DETECTING TRENDING EVENT TOPICS USING SENTIMENT DRIVEN DERIVATIVES METHOD ON TWITTER

Poonam Vijay Tijare

Research Scholar, Department of Computer Science and Engineering (VTU RC),
CMR Institute of Technology, Bangaluru, Karnataka, India
poonamtijare@gmail.com

Jhansi Rani P.

Professor, Department of Computer Science and Engineering (VTU RC),
CMR Institute of Technology, Bangaluru, Karnataka, India
jhansirani.p@gmail.com

Abstract

In the era of social media, Twitter has gained a lot of attraction for depicting public opinion about various activities in the world. There is a close correlation between news, the public opinion and behavior on social media platforms; all of these activities are mutually dependent on each other. The analysis of this influence can provide perception into events happening around. The research article presents a novel framework to detect an event topic on a social media platform independent of the domain. The proposed framework uses sentiment analysis, K-means and a second order derivative algorithm to find trending event topics. The derivatives on the timeline help in amplifying the trends.

The proposed work is tested on four datasets with different domains like Scientific, Social, Sports and Political. The accuracy of the proposed framework turned out in the range of 88% to 95% showing affirmative results. The proposed framework can be further extended for streaming data analysis to predict trending events in real time.

Keywords: Twitter; Sentiment analysis; K-means; Derivative; Trending events.

1. Introduction

The social media (SM) platforms are flooded by events and discussions related to those events. The news pops up and discussion begins with online platforms and within a few seconds, the media gets piled up with public reactions. There is a very close relationship between people mentioning their opinion and events happening in reality, sometimes, the public stance can start the event at inception. Due to the wide reachability of SM, that event can turn into reality. Twitter had a total of 365 million active user accounts in the first quarter of 2020 that explains the velocity and variety of users. Out of every 10 users, 6 users of Twitter are in the age group between 35 and 65 years, which holds for nearly 63% of the total user base reflecting the maturity of the audience (Chen, 2019).

The humongous information generated on SM gives lots of inputs about social behaviour. Sentiment analysis on these texts helps in digging out hidden details of the opinions of people. The use of sentiment analysis is applied in the area of event analytics. Events or news around us triggers discussion. The discussions further can lead to advancements in that event, in a positive or negative direction. These advancements through Twitter discussions can be analysed to understand upcoming alarming situations. Adverse situations can be better managed by authorities if this information is used properly. Research happening in this area is mostly proposed for a single type of events such as natural disaster, traffic, or election debates. There are hardly any labelled benchmark datasets available in this area. This brings challenges in testing the proposed models in the event detection domain (Bondielli, 2019) (Zhou,2020).

We are proposing a model based on sentiment derivatives that can amplify even small trends and detect events. The novel framework was tested on four diverse domains like scientific events, political events, social events, and sports events that happened in India. The events happened in the outline of January 2019 to June 2019. The tweets are extracted using the Twitter API. The proposed model works in the following phases:

(1) Apply the K-means algorithm to the sentiment of tweets.
(2) Daily analysis of tweets.
(3) Apply a novel algorithm based on derivatives.
(4) Confirm detected events using actual news sources.

