ICIC Express Letters
Part B: Applications
Volume **12**, Number **9**, September **2021**

ICIC International ©2021 ISSN 2185-2766
pp. **789–796**

SENTIMENT PIVOT APPROACH FOR EVENT DETECTION FROM TWITTER STREAM

K. Victor Rajan[1,*], V. Rhymend Uthariaraj[1] and Kolin Paul[2]

[1]Department of Computer Science and Engineering
College of Engineering Guindy
Anna University
Chennai 600025, India
*Corresponding author: victor@jts.co.in

[2]Department of Computer Science and Engineering
Indian Institute of Technology
Delhi 110016, India

Received December 2020; accepted March 2021

Abstract. *Gathering public opinion for decision making is normally conducted through traditional polling and surveys. With the increased use of technology and social media, people respond to government policies, election results and other public events through micro blogging sites like Twitter. Often these messages reflect the emotion, opinion and sentiment of the public towards an event. If we device a method to measure the momentum of public opinion along with sentiment associated with the text, this can be used to identify events of potential impact to decision makers like Government agencies, and Commercial organizations. In this paper, we propose a probabilistic model to analyze short messages posted by users in Twitter and predict emerging hot topics and events. The sentiment of the people in their messages can continue to progress in a given direction (positive or negative) or continue to be neutral. We combine the trajectory of people's sentiment along with the momentum of messages to identify high impact events. These high impact events originating from a particular region can be used by online marketing companies, government agencies for decision making. Our experiment results showed prediction accuracy as high as 80% and captured important large-scale events. This highlights the potential of text streams as a substitute and supplement for traditional polling.*

Keywords: Social media, Twitter analytics, Sentiment polarity, Trending hashtags

1. **Introduction.** Event detection from social media would help us to understand the public opinion/sentiment with respect to the social events, making it possible for a company/organization to make fast response to any emerging crisis. If we need to find out the public opinion about a government policy or a newly launched consumer good, the traditional approach is to conduct a random poll of the public. A traditional opinion poll would involve surveys with many people traveling, making phone calls etc., resulting in thousands of dollars to run. On the other hand, people are pro-active in expressing their opinions, sentiments and stories in social media forums like Twitter and it gives us an opportunity to gather and analyze public opinion without conducting a manual opinion poll. With the dramatic increase of social media usage nowadays, we can use the opportunity to analyze the text rich blogs and messages in a faster and cost effective way than traditional polls. Social media also gives the facility to collect messages pertaining to a specific region, language, a historical period, etc. Online advertisers could use this analysis for efficiently targeted marketing campaigns. Government organizations can know how the society is influenced by a decision or policy and then determine how to respond

DOI: 10.24507/icicelb.12.09.789

to the public opinion. A high impact event is defined as a story about which several news articles, tweets are written during short duration. Similar to relevance, the definition of what impact an event may have is highly subjective and open to multiple interpretations. Extracting the public opinion from social media text is still a challenging task due to the complexity of natural language processing of tweet language. The unique characteristics of tweets are short and noisy content, diverse and fast changing topics, large data volume, etc. The research on social media analysis is still evolving. Without getting into the complexities of understanding the language, we first try to extract trending hashtags in Twitter. Then we measure their momentum and sentiment polarity at regular intervals. This is used to predict whether such hashtags will continue to be discussed in the near future. The events and topics associated with the high-impact hashtags are identified and presented to decision makers.

Impact of sentiment on public events. We analyzed tweets related to a region (Kashmir) for two months (July and Aug. 2020). After careful analysis, we found that among the tweets generated in a day (500 million), 14% have hashtag along with sentiment. We use popular hashtags and sentiment polarity to identify high-impact events. The main contributions of our work can be summarized as follows.

1) We detect bursty events that are gaining momentum and will continue to be hot in the near future.

2) We propose a spread model based on the analysis of both event growth and users sentiment. The major advantage of our model is that it distinguishes users' contributions according to sentiment.

3) Our predictions when compared with the newspaper topics showed that our approach is really fruitful. We evaluated our approach in real-life data sets.

