International Journal of Cognitive Informatics and Natural Intelligence
Volume 15 • Issue 3 • July-September 2021

Semrank:
A Semantic Similarity-Based Tweets Ranking Approach

Jagrati Singh, Motilal Nehru National Institute of Technology, Allahabad, India
Anil Kumar Singh, Motilal Nehru National Institute of Technology, Allahabad, India

ABSTRACT

Popular real-world events often create huge traffic on Twitter including real-time updates of important moments, personal comments, and so on while the event is happening. Most of the users are interested to read the important tweets that possibly include important moments of that event. However, extracting the relevant tweets of any event is a challenging task due to the endless stream of noisy tweets and vocabulary variation problem of social media content. To handle these challenges, the authors introduce a new approach for computing the relative tweet importance based on the concept of the Pagerank algorithm where adjacency matrix of the graph representation of tweets contains semantic similarity matrix based on the word mover's distance measure utilizing Word2Vec word embedding model. The results show that top-ranked tweets generated by the proposed approach are more concise and news-worthy than baseline approaches.

KEYWORDS

Semantic Similarity, Summary, Tweets, Word Mover's Distance, Word2vec Model

INTRODUCTION

Online social media networks have become a rich source of news distribution about real-world events of all kinds. Twitter as a social networking site has emerged to be an endless dynamic real-time global stream of news. Many people use Twitter as a source of news content instead of sharing thoughts and emotions. Journalists are also increasingly adopting twitter as a professional tool for the process of news selection and presentation by continuous monitoring of emerging user-generated newsworthy stories from Twitter stream overloaded with the high amount of noise and redundant information. The typical monitoring method is to search the stream with event relevant keywords. However, the search results after satisfying such a Boolean query is formidable. For example, victory of Narendra Modi as a Prime Minister of India in general election 2019 induced millions of tweets on result declaration night. In such sea of tweets on a topic or related to any event, ranking has become an important issue in Twitter not just in Web search. While there exist an extensive research on ranking for web search (Brin and Page (1998), Agichtein et al. (2006), Xiang et al (2010), Aggarwal (2010)), there is little work done for ranking of tweets that generate the need to develop an efficient Tweets ranking model with the following goals:

Relevance: The top-ranked tweets constituting the summary of the event must be relevant to the specific event and contain some important information that can be used in event analysis.

DOI: 10.4018/IJCINI.20210701.oa6

Diversity: The resultant tweets presenting the summary should capture the diverse information and must not be similar in nature-wise.

There are various applications like Twitter Sentiment Analysis (Thelwall et al. (2011), Zhang et al. (2018), Tripathi et al. (2019)) and News Recommender System (Abel et al. (2011), Phelan et al. (2011), Kywe et al. (2012)) also required relevant tweets to process. Some popular text ranking methods like Lexrank (Erkan and Radev (2004)) and Textrank (Mihalcea and Tarau (2004)) are based on Google's Pagerank algorithm (Brin and Page (1998)) to rank sentences which are suitable for traditional news data but fails in case of Twitter. This is primarily because, in contrast to traditional news documents, tweets are inherently noisy with high variance in word frequencies and low word count. Further, tweets have some other challenging features like high redundancy and large diversity in the vocabulary to convey the same information. Hence, extracting important tweets based on lexical similarity is not effective to handle vocabulary variation problem. For example:

(1) Modi speaks to the media in Kashmir.
(2) The prime minister greets the press in Pulwama.

These two tweets convey the same information means that there exists some relationship with no common vocabulary. In this case, similarity measures like the Cosine or Jaccard metric fails because of using the Vector Space Model based on the common occurrence of textual units between two documents. To resolve this problem, word embedding language models (Mikolov et al. (2013), Pennington et a. (2014), Song et al. (2019)) come into picture that capture semantics or meaningful relationships. So that, we calculate the semantic similarity between tweets using Word Mover's Distance (WMD) (Kusner et al. (2015)) based on Word2Vec model (Mikolov et al. (2013)). Main contributions of the proposed approach are as follows:

(1) We utilize several social influence features like re-tweet count, follower count and presence of URL to remove the noisy tweets like personal, fake and so on.
(2) We also utilize tweet content features to rank the tweets by using a semantic graph model where vertices represent tweets and edges represent the semantic similarity between tweets. Top-ranked vertices (tweets) are extracted to represent the event summary by utilizing the Pagerank algorithm.
(3) We compared the results of the proposed approach with four baseline approaches (Lexrank, Textrank, Re-tweet voting, and Follower voting). Our system outperforms the baseline approaches in accuracy measured by ROUGE metric and Human evaluation score.

The rest of the paper is organized as follows: The next section briefly discusses the related work. Section 3 describes the proposed Semrank Twitter ranking model in detail. The experimental dataset, evaluation method and results obtained by comparing the proposed approach with four competitive baseline methods have been presented in section 4. Finally, section 5 concludes with directions for further research.

BACKGROUND

Most of the existing Twitter ranking algorithms are based on traditional text ranking approaches suitable for traditional news text data. These algorithms suffer various challenges of Twitter data like high volume data and the use of informal language. Sharifi et al. (2010a) proposed a Phrase Reinforcement (PR) algorithm to generate the most representative tweet for summarization of any event. They built an ordered acyclic graph where the root node represents the starting phrase of tweet and words that present just after or before the phrase connect the root node. Weights are assigned to

each node based on the frequency of the word. This PR algorithm calculates the most weighted path from the root to the leaf node to generate a summary. The main limitation of this approach is one tweet summary that is not sufficient to cover the whole information of an event. Sharifi et al. (2010b) handle this limitation by generating four most weighted paths from the root to the leaf node. Judd and Kalita (2013) also improved the Pagerank algorithm by handling the drawback of syntactic well-formedness of sentence. Improved algorithm parses the summary by building dependencies using POS tagging. They discard those summaries if any grammatical mistake or incorrect dependency present. The proposed work by Nichols et al. (2012) also used the Pagerank algorithm by adding one more feature like user status updates for domain-specific events, specifically for sports various researchers used Pagerank algorithm Brin and Page (1998) for the summarization of traditional news text data. (Erkan and Radev (2004)) proposed Lexrank graph algorithm in which nodes represent sentences and edges represent the similarity between sentences. The similarity is calculated using the Cosine measure and a score of each sentence is computed using the Pagerank algorithm. Another classical algorithm Textrank (Mihalcea and Tarau (2004)) also used the Pagerank algorithm for ranking the text data. Here, nodes represent the keywords of text documents and edges represent the co-occurrence of keywords in the same document. Xu et al. (2013) extend the Textrank algorithm to make suitable for Twitter data. They extracted bi-grams instead of uni-grams from the tweets to make nodes of a graph where edges represent the co-occurrence of bi-grams within fixed time-window. For each node, a weighted score is calculated using the Pagerank algorithm. Xu et al. (2013) also extend the Textrank algorithm by considering named entities and event phrases to make nodes of the graph. The proposed work by Khan et al. (2013) applies the Pagerank algorithm on the lexical graph of each event. Firstly, they applied the LDA topic model (Blei and Lafferty (2006)) to get the event clusters. For each cluster, a lexical graph is built and high scored lexical units are included in the summaries that are the main discussion points of the tweets.

