International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Realizing Emergent Topics in Twitter by Detection of Abnormality Links

Goutami Chenumalla

Assistant Professor, Department of Bachelor of Computer Applications, East West School of Business Management, Yelahanka, Bangalore, Karnataka-560064, India.

Abstract

In the present paper, we demonstrated the probability model which may capture the usual mentioning behavior of an end user consisting of both the number of mentions for every post and the frequency of users occurring in the mentions. Subsequently, to compute the anomaly of future user behavior this model is needed. While using the proposed probability model, we can quantitatively compute the originality or probable effect of a post resembled in the mentioned behavior of the end user. We aggregate all the anomaly scores obtained in this way above countless end users. The efficiency of the proposed method is demonstrated on four data sets we have obtained through Twitter. We demonstrated that the mention-anomaly-based method combined with the TFIDF method can detect the emergence of a new topic at least as quickly as text-anomaly-based counterparts.

Keywords: Twitter, Anomaly Detection, Probability Detection, Burst method

1. Introduction

In recent years, social media platforms have become integral sources of information, enabling users to share ideas, news, and trends rapidly [1]. Among these platforms, Twitter has emerged as a popular microblogging site, facilitating real-time discussions and interactions on diverse topics [2]. However, with the ever-increasing volume of information shared on Twitter, distinguishing valuable and relevant content from potentially harmful or misleading material has become a formidable challenge. In particular, the spread of abnormal or malicious links poses significant risks, ranging from misinformation and phishing attempts to cyber security threats. Particularly, we are enthusiastic about the issues of discovering emerging topics in social networks, which can able to produce automatic “breaking news” or find out invisible current market desires or secret political actions [3]. Social media will able to catch the initial, unedited tone of common people in comparison with traditional media. Thus, the challenge here is usually to identify the emerging topics quickly as soon as possible at a reasonable number of false positives. We have been fascinated by discovering emerging topics from social network streams depending on monitoring the mentioned behavior of end users. The fundamental supposition is that a new (emerging) topic is anything individuals seem to like discussing, commenting on, or forwarding information even more to their friends. Traditional methods of topic detection have primarily been focused on the frequencies of (textual) words or phrases [4-5]. A term-frequency-based method might experience ambiguity induced using synonyms or homonyms. Perhaps it will furthermore need complicated preprocessing (e.g., segmentation) based on the target language. In addition, it cannot be employed when the contents of the messages are generally non-textual data. However, the “words” produced by mentions

are generally unique, demand small preprocessing to acquire (the data is normally segregated from the contents), and are readily available regardless of the nature of the contents [6]. In the present work, we mean by mentioning links between different end users of the same social network (like Twitter) in the form of reply-to, message-to, retweet-of, or clearly in text messages. A single post might consist of many mentions. Several users can hardly include mentions in their posts; in addition, some other users might be mentioning their buddies at all times. Many users (like celebrities) might receive mentions each minute; for some other users, staying mentioned might be an unusual situation. In particular, mention is similar to a language having the number of terms equal to the number of end users in a social network.

2. Objectives

The objective of this work is to build up a method for modeling bursts in such a way that they can be strongly and proficiently recognized. The aim is to develop a conventional method for modeling such "bursts," to the extent that they can be robustly and efficiently identified. The idea of "bursts" in document streams. An essential trouble with text data mining is extracting meaningful structure from document streams that arrive continuously as time passes. E-mail and news articles are two natural samples of this sort of stream, each characterized by topics that appear, rise in intensity for some period, and then disappear. The appearance of a topic in a document stream is signaled by a "burst of activity," with particular characteristics increasing sharply in frequency as the topic emerges.

In this paper, we recommend

1) The probability model [7] can capture the usual mentioning behavior of an end user, which consists of both the number of mentions for every post and the frequency of users occurring in the mentions. And then, we can use this model to calculate the anomaly of future user behavior.

2) The learned probability distribution to determine an anomaly score for each and every post.

