J Inf Process Syst, Vol.14, No.1, pp.79~100, February 2018
https://doi.org/10.3745/JIPS.02.0079

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A Survey on Automatic Twitter Event Summarization

Dwijen Rudrapal*, Amitava Das**, and Baby Bhattacharya***

Abstract

Twitter is one of the most popular social platforms for online users to share trendy information and views on any event. Twitter reports an event faster than any other medium and contains enormous information and views regarding an event. Consequently, Twitter topic summarization is one of the most convenient ways to get instant gist of any event. However, the information shared on Twitter is often full of nonstandard abbreviations, acronyms, out of vocabulary (OOV) words and with grammatical mistakes which create challenges to find reliable and useful information related to any event. Undoubtedly, Twitter event summarization is a challenging task where traditional text summarization methods do not work well. In last decade, various research works introduced different approaches for automatic Twitter topic summarization. The main aim of this survey work is to make a broad overview of promising summarization approaches on a Twitter topic. We also focus on automatic evaluation of summarization techniques by surveying recent evaluation methodologies. At the end of the survey, we emphasize on both current and future research challenges in this domain through a level of depth analysis of the most recent summarization approaches.

Keywords

ROUGE, Social Media Text, Tweet Stream, Tweet Summarization

1. Introduction

In recent days, the popularity of social media communication like Twitter has grown up at an unrivaled rate due to the increased use of ubiquitous devices. Anyone can post or react on anything instantly on Twitter. Since its inception to till September 2, 2017, total number of Twitter registered users are 390 million [1]. On an average, around 6,000 tweets are posted on Twitter in every second [2]. Statistics shows that Twitter has become one of the ten most visited websites and accepted as “the SMS of the Internet” [3].

In very recent time, Twitter reports first news on some events like earthquake in North East India, missing of M327 flight, the United States Presidential Election, UEFA League status, etc., and many more even before the traditional news media. Regarding recent presidential election of the United States, Twitter has claimed largest source of breaking news with 40 million tweets as on November 8, 2016 [4]. Even though, enormous number of tweets shared for any event, very few are topic relevant

Manuscript received November 8, 2017; accepted December 26, 2017.

Corresponding Author: Dwijen Rudrapal (dwijen.rudrapal@gmail.com)
* Dept. of Computer Science and Engineering, National Institute of Technology, Agartala, India (dwijen.rudrapal@gmail.com)
** Dept. of Computer Science and Engineering, Indian Institute of Information Technology, Sricity, Andhra Pradesh, India (amitava.santu@gmail.com)
*** Dept. of Mathematics, National Institute of Technology, Agartala, India (babybhatt75@gmail.com)

and resourceful. The reasons are mostly regarding the nature of Twitter text. In the following subsections we discuss the distinctiveness of tweet text and current overview of Twitter topic summarization problem.

1.1 Tweet Characteristics

A tweet often publishes without proofreading and consequently it includes self created abbreviations, acronyms, out of vocabulary words and grammatically wrong expressions. Sometimes, users post completely irrelevant information or continue online discussion towards an event by posting one/two words tweet. Due to the diverse nature of text, several obstacles stand on the way of processing tweets.

Firstly, users often misspell words either deliberately or accidentally by expanding (sooooo, Goooooood) and abbreviating words (alrdy, c, congo.), using lexical/numeric substitutions (b4, 4u), nonstandard acronyms such as (lol, smh). User also use of Hashtags (#demonetization2016), user tags (@gamini1980) or Twitter specific terminologies (RT mentioning retweet) and trending topics (TT) frequently. This writing style poses problems for standard natural language processing (NLP) techniques for preprocessing tweets. Normalization of tweets into standard text as proposed in the work [5,6] makes standard NLP tools applicable on social media text up to a certain extent. But these approaches do not make the performance as effective as on traditional text. Thus, retrieval of actual content from tweet remains a difficult problem.

Secondly, tweets are limited to 140 characters. Thus, tweets are feature sparse and challenging for clustering where features are highly perceptive for various measures like string comparison, topic modeling where sufficient bag of words need to be corresponded the topic texts.

Thirdly, sentence boundary detection (SBD) in a tweet [7] is also a challenging task. In many NLP applications like tweet summarization, SBD is an essential prerequisite. While tweeting, user applies punctuation in creative and nonstandard ways, which render great challenges for SBD. Sometimes, tweet includes no punctuation at all even though it includes multiple sentences. For example, below given tweet includes two sentences.

Tweet: *This's the LAST SUNDAY before #USElections By next weekend we'll know whether the orange grinch or the drone loving woman will be president*

Sentence 01: *This's the LAST SUNDAY before #USElections*

Sentence 02: *By next weekend we'll know whether the orange grinch or the drone loving woman will be president*

In spite of these challenges, any event on Twitter is very important and attracts NLP researchers towards Twitter topic summarization problem due to its fast and rich information.

1.2 Twitter Topic Summarization: An Overview

Summarization is the process of representing a text document with desired length of text that retains the most important points of the original document. A summary includes original sentences or reinterpreted sentences from the source document. Twitter topic summarization task is to produce a summary of a topic with desired number of tweets. Say, a stream of tweets $T = [t_1; t_2; .....; t_m]$ are related to a topic, then set of tweets $S = [s_1; s_2; ...; s_n]$ represents the summary of the topic with maximum amount of essential information.

It is very difficult to keep track of one event in Twitter due to enormous volume of tweets. Nevertheless, Twitter.com provides tools to find the most important topics and related tweets for a particular moment or time period. Any user can see tweets on desired event through search by topic related words. Tool fetches all the query related tweets in order of posting time. The retrieved list includes many nonrelated or other language tweets also. To get the desired tweets, a rigorous manual filtration process is required. Eventually, researchers propose various summarization algorithms to retrieve the most important information from the topic relevant tweets to generate a robust summary.

In this survey work, we primarily aim to investigate how empirical methods have been used to build summarization systems for tweet streams. We also highlight on the evaluation approaches by different research works to establish quality summary generation. To the best of our knowledge, this is the first comprehensive study that investigates Twitter topic summarization approaches. The paper is organized as follows: Section 2 describes taxonomy followed by Section 3 which describes different state-of-the-art tweet summarization approaches and Section 4 highlights evaluation approaches. In Section 5, we discuss the research challenges towards Twitter topic summarization and finally conclude the survey in Section 6.

2. Taxonomy

The solution to the problem of Twitter topic summarization, involves some basic theories and assumptions. The basic terms and definitions which are frequently used in different tweet summarization algorithms and evaluation methodologies are discuss in this section.

Event: Event is a real time or historical topic in Twitter, which attracts set of tweets related to the topic during a short span of time that is at the moment when the topic is issued or raised in Twitter.

Sub-event: A sub-event corresponds to a group of tweets that emerge from the relevant tweet stream, being exhaustively discussed during a short period time and then slowly fades away.

Generic summaries: The source of generic summaries are not related with any specific class of event like politics, crisis etc. and hence all the tweets are homogeneous texts.

Domain specific summaries: Domain specific summaries are related with the events like sports, crisis events etc. and others. These Twitter events share common characteristics and easier to gather knowledge from past similar type of events for prediction of important moment or detection of similar tweets towards clustering process.

Tweet graph: A tweet graph is denoted by $G = (V, E)$, constructed from the related tweet streams V which represents a word or a phrase or the complete tweet while E represents the relation between words or phrases or tweets. Fig. 1 represents a unigram tweet graph where each node except the start and end node represents unigram in a tweet and edges represent the relation of a unigrams with co-occurred unigrams in the tweet.