Inouye (2010) proposed two approaches to generate Twitter event summary. In the first approach, event relevant tweets are classified into sub-topics by using the hybrid form of k-means++ algorithm (Arthur and Vassilvitskii (2007)). For each cluster, the weight of each tweet is computed by using a hybrid TF-IDF weighting scheme. Maximum scored tweet of each cluster represents the summary of an event. In the second approach, the hybrid TF-IDF summarization algorithm as proposed in the work (Sharifi et al. (2010b)) is modified to produce four tweet summaries. Yang et al. (2012) proposed a batch summarization algorithm (SPUR) from incoming tweets by dividing the stream into clusters based on one-hour time-window. From each time-window, tweets are ranked based on the frequency of used phrases. This work also introduced dynamic SPUR algorithm (DSPUR) based on a pyramidal time window approach (Aggarwal et al. (2003)). Zhao et al. (2016) detected bursty phases of the event to gather the user's collective interests during the event happening. The features; like informativeness, interestingness, and diversity of tweets are the basis of selection of tweets. The system applies the Lexrank algorithm to extract informative sentences that are used to know the user's collective interests and diversity ranking algorithm Marginal Relevance Ranking (MRR) is used to reduce the redundant information from each cluster. Top-ranked tweets among these three features are extracted to represent the result summary. The traditional TF-IDF weighting scheme is used by various researchers of this field to extract top event relevant tweets. Some of the state-of-the-art approaches are discussed here. Alsaedi et al. (2016) proposed three statistical methods to summarize Twitter events: Temporal TF-IDF, Re-tweet voting method, and temporal centroid representation approach. All these methods divide the Twitter stream into a one-hour time-window, apply the clustering algorithm and combine the most important tweets from each cluster to make a summary. The temporal TF-IDF method ranks tweets based on the term frequency within the time frame. While the voting approach method is based on re-tweet count and temporal centroid method chooses tweets that present in the cluster centroid within the time frame to make the summary. Belkaroui and Faiz (2017) explored many social features to generate good event summary. The system uses a set of hash-tags as a search query to extract event-related conversations. Three features; tweet influence, tweet relevance regarding

initial text, and tweet relevance regarding URL are used to generate a summary. The tweet influence is measured by reply-count, re-tweet count, and favorite count. Relevance score is computed by the Cosine similarity between the initial tweet and all other relevant tweets. The final score would be the linear combination of these feature scores and ranks the tweets based on the final score to form the summary. He et al. (2018) proposed a Twitter summarization framework named SNSR based on the integration of social network and sparse reconstruction. The sparse reconstruction process is based on selecting tweets that cover the whole topic with maintaining diversity and free from sparsity problem. Naik et al. (2018) proposed a summarization system for tweets that makes use of the Particle Swarm Optimization (PSO) algorithm (Kennedy et al. (1997)). Chin et al. (2019) designed tweets summarization engine for mobile devices based on Latent Dirichlet Allocation (LDA) topic modeling.

PROPOSED WORK

There are a number of steps required for ranking the tweets to extract an important summary of an event that is shown in Figure 1. These steps are as follows:

Aggressive Filtering of Non-Informative Tweets

As a first step, the proposed approach keeps only those tweets that are potentially informative. We use social network features, which say a post is more informative if it has been retweeted so many times and published by a user with more number of followers. Importantly, tweets containing multimedia content URL give the live and rich coverage of the event. So, scores are given to each tweet based on the following features.

Re-tweet Score: Count of re-tweets is an indication of popularity but only popularity is not sufficient to decide the tweet is informative or not because many times users re-tweet celebrities post without even reading. Re-tweet score of a tweet can be defined as follows in Equation (1):

$$R_{score}\left(t_i\right) = \frac{RT_i}{\sum_{i=1}^{n} RT_i} \tag{1}$$

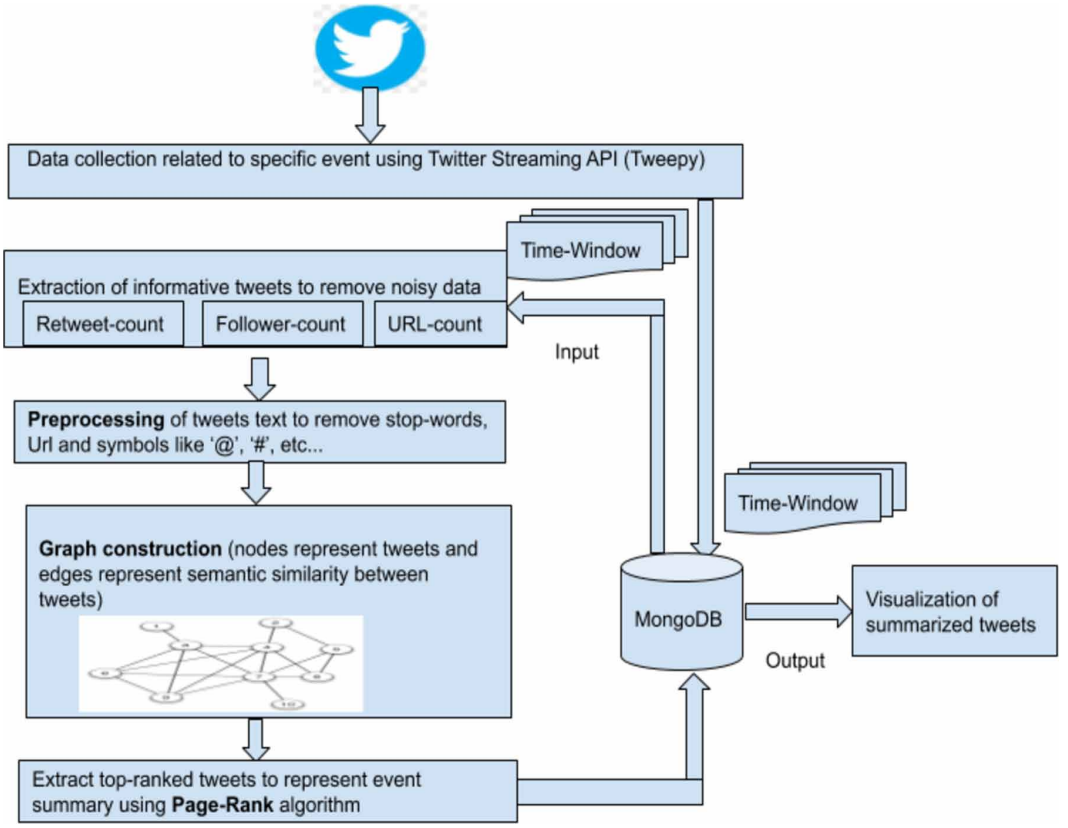
Where, t_i represents the i^{th} tweet in the current time-window and RT_i represents the re-tweet count of the i^{th} tweet and n is the total number of tweets in the current window.

Follower Score: A user with more number of followers is likely to be a more authentic user and their tweets are also likely to be more informative. The follower score is defined as follows in Equation (2):

$$F_{score}\left(t_i\right) = \frac{FL_i}{\sum_{i=1}^{n} FL_i} \tag{2}$$

Where, FL_i represents the follower counts of the i^{th} tweet.

Figure 1. Overview of Semrank approach



URL Score: The presence of URL in the tweet gives more information regarding the event that makes it easy to understand. It is defined in Equation (3):

$$URL_{score}\left(t_i\right) = \frac{freq\left(URL_i\right)}{N} \tag{3}$$

Where, $freq\left(URL_i\right)$ represents the frequency count of i^{th} tweet URL and N is the total count of URL in the current time-window.

We compute the final informative score of a tweet as a linear combination of the above scores by using Equation (4). Low-scored tweets are filtered to keep only informative tweet.

$$tweet_{score} = R_{score} + F_{score} + URL_{score} \tag{4}$$

The Semrank Approach

Graph-based ranking algorithms are based on the global importance of a node within the entire graph. The basic idea of a graph-based ranking algorithm is voting or recommendation. When one

node links to another node, it is vote casting for the other node. The higher number of vote casting for a node means more importance of that node. In the context of Web searching, original Pagerank algorithm is made for un-weighted and directed graphs. However, the proposed graph model is built from micro-blogging twitter texts as follows: Firstly each tweet text is pre-processed to remove media-Url, useless symbols like "@", "#", punctuations and digits from the text. The remaining clean text is tokenized to remove stop words. The tweets are now filtered based on the tweet structure. Tweets containing more than two user mentions or more than three hash-tags or less than five text tokens are removed. This structure-based filtering helps to filter those tweets which do not carry enough news-like content. These pre-processed tweets represent the vertices of the graph. We build edges between vertices based on the semantic similarity between tweets using a recently developed distance measure called Word Mover's Distance (Kusner et al. (2015) based on the word embedding Word2Vec model (Mikolov et al. (2013)).

Word2Vec Model

Word embeddings are language models in which each keyword is represented as a location in an n-dimensional continuous space. The model generates a vector for each keyword by using neural network architecture. The model is trained to maximize the log probability of neighboring keywords for each keyword vector in a corpus by using following formula mentioned in Equation (5).

For a given sequence of keywords $k_1, k_2, k_3, \ldots, k_j, \ldots, k_n$

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{j\in neigh(n)}\log \mathrm{p}\left(k_j|k_n\right) \tag{5}$$

Where, *neigh(n)* is the set of neighbors of keyword *n* and $p\left(k_j|k_n\right)$ calculates the hierarchical softmax of the associated keyword vectors V_j and V_n. This equation indicates that keywords that are closer by location are similar.