3) Kleinberg's burst detection method along with the TF-IDF (Term Frequency-Inverse Document Frequency) method [8] determines the frequency of a topic in a given period. This method detects whether the interval of the arriving messages is denser than that in a normal condition through comparison with other document streams.

3. Methodology

In this work, we suggest and experimentally estimated the probability model that can capture the usual mentioning behavior of an end user, which consists of both the number of mentions for every post and the frequency of users occurring in the mentions. And then, we can use this model to calculate the anomaly of future user behavior. The data comes from a social network service via some API. For every single new post, we make use of samples from the past time interval associated with length T with regard to the corresponding user for training the mentioned model. We determine an anomaly score for each and every post using the learned probability distribution. This score will be then aggregated around users and additionally provided by Kleinberg's burst detection method along with the TF-IDF method. The entire flow of the proposed method is demonstrated in the following Figure 1.

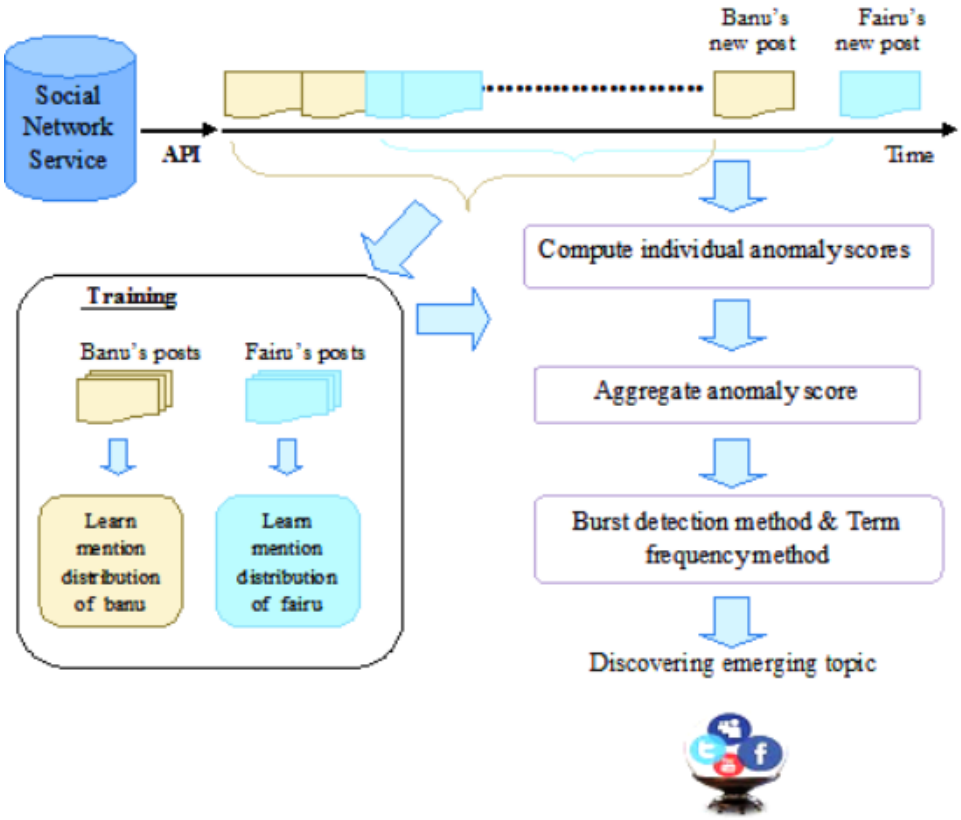

Figure 1. The flow diagram of the proposed method

3.1. Calculating Link-Anomaly Score

A single post *x* within Twitter is characterized by means of the number of mentions *k* it has, as well as the set *V* of names (IDs) of the mentioned. The deviation of the user behavior is calculated based on the anomaly score. The anomaly score is calculated for each post in the social data. For calculating the anomaly, the joint probability distribution and geometric distribution are used.