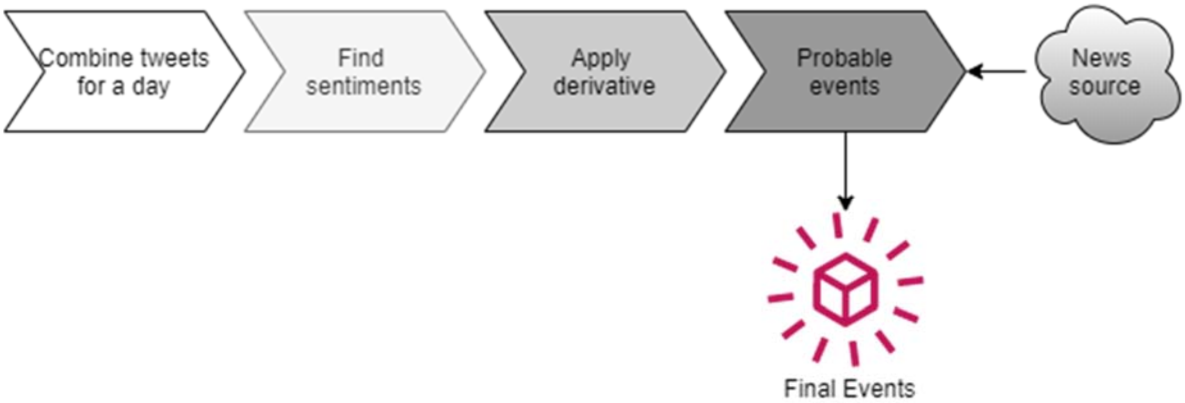

Fig. 1. Stages of proposed framework

The summarised stages of the overall methodology are shown in figure 1. The organisation of the paper is as follows. Related words and shortfalls are presented in section 2. The proposed framework and algorithm are presented in section 3. The results and performance of the model are put on view in section 4 followed by the conclusion and future work.

2. Related Work

The relationship between Twitter feeds and trending event topics have been explored by many researchers. The approaches based on sentiment analysis, incremental Markov chain with temporal relationship, a neural network for classifying the events along with term frequencies were used (Acharya, 2017) (Repp, 2018). The accuracy of these models may vary based on the kind of dataset they are using. The maximum models were not applied to diverse datasets. The datasets were labelled manually to find accuracy.

The investigators also exploited topic modelling based approaches like LDA (Latent Dirichlet Allocation), NMF (Non-negative Matrix Factorization), and variations of them (Hu, 2017) (Nolasco and Oliveira, 2019) (Fan, 2019). Various trend detection and ranking algorithms were proposed using TF/IDF and the Biterm topic model (Alomari,2020). The graph-based frameworks are also used to discover the influential nodes in the network (Jastania,2020). The Temporal and spatial approaches are leveraged for better accuracy using clustering methods like Density-based spatial clustering (Chong, 2017) (Nguyen,2017) (Gupta,2020). The accuracy of these models is primarily based on location coordinates. For most of the datasets, the spatial coordinates will not be available. Many times the Twitter API does not allow location information due to API restrictions.

The shortage of publicly available datasets in this area of work leads the work to investigate more about benchmark datasets. Some of the event detection frameworks (Hasan, 2016) (Hasan, 2019) implemented on Event 2012 dataset. These frameworks can detect fewer busty events using locality-based hashing along with incremental clustering. The other event detection models used public datasets like Super Tuesday, US election datasets to test the frameworks (Saeed, 2019) (Fan,2019) (Papadopoulos, 2013).

The SM platforms like Twitter impose some restrictions on the attributes. To design the event detection model based on the use of easily available attributes like tweets with date and time information pose challenges. The other challenge is the absence of a common framework to detect diverse sets of events. This motivated us to propose the derivative-based model. This model can work with tweets, date, and time attributes with an average detection accuracy of 86%. The next section describes the proposed framework and algorithm.

3. Proposed Framework

The state-of-the-art models proposed by various researchers dominated using spatial coordinates, use of topic modelling methods. The proposed framework can find events even with the trivial and massive sweep on event timeline. The architecture and algorithm of the research presented are shown in the coming section.

3.1. *Architecture*

The proposed framework shown in figure 2 works as follows. The tweets downloaded using the Twitter API, then cleaned and stored in a dataset. The cleaned dataset has to be used for further analysis. Step one involves finding sentiment cluster centers. The K-means algorithm is used on the sentiment of each tweet. The algorithm had been applied by taking three clusters in the account as positive, negative, and neutral. The first cluster center is the most