Word Mover's Distance (WMD)

WMD is a good technique of semantic distance measure derived from earth mover's distance (EMD) (Rubner et al. (2000)) based on the Transportation problem. As in WMD, the distance between two tweets t_1 and t_2 is the minimum cumulative distance that keywords from tweet t_1 need to travel to match the keywords of tweet t_2. The transportation matrix between tweets t_1 and t_2 is shown in following Equation (6):

$$\begin{array}{cc} & \begin{array}{ccccccc} keyword_1 & \cdots & keyword_i & \cdots & keyword_n \end{array} \\ & \begin{array}{ccccccc} rf_1 & \cdots & rf_i & \cdots & rf_n \end{array} \\ \begin{array}{c} keyword_1' \\ \vdots \\ keyword_j' \\ \vdots \\ keyword_m' \end{array} & \begin{array}{c} rf_1' \\ \vdots \\ rf_j' \\ \vdots \\ rf_m' \end{array} \end{array} \begin{bmatrix} ed_{1,1} & \cdots & ed_{1,i} & \cdots & ed_{1,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ ed_{j,1} & \cdots & ed_{j,i} & \cdots & ed_{j,i} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ ed_{m,1} & \cdots & ed_{m,i} & \cdots & ed_{m,n} \end{bmatrix} \tag{6}$$

Where, $keyword_i$ and $keyword'_j$ represent the keywords of tweet t_1 and t_2 respectively. rf_i and $rf_j^{'}$ denote the relative term frequency of each keyword in the corresponding tweet. Euclidean distance between $keyword_i$ and $keyword_j^{'}$ is represented as $ed_{i,j}$. To mea t_2 sure the distance between two keywords, they used Euclidean distance formula given in Equation (7) where each keyword is represented as vector V learned from the trained Word2Vec model.

$$ed_{i,j} = \left| V_i - V_j \right|_2 \tag{7}$$

Let T be a Transportation matrix where $T_{ij \geq 0}$ denotes how much of $keyword_i$ in t_1 travels to $keyword_j^{'}$ in . To transform t_1 entirely into t_2, we ensure that the entire moving quantity from $keyword_i$ equals rf_i, i.e. $\sum_{j=1}^{m} T_{ij} = rf_i$. Further, the amount of moving quantity from $keyword'_j$ must match $rf_j^{'}$, i.e. $\sum_{i=1}^{n} T_{ij} = rf'_j$. Finally, distance measured formula is defined in Equation (8) as the minimum distance required to move all keywords from t_1 to t_2 tweet based on the Transportation optimization formula.

$$\begin{aligned} &\min \sum_{i,j=1} T_{ij} ed_{i,j} \\ &subject\,to : \sum_{j=1}^{m} T_{ij} = rf_i \qquad \forall\ i \in \{1, \ldots, n\} \\ &\sum_{i=1}^{n} T_{ij} = rf_j^{'} \quad \forall\ j \in \{1, \ldots, m\} \\ &\qquad T_{i,j} \geq 0 \end{aligned} \tag{8}$$

Here, we explain the WMD algorithm in detail on two tweets t_1 and t_2.

$t_1 =$ *Fights over the nation against terrorism.*

$t_2 =$ *Protests across country after terrorist attack in Kashmir.*

As we mentioned that WMD algorithm is similar as the Transportation optimization problem with the objective is to minimize the travel cost of moving a product from several sources to several destinations whereas our objective is to minimize the cumulative distance that keywords from tweet t_1 need to travel to match the keywords of tweet t_2. Table 1 shows the transportation matrix regarding our problem. Here, keywords of one tweet are considered as the number of sources and keywords of another tweet are considered as the number of destinations. The relative frequency of each keyword in the corresponding tweet is assumed as the supply or demand quantity of product. Each cell value contains the Euclidean distance between keywords of different tweet which denotes the travel cost between source and destination. The problem is solved the same as the Transportation problem using the least-cost method.

Table 1. Sample example of Word Mover's Distance represented as Transportation matrix

(Destination) t_j / (Source) t_i	fights	nation	Terrorism	
protest	0.36	0.39	0.34	0.166
across	0.48	0.39	0.46	0.166
country	0.43	0.20	0.33	0.166
terrorist	0.40	0.34	0.19	0.166
attack	0.40	0.40	0.34	0.166
Kashmir	0.51	0.42	0.39	0.166
	0.333	0.333	0.166	(Supply) rf_i / (Demand) rf_j

In the first step, we choose minimum distance cell (0.19) and assign a minimum of (rf_i, rf_j) to this cell which turns out to be in (0.333, 0.166) as 0.166. After that rf_i value corresponding to this row becomes 0 and rf_j value becomes 0.166. In further processing, this row is not considered as the rf_i value becomes 0 which is stripped out in Table 2. In the next step, iteratively we choose the next minimum distance cell (0.20) and assign a minimum of (0.333, 0.166) to this cell as 0.166. Now, rf_i value corresponding to this row becomes 0 and rf_j value becomes 0.166. This row is also stripped out in Table 3. We repeat the same process until all values of rf_i and rf_j becomes zero means supply fulfill the demand. All the remaining steps to solve the problem are shown in Table 4, 5, 6 and 7.

$$\mathrm{WMD}\left(t_1, t_2\right) = 0.34*0.166 + 0.39*0.166 + 0.20*0.166 + 0.19*0.166 + 0.40*0.166 + 0.51*0.166$$

Finally, 0.33 semantic distance value is obtained between tweet t_1 and t_2 means that the context of both tweets are the same whereas the Cosine distance is 1 means that there is no similarity. It concludes that WMD measure is able to capture the semantics of tweets and other measures like Cosine or Jaccard fails to measure this. Now, semantic similarity can be defined as the inverse of semantic distance in the following Equation (9) which is considered as the semantic weight sw_{ij} of an edge between each pair of vertices in the graph. We add 1 in the denominator to handle divide by zero error in case of similar tweets.

$$sw_{ij} = \frac{1}{WMD\left(i, j\right) + 1} \tag{9}$$

Extraction of Top-Ranked Tweets Based on The Pagerank Algorithm

Formally, let G (V, E) be a directed graph with the V set of vertices and E set of edges. For a given vertex V_i, let $\mathrm{In}\left(V_i\right)$ be the set of vertices that point to it called as predecessor and $\mathrm{out}\left(V_i\right)$ be the set of vertices pointed by vertex $\left(V_i\right)$ called as successors. The Pagerank score (Brin and Page (1998)) of a vetex V_i is defined as $PR\left(V_i\right)$ in Equation (10):

$$PR\left(V_i\right)=\left(1-d\right)+d*\sum_{j\in In\left(V_i\right)}\frac{1}{\text{out}\left(V_j\right)}PR\left(V_j\right) \tag{10}$$

Where, d is a damping factor which gives the probability of switching from a given node to another random node in the graph. The default value of d is taken as 0.85 as is used in computing Pagerank over WWW (Brin and Page (1998)). In the starting, arbitrary values are given to each node in the graph and algorithm iterates until convergence is achieved. Notice that after the final scores obtained, completion of Pagerank runs is not affected by the choice of any initial value, only the iterations count may be different for convergence. We introduce a modified Pagerank formula named Semrank formula for ranking of tweets that considers the undirected weighted graph G(V, E). Let Neigh (V_i) be the set neighboring vertices of (V_i) and sw_{ii} be the semantic weight of an edge given in Equation (9). A similar Simrank score of a vertex V_i is defined as $SR\left(V_i\right)$ in Equation (11) to give ranking to the vertices (tweets):

$$SR\left(V_i\right)=\left(1-d\right)+d*\sum_{V_j\in Neigh\left(V_i\right)}\frac{sw_{ij}}{\sum_{V_{k\in Neigh\left(V_j\right)}}sw_{jk}}SR\left(V_j\right) \tag{11}$$

Extraction of Top-Ranked Tweets Based on The Pagerank Algorithm

Formally, let G (V, E) be a directed graph with the V set of vertices and E set of edges. For a given vertex V_i, let In (V_i) be the set of vertices that point to it called as predecessor and out (V_i) be the set of vertices pointed by vertex (V_i) called as successors. The Pagerank score (Brin and Page (1998)) of a vetex V_i is defined as $PR\left(V_i\right)$ in Equation (13):

$$PR\left(V_i\right)=\left(1-d\right)+d*\sum_{j\in In\left(V_i\right)}\frac{1}{\text{out}\left(V_j\right)}PR\left(V_j\right) \tag{13}$$