Multiview tweet graph: A multi-view tweet graph [8] stores semantic and temporal information among relevant tweets. Each graph comprises with 4 tuples, such as set of vertices (nodes), weights of nodes, set of undirected edges which represent the similarities between tweets and set of directed edges (arcs) which represent the time continuity of the tweets.

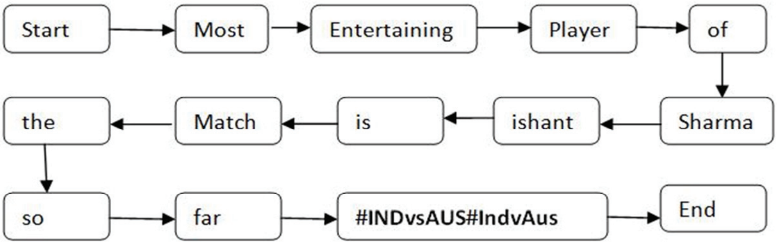

Fig. 1. Unigram words graph of a tweet.

Burstiness: Burstiness is the measure to quantify the burst nature of tweet stream that represents rapid tweets posting in short time periods alternating with long inactive periods.

Spike: A spike corresponds to an important moments of an active phase of an event. Each spike can be defined with start time, peak time and end time. Tweets segments are prepared based on the spike list of events.

Tweet cluster vector: Tweet cluster vector (TCV) [9] is a data structure to keep important tweet information in clusters. TCV for a cluster is defined with 6 tuples. These are sum of normalized textual vector, sum of weighted textual vector, sum of timestamps, the quadric sum of timestamps, number of tweets in the cluster and the number of tweets which are closest to the cluster centroid.

3. Tweet Summarization Approaches

Most of the Twitter topic summarization approaches are based on conventional summarization techniques formerly applied on traditional text documents like newswire text. These algorithms modified accordingly to deal with the challenges of informal text like tweets. Twitter topic summarization approaches can be categorized into two ways: one, based on summary content and another based on event category. Content based approaches again classified into two types: extractive (Extr.) and abstractive (Abstr.) summaries, event category based approaches is also classified into two types: generic (Gen.) and domain specific (Dom.) summaries. Fig. 2 shows the structure of different categories. Both the content based approaches summarize generic and domain specific events. On the

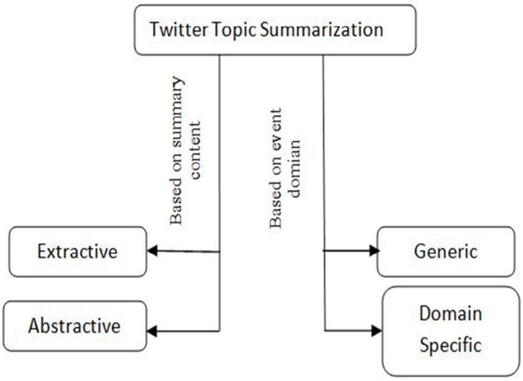

Fig. 2. Tree structure of Twitter topic summarization approaches.

other way, a generic or domain specific event also produces extractive or abstractive summary. In this survey work, we mainly focus on content based summarization approaches. Every approach applies different techniques like graph based (Gr), cluster based (Cl), statistical based (St) and other techniques (Ot) to generate summaries. The following subsections describe various techniques under each approach.

3.1 Extractive Summaries

Extractive summaries focus on identifying the important tweets to represent the event summary. Extractive summaries include original tweets from the source document without any changes. Initial research works on tweet summarization produces one tweet length extractive summary. With time, the research focuses on more information coverage of the event through multiple tweet summaries. In this section we discuss different techniques used to select most informative tweets from a stream of tweets for a given event.

3.1.1 Graph based approaches

Most of the tweets for any specific topic hold a relationship among them towards the topic theme. This relationship of tweets can be modeled through a graph where each tweet/phrase/word represents as node and the relation between them represents as edge. Nodes are weighted to measure the importance and accordingly incorporated into the generated summary.

In 2010, Sharifi et al. [10] introduced a “phrase reinforcement (PR) algorithm” to generate summary of one tweet length for any Twitter event. The algorithm chooses topic name or user specified phrase to retrieve all the related tweets from Twitter. These extracted tweets are preprocessed by filter out spam tweets, duplicate and non-English posts to obtain candidate tweets for summarization. These candidate tweets construct an ordered acyclic graph by representing starting phrase as root node and words that occur either immediately before or after the starting phase as adjacent to the root node. Every node weight measures by weighing the frequency of occurrence. The algorithm sum up the weight of every unique path from the root to each of the leaf nodes in order to find most weighted path as partial summary. Considering the partial summary as root, the algorithm builds another graph at opposite direction from the path found earlier with greatest weight. This new graph successfully constructs the final summary having maximum total weight path in the graph. Fig. 3 shows phrase graph of tweets where *P* represents the root phrase, W_i represents other phrases and n_j is the weight of each node.

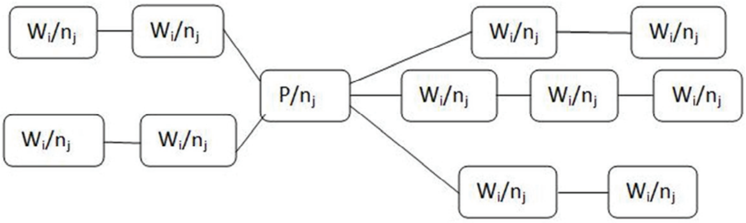

Fig. 3. Phrase graph of tweets using PR algorithm.

The PR algorithm generates summary with most weighted path or phrase. However, the most weighted path does not represent most important tweet always. Moreover, summary of 1 tweet length often does not cover the whole topic. To cover an event with more information as summary, the work by Sharifi et al. [11] extends PR algorithm by generating a summary of 4 tweets length. The work chooses 4 most weighted path, starting from the root node to leaf node from the generated phrase graph to produce summary of 4 tweets of an event. Another enhancement work [12] over PR algorithm focuses on the drawbacks of the algorithm like syntactic well formedness of sentence. Proposed work parse the summary produced by PR algorithm and build dependencies between the dependent and governor words in the summary. The dependencies are built by POS tagging of each word and its relation with other words. If any word does not have proper grammatical sense or not close to other related words then the word eliminated from PR algorithm output summary to make the summary grammatically more correct. The work by Nichols et al. [13] added one additional feature like user status updates with PR algorithm for domain specific Twitter event summarization, specifically for sports events. For any sports event, user use to update status at every important moment of the ongoing event. So, some specific time duration hold high volume of tweets. This time duration is called spike. Proposed system generates a list of spikes and constructs a phrase graph from the longest sentence in each spike. Top scored nodes in the generated graph produce the summary. Proposed algorithm evaluated on 36 games of the 2010 FIFA World Cup football and other type of games like baseball and shows better performance over existing algorithms. The approach by Harabagiu and Hickl [14] produces a fixed length extractive summary of 250 words. The system collects event related tweets by querying through Twitter. Features like named entities, event mentions, and interevent relationships retrieved from relevant tweets. An event mention is any predicate which makes reference to real world event while an event relation is semantic belongings to two or more event mentions. System calculates a relevance score for each tweet based on the number of times it appeared in the relevant tweet collection. From this relevance score of tweets, top 250 words produce the summary of the event. But, fixed length summary for Twitter event does not incorporate cohesive information. Due to the character limitation in twitter posts, some expressions are very sensitive to the context of its earlier tweets or related to another tweet. To overcome this issue, proposed summarization system [15] produces more cohesive summary based on the analysis of social network, tweet readability and user diversity. Social network analysis represents features like number of re-tweet and followers of the account that published the tweet. Readability of tweet measures the presence of out-of-vocabulary (OOV) words, abnormal symbols like non-ASCII character and many more. User diversity focuses on biasness of a particular user on the topic by discarding tweets after a certain limit. Based on these features, every tweet weighted using Eq. (1):

$$a(j) = retw(j).foll(j).readability(j) \tag{1}$$

For a tweet *j*, *retw*(*j*) counts retweets, *foll*(*j*) counts followers and *readability*(*j*) determines the readability score. Weighted tweets form a graph where vertex represents tweet and edges represents weight. The most weighted tweets select for the summary generation by using maximal marginal relevance (MMR) algorithm [16].

