having mistake no more prominent than e/B with likelihood $e^{(-N/eH)}$. The subtle elements of confirmation are present in [21].

3.3.3 Efficient-Sketch Maintenance

The center piece of portray support in the answer is quickening count. As introduced in previous section, equation 3.1 is received to compute acceleration. Nonetheless, straightforwardly applying equation 3.1 is a long way from productive. Watched that the speed v(t) in equation 3.1 can be ascertained in an incremental route. So all things considered of specifically putting away one quickening, we incrementally keep up two speeds $v\Delta T1$ (t) and $v\Delta T2$ (t), using which, increasing speed can be determined on the go. Both space intricacy for keeping up one quickening & the time multifaceted nature for refreshing one increasing speed are in this way O(1).

Also there is one more issue that is, there are $O(N^2)$ increasing velocities to be refreshed whenever any tweet arrives, and $O(H \cdot B^2)$ in case of the measurement diminished case. Here, we take sluggish support system to decrease the processing cost. Indeed, for every speeding up, take (w_1, w_2) cell in M2 for instance, we store a speed match ($v\Delta T1$ (t), $v\Delta T2$ (t)) and also a timestamp denoted as t* speaking to the last change time. At the point whenever a tweet di arrives, if $C_i,w_1 \cdot C_i,w_2 > 0$, we refresh the match ($v\Delta T1$ (t), $v\Delta T2$ (t)) and refresh timestamp t* to t_i, else we receive sluggish technique and just do nothing. At the point when a bursty subject is activated at time t, take a depiction of the draw (duplicate on compose), at that point recoup all the speed matches up to time t by determining the increasing velocities from the speed sets as per Equation 6. Figure 2 outlines this refreshing methodology: When a tweet arrives, its assertion vector is appeared. The dark cells present in word vector indicates the event of significant words in tweet. In the particular case, we have three words in the present tweet. Dark cells in portray speaks to the refreshed components in the portray. $O(|d|^2)$ cells are refreshed altogether. Figure 5 shows this refreshing system for measurement diminished case.

Then, after hashing, the three words in present tweet are then mapped to two basins for three of the hash works separately. Altogether, $O(H \cdot |d|^2)$ cells are refreshed.

3.3.4 Topic-Inference

When once the framework is activated, we then take a depiction of the outline, and after that induce the bursty themes from the draw as appeared in previous section. And as talked about later, of straightforwardly keeping up an outline with the size of N * N, we keep up H portrays with the size of B * B, here B is much littler than N. Again we demonstrate to construe the bursty points from the draw after measurement diminishment. Initially, we get points from the H distinctive outlines by Tensor Decompose (equation 3.1). It is to be noted that this progression can be actualized in parallel. One inconspicuous issue again is that a bursty theme may also have distinctive theme lists in various representations (review we now have H diverse draws). For example, portray h = 1 might catch a bursty theme in point φ (1) with file 1, in any case outline h = 2 may catch the same bursty point in subject φ (2) with list 3.So,to adjust the subjects which speak to same bursty subject, these points are sorted by comparing the increasing speeds, and afterward re-list them as per the request. Finally, we get the points recuperated by Topic Recover. With the end goal of strong, we likewise propose a variation of this calculation by utilizing the second least an incentive rather than the base esteem. We will examine this more in the test part.

Subsequent to getting the points from topic inference method, a heuristic advance may be then added to refine the subjects, due to the accompanying reasons. In the first place, in light of the first increasing speed based outline, the answer is delicate when confronting spam profiles and accounts. They as a rule infuse a considerable measure of copy or then again comparable tweets seriously in a brief timeframe, which would deliver words with noteworthy speeding up, furthermore, in this way trigger our framework. Second, because of hashing impact, it becomes conceivable that few uncommon words show up in the recuperated theme. The procedure of refining is as per the following.

- Trivial point sifting: sift through the themes with little number of important tweets.
- Noisy point sifting: sift through the themes whose entropy is high, which do not have any high likelihood words. These type of points look like clamours.
- Spam separating: Done by checking the tweets related to the subject in the late five minutes. On the off chance that there is a record that posts noteworthy many tweets, sift through this theme.

- Rare- word pruning: the words which don't show up in the ongoing tweets are pruned
- A rearranged file of ongoing tweets is actualized, so that this progression can be performed effectively.

3.3.5 Analysing the Detected Bursty Topic

The analysis of the detected bursty topic is a step forward to provide more meaningful information regarding the detected topic. We know the detected topic in itself provides no useful information as to what is actually being discussed about it, that is to say, if the general trend of the tweets towards the topic are positive negative or have no sentiments (neutral).