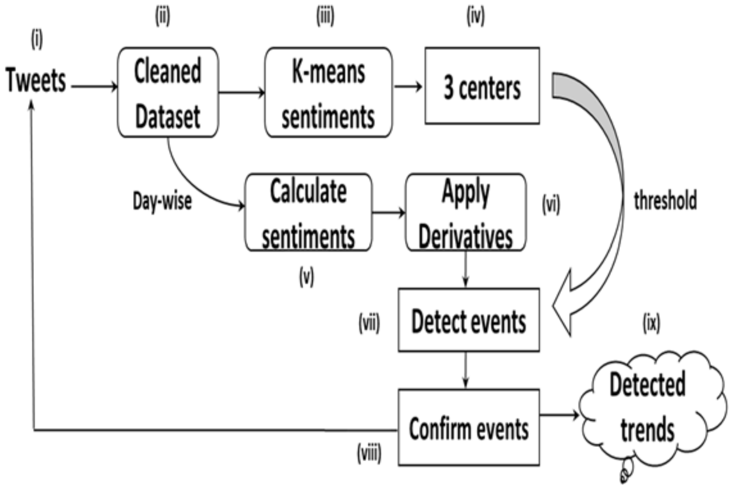

Fig. 2. Proposed framework for detecting event trend

dominated and dense. Hence, this value is used as a threshold for comparison in the last phase of the model. In the second step, the user tweets are combined for each day and find the sentiment for that day using combined tweets. Later, a second-order derivative had applied by considering the x-axis as dates and the y-axis as sentiment function. If the derivatives result in more than the threshold value, then it gets added to the list of detected events. The detected events in the second step are then verified using the various news sources published for the same.

3.2. *Algorithm*

It is observed that there is a very high probability of an event occurrence when there is a sudden drift in sentiments (Tijare, 2020). On the other hand, it is equally difficult to detect these drifts with a simple threshold. If the threshold to be set is too low then, there are many sentiment drifts left undetected that may lead us to miss on the actual events, or setting the threshold to a very high value might eliminate some of the important events. Hence, choosing an appropriate threshold value to detect drifts becomes a real challenge. The Proposed algorithmic approach using derivatives will help in solving these challenges.

The derivatives of function $y = f(x)$ can be found using (Kwok, 2008),

$$\text{slope} = \frac{\text{changes in y}}{\text{changes in x}} = \frac{\Delta \text{y}}{\Delta \text{x}} \tag{1}$$

where, x indicates day/date on the timeline change to $x + \Delta x$. The interval between days is 1 hence Δx is 1. Δy indicates changes in sentiment for the tweets on that day, considered as $f(x + \Delta x)$. Therefore, slope is,

$$\Delta y1 = \frac{\Delta y}{\Delta x} = f(x + \Delta x) - f(x) \tag{2}$$

Equation 2 will give the first order derivative which helps in amplifying the peaks. Still, many small trends will be left undetected. Hence, a second order derivative calculated using equation 3 is as follows.

$$\Delta y2 = \Delta y1(x + \Delta x) - \Delta y1(x) \tag{3}$$

where, $\Delta y2$ is the second order derivative, $\Delta y1(x + \Delta x) - \Delta y1(x)$ is the changes in first order derivative results. The results obtained from the second order derivative will further amplify trends which can be indicators of the trending event. These results are then compared with subjectivity threshold and sentiment threshold which will give the keywords and event dates. The Proposed algorithmic approach is as follows:

<u>*Proposed algorithmic approach:*</u>

```
Dataset : Collection of four datasets (MissionShakti, Sabarimala verdict, Rafale deal, ICCWorldCup
Day : Dataset generated by combining for tweets for each date
for Day in each dataset :
step 1 : Initialise $\Delta y1$=[ ], $\Delta y2$ =[ ]
```

```
        sent_threshold = K-means center for first cluster
        subj_threshold = 0.45
    step 2 :  for i in Day:
            $day\_tweet\_sentiment = VADER(Day[‘tweet’])$
            $\Delta y1 = day\_tweet\_sentiment[\mathrm{i}+1] - day\_tweet\_sentiment[\mathrm{i}]$ // first derivative
            $\Delta y2 = \Delta \mathrm{y1}[\mathrm{i}+1] - \Delta \mathrm{y1}[\mathrm{i}]$  // second derivative
    step 3:  if $\Delta y2$>sent_threshold  and  subjectivity >subj_threshold :
            Display most frequent unigrams with date
```

The proposed algorithm detects the events with a date. The algorithm works as follows: The ‘subj_threshold’ is the threshold value for tweet subjectivity is set to 0.45 (Tijare, 2020). The subjectivity varies from 0 to 1, stating that the tweet is subjective. The subjectivity needs to be checked as the datasets are collected using keyword and hashtags based search on a daily basis for the stated timeline. This helps in filtering out the irrelevant tweets. ‘sent_threshold’ is the sentiment threshold which is needed to decide whether on the date the trend reflects an event. This value is taken from K-means algorithm first cluster center, which is the most dominating cluster. This indicates that the maximum number of tweets belong to this cluster.