2. **Related Work.** A simple statistics in our study shows that the number of distinct bursty segments is about 75% of the number of distinct tweets in a randomly chosen time window. Among the bursty segments detected, many contain misspelled words and informal abbreviations. We therefore source for the wisdom of the crowds to filter the hashtags. In this section, we present an overview of the previous works done in the field of event detection and extraction. Previous research work can mainly be grouped under two topics namely topology-based approach [4] and feature pivot approach [5]. The topology-based approaches are not capable of analyzing the content of the propagated information. Feature-pivot method combines the social relations of users and the frequency distribution of words to detect burst words. However, it does not weigh the sentiment of the user. Recently, event detection on Twitter stream becomes a hot research topic. Mathioudakis and Koudas [6] proposed a trend detection system over Twitter stream by identifying bursty keywords. A trend is identified as a set of bursty keywords that occur frequently together in tweets. Petrović et al. [7] proposed First Story Detection (FSD) by applying Locality Sensitive Hashing (LSH). Popescu et al. [8] proposed a method for entity-based event detection on Twitter streams. Set of tweets containing the predefined target entity are processed and machine learning techniques are used to predict events of interest. Li et al. [9] proposed to detect Crime and Disaster related Events (CDE) from tweets. They used conventional text mining techniques to extract information. To summarize, most existing approaches for detecting events from tweets are applicable to certain types of tweets (e.g., having a specific hashtag, containing a predefined entity, or related to crime and disaster). Instead of solely predicting the popularity of a detected event, we model the spread of an event by combining the use of the hashtags with the sentiment of infected users. An event having frequent sentiment oscillations (negative during one interval and positive during another interval) may not have much impact on public compared to an event having high negative sentiment all the time. The proposed approach utilizes a

linear spread prediction function to predict the future popularity of the events. The linear function combines the rate of growth with sentiment polarity to identify the growth of the event. We identify events with upward momentum with positive or negative sentiment polarity as high-impact events. To combine the above factors, the growth of an event is assumed to be a linear function of the volume at regular intervals combined with a local weight of sentiment polarity during the interval. The main contributions of this paper can be summarized as follows.

1) We devise a three state model to detect bursty event by combining hashtag volume and users sentiment information to deal with real life situation.

2) Our model distinguishes users' contributions according to sentiment which is different from other approaches that use the rate of growth or user influence.

After ranking the hashtags by their momentum, we then retain the top-k hashtags as potential event-related hashtags for further processing.

3. Event Detection Using Sentiment.

Our objective is to identify events which are closely related to the sentiment of social media users, mainly Twitter. We follow a two-step approach. The first step identifies the hashtags which are exhibiting sudden growth within a given time period with consistent sentiment polarity. The second step involves analyzing events associated with these bursty hashtags that may have high impact in the society.

3.1. Bursty hashtag identification.

Consider a tweet stream $T = \{tw_1, tw_2, \ldots\}$ from a geographical region G, where tw_i is a tweet. We divide the tweet stream T, into i non-overlapping time windows, w_j of the same length. Each tweet may contain a set of hashtags $H = \{ht_1, ht_2, \ldots\}$. If a tweet tw_i has user's sentiment, then we assign the sentiment polarity to the tweet as positive or negative or neutral. The sentiment polarity of a hashtag ht_i during a window w_j is the average sentiment polarity of all tweets containing hashtag ht_i during the window w_j. We define the event detection state machine M as a quintuple $M = \{T, H, W, P, S\}$ where

$T = \{tw_1, tw_2, \ldots, tw_n\}$ is a set of tweets

$H = \{ht_1, ht_2, \ldots, ht_m\}$ is a set of hashtags

$W = \{w_1, w_2, \ldots, w_l\}$ is a set of non-overlapping time windows

$P = \{p_j(ht_i), 0 \leq i \leq m, 0 \leq j \leq l\}$ is a set of sentiment polarities and

$S = \{q_r, q_i, q_d\}$ is a set of states.

q_r is the state corresponding to hashtag growth along with sentiment polarity retention,

q_i is the state corresponding to hashtag growth along with sentiment polarity inversion and

q_d is the state corresponding to hashtag decline (polarity change may or may not happen).

The tweets are not generated at regular rate in a day. People will be active during the daytime and more tweets will be generated during the day compared to night. We normalize the occurrences of hashtag ht_i for every time window as follows:

$$n(ht_i) = tw(ht_i) * 100/T$$

where $tw(ht_i)$ is the number of tweets containing hashtag ht_i during the window and T is the total tweets generated during the window i.