Taking an example to explain clearly. We all know that bursty topic is a topic which gains more popularity in a short span of time due to the large number of tweets related to that topic. So if for example we detect a bursty topic as "BJP", now we know that we have many tweets for "BJP" in a short span of time. But we don't know if those tweets are in positive context or negative. It may happen that some people are tweeting positively about the party, it may happen that some are tweeting negatively or again some may have absolutely neutral opinion for the party. So we need to find out what is the general or most common opinion of people regarding the party which they express through the tweets. Now after knowing this opinion, we have more productive and meaningful information regarding the detected bursty topic which can be used further for various analysis purposes.

Now this step can basically be termed as sentiment analysis. Sentiment mining (may be also called as supposition investigation) alludes to utilizing of content examination ,characteristic dialect preparing, biometrics and computational etymology to distinguish, separate, measure, and concentrate subjective data and emotional states. As a rule, feeling examination plans to decide what is actually the mentality of an essayist, a speaker or some other subject as for some theme or the general extremity which is relevant or enthusiastic response to a report, occasion, or cooperation. This disposition might be some judgment or some sort of assessment, full of feelingful state (or, the passionate state of the creator or speaker), or the suggested enthusiastic correspondence (or, the impact which is enthusiastic in nature expected by the creator or conversationalist)

Now, we detect the sentiment of the bursty topic by using the technique of Naïve Bayes classification.

In machine learning, naive Bayes classifiers are actually a group of simple probabilistic classifiers as in the view of applying the Bayes' hypothesis with solid (gullible) freedom presumptions.

Naive Bayes is contemplated widely from 1950s. Actually, it was firstly brought with a different name in the field of content recovery network in mid 1960s, and still remains a mainstream (gauge) technique for classification of content, issue of records judging as they might have a solution as either one classification or the other one, (as for example, games or governmental issues, spam or genuine, etc.) having the highlights as word frequency. With proper pre preparing, it is focused in this particular area with many further strategies which are more developed as like bolster vector machines. In addition, it also has application in the field of programmed medicinal diagnosis.

Bayes classifiers are very adaptable, and require several parameters straight in the amount of factors is hence a learning issue. Greatest preparation of probability is possible with the help of assessing a shut shape expression that takes straight time, instead of expensive iterative estimate as used in many other sorts of classifiers.

The field of software engineering give different names to the Bayes names, including simple Bayes and independence Bayes. Actually all of these names point how Bayes' hypothesis is used in the choice of classifier's choice, yet naive Bayes cannot actually be called a Bayesian method.

So, we know the last step is to analyse the detected bursty topic for their sentiment or polarity which is done using Naïve Bayes classifier. We have the stored tweets comprising the bursty topic. These stored tweets are already in pre-processed form and hence directly Naïve Bayes classifier can be applied on those tweets for detecting the sentiment of the detected bursty topic in the previous step.
The main reason for using Naïve Bayes classifier is that it gives best accuracy in the case of text analytics.

CHAPTER 4

Experimental Setup

This work is implemented in Python language. Hence various python libraries need to be imported as required by the system. These also include the nltk package which is a natural language processing toolkit package which is mainly required in the analysis stage of the bursty topic.

Visual Studio Supporting tools are also required for supporting our work.

For the purpose of detecting the bursty topics over time, live tweets are used, which are saved in a MySQL database. This database saves the tweets in processed form which is required for further calculations.

These tweets are then used for classifying the bursty topics using the algorithms explained in chapter 3 which mainly include comparing them with the predefined threshold.

The sketch obtained on the screen shows us the bursty topics obtained over time.

This topic is then analysed, i.e., the tweets related to the topic are analysed using the Naïve Bayes classifier for detecting the polarity of the bursty topic.

At last another graph is obtained which shows the polarity of the detected bursty topic.

CHAPTER 5

Result and Analysis

After performing the steps as discussed in chapter 3, we obtain our results. The results are obtained in graphical form.

We obtain two graphs which show the following:

1) The first graph shows the detected bursty topics and their changing frequency with respect to time.
2) The second graph obtained shows the sentiment score of the detected bursty topics obtained in the first graph. A value of sentiment score greater than zero shows that the detected bursty topic is positive in nature that means people are tweeting about that topic more in a positive manner, whereas a negative sentiment score means that people are tweeting about that topic more in a negative way. Whereas a sentiment score of zero suggests the topic to be neutral, that means there are no positive or negative reaction of people tweeting about that topic.