Table 1. Event datasets

Dataset	Size (Number of rows)
MissionShakti	80256
Sabarimala verdict	52498
Rafale deal	174166
ICCWorldcup	75274

Combine all tweets collected for that day, find sentiment using VADER sentiment analysis. VADER is aspect based sentiment analyser. Apply the first order derivative to the daily tweets. Here, the x-axis is the date and the y-axis is the sentiment on that date. Since the difference between the x-axis values is always 1, store results in $\Delta y1$. Apply a second derivative and store results in an array $\Delta y2$. Repeat this process for all the dates in the dataset on which tweets are collected. If the second derivative $\Delta y2$ is greater than sent_threshol and subjectivity greater than subj_threshold, then resulting dates are considered as trending event dates.

4. Results and Understandings

4.1. *Datasets*

The tweets with date and time information collected using the Twitter API Tweepy using hashtag and keyword based search on keywords like ‘MissionShakti’, ‘Sabarimala verdict’, ‘Rafale deal’, ‘ICCWorldcup’ or ‘Cricket World Cup’ in the timeline of January 2019 to July 2019 depending on the time the event is active. These events cover ‘Scientific’, ‘Political’, ‘Social’ and ‘Sports’ domains give the chance to test the proposed framework covering all dimensions. The Missionshakti event was about the launch of an anti-satellite missile on March 27th, 2019. Sabarimala verdict event was related to the Supreme Court’s verdict against allowing only male gender devotees into temple premises. The Rafale deal event was on the political agenda of many parties and used throughout the period of Loksabha elections by Congress and BJP in 2019. Finally, the ICC Cricket World Cup event in 2019 is of supreme importance as cricket is the most popular sport in India.

The events used to test the proposed algorithm were in continuous news and created a buzz on all platforms, including Twitter. After the cleaning and pre-processing, the size of datasets in terms of ‘number of tweets’ for each event is shown on table 1. This process is reflected in phase (ii) of figure 2.

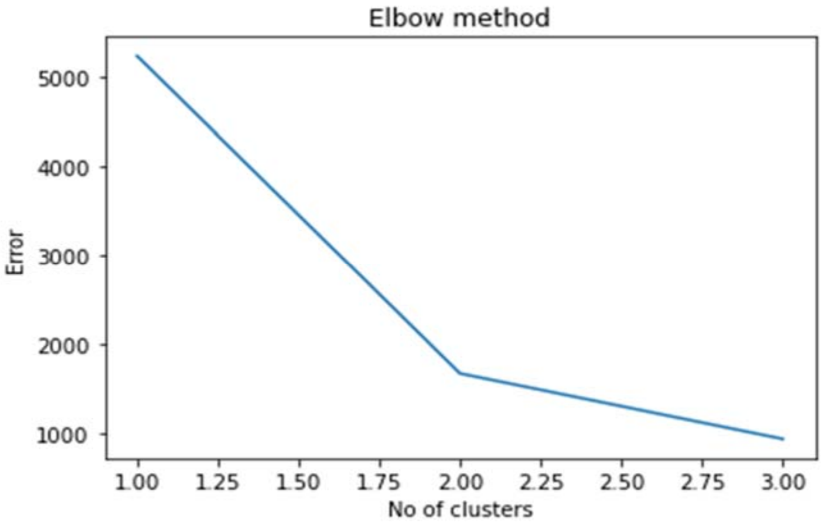

Fig. 3. K-means applied to Cricket World Cup dataset

4.2. *Experimental setup on World Cup dataset*

The proposed algorithm is applied to four datasets discussed in section 4.1. The experimental set up on the Cricket World Cup dataset is discussed here. K-means is applied to tweet sentiments as shown in phase (iii) and (iv) of figure 2. The numbers of clusters are verified by computing mean square error is shown in figure 3 where, x-axis indicates the number of clusters and y-axis indicates the error value. Error value on the y-axis indicates the sum of the mean squared distortion calculated using Euclidean distance. The value of k selected at the elbow where error starts decreasing linearly. The graph in figure 3 shows that error settling down with a value of k as 3. The first cluster of k-means algorithm is the most dominating cluster. Hence, it is taken as a threshold for detecting probable events. The k-means first cluster centers detected at 0.0321235. This value is taken as the threshold value as shown in figure 2.