The anomaly score for a post *x= (t, u, k, V)* by user *u* at time *t* containing *k* mentions to user *V*, is calculated by using the formula:

$$S(x) = -log\left(\left(P(k|T_u^{(t)}\right)\prod_{v\in V}\left(P(v|T_u^{(t)}\right)\right) \quad (1)$$

$$= -logP\left(k|T_u^{(t)}\right) - \sum_{v\in V}\log P\left(v|T_u^{(t)}\right) \quad (2)$$

Here $T_u^{(t)}$ is the training set which is a collection of posts by user *u* in the time period [*t-T, t*].

The both the terms in the previously mentioned equation is usually calculated with the predictive distribution of the number of mentions and the predictive distribution of mentionee.

3.2. Burst detection method

Further, we use a burst-detection method to determine the frequency of a topic in a given period. The method has been proposed in. This method detects whether the interval of the arriving messages is denser than that in a normal condition through comparison with other document streams.

The burst-detection method defines a probabilistic automaton (*A*) consisting of two states: (1) whenever *A* is in state q_0, messages arrive at a slower rate. (2) Whenever *A* is in state q_1, messages arrive at a faster rate. The period (T) is defined as the interval between the arrival of the first message and that

of the last message, $n + 1$. If the arrival time is random, a gap time x between the messages i and $i + 1$ follows an exponential distribution. According to Poisson distribution, in state q_0, the probability of arrival of the next message at interval x is $f_0(x) = \alpha_0 e^{-\alpha 0 x}$, where $\alpha_0 = n/T$. In state q_1, the gap time is shorter than in state q_0. Consequently, the probability of interval x between any two consecutive messages is $f_1(x) = \alpha_1 e^{-\alpha 1 x}$, where $\alpha_1 > \alpha_0$. In addition, we determine a given number of n messages with a particular arrival time as inner arrival gaps $x = (x_1, x_2 \ldots\ldots, x_n)$, where $x_i > 0$. Similarly, we set the conditional probability of a state sequence $q = (q_{i1}, q_{i2} \ldots\ldots, q_{in})$. Each state sequence q derives a density function f over sequences of gaps, which is represented by the following formula.

$$f_q(x_1, \ldots x_n) = \prod_{t=1}^{n} f_{i_t}(x_t) \qquad (3)$$

3.3. TF-IDF method:

TF-IDF is short for term frequency-inverse document frequency, and the TF-IDF weight is a statistical measure helpful to estimate how essential a word is to a document in a collection or corpus.

$$\text{TF-IDF}(w, t, d) = \text{tf}(w, t) * \text{idf}(t, d) \qquad (4)$$

$$\text{idf} = (t, d) = \log \frac{T}{|\{t \in d : w \in t\}|} \qquad (5)$$

The equation of TF-IDF for the word w is defined in (1). In (1) t represents one tweet and d stands for a document, which would be the corpus of tweets in our work. $tf(w,t)$ is the term-frequency, $idf(t,d)$ is defined in(2), in which T denotes the total number of tweets in document d.

3.4. Clustering Procedure for Proposed Method

1) Select topic $t_i \in$ TOPIC, where i=1…..n
2) For each topic t_i, sort the link-messages, images and videos in ascending order of their score.
3) Score= Anomaly score + score generated by retweets, mentions and forwards, for the categories of messages, link-messages, images and videos.
4) We combine these score to generate one single list called EMERGING TOPIC list.

The proposed method may be useful to the both the cases where topics are concerned with information like texts and the non-textual information like images, video, and so on.

4. System Design

4.1 Use Case Diagram

A use case diagram as shown in Figure 2 is a kind of behavioral diagram distinct by and produced from a Use-case study. Its principle is to present a graphical impression of the functionality provided by a system in conditions of actors, their objectives (represented as use cases), and any dependencies among those use cases.