Where, d is a damping factor which gives the probability of switching from a given node to another random node in the graph. The default value of d is taken as 0.85 as is used in computing Pagerank over WWW (Brin and Page (1998)). In the starting, arbitrary values are given to each node in the graph and algorithm iterates until convergence is achieved. Notice that after the final scores obtained, completion of Pagerank runs is not affected by the choice of any initial value, only the iterations count may be different for convergence. We introduce a modified Pagerank formula named Semrank formula for ranking of tweets that considers the undirected weighted graph G(V, E). Let Neigh (V_i) be the set neighboring vertices of (V_i) and sw_{ij} be the semantic weight of an edge given in Equation (9). A similar Simrank score of a vertex V_i is defined as $SR\left(V_i\right)$ in Equation (14) to give ranking to the vertices (tweets):

$$SR\left(V_i\right)=\left(1-d\right)+d*\sum_{V_j\in Neigh\left(V_i\right)}\frac{sw_{ij}}{\sum_{V_{k\in Neigh\left(V_j\right)}}sw_{jk}}SR\left(V_j\right) \tag{14}$$

Table 2. Step 1 considers dist(terrorist, terrorism)

t_i \ t_j	fights	nation	terrorism	
protest	0.36	0.39	0.34	0.166
across	0.48	0.39	0.46	0.166
country	0.43	0.20	0.33	0.166
terrorist	~~0.40~~	~~0.34~~	~~0.19~~ **(0.166)**	0
attack	0.40	0.40	0.34	0.166
Kashmir	0.51	0.42	0.39	0.166
	0.333	0.333	0.166	rf_i / rf_j

Table 3. Step 2 considers dist(nation, country)

t_i \ t_j	fights	nation	terrorism	
protest	0.36	0.39	0.34	0.166
across	0.48	0.39	0.46	0.166
country	~~0.43~~	~~0.20~~ **(0.166)**	~~0.33~~	0
terrorist	~~0.40~~	~~0.34~~	~~0.19~~ **(0.166)**	0
attack	0.40	0.40	0.34	0.166
Kashmir	0.51	0.42	0.39	0.166
	0.333	0.166	0.166	rf_i / rf_j

EXPERIMENTAL STUDIES

Datasets

Two event datasets are used for the experiments; the first one is related to "Terrorist attack in Pulwama, Kashmir on 14th Feb in India" that was collected from 15th Feb 2019 to 29th Feb 2019 around 14,511,78 tweets using Twitter Streaming API. For this dataset, we make day-wise time window for the experiment. Our second dataset contains 15,04,348 tweets related to "General election 2019 in India" event that was collected on result declaration date 23rd May 2019. Here, we consider hour-wise time-window for the experiment due to the high burst of tweets in one day. Statistics of data is presented in Table 9 for both event dataset. MongoDB database is used for storing data in JSON format due to the support of storing document data. Due to the lack of ground-truth data to validate

Table 4. Step 3 considers dist(protest, terrorism)

t_i \ t_j	fights	nation	terrorism	
protest	~~0.36~~	~~0.39~~	**~~0.34~~ (0.166)**	0
across	0.48	0.39	~~0.46~~	0.166
country	~~0.43~~	**~~0.20~~ (0.166)**	~~0.33~~	0
terrorist	~~0.40~~	~~0.34~~	**~~0.19~~ ~~(0.166)~~**	0
attack	0.40	0.40	~~0.34~~	0.166
Kashmir	0.51	0.42	~~0.39~~	0.166
	0.333	0.166	0	rf_i / rf_j

Table 5. Step 4 considers dist(nation, across)

t_i \ t_j	fights	nation	terrorism	
protest	~~0.36~~	~~0.39~~	**~~0.34~~ (0.166)**	0
across	~~0.48~~	**~~0.39~~ (0.166)**	~~0.46~~	0
country	~~0.43~~	**~~0.20~~ (0.166)**	~~0.33~~	0
terrorist	~~0.40~~	~~0.34~~	**~~0.19~~ ~~(0.166)~~**	0
attack	0.40	~~0.40~~	~~0.34~~	0.166
Kashmir	0.51	~~0.42~~	~~0.39~~	0.166
	0.333	0	0	rf_i / rf_j

our results, we manually make the summary of the most important tweets which contain important facts about an event with diverse and non- redundant information. To make the summary from the whole data of each time-window turned out to be very difficult for humans due to the high volume of data. To handle this difficulty for making a ground-truth summary, we manually extract the 10 most relevant tweets from the top 500 informative tweets by applying the first step (aggressive filtering of noisy tweets) of proposed approach instead of considering all tweets of each window.

Table 6. Step 5 considers dist(attack, fights)

t_i \ t_j	fights	nation	terrorism	
protest	~~0.36~~	**~~0.39~~ (0.166)**	**~~0.34~~ (0.166)**	0
across	~~0.48~~	~~0.39~~ (0.166)	~~0.46~~	0
country	~~0.43~~	**~~0.20~~ (0.166)**	~~0.33~~	0
terrorist	~~0.40~~	~~0.34~~	**~~0.19~~ ~~(0.166)~~**	0
attack	**~~0.40~~ (0.16)**	~~0.40~~	~~0.34~~	0
Kashmir	0.51	~~0.42~~	~~0.39~~	0.166
	0.166	0	0	rf_j \ rf_i

Table 7. Step 6 considers dist(Kashmir, fights)

t_i \ t_j	fights	nation	terrorism	
protest	~~0.36~~	~~0.39~~	**~~0.34~~ (0.166)**	0
across	~~0.48~~	**~~0.39~~ (0.166)**	~~0.46~~	0
country	~~0.43~~	**~~0.20~~ (0.166)**	~~0.33~~	0
terrorist	~~0.40~~	~~0.34~~	**~~0.19~~ ~~(0.166)~~**	0
attack	**~~0.40~~ (0.166)**	~~0.40~~	~~0.34~~	0
Kashmir	**~~0.51~~ (0.166)**	~~0.42~~	~~0.39~~	0
	0	0	0	rf_j \ rf_i

Evaluation Methods

ROUGE Metric: We evaluate summarization quality by using Rouge metric (Lin (2004)) based on the counting of overlapped units such as uni-gram, bi-gram, n-gram, and word sequences between the systems generated summary S and ground-truth summary G. ROUGE-N metric can be defined with three sub-metrics given in Equation (15), (16) and (17):

$$ROUGE-N\left(R\right)=\frac{\sum_{I\in G}\sum_{n-gram\in I}Matched_{N-gram}}{\sum_{I\in G}\sum_{n-gram\in I}Total_{N-gram}} \tag{15}$$

$$ROUGE-N\left(P\right)=\frac{\sum_{I\in S}\sum_{n-gram\in I}Matched_{N-gram}}{\sum_{I\in S}\sum_{n-gram\in I}Total_{N-gram}} \tag{16}$$

$$ROUGE-N\left(F\right)=\frac{2*ROUGE-N\left(P\right)*ROUGE-N\left(R\right)}{ROUGE-N\left(P\right)+\ ROUGE-N\left(R\right)} \tag{17}$$

N-gram $\in$ G denotes the N-grams in ground-truth while N-gram $\in$ S denotes the N-grams in the system generated summary. $Matched_{N-gram}$ is the count of matched N-gram.

$Total_{N-gram}$ is the total number of N-grams in ground-truth or generated summary. We use the F-scores of ROUGE-1 for unigrams, ROUGE-2 for bi-grams and the ROUGE-L for longest common subsequence as metrics.

Human judgment score: Manually creating ground-truth summary is difficult and time consuming job. To reduce the difficulty and time, human judgment is used in which two annotators having domain knowledge were asked to give the score to the system generated summary (top representative tweets) on a one-point scale for compactness. Compactness refers to how much important fact and non-redundant information present in the summary. The Human scores given to each tweet of summary for the first time window of General election 2019 dataset are shown in Table 12 and the average scores given to each window summary for both datasets are displayed in Table 13.

Compared Approaches

Before comparing the proposed approach with other existing approaches, we compare the existing following pre-trained Word2Vec models to decide which model fits to our dataset.

GoogleNews-Vectors-Negative300.Vec: This Word2Vec model is trained on Google news data, which includes word vectors for 1 million words and phrases. The length of the word vector is 300 features.