As per phase (v) and (vi), the combined sentiment needed to be calculated for all the tweets on a given day. This process is needed to be repeated on all the dates in the dataset. The graph of sentiment flow is shown in figure 4(a). The x-axis reflects the numbers indicating dates mapped to ‘day numbers’. The y-axis indicates the sentiment flow between minimum -1 to maximum +1. The first order derivatives and second order derivatives are calculated as shown in equation 2 and 3. The graphs of sentiment flow with first order and second order derivatives are shown in figures 4(b) and 4(c). The graphs in figure 4 show the clear improvements in the sentiment trend after application of first and second order derivatives. The developments of the minor peaks to major peaks can be seen in figures 4(a) to 4(c). This helps in identifying minor trends in a timeline.

The second order derivative is then compared to the sentiment threshold as 0.0321235 first cluster center which gives the trending events as shown in phase (vii) of figure 2. These trending events are then compared with actual news sources on that event to confirm the event shown in phase (viii) in figure 2. Table 2 shows some of the results detected by the proposed model on the World Cup dataset. The first column shows the flow of trend, ‘Date’ indicates middle date, and sentiment is shown for three days: day before ‘Date’, ‘Date’ and day after ‘Date’ showing a transition in sentiments. Subjectivity indicates the tweets are related to subjects with a score of more than 0.45. These results are mapped with the results of the World-Cup matches happening on the dates and the news source as shown in the last column. The results show that the trends are getting mapped to major events. The trend on 10th June 2019 was positive. The sentiment scores on the day and day after are positive i.e. on 11th June. Result shows a negative trend between 12 and 14 June, a match was cancelled due to rain showing the flow of the event and the trend matching with the results returned by proposed model. Similarly, results on July 9th show the shifting of trend from negative to positive.

Table 3 shows mapping of tweets with keywords detected by the model. Table shows date detected by model, event related words detected by proposed model and Tweets posted after the match. These keywords are mapped with the tweets on actual events posted on Twitter. The trends detected are matching with trends as positive if India won the match and Negative when the match was drawn due to rain or India lost the match. Keywords are matching with the after match reactions by the public. The labeling process on datasets and computation on accuracy is presented in following section.