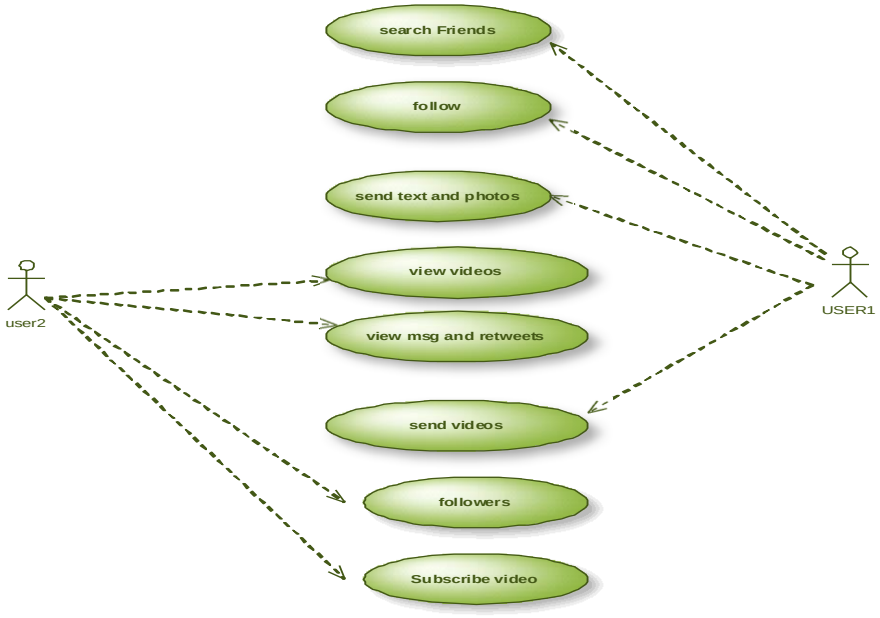

Figure 2. Use case Diagram

4.2. Data flow diagram

The DFD (Data Flow Diagram) is also called as bubble chart (Figure 3). It is a simple graphical formalism that can be used to represent a system in terms of the input data to the system, various processing carried out on these data, and the output data is generated by the system.

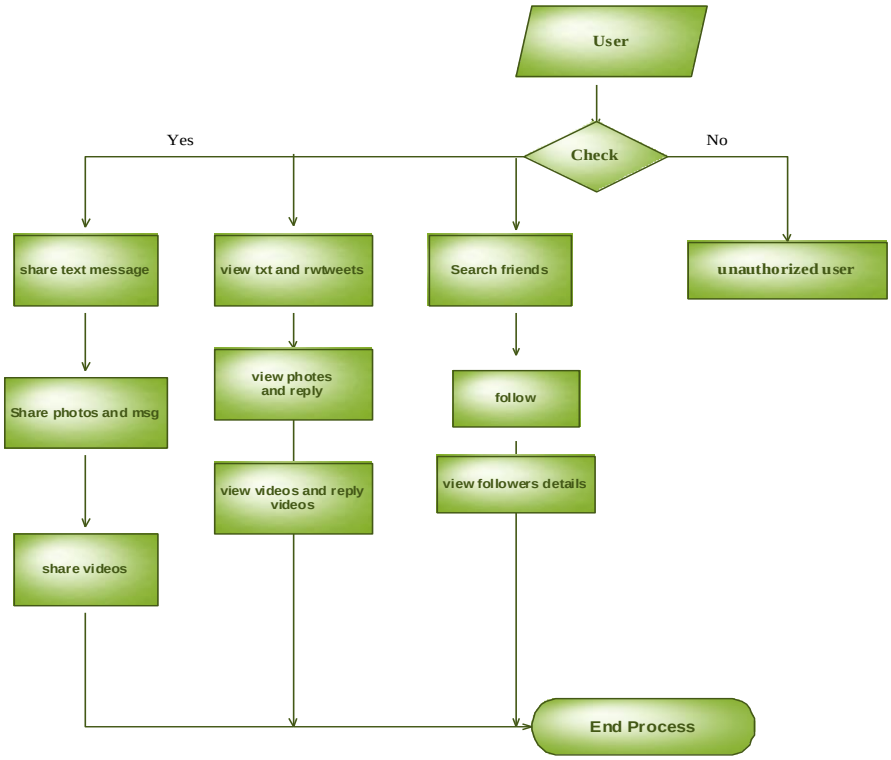

Figure 3. Data Flow Diagram

5. Results and Discussion

5.1. User module

Twitter will provide the following options for a user. Twitter user will undergo one of the following options definitely.

