Multi-Tweet Summarization of Real-Time Events

Muhammad Asif Hossain Khan*, Danushka Bollegala*, Guangwen Liu† and Kaoru Sezaki†
*Graduate School of Information Science and Technology
The University of Tokyo, Tokyo 113–8656, Japan.
Email: asif@mcl.iis.u-tokyo.ac.jp, danushka@iba.t.u-tokyo.ac.jp
†Center for Spatial Information Science
The University of Tokyo, Tokyo 153–8505, Japan.
Email: liugw198209@mcl.iis.u-tokyo.ac.jp, sezaki@iis.u-tokyo.ac.jp

***Abstract*—Popular real-time public events often cause upsurge of traffic in Twitter while the event is taking place. These posts range from real-time update of the event's occurrences highlights of important moments thus far, personal comments and so on. A large user group has evolved who seeks these live updates to get a brief summary of the important moments of the event so far. However, major social search engines including Twitter still present the tweets satisfying the Boolean query in reverse chronological order, resulting in thousands of low quality matches agglomerated in a prosaic manner. To get an overview of the happenings of the event, a user is forced to read scores of uninformative tweets causing frustration. In this paper, we propose a method for multi-tweet summarization of an event. It allows the search users to quickly get an overview about the important moments of the event. We have proposed a graph-based retrieval algorithm that identifies tweets with popular discussion points among the set of tweets returned by Twitter search engine in response to a query comprising the event related keywords. To ensure maximum coverage of topical diversity, we perform topical clustering of the tweets before applying the retrieval algorithm. Evaluation performed by summarizing the important moments of a real-world event revealed that the proposed method could summarize the proceeding of different segments of the event with up to 81.6% precision and up to 80% recall.**

***Keywords*—Tweet summarization, Twitter search, Social network analysis, Text mining**

I. Introduction

Twitter is increasingly becoming an ideal platform for getting access to real-time response from the crowd about ongoing public events. It restricts its users to express what they see, hear and feel around them in a concise form of 140 characters. Large public events with different possible outcome, that take place in a pre-defined period of time, enjoy real-time coverage by interested Twitter users. These events range from popular sports events such as soccer matches, public debates such as presidential debates, annual award declaration ceremony such as Academy award etc. Twitter also has another user group, typically its search users, who seek these live updates [1]. The typical method is to search the social stream with event relevant keywords and *hashtags*. However, the volume of search results, satisfying such Boolean query, is formidable. For example, President Obama's 2012 election victory induced 237,000 tweets per minute with over 20 million tweets sent on election night.

A typical search user is generally not inclined to read beyond the first few tens of tweets matching the query terms. Unfortunately, state-of-the-art social search engines still use a recency based retrieval algorithms; i.e. tweets satisfying the Boolean query are presented in reverse chronological order. Hence, Twitter search user have to struggle to find informative tweets with non-repetitive content covering diversified aspect of the event of their interest among the huge bulk of returned search results, which often causes frustration [2].

In this paper, we present a method for identifying a small set of tweets from a large bulk of event relevant tweets, which can delineate the proceedings of the event and thus act as a multi-tweet summarization of the real-time event. We try to satisfy three objective functions while selecting the set of tweets: a) they represent popular discussion points discussed in the event-relevant tweets, b) they represent the topical diversity present in the collection of relevant tweets, and c) they do not repeat the same information. For simplicity, we decoupled the problem; i.e. we first perform topical clustering of relevant tweets, then apply the proposed retrieval algorithm on each cluster independently and finally make sure that tweets recommended to the search users do not overlap in terms of information content up to a certain threshold.

Our proposed method would work for those events that satisfy two criteria: a) they elicit large response from Twitter users, and b) they are real-time events taking place within a limited time span in certain pre-defined period of time. The proposed solution is based on a hypothesis that the discussion points that are common in majority of event-relevant tweets are motivated by the proceedings of the event. Thus by indentifying and quantifying these popular discussion points and retrieving tweets composed of maximum number of such highly popular discussion points, it is possible to summarize the occurrences of an ongoing event even with a very small set of tweets. The experimental evaluation performed on real-time tweets relevant to the first presidential debate between President Obama and Governor Romney corroborates the hypothesis.