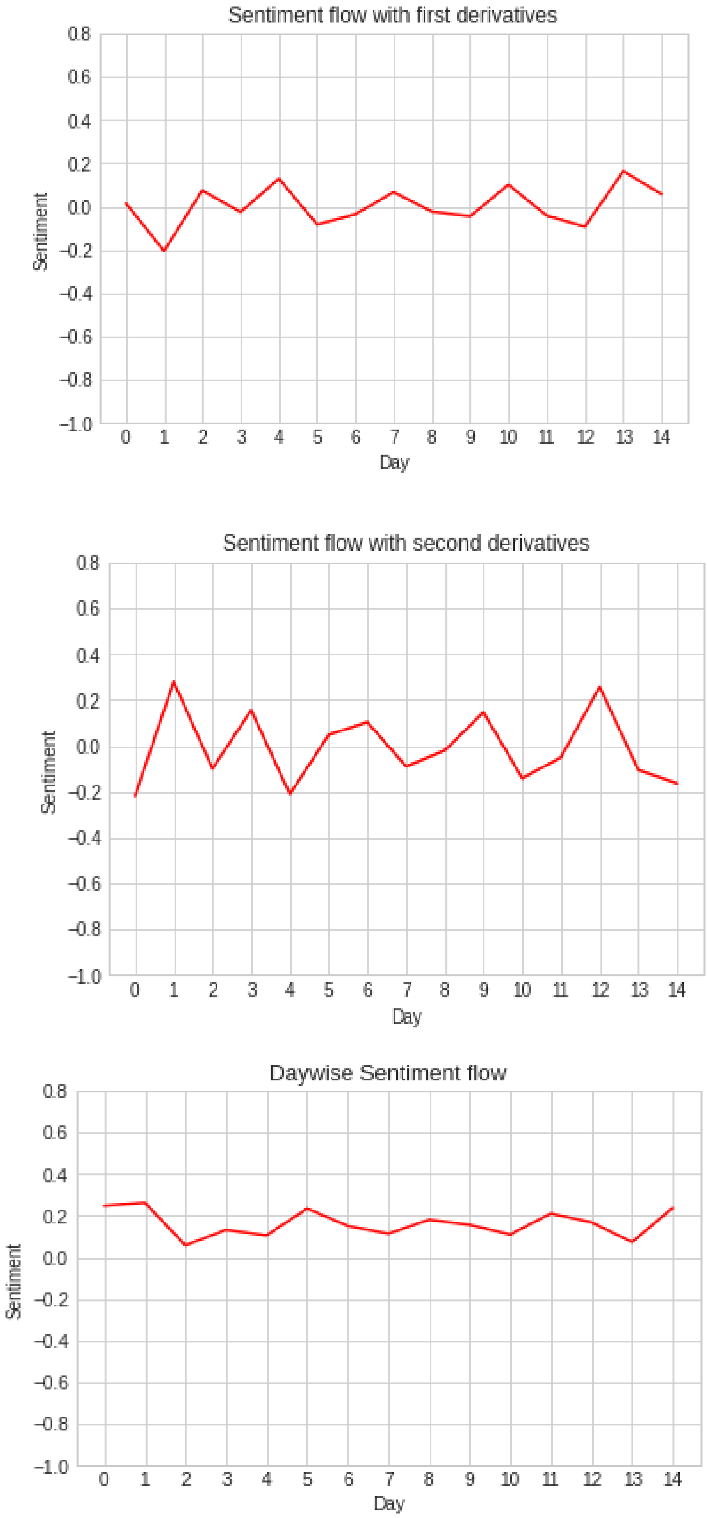

Fig. 4. Sentiment flow a) without derivatives b) with first derivatives c) with second order derivative

Table 2. Results on Cricket World Cup 2019

Detected trend	Date	Sentiment (day before --->Event Day---> day after)	Subjectivity	Matches and Results	News Source
Positive trend	2019-06-06	-0.21984246395527646--> 0.2790147254753345--> -0.10035400311144971	0.50618	India match on 5th June 2020, INDIA Won	https://www.cricketworldcup.com/match/8199
Positive trend	2019-06-10	-0.2120837650423862--> 0.04702221237347469--> 0.10291301646441087	0.47768	India match on 9th June 2020, INDIA Won	https://www.cricketworldcup.com/match/8205
Negative trend	2019-06-13	-0.09072366532022212--> -0.02130272743323723--> 0.14582832716393673	0.46814	Match abandoned without a ball bowled	https://www.cricketworldcup.com/match/8209
Positive trend	2019-06-23	-0.2243288560659627--> 0.27500857886542346--> -0.021729113525383995	0.50357	India won, match on 22nd June	https://www.cricketworldcup.com/match/8219
Positive trend	2019-06-23	-0.2243288560659627--> 0.27500857886542346 --> -0.021729113525383995	0.50357	India won, match on 22nd June	https://www.cricketworldcup.com/match/8219
Positive trend	2019-06-26	-0.1368892672167114 --> 0.025809138150608785 --> 0.0601326490124551	0.64385	India won , match on 27th June	https://www.cricketworldcup.com/match/8225
Negative trend	2019-07-01	0.14033696757695086 --> -0.12405856013630424 --> 0.10918125684299437	0.47788	India lost , match on 30th June	https://www.cricketworldcup.com/match/8229
Positive trend	2019-07-02	-0.12405856013630424 --> 0.10918125684299437 --> -0.06631671794263344	0.49424	India won with huge margin, match on 2nd July	https://www.cricketworldcup.com/match/8231
Positive trend	2019-07-09	-0.014368020189210184 --> 0.05400281833148651 --> -0.03549183493391782	0.60020	India won with Srilanka, match on July 6th	https://www.cricketworldcup.com/match/8235
Negative trend	2019-07-10	0.05400281833148651 --> -0.03549183493391782 --> 0.026855436152711284	0.59213	India lost with New Zealand, match on 10th July	https://www.cricketworldcup.com/match/8237