We model a tri-state state machine as shown in Figure 1 to analyze the hashtag growth/decline.

This state machine accepts sequences $\{q_d(q_r|q_i) * q_d\}$. The event detection is to identify the set of state transitions where $n(q_r) > n(q_i)$, i.e, we identify hashtags which have the state transition sequence like $\{q_r q_i q_r q_r q_i q_d\}$ where the sentiment polarity inversion is minimal. In general, the rate of arrival of a hashtag is very 'rugged': it does not typically rise smoothly to a crescendo and then fall away, but rather exhibits frequent alternations

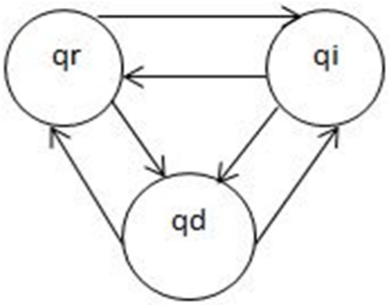

FIGURE 1. State machine for event detection

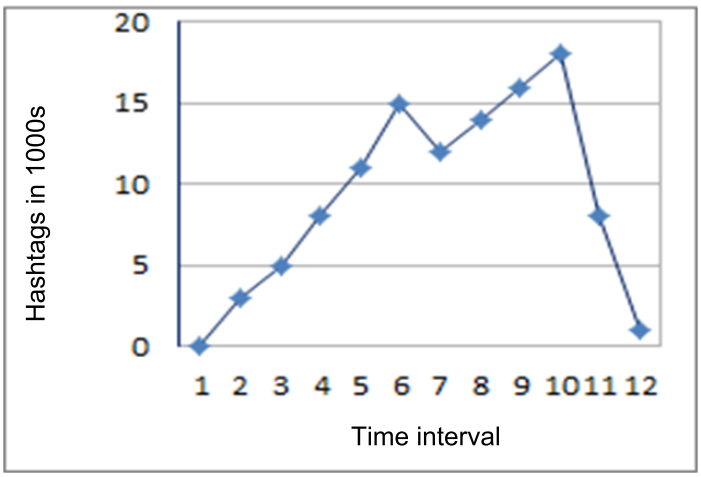

FIGURE 2. Arrival of hashtags

of rapid flurries and longer pauses in close proximity. It reaches to a maximum and falls smoothly or suddenly fades away as shown in Figure 2.

For a given window w_i from Twitter stream, let r_i be the number of tweets containing hashtag ht. Then, the probability of transition to state q_r can be calculated using binomial distribution $P(q_r, ht) = \begin{pmatrix} n_i \\ r_i \end{pmatrix} p^{r_i}(1-p)^{(n_i - r_i)}$ where p is the expected probability of tweets containing hashtag ht with polarity retention. Using Bayes' theorem, we get $p = P(ht|E_r) = P(ht \cap E_r)/P(E_r)$ where E_r denotes the event corresponding to polarity retention (no polarity direction change during the interval w_i).

Considering the large volume of tweets published at any time, it is reasonable to approximate this to normal distribution. However, the curve is not perfectly bell-shaped (as shown in Figure 2) due to external factors affecting an event and sudden fall of tweets (for example, end of football match). It is not possible to predict the growth of hashtags using standard probability models.

3.2. **Sentiment polarity.** The random behavior of hashtag growth combined with users' sentiment makes the decision making process a complex activity. In this paper, 'sentiment polarity' takes on a specific meaning, that is, the net of positive and negative opinion expressed about an event. We derive day-to-day sentiment scores by counting positive and negative messages. Positive and negative words are defined by the subjectivity lexicon from Opinion Finder, a word list containing about 2000 and 4700 words marked as positive and negative, respectively. A tweet message is defined as positive if it contains any positive word, and negative if it contains any negative word. This gives primitive but efficient results since Twitter messages are so short (about 140 characters). We define the net sentiment polarity score (S_p) of a tweet as the difference between positive and negative

words. This in turn is converted as ratio to the sum to get a rational value between -1 and 1. Hence,

$$S_p = (W_p - W_n)/(W_p + W_n)$$

where W_p is the number of positive words in the tweet and W_n is the number of negative words in the tweet.