   So explaining this by taking an example of a result obtained over time:

	burstyWord	startTime	...	sentiment	frequency
0	#CongBetrayedVemula	2018-06-20 09:56:53	...	-5	10401
1	UNHRC	2018-06-20 09:57:11	...	7	24445
2	Gulf	2018-06-20 09:56:41	...	7	12467
3	The US	2018-06-20 09:57:30	...	2	484679
4	#WednesdayWisdom	2018-06-20 09:57:38	...	5	31550
5	#WorldRefugeeDay	2018-06-20 09:57:51	...	-1	49878
6	#Thalapathy62FLTomorrow	2018-06-20 09:58:03	...	19	15783
7	#ARMYHiveStreamingParty	2018-06-20 09:58:15	...	2	428881
8	#ENGvAUS	2018-06-20 09:58:12	...	3	39903
9	#RUSEGY	2018-06-20 09:58:54	...	2	161006
10	#CongBetrayedVemula	2018-06-20 09:59:46	...	1	10639
11	UNHRC	2018-06-20 10:00:39	...	10	24534
12	Gulf	2018-06-20 10:00:55	...	0	12533

[13 rows x 6 columns]

Figure 5.1 Output showing the detected bursty topics with the details

As seen in figure 5.1, we obtain the output as the set of bursty topics, their frequency and their average sentiment score over the given period of time. Hence, the date and start time is also

taken into consideration. More explained output is given by the graphs of the bursty topics and their sentiments obtained respectively.

So as, seen in figure 5.1 the first result we obtain is the graph showing the detected bursty topics. The X axis shows the date and time for which the tweets are considered for detecting the bursty topics. The Y axis shows the frequency or changing acceleration of those topics with respect to time. In other words this graph shows the bursty topics in the given time and their changing trend with respect to that time.

In figure 5.2, we can see we obtain 10 bursty topics for the given time period, which is represented on the X-axis .Hence, the X-axis shows the date and the time (which is in GMT). Hence, we get the detected bursty topics in the given time period , that is to say that people are tweeting more and more about these topics in the given time period or in other words , it can be said that these are the topics which are trending or have gained popularity in the given time period. Also the given graph shows the frequency of these topics which is shown in the Y-axis. The frequency of these topics mean the number of tweets associated with the given topic in the specified time interval. Hence, we know the trending topics as well as the number of tweets associated with those topics.

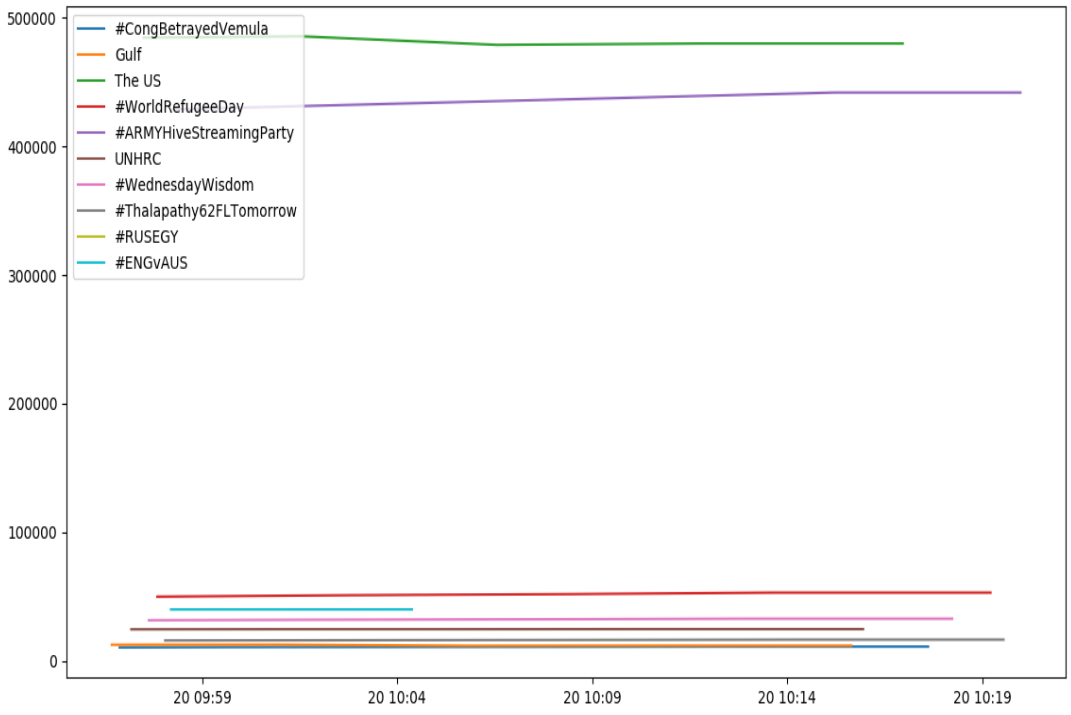

Figure 5.2 Graph showing detected bursty topics and their frequency

The slope of the lines in the graph can be calculated to find the acceleration of the topics that means how the trend of the tweets of that topic change with respect to time.