The rest of the paper is organized as follow: section II briefly discusses on the related work. Section III describes the proposed method in detail. The real-world event and the relevant dataset, evaluation procedure and experimental results obtained by comparing the proposed method with two competitive baseline methods have been presented in section IV. Section V gives insights about the obtained results. Finally, section VI concludes with directions for further research.

II. Related Work

Numerous efforts for characterizing an event using relevant tweets have been made by researchers in recent years. Chakrabarti et al. [3] proposed a method for summarizing highly structured and recurring events such as football matches. They assumed that new events had already been detected by some other methods. Their proposed method tried to extract a few tweets that best describe the interesting occurrences in the event. They trained an HMM based model to identify occurrences of sub-events based on tweet "activity threshold" within a time segment. For retrieving tweets that are close to other tweets in the corpus they used a "*tf-idf* with *cosine similarity*'' based model that we have used as a baseline model in this paper. Availability of highly structured recurring events is scant in reality and hence their approach would not be able to handle vast majority of real-world events. Our proposed model does not rely on the structure or recurrence of events. Moreover, our method is completely unsupervised. On top of that, we have shown in the evaluation section that our proposed tweet retrieval algorithm outperforms the *tf-idf-cosine* model.

Sharifi et al. [4] also proposed a method for microblog summarization. Their model outputs a single sentence that serves as a journalistic summary of the event. They proposed two models for measuring relevance of co-occurrences in the tweets — one similar to the *tf-idf-cosine* model used in [3] and the other is a graph based model. The later model makes a graph of words around the "key phrase" based on the top N tweets returned by Twitter given the same *key phrase* as query. Their method returns a single sentence as a summary of the corpus. They used a frequency based approach to rank collocations around the key phrase and picked up tweets containing longest phrase obtained in this way. Our proposed model also identifies word co-occurrences that are popular among the relevant tweets and in the "Proposed Method" section we have analytically shown that the proposed method can distinguish co-occurrences with higher association strength better than the frequency based approach. Nichols et al. [5] used a slight variation of the phrase graph model proposed in [4] to generate a three-sentence summary for important moments in an event. Sudden upsurge of tweet traffic is used for detecting important moments. Finally, they added up the scores of each phrase encountered in the longest sentence of a tweet for obtaining a tweet score and output the top three tweets with the highest score.

Hu et al. [6] used a topic model to extract sense from Twitter feed relevant to public and televised events. Their model enables auto-segmentation of the events and characterization of tweets into two categories: *episodic* and *steady* tweets. However, they need a transcript of the event to acquire topical knowledge about the event, which can only be obtained after the event. Hence, their model serves as a post-event analysis tool that can measure how much attention each segment of a public event received in Twitter feed. In contrast, our method needs no transcript of any sort of external knowledge about the event and we identify the relevant tweets while the event is taking place.

Some efforts have been made for generating visual summaries of tweets on a topic [7], [8]. However, they do not offer sentence-level summaries and their recommended word clouds or word labels must be interpreted by users themselves.

A common requirement in [3] and several other works on summarization is the need to detect important moments in the tweet collection by some third party system. Our proposed solution has no such prerequisites. Hence, unlike our approach none of the aforementioned endeavors propose an unsupervised model which requires no external knowledge about the event to group event-relevant tweets in topical clusters and retrieve tweets that comprise popular discussion points within each cluster so that the recommended tweets serve as a journalistic summary of the event while the event is taking place.