Wiki-news-300d-1M.Vec: This model is trained on Wikipedia 2017, UMBC web base corpus (three billion words included by Stanford WebBase project) and statmt.org (statistical machine translation website) news dataset containing 16B tokens which includes word vectors for 1 million words and phrases.

Wiki-News-300d-1M-Subword.Vec: This model is trained on 1 million word vectors with sub-word information on Wikipedia 2017, UMBC web base corpus and statmt.org news dataset (16B tokens). This model trains fast and consider sub-word features as compared to simple Wiki-news-300d-1M.vec model.

Crawl-300d-2M-Subword.Vec: This is trained on 2 million English word vectors with 300-dimensions containing sub-word information on Common Crawl (600B tokens) released by Facebook.

We compare the proposed SemRank approach with several text ranking methods. Note that we applied the same preprocessing and filtering steps to all approaches. More specifically, we evaluated the following ranking methods:

Textrank (Mihalcea and Tarau (2004)): Algorithm exploits the structure of the text by applying graph modeling approach where vertices represent key-phrases and edges represent co-occurrence of key-phrases to determine important key-phrases in the same way that Pagerank algorithm extract most relevant web pages as a search query result.

Lexrank (Erkan and Radev (2004)): This approach is also based on Pagerank algorithm to rank the sentences. In Graph, vertices represent sentences and edges represent cosine similarity between sentences.

Re-Tweet Voting (Alsaedi et al. (2016)): This method ranks the tweets based on the Re-tweet-count which indicates the popularity of the tweets.

Follower Voting (Cha et al. (2010)): This method rank the tweets based on the Follower-count that is the indication of the tweet authenticity.

Results

All the experiments are carried out on a machine with Intel Core i7@4.0GHz quad-core processor and 16GB memory running on Linux machine. Python 2.7 programming language is used due to the greater support of machine learning libraries. Some libraries like Tweepy for Twitter data collection, Nltk for text preprocessing, Gensim for creating Word2Vec model and Scipy for computing distance matrix are used to implement the proposed work. Several experiments are performed to evaluate different aspects of the proposed ranking method. In the first experiment, we compare the pre-trained Word2Vec models (described in the previous subsection) on 100 pairs of keywords in which each pair is almost similar in meaning but the vocabulary is different. We have shown Euclidean distance between 10 pairs of keywords in Table 8 and found that "wiki-news-300d-1M-subword.vec" model performed best by giving the minimum distance between semantically similar keywords like (win, victory), (press, media), etc. So, we use the best performing model to evaluate the results from the proposed approach. In the second experiment, we compute the ROUGE-1, ROUGE-2 and ROUGE-L metric to compare the results of the proposed approach with baseline approaches for both datasets that is shown in Table 10 and 11. Table 10 shows the result of three days summary (15, 16 and 17th Feb) of Pulwama attack event in which proposed approach obtained highest value of R-1 (0.50, 0.71, 0.58) for all three days compared to all other compared approaches but gets R-2 score (0.40) on 15th Feb lower than the R-2 score (0.45) of Follower voting approach and also gets highest R-L scores (0.47, 0.70, 0.57). Overall, proposed approach gets highest scores for all ROUGE metrics and Re-tweet voting approach gets lowest scores R-1 (0.20, 0.12, 0.16), R-2 (0.17, 0.08, 0.13) and R-L (0.19, 0.09, 0.15) because news containing negative sentiments propagate very rapidly. The Follower voting approach gives better scores compared to Lexrank and Textrank approach because more popular users can not give any random information on the public social platform. The proposed approach achieved better performance for both datasets compared to all baseline ranking approaches because of the capability of capturing semantically similar tweets.

Resultant top-ranked tweets of proposed approach for the first time-window of second dataset is presented in Figure 2 and compared approaches are presented in Figure 3, 4, 5, and 6 respectively. In Figure 7, a manual ground-truth summary is presented. The third experiment compares the Human judgment score given to the proposed approach summary and compared approaches summary. Detailed view of Human judgment score given to each tweet corresponding to each resultant summary is shown in Table 12. In Figure 2, the first tweet obtains the (0.8, 0.7) Human score because it contains winner candidate information but the second tweet gets the low score (0.4, 0.5) because it's a congratulation message. So, tweets containing vote statistics information get the higher score in compared to those tweets which contain normal statement or redundant information related to election. The proposed approach contains 8 vote statistics messages whereas Textrank contains 7, Lexrank contains 5, Follower voting contains 2 and Re-tweet voting contains no relevant message related to statistics of general election 2019. So, Re-tweet voting approach gets the worst Human judgment scores. The average Human judgment scores of the proposed approach corresponding to three time-windows for both datasets is (0.59, 0.70, 0.67) and (0.65, 0.55, 0.42) respectively which are higher than the compared approaches shown in Table 13. We achieve the best score in terms of both metric's (ROUGE and Human judgment score) for the proposed approach compared to other approaches.

Table 8. Euclidean distance between keywords by using four pre-trained Word2Vec model

	Google-News	Wiki-News	Wiki-News-Subword	Crawl-News-Subword
(BJP, Modi)	2.55	2.20	**0.58**	2.04
(Election, Voting)	3.08	1.9	**0.31**	0.73
(Win, Victory)	1.85	1.53	**0.54**	1.14
(Attack, Protest)	3.6	2.08	**0.35**	0.84
(Rahul, Congress)	3.4	2.9	**0.49**	1.55
(Speak, Deliver)	3.27	2.07	**0.39**	0.91
(Media, Press)	2.13	1.55	**0.34**	0.90
(Terrorist, Terrorism)	2.44	1.58	**0.19**	0.52
(Killed, Died)	2.44	1.67	**0.46**	0.95
(Terrorist, Attack)	4.00	2.33	**0.34**	0.81

Table 9. Data statistics

	Number of tweets 14, 511, 78			Number of tweets (23rd May) 15, 04, 348	
	14 Feb	1, 49, 924			
	15 Feb	3, 77, 925		1 hour	1, 72, 673
	16 Feb	2, 41, 465		2 hour	2, 30, 202
	17 Feb	2, 26, 489		3 hour	2, 75, 728
	18 Feb	79, 447		4 hour	2, 82, 158
Terrorist attack in	19 Feb	51, 178	**General election**	5 hour	2, 33, 232
Pulwama, Kashmir,	20 Feb	41, 013	**2019 in India**	6 hour	2, 37, 807
India	21 Feb	53, 358		7 hour	72, 548
	22 Feb	47, 957			
	23 Feb	47, 217			
	24 Feb	46, 922			
	25 Feb	47, 104			
	26 Feb	47, 133			
	27 Feb	50, 078			
	28 Feb	47, 117			
	29 Feb	46, 775			

CONCLUSION AND FUTURE WORK

This paper has proposed Semrank: a graph-based ranking approach based on semantic similarity instead of lexical similarity for ranking of tweets to generate an event summary from the real stream of tweets. The proposed approach can effectively handle issues like vocabulary mismatch and diversity of the tweets for the specific event while existing techniques like Lexrank and Textrank fail to do so. Another main contribution is aggressive filtering of noisy tweets based on social conversational features like re-tweet count, followers count, and URL frequency features which help users to get more popular and authentic information when using Twitter. The experimental results obtained show that the proposed approach can generate more news-worthy summaries in compared to other baseline approaches. Although the proposed approach has been applied on Twitter news stories, however, it can also be used for news stories coming from other sources. One important aspect of this method is the power of handling multiple languages by using language-specific trained Word2Vec model. In the future, this approach could be applied to other language tweets by using the word embedding model trained on the corpus of that language.