Table 3. Results of world cup event keywords detected mapping to tweets posted

Date	Event Words Detected	Tweets Post Match
2019-06-10	'indvau', 'reggaegirlz', 'india', 'cwc', 'cricket', 'australia', 'jamvbrakeeper'	Chahal gets another and #TeamIndia are closing in on their second victory! Substitute Ravindra Jadeja takes a fine catch running in off the boundary. India win by 36 runs.
2019-06-13	'cwc', 'england', 'match', 'auvpak','indvnz','rain'	"India win by seven wickets and for the time being go top of the standings!" ICC Tweeted.
2019-06-23	'tournament', 'cwc', 'afghanitan', 'indvafg', 'teamindia', 'india'	The wicket looks slow but well bowled Afghanistan The Indian team will come hard with the ball, we've got the team to defend this total #MeninBlue #CWC19
2019-07-01	'india', 'cwc', 'dhoni', 'england', 'cricket', 'indveng'	Gonna be fireworks on the pitch in Edgbaston today! MASSIVE game for England! Where’s your money going? England for me! 🔥🔥 #EngvInd #CricketWorldCup19

4.3. *Accuracy on Un-Labelled Dataset*

The accuracy of the proposed model is tested by creating a confusion matrix. The dataset is labeled to find the TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative). The following formulas (Han, 2011) used to prove the accuracy of the model.

$$\text{Precision} = \text{TP}/\ (\text{TP} + \text{FP}) \qquad (4)$$

$$\text{Recall} = \text{TP}\ /\ (\text{TP} + \text{FN}) \qquad (5)$$

$$\text{F1} - \text{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (6)$$

Where,

TP: Positive trends classified as positive

TN: Negative trends classified as negative

FP: Trend wrongly classified as positive

FN: Trend wrongly classified as negative

Recall: Fraction of positive trends among total positive trends detected

F1-Score: The weighted mean of Precision and Recall

The proposed model was applied on event datasets created as per section 4.1. The first task was to label the datasets. The datasets were labeled on the basis of ‘Date’. Each date is labeled manually for the possibility of the event by taking reference of media news sources published on that event. Table 4 shows the datasets with their accuracy, recall, precision and F1-Score. The results on an accuracy show that the mode shows accuracy in the range of 0.88 to 0.95. This shows that the model is not missing many events and with good precision. The F1-score of the model lies in the range of 0.86 to 0.95.

Table 4. Dataset accuracy using proposed approach

Dataset	WorldCup	Mission Shakti	Sabarimala	Rafale
Accuracy	0.95	0.92857	0.88	0.92593
Recall	0.90909	1	0.90909	1
Precision	1	0.88889	0.83333	0.84615
F1-score	0.95238	0.94118	0.86957	0.91667

5. Conclusion and Future Work

The proposed derivative based model is applied on four datasets belonging to different domains. The model works by computing second order derivatives on sentiments of the tweet. It is observed that the trends are amplified due to derivatives. The amplified trends give probable event dates. The accuracy and F1-score of the model shows that the results are precise and have good accuracy. The proposed approach can be applied to any event and it is not restricted to a particular event, unlike the models proposed by researchers in this area. The model adopts simple methodology to detect the trending event topics. The model can also be extended to predict events with minor modifications. This model can also be used to predict events in real time analysis of the data.

Declarations:

Funding: Non-funded project

Conflict of Interest: Author Poonam Tijare declares that she has no conflict of interest. Author Jhansi Rani P. declares that she has no conflict of interest.