The second graph is the obtained which shows the sentiment score of the detected bursty topics. This sentiment score ranges is represented on the Y-axis and same as the previous graph, time is represented on the X-axis. A positive sentiment score shows the positive polarity expressed

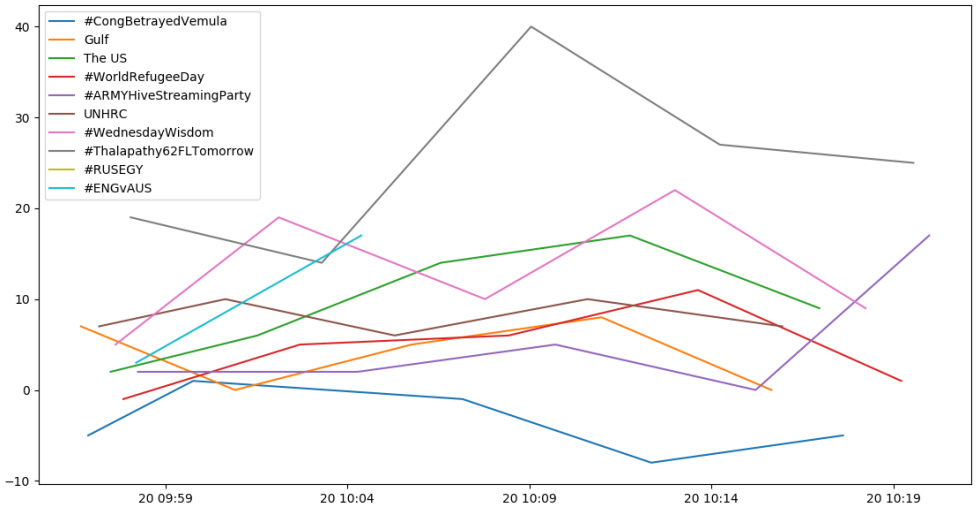

Figure 5.3 Graph showing sentiment score of the bursty topics

by the bursty topic that is the related tweets are positive in nature. Similarly, a negative sentiment score shows the negative polarity expressed by the bursty topic. Or a score of zero shows neutral polarity. Here, this graph also shows the changing sentiments associated with the topic with changing time. It means it may happen that at some time the tweets related to the topic are more of positive nature or at some other point of time, this sentiment may change to negative or neutral. Hence the information of the changing sentiments associated with the detected bursty topic is also obtained though this graph.

CHAPTER 6

Conclusion and Future Work

6.1 CONCLUSION

In this research work, the TopicSketch approach, a system for continuous recognition of bursty subjects from Twitter, is used. Because of the colossal volume of stream of tweets, existing subject models can scarcely scale to information of these sizes for ongoing subject displaying assignments. We built up a "draw of theme", which gives a "depiction" of the present tweet stream and can be refreshed proficiently. When blasted identification is activated, bursty themes can be construed from the outline productively. Contrasted and existing occasion location framework, from an alternate point of view – the "increasing speeds of themes", our arrangement can identify bursty subjects progressively, and exhibit them in better granularity.
We see this approach is very effective and efficient to detect the bursty topics from twitter in the real time , that is to say detecting the bursty topics continuously from the live tweets , which the methods used before were incapable to do. They did not identify the bursty topics in the ongoing tweets or otherwise were very inefficient in doing so.
Moreover, this work of detecting the bursty topic was extended to analyse the bursty topic as well which helps in providing meaningful information related to the detected bursty topic.
In other words the approaches of detecting the bursty topic and sentiment analysis is combined in this research work to gain more insight on the detected bursty topic so that it can be helpful and be used for analysis in various fields. This information can be used for data mining. As for example in elections the positive or negative talks about a popular party can help them analyse their current situation in the country. Same way there are many practical applications where this combined approach of analysing the sentiment of bursty topic can be used.
Hence this approach as a whole provides an effective and efficient solution to detect the bursty topics from the ongoing tweets and finally analysing them based on the opinions of the general people who tweet about that topic.