III. Proposed Method

A. Assigning Tweets into Topical Clusters

We have leveraged the popular topic model LDA [9] for topical clustering of tweets. LDA assumes that each tweet in a given collection $\mathcal{T}$ is generated using a multinomial distribution, θ, over k topics. Each topic on the other hand is associated with a multinomial distribution, φ, over the vocabulary. Topic assignment for each word in $t \in \mathcal{T}$ is performed by sampling a particular topic z from multinomial distribution θ_t associated with the tweet. A particular word $w \in t$ is generated by sampling from the multinomial distribution φ_z associated with the topic z. This generative process is repeated n_t times (n_t is the total number of words in tweet t) to produce t. α and β are hyper–parameters for the dirichlet priors of θ and φ respectively. Blei et al. [9] used variational inference based algorithm for obtaining approximate maximum-likelihood estimates for θ and φ. However, in this experiment we have used a learning algorithm based on collapsed Gibbs sampling proposed in [10], which is arguably more accurate since it systematically approaches the correct distribution. Following [11] we have used symmetric dirichlet priors $\alpha = \beta = 0.01$.

The outcome of LDA is a soft-clustering of tweets in the collection. However, for each tweet we need to know the topic towards which it is most inclined to. We measure that by using the word token distribution of each tweet over the topic dimensions. At convergence of LDA, let $\mathbf{DP'}$, a $|\mathcal{T}| \times k$ matrix, holds the number of times a word-token in $t_i \in \mathcal{T}$ has been assigned to topic $j \in k$. We convert each row of $\mathbf{DP'}$ into a probability distribution over the k topics and produce a new matrix $\mathbf{DP}$; i.e. for each row of $\mathbf{DP}$, $\|\mathbf{DP}_{i.}\|_1 = 1$. For each tweet $t_i \in \mathcal{T}$, $\mathbf{DP}_i$ represents the *topic-inclination distribution* of t_i. Each cluster is represented by a vector, called *topical-signature*, which is the average of the topic-inclination distribution of the tweets already assigned in the cluster. Topical-signature vector for each cluster is initialized as a unit-vector in the direction of the topic dimension; i.e. the i-th cluster's initial topic-signature is a vector v, where $v_j = 0$ if $i \neq j$ and for $i = j$, $v_j = 1$. A tweet is assigned to the cluster whose topical-signature is closest to its topic-inclination distribution. We measured the distance between the two distributions using Jensen-Shannon divergence.

Determining Optimum Number of Topics

LDA takes the number of topics, k, in the tweet collection as one of its input parameters. However, determining the most appropriate number of topics in any text collection is a classical model selection problem. There are numerous ways to tackle this problem. Following the approach of [12], [13], [14], we used a *cluster validation* method for determining the most

appropriate value of k for the tweet collection $\mathcal{T}$. Given a range $\mathcal{I}$ of possible number of topics in the tweet collection $\mathcal{T}$, the cluster validation method estimates the most appropriate value $k^* \in \mathcal{I}$ for which the cluster structure estimated from $\mathcal{T}$ is most stable against re-sampling. The assumption behind the validation method is that for the ideal value k^*, a clustering algorithm applied to a tweet collection $\mathcal{T}$ and that applied to another tweet collection $\hat{\mathcal{T}} \subset \mathcal{T}$, would result in identical clustering of the tweets in $\hat{\mathcal{T}}$; i.e. if $t_i, t_j \in \hat{\mathcal{T}}$ are put in the same cluster in the former case, then they will be placed together in the later case too.

B. Identifying Representative Tweets from Topic Clusters

After dividing tweets into topic clusters, next we focus on identifying the tweets that comprise key discussion points within the clusters. For the tweets in each topic cluster, we first construct a lexical graph and then apply a variant of the PageRank algorithm [15] to determine the score of individual lexical units in the graph. Tweets comprising higher proportion of high-scored lexical units are recommended to the users. The following subsections describe the procedure.