Table 10. ROUGE score metric comparison on Pulwama terrorist attack event dataset

	SemRank			TextRank			LexRank			Follower Voting			Re-tweet Voting		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
15 Feb	**0.50**	0.40	0.47	0.36	0.28	0.34	0.30	0.23	0.27	0.49	**0.45**	**0.48**	0.20	0.17	0.19
16 Feb	**0.71**	**0.67**	**0.70**	0.41	0.26	0.36	0.42	0.34	0.38	0.63	0.55	0.59	0.12	0.08	0.09
17 Feb	**0.58**	**0.52**	**0.57**	0.20	0.08	0.18	0.43	0.35	0.39	0.28	0.21	0.27	0.16	0.13	0.15

Table 11. ROUGE score metric comparison on General election 2019 event dataset

	SemRank			TextRank			LexRank			Follower Voting			Re-tweet Voting		
23 May	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
1 hour	**0.39**	**0.27**	**0.37**	0.26	0.13	0.34	0.26	0.13	0.23	0.32	0.23	0.29	0.13	0.07	0.12
2 hour	**0.41**	**0.35**	**0.39**	0.41	0.29	0.35	0.32	0.24	0.28	0.23	0.25	0.29	0.10	0.06	0.07
3 hour	**0.35**	0.26	**0.34**	0.29	0.20	0.25	0.31	**0.27**	0.30	0.24	0.21	0.22	0.13	0.05	0.09

Table 12. Human judgment score comparison on one-hour time-window of General election 2019 dataset

Id	SemRank		TextRank		LexRank		Retweet Voting		Follower Voting	
	H-1	H-2	H-1	H-2	H-1	H-2	H-1	H-2	H-1	H-2
0	0.8	0.7	0.8	0.8	0.5	0.5	0.1	0.1	0.1	0.1
1	0.4	0.5	0.8	0.6	0.3	0.4	0.1	0.2	0.9	0.8
2	0.6	0.6	0.7	0.6	0.2	0.2	0.1	0.1	0.3	0.2
3	0.7	0.6	0.3	0.4	0.6	0.7	0.2	0.1	0.1	0.2
4	0.4	0.3	0.5	0.5	0.4	0.5	0.1	0.1	0.2	0.3
5	0.4	0.4	0.6	0.4	0.4	0.3	0.1	0.2	0.4	0.5
6	0.7	0.7	0.2	0.3	0.8	0.8	0.1	0.3	0.5	0.5
7	0.8	0.7	0.7	0.8	0.6	0.6	0.3	0.2	0.2	0.3
8	0.8	0.6	0.5	0.5	0.8	0.6	0.1	0.2	0.5	0.5
9	0.7	0.6	0.7	0.7	0.3	0.4	0.2	0.3	0.5	0.5
10	0.9	0.8	0.2	0.3	0.6	0.7	0.3	0.1	0.7	0.6

Table 13. Average human judgment score comparison

		SemRank H-1	SemRank H-2	TextRank H-1	TextRank H-2	LexRank H-1	LexRank H-2	Re-tweet voting H-1	Re-tweet voting H-2	Follower voting H-1	Follower voting H-2
Pulwama Attack (Feb)	15	**0.59**	0.57	0.49	0.48	0.51	0.52	0.29	0.31	0.40	0.37
	16	0.69	**0.70**	0.61	0.60	0.45	0.47	0.31	0.29	0.37	0.41
	17	**0.67**	0.64	0.30	0.35	0.52	0.48	0.44	0.41	0.32	0.25
Election 2019 (23rdMay)	1 h	0.60	**0.65**	0.54	0.53	0.50	0.51	0.14	0.17	0.40	0.40
	2 h	**0.55**	0.52	0.42	0.44	0.51	0.50	0.25	0.21	0.37	0.35
	3 h	**0.42**	0.41	0.38	0.39	0.40	0.40	0.31	0.33	0.34	0.32

Figure 2. Result of proposed Semrank approach

	Top ranked tweets	SemRank-score
0	#Election2019Results | BJP candidate from #Dharwad, Pralhad Joshi, who has a huge margin against #Congress' Vinay Kulkarni has begun celebrations with other party leaders. Follow LIVE updates of #Verdict2019 in the south states here: https://t.co/5tqcrlEwuG https://t.co/r1WyhhFowi	1.236584
1	Prime Minister of Israel, Benjamin Netanyahu congratulates PM @narendramodi says, "will continue to strengthen our friendship between India and Israel". #VerdictWithTimes #ElectionResults2019 #Verdict2019 Follow live updates: https://t.co/AizmtYeI0W	1.232031
2	#BJP leaders in #Gujarat celebrated early trends which indicated a BJP victory. Amit Shah maintained a lead from Gandhinagar seat since beginning of vote counting for #LokSabhaElections2019. Follow LIVE updates on polling in the western states: https://t.co /AVHk4XPKf1 https://t.co/fdsJbcwcqG	1.229345
3	#Verdict2019 | Rahul Gandhi's margin in #Wayanad has widened to a record 2 lakh, This is the biggest margin by which an MP will be elected in this relatively new #Kerala constituency. Follow LIVE updates of the #Election2019Results in the southern states: https://t.co /uIV2vdMRML https://t.co/PjH3vjH7UZ	1.225631
4	Congress candidate Urmila Matondkar alleges EVM change, says signatures and the machine numbers different Catch all LIVE #LokSabhaElectionresults updates here #Verdict2019 #ElectionsWithBS #ElectionResults2019 https://t.co/Yh4QGlNfTq	1.225572
5	Sri Lanka PM Ranil Wickremesinghe congratulates PM @narendramodi #VerdictWithTimes #ElectionResults2019 #Verdict2019 Follow live updates: https://t.co/AizmtYeI0W https://t.co/Gn1Io3GYMq	1.215700
6	#Medinipur Lok Sabha Election Results 2019 West Bengal: BJP's #DilipGhosh takes lead over TMC's #ManasBhuniya #ElectionResults2019 #Verdict2019 #loksabhaElections2019results https://t.co/Zj2yTJX2cc	1.215697
7	#ResultsOnIndiaToday Manoj Tiwari leading by 1,98,524 in North Delhi Read live updates: https://t.co/wIIZ05HDhm #ElectionResults2019 #LokSabhaElections2019 https://t.co/RpkQdBkdih	1.213114
8	BJP leads inch closer to 300, leaders congratulate PM Narendra Modi Read live updates: https://t.co/1RW6pnHsfX #ResultsOnIndiaToday #ElectionResults2019 https://t.co/4onUfTTS99	1.212956
9	INC's Urmila Matondkar finally crosses 1 lakh mark. She's currently trailing with 1.03 lakh votes. BJP's Gopal Shetty is leading with 2.83 lakh votes Follow live updates: https://t.co/AizmtYeI0W #VerdictWithTimes #ElectionResults2019 #Verdict2019 https://t.co/jqJ6IhS4h1	1.207130
10	election commission reports that Prime Minister Narendra B.J.P. is leading the vote count in 292 of the 543 available seats in Parliament. At this pace, it would not need its other coalition partners to stay in control. https://t.co/3kcoh7g4Fo	1.206917

Figure 3. Result of Textrank approach

	Top ranked tweets	TextRank-score
0	#ElectionsWithHT | BJP is leading in 28 of the 29 seats in Madhya Pradesh Follow LIVE updates here: https://t.co/hqQSkilnEM #ResultsWithHT #Verdict2019 #ElectionResults2019 https://t.co/Y1zmr7A07f	1.315895
1	BJP leads inch closer to 300, leaders congratulate PM Narendra Modi Read live updates: https://t.co/1RW6pnHsfX #ResultsOnIndiaToday #ElectionResults2019 https://t.co/4onUfTTS99	1.283521
2	#Kangra Lok Sabha election results 2019 Himachal Pradesh: BJP's Kishan Kapoor leading by hefty numbers #ElectionResults2019 #Verdict2019 #loksabhaElections2019results https://t.co/G6uBSpBwHQ	1.245963
3	BJP-led NDA set to sweep Lok Sabha polls with over 300 seats, trends show #VerdictWithTimes #ElectionResults2019 #Verdict2019 Follow live updates: https://t.co/mrsFOXFmTg https://t.co/cjaR8i6ERM	1.245354
4	Before I leave for HILLS, let me tell u what my estimate is Modi Ji is PM UP - 65 to 70 could be even 75 Bengal, https://t.co/hsILEemT27	1.234824
5	Facts: Public approval rate of President Trump: 51% Public approval rate of Congress: 20% Only 4% of Americans https://t.co /7ZhrDbefGr	1.200817
6	#VerdictWithTimes #ElectionResults2019 #Verdict2019 PM Modi's work has defeated 'negative' politics of opposition: BJP Read: https://t.co/AHzHfdIX45 https://t.co/7CRP9M6sOA	1.191762
7	Patna Sahib, Bihar: Ravi Shankar Prasad (BJP) is leading ahead of Shatrughan Sinha (Congress) by 1,33,959 votes Follow live updates: https://t.co/AizmtYeI0W #VerdictWithTimes #ElectionResults2019 #Verdict2019	1.187770
8	#ResultsOnZee: Wishes pouring in from other countries as well. Sri Lankan PM Ranil Wickremesinghe tweets his congratulatory message to PM Modi #AbkiBaarKiskiSarkar https://t.co/kFUKcEURXM	1.168506
9	BJP's Jay Panda is trailing by over 2,000 votes from Odisha's Kendrapara #VerdictWithTimes #ElectionResults2019 #Verdict2019 Follow live updates on: https://t.co/AizmtYeI0W https://t.co/lmxV8ZFVNo	1.165838
10	I join my colleagues in calling on House Democrats to work with us to enact #USMCA. Implementing USMCA means https://t.co /Tj0tmUsJ36	1.139118