3.3. **Markov process.** Our objective is to identify events which remain hot and retain the direction of sentiment polarity at every interval. A hot discussion on Cricket match might suddenly stop if the match is canceled due to rain or bad weather. Since the system changes randomly, it is generally impossible to predict with certainty of an event from tweets using standard linear statistical models. After careful analysis of the tweet behavior, we decided to use Markov chain to predict the future. The sentiment polarity of Twitter users resembles a Markov chain called 'drunkard's walk', a random walk on the number line. At each tweet, the sentiment polarity may change between positive and negative with equal probability. We model the Twitter event burst as Markov chain. Following Markov chain on a countable finite state space represents our stochastic prediction model.

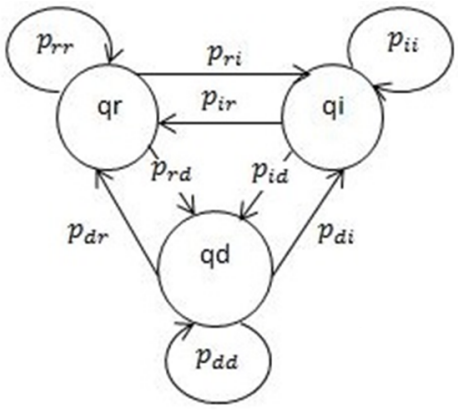

FIGURE 3. Transition probabilities of the state machine

When M is in state q_r, hashtags exhibit growth and polarity retention.
When M is in state q_i, hashtags exhibit growth and polarity inversion.
When M is in state q_d, hashtags are declining.
The transition matrix Mt for stage t of Markov chain is given by

$$Mt = \begin{pmatrix} p_{rr} & p_{ri} & p_{rd} \\ p_{ir} & p_{ii} & p_{id} \\ p_{dr} & p_{di} & p_{dd} \end{pmatrix}$$

The probability for transition from state q_r to q_r denoted by (p_{rr}) is given by the formula

$$p_{rr} = \sum_{i=1}^{n} Pi(ht|E_r)/n$$

The probability of a hashtag ht to be burst after n intervals is predicted using Markov chain

$$Mt^{(n)} = (100)Mt^{(n-1)}$$

However, along with the Markovian transitions if the full history of the previous transitions is taken into account for prediction, it will provide powerful clues about the likely next stage. We combine learning with Markov transition and develop an Additive Learning Markov chain. In this process, at every interval a deviation of estimated $E(p_{rr})$ from

the actual $A(p_{rr})$ probability is calculated. We add a correction factor to p_{rr} at every step to smoothen the transition probability.

$$Et(p_{rr}) = \sum_{i=1}^{n} \{Ai(p_{rr}) - EiA(p_{rr})\}/(t-1)$$

The correction factor is the average of deviations between estimated and actual probabilities of state transitions of all previous intervals assuming that we have training data set with history of n intervals. The learned probability will not in general be different from the empirical probability, as the model might choose to converge. As the learning continues, the estimated and actual probabilities converge to a common value when we have sufficiently enough data to train. The steady state transition matrix using the Markov process is used to predict values for future intervals.

4. **Experiment and Results.** In this section, we describe the evaluation tasks, the data sets used and the experimental results of the proposed approach.

Evaluation Tasks:

i) We evaluate our approach on Twitter data set.

ii) We evaluate our approach on several tasks.

Our goal is to prove the real life application of our approach. We aim

• To evaluate the quality of bursty hashtags identification. The empirical probability is compared with real-time data to harness the accuracy. The results show promising output after weeding out co-occurrences and collision of events from multiple hashtags.

• To evaluate the correctness of story finding where event embedding is translated into social activity.

4.1. **Data set.** Tweets are generated by people around the world and this text rich social media platform serves as a desirable platform to study the spreading information from many perspectives like politics, elections, consumer products and many more. Twitter messages are short in size. They spread very fast but short lived. For example, many users would discuss about a football match during the match or within few hours right after the match but not for many days after the match. Our data set for evaluation was set of tweets from July 2020 to August 2020 related to Kashmir geographic region. After prediction of events, we verified the occurrence of such events by doing a lookup in newspaper passages.