5.1.1. Registration

To create a Twitter account, we need to follow the instruction as given below:

1. Enter your full details of the signup page.
2. Enter your phone number when prompted.
3. Twitter will send a verification code to your phone. When you get that text message, enter the code.
4. You'll then be asked to create an account by entering your email address, a password, and a name for your account. Your username will already be entered, since you created this when you signed up via SMS.
5. Click Create my account as shown below, and you're all done! Twitter will walk you through finding some friends you may know on Twitter and then will direct you to your home page.

5.1.2. Login

Twitter has evolved from a microblogging service into a popular social messaging platform; it has been instrumental in providing the "pulse" on news and events across the globe. In addition to its widespread acceptance among the news media and entertainment industry, Twitter has also become a popular social media marketing tool and a great way to communicate with both friends and co-workers. It is many different things to many different people. It can be used by a family to keep in touch, a company to coordinate business, the media to keep people informed, or a writer to build up a fan base. Twitter is micro-blogging. It is social messaging. It is an event coordinator, a business tool, a news reporting service, and a marketing utility.

5.1.3. Follow

Twitter can be a great way to promote you, your brand or your company. Actors, writers, sports stars and others are turning to Twitter as a way to connect with fans and promote themselves to millions of people across the globe. Twitter is also a great way for bloggers to gain more traction in the blogosphere and get noticed.

But how do you build up a Twitter following? These simple tips will help you get started.

- **Follow People:** The number one way to gain followers is to follow people. Many Twitter users follow anyone that follows them, while others check to see if the profile is active before deciding to follow it. You can find interesting people to follow by doing a twitter search.
- **Be Active:** Don't just link your web feed to your Twitter profile. Post messages and make your Twitter page a place to connect with your readers.
- **Promote your Profile:** Add a link to your Twitter profile to your blog, website and as part of your signature in email messages and discussion forums.
- **Respond to Messages:** Always pay attention to who mentions you. If you aren't following them, do so. And when you get a direct message, make sure you respond.

5.1.4. Followers

Several companies have sprung up offering to help people and companies grow their Twitter following the easy way by forking over a little cash. The usefulness of a large Twitter following isn't simply a number. Having 10,000 followers won't do you much good if only half of them are real people and most of them aren't interested in anything you tweet. It may change the number of followers at the top of the screen, but it will have little effect on how many people follow what you say. The entire point of growing a large Twitter following is to communicate with people who are interested in your topic, your company, or your brand. They want to read what you write, retweet what you tweet, and check out the links you post. This is what makes a Twitter following valuable. And a Twitter following that simply ignores your tweets simply isn't worth it.

5.1.5. Tweets

A tweet is a post or status update on Twitter, a micro-blogging service. Because Twitter only allows messages of 140 characters or less, "tweet" is as much a play on the size of the message as it is on the audible similarity to Twitter.

5.1.6. Retweets

A "retweet" is a reply to a tweet that includes the original message or a tweet that includes a link to a news article or blog post that you find particularly interesting. Like hashtags, re-tweets are a recent community-driven phenomenon on Twitter to make the service better and allow people to follow discussions easier.

5.2. Admin module

Admin will view all Twitter user details like email id, phone number, profile photo, username, and the captcha entered while the registration. And can view the topics list that the users discussed for example YouTube, BBC News, Job hunting, and NASA. And can view the comparison chart for the topics the user had discussed and will discover the emerging topic from this chart.