The proposed method [8] generates summary for an ongoing event through graph optimization.

Based on the burst periods in active phases of an event, system extracts relevant tweets by using a language model with dynamic pseudo-relevance feedback (DPRF) algorithm. For any observation time T', the burst score ($B(w, T)$) of a term (w) occurs for ($tf(w, T)$) times in time interval (T) is defined as in Eq. (2):

$$B(w,T) = \frac{tf(w,T)}{\sum_{T'} tf(w,T')} - \frac{|T|}{|\sum T'|} \tag{2}$$

Extracted relevant tweets constructs a multi-view tweet graph based on semantic and temporal information of the tweets. The system selects representative tweets from the graph by using minimum weight dominating set problem. Minimum weight dominating set problem finds smallest total weight path from start to leaf node in a weighted undirected graph. Selected tweet from each graph connects to produce summary in each phase of the event using minimum Steiner tree algorithm.

Classical PageRank algorithm [17] which is shown to be effective for the summarization of newswire domain text extends in the proposed work [18] by partitioning event graphs and detecting fine grained aspects of the event to generate event summary. The system extracts named entities and event phrases from tweets and constructs an undirected event graph by representing events as node and edges as the relationship between events. Each node (V_i) weight is computed by using Eq. (3). Here $ad(V_i)$ is the set adjacent vertices and w_{ij} is the weight between (V_i) and (V_j).

$$S(V_j) = (1-d) + d \,.\sum_{V_j \in ad(V_j)} \frac{w_{ij} S(V_j)}{\sum_{V_k \in ad(V_k)} w_{jk}} \tag{3}$$

Based on computed score, nodes are ranked and the graph is partitioned into sub-events. Finally, a tweet selects from each sub-event to produces the event summary. The work [19] combines popular discussion points of tweets with PageRank algorithm [17] to obtain event summary. All the event relevant tweets are divided into topic cluster by using latent Dirichlet allocation (LDA) [20]. For every cluster, a lexical graph is constructed with relevant tweets and then applies PageRank algorithm to determine the score of individual lexical units in the graph. Higher co-occurrences of terms make high scored lexical units. Maximum scored lexical units are the key discussion points of tweets and included into summary. Another traditional text summarization approach, TextRank algorithm [21] is extended in the work by Xu et al. [22]. System extracts bigrams from the relevant tweets and constructs a weighted graph with bag of bigrams. Each vertex in the graph is a bigram and each edge is the co-occurring times of two ordered bigrams within a fixed time window. For each vertex, a weight score is computed by using TextRank algorithm and rank the nodes in descending order to retrieve a set of candidate bigrams. From these bigrams, key bigrams extracted based on their local densities using LDA. Selection of most important tweets for summary made based on a ranking method applied by combining overlap similarity (OS) and mutual information (MI) techniques. OS technique is the count of overlap bigrams between the sentence and the key-bigram set divided by the key-bigram set size and MI measures the relevance between the words. Finally, by using Eq. (4)–(6), a preference score for each tweet is computed.

$$prior\ (s_i) = \lambda.mean\ (S_i) + (1-\lambda)\operatorname{var}(S_i) \tag{4}$$

$$mean\ (s_i) = \delta.ros\ (S_i) + (1-\delta) rmi\ (S_i) \tag{5}$$

$$\operatorname{var}(s_i) = \begin{bmatrix} \delta\,(ros\ (s_i) - mean\ (s_i))^2 \\ +\ (1-\delta)(rmi\ (S_i) - mean\ (s_i))^2 \end{bmatrix}^{-\frac{1}{2}} \tag{6}$$

In the equations, λ and δ are two parameters with values between and 1, *mean*(S_i) and *var*(S_i) are the mean and variance ranking values of sentence S_i, and *ros*(S_i) and *rmi*(S_i) are the ranks in OS and MI ranking results. The tweets holding more key bigrams selects for summary.

The summarization approach [23] constructs a graph where keywords from the relevant tweets form nodes and weight ($W_{u,v}$) between nodes forms edges. Each node is weighted by using Eq. (7).

$$W_{u,v} = \sqrt{\frac{n_{u,v}}{\max[\ n_{x,y}\ |\ x \in V,\ y \in V\ ]}} \tag{7}$$

where $n_{u,v}$ is the co-occurrence of words *u* and *v* in tweet. The graph is clustered into connected sub-graphs using maximal k-clique clustering algorithm. A sub-graph is a set of tweets which contains all the words in each maximal clique. Clusters are merged together to form final summary by removing redundant tweets.

User shared images are also focused in the proposed work [24] to generate visual summary of Twitter event. A visual summary includes highly relevant to the event and key images. The framework employs text and visual filter to discard low-quality posts and small images to build a multi-graph. The graph help to remove text duplicates while visual redundancy is eliminated by using the clique percolation method (CPM). Then the system detects the main event topics using structural clustering algorithm for networks (SCAN) and computes a selection score for each message. The significance score S_{sig} captures the social attention (S_{att}) that a message receives over time and the coverage (S_{cov}) of the corresponding topic by using Eq. (8).

$$S_{sig}(m) = S_{att}(m).S_{\mathrm{cov}}(m) \tag{8}$$

Finally, a graph based ranking algorithm selects images of the top ranked posts to produce summary. More user interactions and more long active conversations among a group of users are common features to identify an interesting topic in social media. Proposed summarization system by Qu et al. [25] detects hot topics and generates topic summary. The work propose a spreading tree model for online summarization by building a spreading tree with time-stamped tweet as node and the count of user interactions as node degree. A stream of interactions from one node to another node represents as associated time-stamped tweets. Information is diffused in dispersal tree from the root to the leave nodes. The work uses entropy to generate an informative summary from the tree.

The proposed work [26] generates summaries by using hybrid ant colony optimization with a mechanism of local search called “Ant Colony Optimization-Local Search-short text summarization

(ACOLSSTS)". Initially, the system uses graph coloring (GC) algorithm (GCISTS) to contract the solution area of ants into small sets. Proposed ACOLSSTS algorithm extracts the most interactive comments by following three steps. First, NLP module is employed to transform the comments in to a set of n-terms. Second, a graph coloring GCISTS module is exploited to gather the dissimilar or less similar comments together into the same color. Third, a cyclic semi-graph is constructed within each color by considering extremely lengthy comments which are isolated from the graph. Finally, the hybrid ACOLSSTS algorithm is applied to all colors in parallel form to discover important comments as summary.

3.1.2 Cluster based approaches

Since the inception of any event, flow of tweets varies with time. During important moments of the event, more tweets are post comparing the events whole life. Tweets posted during important moments share important information. Various summarization approaches use this trait to cluster tweet stream for an event. Clustering process involves multiple segments and extracts most important tweet from each cluster.

Tweets from all the clusters are combined together in order to produce summary. In this subsection, we discuss cluster based approaches for twitter event summarization.