Table 1. Availability of hashtags in Twitter stream

Description	July 2020	August 2020
Total number of tweets	1,429,669	1,399,320
Tweets with hashtag	467,402 (33%)	405,664 (29%)
Tweets with hashtag and sentiment	198,533 (14%)	183,286 (13%)

From the above table, we observe that appreciable number of hashtags with sentiment is available for analysis.

4.2. **Results.** Figure 4 shows the spread of tweets and hashtags on a particular day (17 July 2020), the time window (wt) being 2 hours.

The linear relationship between tweets and hashtags ($ht = 0.33 * tw$) is a good indicator that prediction based on hashtags is good representation of tweets.

Figure 5 shows the popular hashtags during a period of 24 hours. Analysis shows that the tweets with sentiment are more than neutral tweets for any hashtag. Majority of hashtags having sentiment is an indication that people always express their opinions and emotions in tweets. Sentiment based prediction is an effective approach to detecting events.

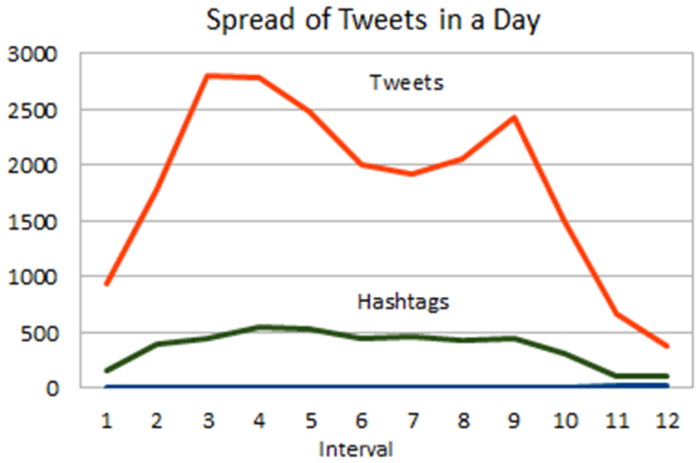

FIGURE 4. Tweet and hashtag volume distribution

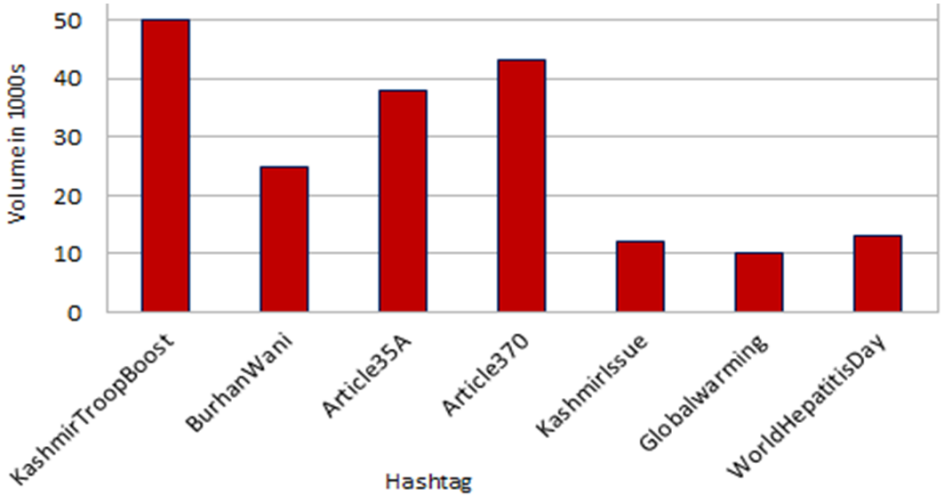

FIGURE 5. Popular hashtags in a day

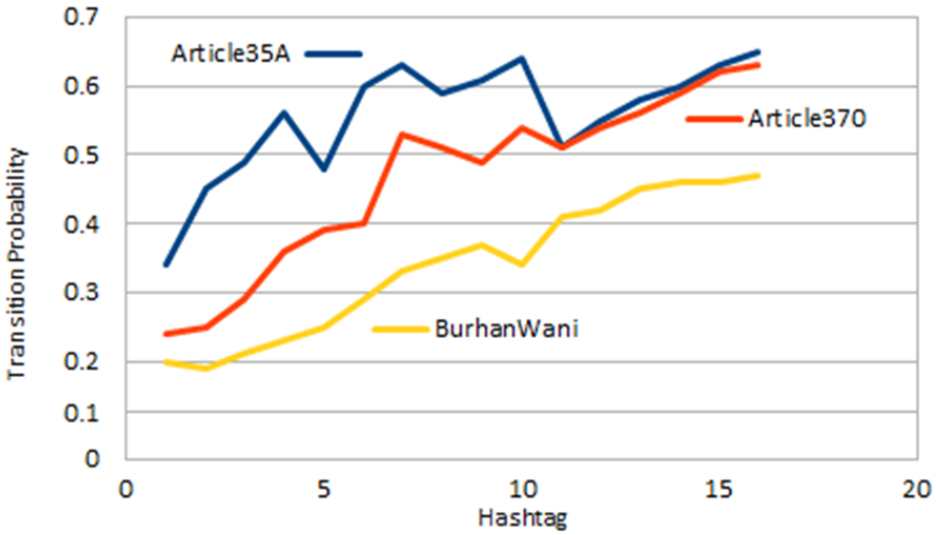

FIGURE 6. Bursty hashtags identified using Markov process

After identifying the hashtags which are expected to grow in the next few intervals, we use text searching to detect events around them. All tweets containing a hot hashtag are combined into a news incident. After grouping similar messages, a query finder is used to find events among tweets. Following is a set of events identified.

Table 2. Events identified from Twitter stream

Hashtag	Sentiment	Event identified
#BurhanWani	Negative	1) Pay tribute to the martyrs of Kashmir 2) Mark the martyrdom anniversary 3) Shutdown call by separatists
#KashmirTroopBoost	Negative	1) Fresh Troops Sparks fear in Kashmir 2) War like situation
#Article35A #Article370	Negative	1) Panic among people 2) Democracy under detention
#KashmirUnderCurfew	Negative	1) No food, no medicine and no communication 2) Torture everywhere

From the above table, we observe that people talk about various events and express their sentiment in social media. This is really an alternate to traditional polling and cost effective solution for decision makers to understand the situation and respond to any emerging crisis.