Figure 4. Result of Lexrank approach

	Top ranked tweets	LexRank-score
0	#ResultsOnZee: Wishes pouring in from other countries as well. Sri Lankan PM Ranil Wickremesinghe tweets his congratulatory message to PM Modi #AbkiBaarKiskiSarkar https://t.co/kFUKcEURXM	0.167082
1	This was six days before the very same EVMs delivered him 67 seats out of 70. He accepted the verdict gleefully. https://t.co/IsEAk64BQp	0.166305
2	#ResultsWithNDTV | Election trends at 12:03 pm. https://t.co/oOz20mcmkH https://t.co/2a4ttsFOwJ	0.165805
3	New Monmouth poll numbers: * 67% say Don McGahn should testify before Congress * 73% say Mueller should testify https://t.co /ZtDCt2nUYj	0.165712
4	PM of Israel @netanyahu congratulates PM @narendramodi #VerdictWithTimes #ElectionResults2019 #Verdict2019 Follow live updates: https://t.co/AizmtYeI0W https://t.co/TkvnuxQVXZ	0.165185
5	BJP candidate from Begusarai @girirajsinghbjp speaks to India Today. Listen in to what he said #ResultsOnIndiaToday #ElectionResults2019 (@rahulkanwal/@sardesairajdeep) Watch LIVE at https://t.co/4fqxBVUizL https://t.co/RzaDRWxyi5	0.164759
6	#ElectionsWithHT | BJP is leading in 28 of the 29 seats in Madhya Pradesh Follow LIVE updates here: https://t.co/hqQSkilnEM #ResultsWithHT #Verdict2019 #ElectionResults2019 https://t.co/Y1zmr7A07f	0.164491
7	#UttarPradesh #ElectionResults2019 LIVE: Will Smriti Irani trounce Rahul Gandhi in Amethi this time? https://t.co/X1iOBNErO5 https://t.co/PG0Efwmuln	0.161962
8	BJP-led NDA set to sweep Lok Sabha polls with over 300 seats, trends show #VerdictWithTimes #ElectionResults2019 #Verdict2019 Follow live updates: https://t.co/mrsFOXFmTg https://t.co/cjaR8i6ERM	0.161170
9	BJP leads inch closer to 300, leaders congratulate PM Narendra Modi Read live updates: https://t.co/1RW6pnHsfX #ResultsOnIndiaToday #ElectionResults2019 https://t.co/4onUfTTS99	0.160495
10	#Verdict2019WithNews18 @AamAadmiParty leading on 1 seat in Punjab. #ElectionsWithNews18 LIVE updates: https://t.co/dbbVGq9WBt https://t.co/uSTzhcpoo3	0.159976

Figure 5. Result of Follower voting approach

	Top ranked tweets	Follower-score
0	What does Europe mean to Europeans today? As Europeans go to the polls in a Continental election, our reporter set out on a 10-day journey to find out. https://t.co/QLaEVmJBKg	0.027321
1	election commission reports that Prime Minister Narendra B.J.P. is leading the vote count in 292 of the 543 available seats in Parliament. At this pace, it would not need its other coalition partners to stay in control. https://t.co/3kcoh7g4Fo	0.027321
2	India's election has left the country more divided than ever | Analysis by CNN's @nkreports https://t.co/o8J0cpFIj7	0.026389
3	Judge rules against Trump to pave way for banks to provide his business records to Congress https://t.co/ZOvIIAHol8 via @ReutersTV https://t.co/eqHyMSZyST	0.012866
4	#VerdictWithTimes #ElectionResults2019 #Verdict2019 PM Modi's work has defeated 'negative' politics of opposition: BJP Read: https://t.co/AHzHfdIX45 https://t.co/7CRP9M6sOA	0.007417
5	PM of Israel @netanyahu congratulates PM @narendramodi #VerdictWithTimes #ElectionResults2019 #Verdict2019 Follow live updates: https://t.co/AizmtYeI0W https://t.co/TkvnuxQVXZ	0.007417
6	#VerdictWithTimes #ElectionResults2019 #Verdict2019 Narendra Modi's ministers congratulate him for BJP's landslide victory Read: https://t.co/op7tSTJxiG https://t.co/t6AIAoqplg	0.007417
7	Prime Minister of Israel, Benjamin Netanyahu congratulates PM @narendramodi says, "will continue to strengthen our friendship between India and Israel". #VerdictWithTimes #ElectionResults2019 #Verdict2019 Follow live updates: https://t.co/AizmtYeI0W	0.007417
8	Sri Lanka PM Ranil Wickremesinghe congratulates PM @narendramodi #VerdictWithTimes #ElectionResults2019 #Verdict2019 Follow live updates: https://t.co/AizmtYeI0W https://t.co/Gn1Io3GYMq	0.007417
9	"Congratulations to the winners. But all losers are not losers", tweets West Bengal Chief Minister @MamataOfficial on #LokSabhaElections2019 results #VerdictWithTimes #ElectionResults2019 #Verdict2019 Follow live updates: https://t.co/AizmtYeI0W https://t.co/WvDawchcOy	0.007417
10	INC's Urmila Matondkar finally crosses 1 lakh mark. She's currently trailing with 1.03 lakh votes. BJP's Gopal Shetty is leading with 2.83 lakh votes Follow live updates: https://t.co/AizmtYeI0W #VerdictWithTimes #ElectionResults2019 #Verdict2019 https://t.co/jqJ6IhS4h1	0.007417

Figure 6. Result of Re-tweet voting approach

	Top ranked tweets	retweet-score
0	Modi Ji, The battle is over. Your Karma awaits you. Projecting your inner beliefs about yourself onto my father https://t.co/Q8cxPQtToy	0.013608
1	Modi's win was a rejection not only of the dynasty but of the colonial Brits, who tried to teach Hindus to be ashamed of their culture.	0.009783
2	I subsequently learned from the Mueller report that there was a lot more evidence of collusion and obstruction of https://t.co/lYP3cC7y0G	0.009501
3	Pre shading of ballots in Lanao del sur IF THIS SHIT AINT ENOUGH TO CONVINCE THAT THIS ELECTION WAS A https://t.co /cUkpOS1vdw	0.008099
4	Do you ever wonder whether PM Modi manages to laugh during the heat of the election campaign? get the https://t.co/JfqCHyOa2M	0.007813
5	John Brennan on the Mueller probe, know if I received bad information, but I THINK I SUSPECTED THAT https://t.co/51a2Gnw9CG	0.007796
6	Megan King, who is running for Superior Court Judge in the Pennsylvania election, has my Full and Total https://t.co/wb7aIetiMf	0.007794
7	Barr says campaign was upon. Trump claims treason. Both are incendiary. Neither is true. Barr https://t.co/gGfGOuSgVu	0.007723
8	Trump just ordered Don McGahn not to testify before the American people. Duplicitous as ever, the White House https://t.co/TFySzgpgN2	0.007400
9	My Administration will be fighting for $200 million for the Army Corps Everglades restoration work this year. https://t.co/XTPBS8T6ox	0.007066
10	Every day Trump prevents anyone from his administration from testifying to Congress is another day he commits https://t.co /I524voHYNz	0.006882