Ethical Approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] X. Chen, S. Wang, Y. Tang, and T. Hao. (2019). A bibliometric analysis of event detection in social media. Online Information Review.

[2] A. Bondielli and F. Marcelloni. (2019). A survey on fake news and rumour detection techniques. Information Sciences. vol. 497. pp. 38–55.

[3] H. Zhou, H. Yin, H. Zheng, and Y. Li . (2020). A survey on multi-modal social event detection. Knowledge-Based Systems. pp. 105695.

[4] S. Acharya, B. S. Lee, and P. Hines. (2017). Causal prediction of top-k event types over real-time event streams. The Computer Journal. vol. 60. no. 11. pp. 1561–1581.

[5] O. Repp and H. Ramampiaro. (2018). Extracting news events from microblogs. Journal of Statistics and Management Systems. vol. 21. no. 4. pp. 695–723.

[6] J. Hu, Y. Wang, and P. Li. (2017). Online city-scale hyper-local event detection via analysis of social media and human mobility. in 2017 IEEE International Conference on BigData (Big Data). IEEE. pp. 626–635.

[7] D. Nolasco and J. Oliveira. (2019). Subevents detection through topic modeling in social media posts, Future Generation Computer Systems. vol. 93. pp. 290–303.

[8] C. Fan and A. Mostafavi. (2019). A graph-based method for social sensing of infrastructure disruptions in disasters. Computer-Aided Civil and Infrastructure Engineering. vol. 34. no. 12. pp. 1055–1070, 2019.

[9] E. Alomari, I. Katib, and R. Mehmood. (2020). Iktishaf: A big data road-traffic event detection tool using twitter and spark machine learning. Mobile Networks and Applica-tions. pp. 1–16.

[10] Z. Jastania, R. A. Abbasi, K. Saeedi, and M. A. Aslam. (2020). Using social network analysis to understand public discussions: The case study of# Saudiwomencandrive on twitter, International Journal. vol. 11. no. 2. pp.223–231.

[11] Y. S. Chong and Y. H. Tay. (2017). Abnormal event detection in videos using spatiotemporal autoencoder. in International Symposium on Neural Networks. Springer. pp. 189–196.

[12] D. T. Nguyen and J. E. Jung. (2017). Real-time event detection for online behavioral analysis of big social data. FutureGeneration Computer Systems. vol. 66. pp. 137–145.

[13] S. Gupta and B. Banerjee. (2020) Unsupervised event detectionusing self-learning-based max-margin clustering: Anal-ysis on streaming tweets. IETE Journal of Research.vol. 66. no. 4. pp. 569–578.

[14] M. Hasan, M. A. Orgun, and R. Schwitter. (2016). Twitternews: real time event detection from the twitter data stream. PeerJ PrePrints. vol. 4. p. e2297v1.

[15] M. Hasan, M. A. Orgun, and R. Schwitter. (2019). Real-time event detection from the twitter data stream using the Twitternews+ framework. Information Processing & Management. vol. 56. no. 3. pp. 1146–1165.

[16] Z. Saeed, R. A. Abbasi, O. Maqbool, A. Sadaf, I. Razzak,A. Daud, N. R. Aljohani, and G. Xu. (2019). Whats happening around the world? a survey and framework on event detection techniques on twitter. Journal of Grid Com-putting. vol. 17. no. 2. pp. 279–312.

[17] S. Papadopoulos, E. Schinas, V. Mezaris, R. Troncy, and I. Kompatsiari. (2013). The 2012 Social Event Detection dataset, in Proceedings of the 4th ACM Multimedia Systems Conference, ser. MM Sys '13. New York. NY. USA: Association for Computing Machinery. p. 102107. Available: https://doi.org/10.1145/2483977.2483989.

[18] P. Tijare and P. J. Rani. (2020). A sentiment driven approach to detect an offline event on social media platform. 2nd PhD Colloquium on Ethically Driven Innovation and Technology for Society (PhD EDITS).IEEE. pp.1–2.

[19] Y.K. Kwok. (2008). Mathematical models of financial derivatives. Springer.

[20] J. Han, M. Kamber, and J. Pei. (2011). Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems. vol. 5. no. 4. pp. 83–124.