The work by Inouye [27] introduced two cluster based approaches to generate multiple tweet summaries for a Twitter event. In one approach, event relevant posts are clustered into 4 subtopics by using a clustering algorithm. The clustering algorithm is the hybrid form of k-means++ [28] and bisecting k-means algorithm. Feature vectors of each post in a cluster computed by using Hybrid TFIDF weighing of words. Maximum weighted tweet selects as most important to summarize that cluster and the entire set of clusters generate 4 tweets length summary. In another approach, the hybrid TFIDF summarization algorithm as defined in the work [11] is modified to produce 4 post summaries. The modified approach selects maximum weighted 4 tweets for summary instead of selecting the most important tweet for an event. Clustering of micro-blog data is scandalously difficult and number of clusters varies from topic to topic due to nonstandard orthography, noise data. Variations in cluster numbers highly affect the cluster representative tweets as well as final summary of the event. The research work [29] optimizes the clustering process by utilizing the system [30]. The work also uses normalization tool proposed by [5] and expands tweet terms as proposed in [31]. Evaluation shows that even though only normalization of posts does not impact the summaries but optimization of clustering and term expansions significantly improves summaries.

A cluster based batch summarization algorithm (SPUR) [32] is proposed to summarize tweets from Twitter feeds by dividing input tweet stream into clusters based on time window. Each cluster is of one hour equalized batches and each batch compressed by replacing individual words with frequently used phrases. Frequently used phrases are extracts from relevant tweets of the event. All the tweets are ranked by their utility values and are to be included into the summary. The work also proposes dynamic version of the SPUR algorithm (DSPUR) considering recent tweets. DSPUR summarizes dynamic tweet streams by using pyramidal time window proposed by Aggarwal et al. [33]. Two SPUR output summaries of two time duration are merged together by removing the redundant tweets to generate updated summary.

In real time scheduled events more specifically such as sports events, Twitter users use to tweet instantly on each important moment of the events. Those moments of the event reports huge number of tweets and includes most important contents of the event. The approach [34] introduces a summarization algorithm for scheduled events such as soccer games based on burstiness of tweets. Proposed method detects sub-events based on sudden increase of tweeting activity. For each sub-event, the system selects important tweets by computing frequency of terms in tweets by using Kullback-Leibler divergence [35] method. The method measures how frequent are a term (*t*) within the sub-event (*H*) and during the game until the previous minute (*G*) by using Eq. (9).

$$D_{KL}(H \parallel G) = H(t) \log \frac{H(t)}{G(t)} \tag{9}$$

The tweet that holds maximum terms score for a given sub-event is returned as the representative tweet for that sub-event. All the representative tweets comprise the event summary. Proposed approach applied to all the matches of Copa America 2011 and creates summaries for three different languages: Spanish, English, and Portuguese. Burstiness of tweets are also applies by the proposed approach [36] to cluster an event participant streams. The participant streams are formed by gathering the tweets of event participating persons, organizations, product lines and so on. Two participant clusters are merged into a global cluster based on their time window and the Jaccard similarity [37] score. The tweets are ranked based on TF-IDF score of the consisting words and maximum scored tweets are extracted for summary. The work by Duan et al. [38] uses Bursti period of words in tweets to segment the topic relevant tweet stream into clusters. Then the tweets in each cluster are ranked by a mutual reinforcement model. The model is builds on words, tweet and users relationship. Words are linked by co-occurrence in the same tweet, Tweets are connected to each other by nonzero cosine similarity and users are associated by following-followee relationship. The approach [39] proposed a modified hidden Markov model (HMM) to segment the event timeline depending on burstiness of the tweet stream and the word distribution in tweets. The modification of HMM is work on the tweets for a time period instead of one symbol (tweet), different tweet rates, combining multiple events of same type. A tweet selects from each segment and comprises the summary. The system [40] utilized bursty phases of the event to infer user's collective interests towards the event. System automatically discovers phases of one given topic based on bursty phases. The selection of tweet for each phase is based on the features like informativeness, interestingness and diversity of tweets. System uses LexRank algorithm to determine informative sentences, topic biased LexRank uses to measure user's collective interests and diversity ranking algorithm MRR applies to reduce information redundancy among the tweets in each cluster. Top tweets for these three features are adjusted into the result summary.

The summarization framework [9] called "Sumblr" summarizes tweet streams with time line generation. The algorithm clusters event related tweet stream and holds cluster information by two data stricture called tweet cluster vector (TCV) and pyramidal time frame (PTF) [33]. Initial clusters are created using k-means clustering with small amount of tweets. Proposed TCVs are then initialized accordingly with the initial cluster statistics. On arrival of a new tweet, it goes to an existing cluster or to form a new cluster based on cosine similarity of tweet with clusters centroid. Cluster centroid value (CV) computes based on Eq. (10). $wsum_v$ is sum of weighted textual vectors ($t\ v_i$), n is total tweet count in a cluster, w_i is user rank value.

$$cv = \left(\sum_{i=1}^{n} W_i . tv_i \right) / \, n = wsum_v \, / \, n \tag{10}$$

Pyramidal time frame is used to store and organize cluster snapshots at different moments allowing historical tweet data to be retrieved by any arbitrary time durations. After cluster formation, most similar two clusters merge based on their centroid score. Historical summary includes most important tweets for a specific period. The approach [41] clusters relevant tweets into subtopics by identifying time-ordered tweets based on stream-based and semantic-based approaches. Stream-based subtopic detection is done based on the change of tweet volume while semantic-based subtopic detection is done using Dynamic Topic Model [20]. In semantic based approach, system sorts the subtopics according to their temporal information. Then the subtopics are ordered by the mean timestamp of the tweets in the resultant subtopics. The tweets with the highest scores are selected for each subtopic as summary. In other approach, the tweets in each subtopic are ranked to generate the sub-summaries. The ranking is done based on three features: position-aware coverage, sequential novelty and sequence correlation. Top ranked tweets from each subtopic incorporated into final summary. An online incremental clustering algorithm is used in the work [42] to divide an event into sub-events. Algorithm clusters the tweets based on the features like noun phrases, verbs, hashtags, URLs, numbers. More often a sub-event includes most of the near-duplicate tweets having highest number of key terms with different meanings. These tweets select as candidate tweets for each sub-event to form the summary of the event.

3.1.3 Statistical approaches

In statistical approaches, every word of a tweet is treated as a term and a tweet is treated as a document. The occurring frequencies of each term and positions are used to weight a tweet and ranked accordingly. The sum of the term weights infers the weight of concern tweet. Top ranked tweets from the relevant tweet stream of a topic produces final summary. Traditional TFIDF approach for weighing words is used in many research work of this category to extract important tweets for an event. Some of the important statistical approaches are discussed here.

The system by Sharifi et al. [11] introduced first statistical approach called hybrid TF-IDF summarization approach to produce a Twitter event summary. In the work, the term TF-IDF of classical TF-IDF method is redefined in view of the short length of tweets where content words are often very less in number. According to the modified definition, a document is a single tweet and during computation of term the document defined as the entire set of tweets. TF and IDF of a term (*i*) in a tweet are calculated by using Eq. (11) and (12). Finally, Weight of a tweet *W*(*T*) is computed based on Eq. (13). The most weighted tweets are chosen as the summary.

$$tf(W_i) = \frac{OccurancesOfAllWordsInAllPosts}{WordsInAllPosts} \tag{11}$$

$$idf(W_i) = \frac{SentencesInAllPosts}{SentencesInWhichWordOccurs} \tag{12}$$

$$W(T) = \frac{\sum_{i=0}^{WordsInSentence} W(w_i)}{nf(T)} \tag{13}$$

In above equations the terms are self-explanatory. In most of the cases, Algorithm shows biasness towards longer tweets due to higher terms frequency. So, a normalization factor *nf* is applied to standardize the weight of tweets and hence to eliminate the biasness. The normalization value is calculated by using Eq. (14).