Figure 7. Ground-truth summary

	Human ranked tweets
0	election commission reports that Prime Minister Narendra B.J.P. is leading the vote count in 292 of the 543 available seats in Parliament. At this pace, it would not need its other coalition partners to stay in control. https://t.co/3kcoh7g4Fo
1	#ElectionsWithHT | BJP is leading in 28 of the 29 seats in Madhya Pradesh Follow LIVE updates here: https://t.co/hqQSkilnEM #ResultsWithHT #Verdict2019 #ElectionResults2019 https://t.co/Y1zmr7A07f
2	#ResultsOnIndiaToday Manoj Tiwari leading by 1,98,524 in North Delhi Read live updates: https://t.co/wIIZ05HDhm #ElectionResults2019 #LokSabhaElections2019 https://t.co/RpkQdBkdih
3	#Kangra Lok Sabha election results 2019 Himachal Pradesh: BJP's Kishan Kapoor leading by hefty numbers #ElectionResults2019 #Verdict2019 #loksabhaElections2019results https://t.co/G6uBSpBwHQ
4	#Medinipur Lok Sabha Election Results 2019 West Bengal: BJP's #DilipGhosh takes lead over TMC's #ManasBhuniya #ElectionResults2019 #Verdict2019 #loksabhaElections2019results https://t.co/Zj2yTJX2cc
5	Patna Sahib, Bihar: Ravi Shankar Prasad (BJP) is leading ahead of Shatrughan Sinha (Congress) by 1,33,959 votes Follow live updates: https://t.co /AizmtYeI0W #VerdictWithTimes #ElectionResults2019 #Verdict2019
6	#Election2019Results | BJP candidate from #Dharwad, Pralhad Joshi, who has a huge margin against #Congress' Vinay Kulkarni has begun celebrations with other party leaders. Follow LIVE updates of #Verdict2019 in the south states here: https://t.co/5tqcrlEwuG https://t.co/r1WyhhFowi
7	In Odisha, Pinaki Misra of BJD is leading from Puri. #ResultsWithAIR | #ElectionResults2019
8	#Almora Lok Sabha 2019 results: Ajay Tamta dominates on counting day for BJP #ElectionResults2019 #Verdict2019 #loksabhaElections2019results https://t.co/xCCdM0RH2K
9	INC's Urmila Matondkar finally crosses 1 lakh mark. She's currently trailing with 1.03 lakh votes. BJP's Gopal Shetty is leading with 2.83 lakh votes Follow live updates: https://t.co/AizmtYeI0W #VerdictWithTimes #ElectionResults2019 #Verdict2019 https://t.co/jqJ6IhS4h1
10	In Azamgarh, SP's @yadavakhilesh leads by 52,009 over BJP's Dinesh Lal Yadav 'Nirahua' #VerdictWithTimes #ElectionResults2019 #Verdict2019 Follow live updates: https://t.co/AizmtYeI0W https://t.co/V32NnuSsUa

REFERENCES

Abel, F., Gao, Q., Houben, G. J., & Tao, K. (2011, July). Analyzing user modeling on twitter for personalized news recommendations. In *International Conference on User Modeling, Adaption, and Personalization* (pp. 1-12). Springer. doi:10.1007/978-3-642-22362-4_1

Aggarwal, C. C., Han, J., Wang, J., & Yu, P. S. (2003, September). A framework for clustering evolving data streams. In *Proceedings of the 29th international conference on Very large data bases-Volume 29* (pp. 81-92). VLDB Endowment. doi:10.1016/B978-012722442-8/50016-1

Aggarwal, S. (2014). *Assessment and Process Automation of Two Success Factors for Websites: Usability and Credibility* (Doctoral dissertation). International Institute of Information Technology, Hyderabad.

Agichtein, E., Brill, E., & Dumais, S. (2006, August). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 19-26). ACM. doi:10.1145/1148170.1148177

Alsaedi, N., Burnap, P., & Rana, O. (2016, March). Automatic summarization of real world events using twitter. *Tenth International AAAI Conference on Web and Social Media*.

Arthur, D., & Vassilvitskii, S. (2007, January). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027-1035). Society for Industrial and Applied Mathematics.

Belkaroui, R., & Faiz, R. (2017). Conversational based method for tweet contextualization. *Vietnam Journal of Computer Science*, *4*(4), 223–232. doi:10.1007/s40595-016-0092-y

Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113-120). ACM.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1-7), 107-117.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010, May). Measuring user influence in twitter: The million follower fallacy. *Fourth international AAAI conference on weblogs and social media*.

Chin, J. Y., Bhowmick, S. S., & Jatowt, A. (2019). On-demand recent personal tweets summarization on mobile devices. *Journal of the Association for Information Science and Technology*, *70*(6), 547–562. doi:10.1002/asi.24137

He, R., & Duan, X. (2018, April). Twitter Summarization Based on Social Network and Sparse Reconstruction. *Thirty-Second AAAI Conference on Artificial Intelligence.*

Inouye, D. (2010). Multiple post microblog summarization. *REU Research Final Report*, *1*, 34–40.

Kennedy, J., & Eberhart, R. C. (1997, October). A discrete binary version of the particle swarm algorithm. *1997 IEEE International conference on systems, man, and cybernetics.* doi:10.1109/ICSMC.1997.637339

Khan, M. A. H., Bollegala, D., Liu, G., & Sezaki, K. (2013, September). Multi-tweet summarization of real-time events. In *2013 International Conference on Social Computing* (pp. 128-133). IEEE.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). From word embeddings to document distances. In *International conference on machine learning* (pp. 957-966). Academic Press.

Kywe, S. M., Lim, E. P., & Zhu, F. (2012, December). A survey of recommender systems in twitter. In *International Conference on Social Informatics* (pp. 420-433). Springer. doi:10.1007/978-3-642-35386-4_31

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81). Academic Press.

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404-411). Academic Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781

Naik, S., Lade, S., Mamidipelli, S., & Save, A. (2018, April). Tweet Summarization: A New Approch. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1022-1025). IEEE.

Nichols, J., Mahmud, J., & Drews, C. (2012, February). Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces* (pp. 189-198). ACM.

Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543). doi:10.3115/v1/D14-1162

Phelan, O., McCarthy, K., Bennett, M., & Smyth, B. (2011, April). Terms of a feather: Content-based news recommendation and discovery using twitter. In *European Conference on Information Retrieval* (pp. 448-459). Springer. doi:10.1007/978-3-642-20161-5_44

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*(2), 99–121. doi:10.1023/A:1026543900054

Sharifi, B., Hutton, M. A., & Kalita, J. (2010, June). Summarizing microblogs automatically. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 685-688). Association for Computational Linguistics.

Sharifi, B., Hutton, M.-A., & Kalita, J. (2010). Automatic summarization of twitter topics. *National Workshop on Design and Analysis of Algorithm*.

Song, X., Srimani, P. K., & Wang, J. Z. (2019, June). HWE: Hybrid Word Embeddings For Text Classification. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval* (pp. 25-29). ACM. doi:10.1145/3342827.3342837

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406–418. doi:10.1002/asi.21462

Tripathi, A. K., Sharma, K., & Bala, M. (2019). Parallel Hybrid BBO Search Method for Twitter Sentiment Analysis of Large Scale Datasets Using MapReduce. *International Journal of Information Security and Privacy*, *13*(3), 106–122. doi:10.4018/IJISP.201907010107

Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E., & Li, H. (2010, July). Context-aware ranking in web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 451-458). ACM.

Xu, W., Grishman, R., Meyers, A., & Ritter, A. (2013, June). A preliminary study of tweet summarization using information extraction. In *Proceedings of the Workshop on Language Analysis in Social Media* (pp. 20-29). Academic Press.

Yang, X., Ghoting, A., Ruan, Y., & Parthasarathy, S. (2012, August). A framework for summarizing and analyzing twitter feeds. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 370-378). ACM. doi:10.1145/2339530.2339591

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *8*(4), e1253. doi:10.1002/widm.1253

Zhao, W. X., Wen, J. R., & Li, X. (2016). Generating timeline summaries with social media attention. *Frontiers of Computer Science*, *10*(4), 702–716. doi:10.1007/s11704-015-5145-3

Jagrati Singh received her B.Tech and M.Tech degree in Computer Science from the Banasthali University, Rajasthan in 2012 and 2014 respectively. In 2015, she was taken admission as the Research Scholar in the Department of Computer Science and Engineering at Motilal Nehru National Institute of Technology Allahabad.

Anil Kumar Singh has done his Bachelor's in Science in 1990 and Masters in Computer Application (MCA) in 1994 from the University of Lucknow. He has also done his Masters in Engineering (Computer Sc. and Engg.) from University of Allahabad in 2001. He got his Ph.D. (Computer Science and Engineering) degree from Indian Institute of Technology Roorkee, India in 2012. He is currently working as Professor in the Department of Computer Science and Engineering. Dr. Singh has his research interests and publications in the areas of Data Mining, Information Retrieval, Machine Learning and Big Data Analytics. He has been awarded Commonwealth Scholarship, 2004 CANADA by MHRD, INDIA. He has also received AICTE fellowship (under QIP) during PhD.