6.2 FUTURE WORK

And as a part of the future work, this approach can be further extended to analyse various other aspects of the detected bursty topic. These aspects may include detecting the sarcasm that is if there are more of sarcastic tweets regarding that topic. Or else this approach may be extended to detect fake ids or spammers who may try to spread some positive or negative sentiment about a particular trending topic. Other aspects include improving the efficiency of the approach by using some better and faster algorithm.

REFERENCES

[1] Wei Xie, Feida Zhu, Jing Jiang, Ee-Peng Lim, and Ke Wang. TopicSketch: Real-Time Bursty Topic Detection from Twitter, In *IEEE Transactions on Knowledge and Data Engineering , Volume: 28, Issue: 8, Aug. 1 2016 ,* pages 2216-2229.

[2] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia,* pages 37–45, 1998.

[3] F. Alvanaki, S. Michel, K. Ramamritham, and G. Weikum. See what's enblogue: real-time emergent topic identification in social media. In *15th International Conference on Extending Database Technology, EDBT '12, Berlin, Germany, March 27-30, 2012, Proceedings*, pages 336–347, 2012.

[4] A. Ahmed, Q. Ho, C. H. Teo, J. Eisenstein, A. J. Smola, and E. P. Xing. Online inference for the infinite topic-cluster model: Storylines from streaming text. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS , Fort Lauderdale, USA, April 11-13, 2011,* pages 101–109, 2011.

[5] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[6] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

[7] T. Brants and F. Chen. A system for new event detection. In *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada,* pages 330–337, 2003.

[8] M. Cataldi, L. D. Caro, and C. Schifanella. Personalized emerging topic detection based on a term aging model. *ACM TIST*, 5(1):7, 2013.

[9] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, pages 288–296, 200.

[10] 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.

[11] Q. Diao, J. Jiang, F. Zhu, and E. Lim. Finding bursty topics from microblogs. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*, pages 536–544, 2012.

[12] N. Du, M. Farajtabar, A. Ahmed, A. J. Smola, and L. Song. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 219–228, 2015.

[13] W. Feng, C. Zhang, W. Zhang, J. Han, J. Wang, C. Aggarwal, and J. Huang. STREAMCUBE: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream. In *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 1561–1572, 2015.

[14] 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.

[15] P. Guttorp. An introduction to the theory of point processes (D. j. daley and d. vere-jones). *SIAM Review*, 32(1):175–176, 1990.

[16] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 197.

[17] D. He and D. S. P. Jr. 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, Washington, DC, USA, July 25-28, 2010*, pages 443–452, 2010.

[18] 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.

[19] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, pages 769–778, 2012.

[20] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing, Dallas, Texas, USA, May 23-26, 1998*, pages 604–613, 1998.

[21] 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.

[22] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[23] J. Leskovec, L. Backstrom, and J. M. 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, Paris, France, June 28 - July 1, 2009,* pages 497–506, 2009.

[24] C. Li, A. Sun, and A. Datta. Twevent: segment-based event detection from tweets. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 155–164, 2012.

[25] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, USA, June 6-10, 2010,* pages 1155–1158, 2010.

[26] S. Petrovic, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA,* pages 181–189, 2010.

[27] M. Platakis, D. Kotsakos, and D. Gunopulos. Searching for events in the blogosphere. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 1225–1226, 2009.

[28] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 851–860, 2010.

[29] E. Schubert, M. Weiler, and H. Kriegel. Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 871–880, 2014.

[30] Y. Takahashi, T. Utsuro, M. Yoshioka, N. Kando, T. Fukuhara, H. Nakagawa, and Y. Kiyota. Applying a burst model to detect bursty topics in a topic model. In *Advances in Natural Language Processing - 8th International Conference on NLP, JapTAL 2012, Kanazawa, Japan, October 22-24, 2012. Proceedings,* pages 239–249, 2012.

[31] X. Wang, C. Zhai, X. Hu, and R. Sproat. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, pages 784–793, 2007.

[32] J. Weng and B. Lee. Event detection in twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

[33] W. Xie, F. Zhu, J. Jiang, E. Lim, and K. Wang. Topicsketch: Real-time bursty topic detection from twitter. In *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*, pages 837–846, 2013.

[34] Y. Yang, T. Pierce, and J. G. Carbonell. A study of retrospective and on-line event detection. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 28–36, 1998.

[35] H. Yin, B. Cui, H. Lu, Y. Huang, and J. Yao. A unified model for stable and temporal topic detection from social media data. In *29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013*, pages 661–672, 2013.

[36] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. Magnenat-Thalmann. Who, where, when and what: discover spatio-temporal topics for twitter users. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 605–613, 2013.