$$nf = \max[\mathit{MinimumThreshold}, \mathit{WordsInTweet}] \tag{14}$$

Three different statistical methods are proposed in the work [43] for summarizing Twitter events. Methods are *Temporal TF-IDF*, *Re-tweet Voting* approach and *Temporal Centroid Representation* method. All the methods use a one-hour time window to group tweets and clubbed most important tweets from each group into summary. The *temporal TF-IDF* method scores tweets based on the words which occur more frequently across documents within the time frame. While the *voting method* computes re-tweet score based on re-tweet count for a post received in the time window by using the *Re-tweet Score* (rt) defined in Eq. (15).

$$rt = \frac{|\, retweet \quad (u_i) \,|}{|\, retweet \quad (u_{all}) \,|} \tag{15}$$

The temporal centroid method selects posts that correspond to each cluster centroid as the summary which has been a centroid for the longest period on average over a time-window. Social influence as well as several social features is explored in the system [44] for summary generation. Author defines Twitter conversation as a set of tweets posted by the users at specific timestamp on the same topic. System uses a set of hashtags as keyword query to retrieve event related conversations. Candidate tweets for summary generation, selects based on three categories of features. The features are tweet influence, tweet relevance regarding initial text and tweet relevance regarding URL. The tweet influence is determined by reply, re-tweet and favorite influence. Relevance score is measured by computing cosine similarity of initial and all other relevant tweets. The linear combination score of these features ranks the tweets and the top-rank tweets are selected to form the context.

Proposed summarization framework [45] uses two topic models, Decay Topic Model (DTM) and Gaussian Decay Topic Model (GDTM) on event relevant tweets to obtain topics those are highly event relevant. The system computes perplexity score of each tweet *d* with respect to its topic *z*, to select the most representative tweet. Perplexity score is calculated by using Eq. (16), where N_d is the number of words in d, ς_z is Gaussian distribution and ϕ_z is topic word distribution. The tweet having lowest perplexity in a topic is selected to form the summary of the event.

$$perplexity\ (d, z) = \exp\left(\frac{-\log P(d, z \mid \theta, \phi_z, \varsigma_z)}{N_d} \right) \tag{16}$$

3.1.4 Other approaches

This section describes some unconventional approaches rather than the above classes of summarization approaches. The query-focused summarization framework [46] generates summaries based on a fuzzy formal concept analysis (FFCA) algorithm. The algorithm forms fuzzy concept lattice from event related relevant tweets. Finally, best tweet with highest degree of membership is included in the resulting summary. Another approach [47] focuses on query generation to extract the most relevant tweets for event summary generation. In query generation phase, all the relevant tweets are indexed by using an IR toolkit 6. Then each tweet is represented in an n-dimensional vector where n is the number of unique terms in the tweet set. Each indexed tweet is re-ranked by a novelty detector system. The novelty detector system calculates novelty score of a tweet by computing the cosine similarity between a current vector and the top 10% of retrieved tweets. The summarizer selects the most novel sentences until 100 words are produced and arranged the tweets as per position order in stream. User's social status is focused in the work [48] to determine a tweet's importance for summary generation. The method detects key time points for an active event and extracts tweets from each key time point. For each time point the tweets are rank with a ranking algorithm. Based on the high ranking score tweets, the system build the event summary.

3.2 Abstractive Summaries

An abstractive summary produces a summary by collecting key information from the tweets rather including original tweets. Abstractive summary is more amalgamated in nature than a collection of tweets. Furthermore, these summaries restrain no redundant information. Various summarization approaches for abstract summary generation are discussed in this subsection.

3.2.1 Graph based approaches

The work [49] introduced an abstractive summarization algorithm for Twitter streams using word graph and optimization techniques. The algorithm named "Twitter Online Word Graph Summarizer" (TOWGS) extracts relevant tweets for an event and forms trigrams to construct a word graph. Each trigram represents a node in the graph and the weight of each node measures based on the occurrence frequencies of that trigram in the related tweets. The graph updated with recent tweets using decaying windows algorithm [50] at every second. Tweets having highest scoring path get selected to generate summary.

3.2.2 Cluster based approaches

Cluster based abstractive summarization framework proposed in [51] summarizes specific Twitter events such as events on crisis scenarios. Proposed approach extracts set of relevant tweets through an integer linear programming (ILP) based optimization technique. These relevant tweets generates list of bigrams to construct a word graph where bigrams represent nodes and edges represent co-occurrences relation in tweet. A weight (W) of each path in the graph is measured by using Eq. (17) based on the features like presence of content words, informativeness and linguistic quality of tweet. The summarization is achieved by optimizing the following ILP objective function, whereby the highest scoring tweet paths are returned as output of summarization process.

$$W = \max \left(\sum_{i=1}^{n} LQ\ (i).\, I\,(i).\, x_i + \sum_{j=1}^{m} y_j \right) \tag{17}$$

$LQ(i)$ is the linguistic score, $I(i)$ is informativeness score, x_i is the length of the summary and y_i is the important content words. A centroid based ranking algorithm is used to measure informativeness and linguistic quality is computed using a trigram language model as proposed by Heafield [52].

3.2.3 Other approaches

Abstract summary for a Twitter event is generated in the work by Zhang et al. [53] using speech act. Users share different kind of information and sentiment through tweets which are all instances of different speech acts. Proposed work defines 5 types of speech act for tweet following the work on Searle's popular taxonomy of speech acts [54]. Defined speech acts are statement, question, suggestion, comment and the miscellaneous type. The algorithm extracts word and symbol based features for each tweet and labels the tweet by corresponding speech act using SVM classifier. Topic words and the salient words/phrases for each major speech act type are selected by using round robin algorithm and insert into proper slots of speech act-guided templates to generate abstractive summary.

4. Summary Evaluation Methodologies

Evaluation of Twitter event summary is a complex task where texts are highly unstructured. Different human has different views and directions regarding a topic. So, it is hard to define a unique or exact summary of a topic. Thus, it is not possible to set a standard summary for any topic. As a result, comparison of different machine generated summaries and to set a baseline summary in absence of a standard human or automatic evaluation metric is a difficult task. Twitter topic summary evaluation follows traditional text summary evaluation methods like intrinsic or extrinsic methods. Intrinsic evaluation focuses on summary content's quality by evaluating quality of grammar, readability and cohesiveness of the content. This approach measures system accuracy [55] by comparing system summary with one or more manual summaries. Extrinsic evaluation method mainly focuses on the measure that how well a user performs the summarization task. In this section, we discuss promising evaluation processes carried out to standardize different summarization methodologies for Twitter events.

4.1 Human Evaluation

Human evaluation (Hu) process, measures the closeness of machine generated summaries against reference summaries on a certain point scale judgment. The reference summary of a topic is standard summary produced by human. The accuracy of the process is fully dependant on human judge understanding. Very often it is observed that different reference summaries are possible by different human for the same topic. However, the system generated summary may be quite different from any of the reference summaries. Human understanding towards the topic and information order makes summary different. Even same human may select different sentences for the same topic after a gap of certain time [56].

To overcome from the difficulties of human evaluation of summaries, Lin and Hovy [57] proposed an automatic summary evaluation system called NAM_s (accumulative n-gram matching score) which is measured using Eq. (18).

$$NAM_S = a_1 . NAM_1 + a_2 . NAM_2 + ... + a_n . NAM_n \tag{18}$$

where NAM_n is the ratio of matched n-grams (between reference summary and system summary) and total n-grams in RS, and only content words are used to form the n-grams. Some of the other promising automatic evaluation measures are based on similar sentences counts among reference and machine summaries [58], sentence worthiness of the summary [59], similarity of vocabulary in reference and system summaries [60]. These measures are applicable for content based summaries and ignore the number of sentences in both system and reference summary. Combined strategy of human and automatic evaluation [61] is also used for Twitter event summary evaluation.