5.3. Emerging topic module

This module will list all the topics that all Twitter users had discussed on Twitter. And will show three different lists of emerging topics based on link messages, images, and videos separately and one final list by combining scores collected from these three lists.

The following tests have been successfully performed through our proposed model

1. Admin login into the account by entering username and password. The admin can view the emerging topics list under the View Emerging Topic, view the comparison chart under View Chart and view all Twitter user details under View User Details.
2. Admin can view the emerging topics list for the various topics that the Twitter users have discussed. In the above figure, three different lists are shown separately by considering link messages, images, and videos individually. And one final list will be shown by combining the scores of individual topics from these three lists.
3. Admin can view all the details of the Twitter user. The details consist of User id, Name, Password, Date of Birth, Email, Phone number, and captcha entered while registering.

4. The comparison chart for the emerging topics is shown in the above figure. From the figure, it is clearly shown that the most discussed topic (emerging topic) is BBC News which is the first on the list.

6. Conclusions

We have recommended the latest method to discover emerging topics on Twitter. The probability model does not rely on the textual contents of Twitter posts; it may be useful to apply to the scenario wherever topics are concerned with information other than texts, like images, video, audio, and so on. The Burst Detection method is used to pinpoint the emergence of a topic. The TF-IDF method is used when the contents of the posts include information like texts only. Individually mention-anomaly model (probability model) and term-frequency-based methods are not going to instantly notify what exactly the anomaly is. The Combination of the mention-anomaly model with the TF-IDF method will get an advantage both from the overall performance of the probability model and the intuitiveness of the text-frequency-based method. The proposed method is applied to four real datasets (NASA, BBC, YouTube, and Job Hunting) that we have obtained from Twitter. In all of these datasets, our proposed method exhibited encouraging performance.

7. Acknowledgement

The author, Goutami Chenumalla would like to express sincere gratitude to Mrs.S.Reddy Mubaraq, for her invaluable guidance and support throughout the course of this research.

8. Authors' Biography

Mrs. Goutami Chenumalla is currently working as an Assistant Professor in the Department of Bachelor of Computer Applications, East West School of Business Management, Yelahanka, Bangalore. Also, she is pursuing her Ph.D. in Computer Science and Engineering from Presidency University, Bangalore. She obtained her Master of Technology in Computer Science and Engineering from JNTU Anantapur.

9. References

1. Kapoor, K.K, Tamilmani, K, Rana, N.P, Pushp. P, Yogesh. K.D, Sridhar. N, “Advances in Social Media Research: Past, Present and Future,” Information System Frontiers, 2018, 20, 531–558.
2. Malik, A, Heyman. S.C, Johri. A, “Use of Twitter across educational settings: a review of the literature,” International Journal of Educational Technology in Higher Education, 2019, 16, 36.
3. Tintomon. P.A, N. Santhana. K, “Discovering emerging topics in social streams via link anomaly detection,” Journal of Computer Science and Engineering, 2006, 2(1), 39-44.
4. James. A, Jaime. C, George. D, Jonathan. Y, Yiming. Y “Topic Detection and Tracking Pilot Study Final Report,” Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.
5. Kleinberg. J, “Bursty and Hierarchical Structure in Streams,” Data Mining Knowledge Discovery, 2003, 7(4), 373-397.
6. Roman. E, Joanne. Y, “A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BER Topic to Demystify Twitter Posts,” Frontiers in Sociology, 2022, 7, 886498.
7. Takahashi, Toshimitsu, Ryota. T, Kenji. Y. "Discovering emerging topics in social streams via link-anomaly detection." IEEE Transactions on Knowledge and Data Engineering, (2012) 26 (1), 120-130.
8. Zhang. W, Taketoshi. Y, Xijin. T. "A comparative study of TF* IDF, LSI and multi-words for text classification." Expert systems with applications, 2011, 38(3), 2758-2765.