5. **Conclusion and Future Work.** We discussed an approach based on sentiment to identify high impact events in Twitter. Our Markov based approach identifies events of interest using stochastic process. The result set can further be used to train a model under unsupervised learning. The data set from our Markov prediction combined with machine learning will definitely be a good event prediction system for the digital world. A machine learning system trained with our events from Twitter can also be used to identify topics of interest in other social media like Facebook, and Instagram.

REFERENCES

[1] A. Iriani, Hendry, D. H. F. Manongga and R.-C. Chen, Mining public opinion on radicalism in social media via sentiment analysis, *International Journal of Innovative Computing, Information and Control*, vol.16, no.5, pp.1787-1800, 2020.

[2] S. Phuvipadawat and T. Murata, Breaking news detection and tracking in Twitter, *WI-IAT*, pp.120-123, 2010.

[3] G. Pui, C. Fung, J. Xu, Y. Philip, S. Yu and H. Lu, Parameter free bursty events detection in text streams, *Proc. of the 31st VLDB Conference*, Trondheim, Norway, 2005.

[4] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. T. Archibald and X. Liu, Learning approaches for detecting and tracking news events, *IEEE Intelligent Systems*, vol.14, no.4, pp.32-43, 1999.

[5] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Hafner Pub. Co., 1949.

[6] P. M. Mathioudakis and N. Koudas, Twitter monitor: Trend detection over the Twitter stream, *SIGMOD*, pp.1155-1158, 2010.

[7] S. Petrović, M. Osborne and V. Lavrenko, Streaming first story detection with application to Twitter, *HLT-NAACL*, pp.181-189, 2010.

[8] A.-M. Popescu, M. Pennacchiotti and D. Paranjpe, Extracting events and event descriptions from Twitter, *WWW*, pp.105-106, 2011.

[9] R. Li, K. H. Lei, R. Khadiwala and K. C. Chang, TEDAS: A Twitter-based event detection and analysis system, *2012 IEEE 28th International Conference on Data Engineering*, pp.1273-1276, 2012.

[10] A. Agarwal, B. Xie et al., Sentiment analysis of Twitter data, *Proc. of the Workshop on Language in Social Media (LSM2011)*, Portland, Oregon, pp.30-38, 2011.