4.2 Recall Oriented Understudy for Gisting Evaluation

The work [62] introduced a set of metrics called Recall-Oriented Understudy for Gisting Evaluation (ROUGE) which has become standard of automatic evaluation of summaries. Most of the Twitter topic summarization algorithm performance evaluated using the ROUGE metrics. The approach counts number of overlapping units such as n-gram, word sequences and word pairs between the computer-generated summary and reference summaries and proposed different ROUGE matrices: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. ROUGE-N measure is an n-gram recall based evaluation, which derives precision, recall and F-scores. ROUGE-L is based on longest common subsequence (LCS) matching. Two comparative summaries are more similar when longer LCS is in both of them. One advantage of ROUGE-L is that it does not require predefined n-gram length as it automatically includes longest in-sequence common n-grams. But non-consecutive sub-sequence matches are included in this measure. Thus, the measure ROUGE-W is introduced which redefines ROUGE-L by including weights that penalize subsequence matches those are not consecutive. Another metric called ROUGE-S represents Skip-bigram co-occurrence between system summary and a set of reference sum-maries. ROUGE-S score becomes zero for diverse representation of semantically same sentence in competitive summaries. So, ROUGE-S extends as ROUGE-SU with the addition of unigram matching as counting unit.

ROUGE metrics are the most commonly used evaluation metrics for tweet summarization. More specifically, ROUGE-1 score is the most reliable as explained in the work by Lin and Hovy [63]. However, system performance for Twitter topic summarization is not as high as single-document formal text summarization.

4.3 Information-Theoretic Evaluation

The work by Lin et al. [64] proposed an information theoretic approach for automatic evaluation of summaries based on the Jensen-Shannon (JS) divergence of distributions between a machine summary and set of reference summaries. Given a set of documents D = d_1, d_2, ... , d_i , where i = 1 to n, there exists probabilistic distribution with parameters θ_R and θ_A. θ_R and θ_A generate reference and system summaries from D respectively. A good summarizer should have its θ_A very close to θ_R and the summary evaluation process estimate the distance between θ_A and θ_R. Summary evaluation score is measured by using the Eq. (19) based on the JS divergence distance.

$$Score_{summary}^{JS}(S_A|S_R)=-JS_{1/2}(p(\theta_A|S_A)||p(\theta_R|S_R)) \tag{19}$$

The score represents negative value to indicate the similarity between two distributions. S_A and S_R are the two candidate summary compared in the evaluation.

4.4 Model Free Evaluation

All the above evaluation matrices are entirely dependent on gold standard human summaries. These evaluation processes cannot evaluate system performance properly when human summaries are not available or single summary is available. The work by Louis and Nenkova [65] proposed two model-free and one modified automatic evaluation technique to measure system performance. Model free techniques expand the available model summaries rather than gold summaries and evaluated system generated summaries. First two model free techniques use input-summary similarity and consensus based evaluation technique respectively while the later one uses gold summary based evaluation called pseudo-model evaluation technique. In input summary similarity technique, the similarity metric uses distribution similarity, summary likelihood and topic signature words for comparing the reference and system summary. Consensus based evaluation assess summary quality based on the knowledge from available system summaries. In pseudo-model technique, higher accuracy evaluation model is chosen as pseudo-models. Summaries from these models are added to the gold-standard summaries and compared with system summary to produce final evaluation scores.

5. Discussion and Research Challenges

We discussed several promising Twitter topic summarization algorithms in this survey report and recapitulated extractive approaches in Table 1 and abstractive approaches in Table 2 with key traits. The in-depth study of key traits reveals that most of the algorithms are either directly apply or build upon on standard summarization approaches formerly efficiently summarizes traditional text documents. At the same time, extractive summarization approaches are more focused than the abstractive summaries due to two reasons. First, affect of noise text or large diversity of tweets on extractive summary is lesser. Second, producing abstractive summary for huge number of real time tweets consumes more processing and time complexity.

This survey work concludes that, summarization problem of domain specific Twitter events like sports or crisis scenario event is less difficult than generic Twitter events. Domain dependant Twitter events share common characteristics and knowledge acquisition from past similar type of events is easier. Prediction of important moment or detection of similar tweets for a specific time is easier task which helps clustering process. Our survey work reveals that, there are no any standard dataset as well as standard evaluation methodology exists for tweet summarization. Almost all the research works on social media summarization, used different dataset and different experiment setup. This creates difficulty for defining the perfect gist of a Twitter event. Most of the Twitter topic summaries are evaluated by using metrics which is used in traditional text summarization evaluation. This survey work shows that ROUGE-N metric is used for the most research work evaluation.

Table 1. Brief report on extractive summarization approaches

Ref.	Year	Event category	Techniques	Sum length	Evaluation metrics
Sharifi et al. [10]	2010	Gen	Gr	01 Tweet	ROUGE, Hu
Sharifi et al. [11]	2010	Gen	St	01 Tweet	ROUGE, Hu
Inouye [27]	2010	Gen	Cl	04 Tweets	ROUGE, Hu
Beverungen and Kalita [29]	2011	Gen	Ot	04 Tweets	ANOVA metrics
Harabagiu and Hickl [14]	2011	Gen	Gr	250 Words	Hu
Chakrabarti and Punera [39]	2011	Dom	Cl	10-70 Tweets	Precision
Liu et al. [15]	2012	Gen	Gr	Variable length	ROUGE-N
Yang et al. [32]	2012	Gen	Cl	Variable length	Precision, recall and F-measure
Zubiaga et al. [34]	2012	Dom	Cl	Variable length	Precision, recall and F-measure
Lin et al. [8]	2012	Gen	Gr	30 Tweets	ROUGE-N
Nichols et al. [13]	2012	Dom	Gr	Variable length	ROUGE-N, Hu
Duan et al. [38]	2012	Gen	Gr	10 Tweets	ROUGE
Gao et al. [41]	2013	Gen	Cl	Variable length	ROUGE, Hu
Shen et al. [36]	2013	Gen	Cl	Variable length	ROUGE-1
Chua and Asur [45]	2013	Gen	Ot	08-10 Tweets	ROUGEN
Judd and Kalita [12]	2013	Gen	Gr	04 Tweet	ROUGE-N
Khan et al. [19]	2013	Gen	Cl	10-15 Tweet	Precision, recall
Xu et al. [18]	2013	Gen	Gr	01-04 Tweets	Hu
Kim et al. [23]	2014	Gen	Gr	15%–30%	Precision, recall
Wang et al. [9]	2015	Gen	Cl	Variable length	ROUGE
Xu et al. [22]	2015	Gen	Gr	10 Tweets	ROUGE-1
Maio et al. [46]	2015	Gen	Ot	Variable length	Precision, recall, F-measure
Yulianti et al. [47]	2016	Gen	Ot	100 Words	ROUGE
Alsaedi et al. [43]	2016	Gen	St	04-05 Tweets	ROUGE-1
Schinas et al. [24]	2016	Gen	Gr	10 Tweets	Precision, mean reciprocal rank
Zhou et al. [42]	2016	Gen	Cl	02-30 Tweets	Precision, recall, F-measure
Qu et al. [25]	2016	Gen	Gr	Variable length	F-measure
Zhao et al. [40]	2016	Gen	Cl	25-30 Tweets	ROUGE
He et al. [48]	2017	Gen	Ot	12 Tweets	ROUGE
Belkaroui and Faiz [44]	2017	Gen	St	05-10 Tweets	Uni-gram, bi-gram
Mosa et al. [26]	2017	Gen	Gr	Variable length	ROUGE, Hu

Table 2. Brief report on abstractive summarization approaches

Ref.	Year	Event category	Techniques	Sum length	Evaluation metrics
Zhang et al. [53]	2013	Gen	Ot	20% of input	ROUGE
Olariu [49]	2014	Gen	Gr	Variable length	Hu
Rudra et al. [51]	2016	Gen	Cl	200–400 words	ROUGE-1, Hu

Twitter topic summarization task poses new research challenges due to its significant difference from traditional text document. One of the critical difference is, hard to find important and crucial information of an online information, because of huge size of data and complicated evolution. In this survey work, we identified the following research challenges towards Twitter topic summarization:

- **Detection of important phase of event:** Detection of important moments of an event is unpredictable. The rate of posts and informative contents vary radically over a short time. So it is a big challenge to achieve timely summarization by tackling dynamic posts.
- **Informativeness, readability and cohesiveness:** A good summary should reveal the flow of the information so that interesting developments of the event can be traced. Most of the above research works produce summary by removing tweet redundancy. But a tweet does not refer a sentence only, rather it includes multiple sentences. So, even though tweets level redundancy has removed at a great extent but still sentence level redundancies among tweets in the summary exists.
- **Tweet characteristics:** The language of tweets is highly noisy, includes spelling and grammatical mistakes, typos, abbreviations, phonetic substitutions and emoticons, etc. Due to the above characteristics, the word graph for tweet words could contain too much noisy data. So, noise elimination and normalization of tweet tokens are also some challenging research problem
- **Standard dataset:** There are no any standard dataset and consequently no standard evaluation methodology exists for tweet summarization as well as evaluation. For events with huge number of tweets, different human generates different summaries, even same person summarize differently for a gap of certain period. So, generation of a standard dataset with gold summaries is a big research challenge.
- **Summary length:** The length of the summary is an important factor and different topics require different summary length. On average summary length varies from 5% to 30%. However, human created summaries are around 15% of the source text. To determine the exact length of summary for information coverage of an event and estimating the increase of summary with dynamic development of the event is a research challenge.

6. Conclusion

Over time, social media like Twitter stands as an important source of information, ahead of any information sharing platform like general blogs or news media. This important phenomenon makes summarization of social media more important research problem. In this work, we made a comprehensive overview of the most prominent recent approaches for automatic Twitter topic summarization. This survey work emphasized on extractive and abstractive approaches and contrasted approaches in terms of text normalization, measuring word or phrase or tweet weights to produce the summary. This survey work also focused on Twitter topic summarization evaluation methodologies. The evaluation metrics shown effective for formal texts are also applied in this genre of informal text for evaluation.

References

[1] C. Smith, “400 Interesting Twitter facts, demographics and statistics (November 2017),” 2017 [Online]. Available: http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/.

[2] Twitter usage statistics [Online]. Available: http://www.internetlivestats.com/twitter-statistics/.

[3] The top 500 sites on the web [Online]. Available: http://www.alexa.com/topsites.

[4] M. Isaac and S. Ember, "For Election Day influence, Twitter ruled social media," The New York Times, 2016 [Online]. Available: https://www.nytimes.com/2016/11/09/technology/for-election-day-chatter-twitter-ruled-social-media.html.

[5] M. Kaufmann, "Syntactic normalization of twitter messages," in *Proceedings of International Conference on Natural Language Processing (ICON)*, Kharagpur, India, 2010.

[6] B. Han and T. Baldwin, "Lexical normalisation of short text messages: makn sens a #twitter," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, OR, 2011, pp. 368-378.

[7] D. Rudrapal, A. Jamatia, K. Chakma, A. Das, and B. Gamback, "Sentence boundary detection for social media text," in *Proceedings of the 12th International Conference on Natural Language*, Trivandrum, India, 2015, pp. 254-260.

[8] C. Lin, C. Lin, J. Li, D. Wang, Y. Chen, and T. Li, "Generating event storylines from microblogs," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, Maui, HI, 2012, pp. 175-184.

[9] Z. Wang, L. Shou, K. Chen, G. Chen, and S. Mehrotra, "On summarization and timeline generation for evolutionary tweet streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1301-1315, 2015.

[10] B. Sharifi, M. A. Hutton, and J. Kalita, "Automatic summarization of Twitter topics," in *Proceedings of National Workshop on Design and Analysis of Algorithm*, Tezpur, India, 2010.

[11] B. Sharifi, M. A. Hutton, and J. Kalita, "Summarizing microblogs automatically," in *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Los Angeles, CA, 2010, pp. 685-688.

[12] J. Judd and J. Kalita, "Better Twitter summaries?," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, GA, 2013, pp. 445-449.

[13] J. Nichols, J. Mahmud, and C. Drews, "Summarizing sporting events using Twitter," in *Proceedings of the ACM International Conference on Intelligent User Interfaces*, New York, NY, 2012, pp. 189-198.

[14] S. Harabagiu and A. Hickl, "Relevance modeling for microblog summarization," in *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM)*, Barcelona, Spain, 2011, pp. 514-517.

[15] X. Liu, Y. Li, F. Wei, and M. Zhou, "Graph-based multi-tweet summarization using social signals," in *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*, Bombay, India, 2012, pp. 1699-1714.

[16] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell, "Summarizing text documents: sentence selection and evaluation metrics," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, 1999, pp. 121-128.

[17] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107-117, 1998.

[18] W. Xu, R. Grishman, A. Meyers, and A. Ritter, "A preliminary study of tweet summarization using information extraction," in *Proceedings of the Workshop on Language Analysis in Social Media*, Atlanta, GA, 2013, pp. 20-29.

[19] M. A. H. Khan, D. Bollegala, G. Liu, and K. Sezaki, "Multi-tweet summarization of real-time events," in *Proceedings of International Conference on Social Computing*, Alexandria, VA, 2013, pp. 128-133.

[20] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006, pp. 113-120.

[21] R. Mihalcea and P. Tarau, "TextRank: bringing order into texts," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004, pp. 404-411.

[22] B. Xu, H. Hao, Y. Wu, H. Zhang, and C. Liu, "TR-LDA: a cascaded key-bigram extractor for microblog summarization," *International Journal of Machine Learning and Computing*, vol. 5, no. 3, pp. 172-178, 2015.

[23] T. Y. Kim, J. Kim, J. Lee, and J. H. Lee, "A Tweet summarization method based on a keyword graph," in *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication*, Siem Reap, Cambodia, 2014, pp. 1-8.

[24] M. Schinas, S. Papadopoulos, Y. Kompatsiaris, and P. A. Mitkas, "MGraph: multimodal event summarization in social media using topic models and graph based ranking," *International Journal of Multimedia Information Retrieval*, vol. 5, no. 1, pp. 51-69, 2016.

[25] Q. Qu, S. Liu, F. Zhu, and C. S. Jensen, "Efficient online summarization of large-scale dynamic networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 12, pp. 3231-3245, 2016.

[26] M. A. Mosa, A. Hamouda, and M. Marei, "Graph coloring and ACO based summarization for social networks," *Expert Systems with Applications*, vol. 74, pp.115-126, 2017.

[27] D. Inouye, "Multiple post microblog summarization," University of Colorado at Colorado Springs, 2010.

[28] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, LA, 2007, pp. 1027-1035.

[29] G. Beverungen and J. Kalita, "Evaluating methods for summarizing Twitter posts," in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*, Hong Kong, China, 2011, pp. 1-6.

[30] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411-423, 2001.

[31] F. Perez-Tellez, D. Pinto, J. Cardiff, and P. Rosso, "On the difficulty of clustering company tweets," in *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, Toronto, Canada, 2010, pp. 92-102.

[32] X. Yang, A. Ghoting, Y. Ruan, and S. Parthasarathy, "A framework for summarizing and analyzing Twitter feeds," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, China, 2012, pp. 370-378.

[33] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in *Proceedings of the 29th International Conference on Very Large Data Bases*, Berlin, Germany, 2003, pp. 81-92.

[34] A. Zubiaga, D. Spina, E. Amigo, and J. Gonzalo, "Towards real-time summarization of scheduled events from Twitter streams," in *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, Milwaukee, WI, 2012, pp. 319-320.

[35] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79-86, 1951.

[36] C. Shen, F. Liu, F. Weng, and T. Li, "A participant-based approach for event summarization using Twitter streams," in *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies*, Atlanta, GA, 2013, pp. 1152-1162.

[37] L. Lee, "Measures of distributional similarity," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, College Park, MD, 1999, pp. 25-32.

[38] Y. Duan, Z. Chen, F. Wei, M. Zhou, and H. Y. Shum, "Twitter topic summarization by ranking tweets using social influence and content quality," in *Proceedings of the 24th International Conference on Computational Linguistics,* Mumbai, India, 2012, pp. 763-780.

[39] D. Chakrabarti and K. Punera, "Event summarization using Tweets," in *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM)*, Barcelona, Spain, 2011, pp. 66-73.

[40] W. X. Zhao, J. R. Wen, and X. Li, "Generating timeline summaries with social media attention," *Frontiers of Computer Science*, vol. 10, no. 4, pp. 702-716, 2016.

[41] D. Gao, W. Li, and R. Zhang, "Sequential summarization: a new application for timely updated Twitter trending topics," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria, 2013, pp. 567-571.

[42] Y. Zhou, N. Kanhabua, and A. I. Cristea, "Real-time timeline summarisation for high-impact events in Twitter," in *Proceedings of 22nd European Conference on Artificial Intelligence*, The Hague, The Netherlands, 2016, pp. 1158-1166.

[43] N. Alsaedi, P. Burnap, and O. Rana, "Automatic summarization of real world events using Twitter," in *Proceedings of the 10th International Conference on Web and Social Media (ICWSM)*, Cologne, Germany, 2016, pp. 511-514.

[44] R. Belkaroui and R. Faiz, "Conversational based method for tweet contextualization," *Vietnam Journal of Computer Science*, vol. 4, no. 4, pp. 223-232, 2017.

[45] F. C. T. Chua and S. Asur, "Automatic summarization of events from social media," in *Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM)*, Boston, MA, 2013, pp. 81-90.

[46] C. De Maio, G. Fenza, V. Loia, and M. Parente, "Online query-focused twitter summarizer through fuzzy lattice," in *Proceedings of 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Istanbul, Turkey, 2015, pp. 1-8.

[47] E. Yulianti, S. Huspi, and M. Sanderson, "Tweet-biased summarization," *Journal of the Association for Information Science and Technology*, vol. 67, no. 6, pp. 1289-1300, 2016.

[48] R. He, Y. Liu, G. Yu, J. Tang, Q. Hu, and J. Dang, "Twitter summarization with social-temporal context," *World Wide Web*, vol. 20, no. 2, pp. 267-290, 2017.

[49] A. Olariu, "Efficient online summarization of microblogging streams," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, 2014, pp. 236-240.

[50] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*. New York, NY: Cambridge University Press, 2011.

[51] K. Rudra, S. Banerjee, N. Ganguly, P. Goyal, M. Imran, and P. Mitra, "Summarizing situational tweets in crisis scenario," in *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, Halifax, Canada, 2016, pp. 137-147.

[52] K. Heafield, "KenLM: faster and smaller language model queries," in *Proceedings of the 6th Workshop on Statistical Machine Translation*, Edinburgh, UK, 2011, pp. 187-197.

[53] R. Zhang, W. Li, D. Gao, and Y. Ouyang, "Automatic Twitter topic summarization with speech acts," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 649-658, 2013.

[54] J. R. Searle, "Indirect speech acts," in *Syntax and Semantics 3: Speech Acts*. New York, NY: Academic Press, 1975, pp. 59-82.

[55] I. Mani, "Summarization evaluation: an overview," in *Proceedings of the 2nd Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization (NTCIR-2)*, Tokyo, Japan, 2001.

[56] G. J. Rath, A. Resnick, and T. R. Savage, "The formation of abstracts by the selection of sentences. Part I. sentence selection by men and machines," *Journal of the Association for Information Science and Technology*, vol. 12, no. 2, pp. 139-141, 1961.

[57] C. Y. Lin and E. Hovy, "Manual and automatic evaluation of summaries," in *Proceedings of the ACL-02 Workshop on Automatic Summarization*, Philadelphia, PA, 2002, pp. 45-51.

[58] R. L. Donaway, K. W. Drummey, and L. A. Mather, "A comparison of rankings produced by summarization evaluation measures," in *Proceedings of the NAACL-ANLP Workshop on Automatic Summarization*, Seattle, WA, 2000, pp. 69-78.

[59] D. R. Radev, H. Jing, and M. Budzikowska, "Summarization of multiple documents: clustering, sentence extraction, and evaluation," in *Proceedings of the NAACL-ANLP Workshop on Automatic Summarization*, Seattle, WA, 2000, pp. 21-30.

[60] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill Inc., 1986.

[61] D. Inouye and J. K. Kalita, "Comparing twitter summarization algorithms for multiple post summaries," *in Proceedings of IEEE 3rd International Conference on Privacy, Security, Risk and Trust and 2011 IEEE 3rd International Conference on Social Computing*, Boston, MA, 2011, pp. 298-306.

[62] C. Y. Lin, "Rouge: a package for automatic evaluation of summaries," in *Proceedings of the ACL-04 Workshop*, Barcelona, Spain, 2004, pp. 74-81.

[63] C. Y. Lin and E. Hovy, "Automatic evaluation of summaries using N-gram co-occurrence statistics," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Edmonton, Canada, 2003, pp. 71-78.

[64] C. Y. Lin, G. Cao, J. Gao, and J. Y. Nie, "An information-theoretic approach to automatic evaluation of summaries," in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, New York, NY, 2006, pp. 463-470.

[65] A. Louis and A. Nenkova, "Automatically assessing machine summary content without a gold standard," *Computational Linguistics*, vol. 39, no. 2, pp. 267-300, 2013.

Dwijen Rudrapal https://orcid.org/0000-0002-9729-277X

He received M.Tech. degree in Computer Science and Engineering from National Institute of Technology Agartala, India in 2012. Currently he is doing PhD under the Department of Computer Science and Engineering from National Institute of Technology Agartala. His research interests are Human Language, Social Media Text and Artificial Intelligence.

Amitava Das https://orcid.org/0000-0002-3818-8227

He presently working as an Assistant Professor at IIIT Sri City, Andhra Pradesh, India. He obtained his Ph.D. (Engineering) from Jadavpur University, India. During his doctoral study he worked for an Indo-Japan collaborative project with the Tokyo Institute of Technology, Japan. In his last endeavor he worked as a Research Scientist in the Human Language Technologies (HiLT) lab at the University of North Texas, USA. Before moving to USA he worked for Samsung Research India, Bangalore as a Chief Engineer. He spent one year working as a European Research Consortium for Informatics and Mathematics (ERCIM) Postdoctoral fellow at the Norwegian University of Science and Technology (NTNU), Norway. His research interests broadly span three areas and more specifically their intersection: human language, mind/cognition and artificial intelligence. Presently he actively working on code-mixing in social media text and computational social sciences.

Baby Bhattacharya https://orcid.org/0000-0001-6053-6067

She completed her Ph.D. from Tripura University in the year 2006. She is working as an Assistant Professor in the Department of Mathematics at National Institute of Technology Agartala, Tripura, India. She is a life member of TMS and FRSA. Her research area of interest includes real analysis, fuzzy mathematics, bitopology and generalized fuzzy topology.