Constructing Lexical Graph for Tweets in Topic Cluster

As we are trying to identify the key discussion points in the tweet collection, using unigram as lexical unit seems to be a reasonable choice. In our lexical graph, an edge between two nodes represents the strength of association between the unigrams in the tweet collection. In the pre-processing step we remove all URLs, user references (@user), numerals, time expression and non printable characters from the tweet collection. We remove the stop words using a list[1] of commonly used English words. We also consider all *track keywords*, the keywords used to get the tweet collection from the Twitter search engine, as stop words. Duplicate tweets and tweets with less than 10 terms are excluded from the corpus. Let $\mathcal{U}$ be the set of all unigrams encountered in the tweet collection after applying the pre-processing steps. Instead of considering all unigrams in the lexical graph, we use a syntactic filter to identify unigrams of specific part-of-speech and consider only those passing the filter. Only unigrams in the set $\mathcal{U}^* = \{w : w \in \mathcal{U}$ and $\mathcal{POS}(w) \in \{$*verb, noun, adjective*$\}\}$ are considered as lexical units, where $\mathcal{POS}(w)$ returns the part–of–speech of a unigram, which we determine using the "Stanford Log–linear Part–Of–Speech Tagger" [16].

If two unigrams appear at a distance of ψ or less in any tweet, we consider the co-occurrence to be worth investigating. As, Twitter users often do not follow the standard grammar due to the imposed length restriction, this approach is more suitable for identifying co-occurrences from tweets. Let, $\hat{\mathcal{B}} = \{(w_1, w_2) : w_1, w_2 \in \mathcal{U}^*$ and $dist_t(w_1, w_2) \leq \psi$ for some $t \in \mathcal{T}\}$. The function $dist_t(u, v)$ returns the distance between unigrams u and v in tweet t. In this experiment we have used $\psi = 3$. To determine whether an identified co-occurrence is statistically significant, we have adopted the "Likelihood Ratio" measure for hypothesis testing of independence proposed in [17].

For each identified co-occurrence we calculate its likelihood ratio. Likelihood ratio a ratio of two hypotheses that tells how much more likely one hypothesis is over another. The hypothesis of independence $\mathcal{H}_1$ states that there is no association between the words in the co-occurrence beyond chance occurrences. The second hypothesis $\mathcal{H}_2$ states that the association between the words in the co-occurrence are statistically significant. The likelihood ratio of the two hypotheses is $\lambda = \frac{L(\mathcal{H}_1)}{L(\mathcal{H}_2)}$. $(-2\log\lambda)$ is asymptotically a χ^2 distribution. Hence, we reject the hypothesis of independence, $\mathcal{H}_1$, for an identified co-occurrence with 95% confidence if $-2\log\lambda \geq 7.88$, which is the critical value for χ^2 distribution with 1-degree of freedom at confidence level $\alpha = 0.005$. Let, $\mathbf{L}$ be a $1 \times |\hat{\mathcal{B}}|$ vector holding the values of $(-2\log\lambda)$ for the co-occurrences in $\hat{\mathcal{B}}$. Therefore, our identified co-occurrences with statistical significance from the tweet collocation are $\mathcal{B}^* = \{b : b \in \hat{\mathcal{B}}$ and $\mathbf{L}_b \geq 7.88\}$. Let, $\mathcal{U}^b = \{w : ((w, w') \in \mathcal{B}^*$ or $(w', w) \in \mathcal{B}^*)$ and $w, w' \in \mathcal{U}^*\}$.

We define the lexical graph for the tweets of each topic cluster as an undirected weighted graph $G = (\mathcal{U}^b, \mathcal{B}^*, \mathbf{W})$, where each unigram in $\mathcal{U}^b$ is a node in the graph and each identified co-occurrence in $\mathcal{B}^*$ defines an edge connecting two nodes. As $-2\log\lambda$ measures the strength of association between two unigrams, we have used it as the edge weight between the nodes corresponding to those unigrams. The weight matrix $\mathbf{W}$ is defined as follow:

$$\mathbf{W}_{ij} = \begin{cases} \mathbf{L}_b & \text{if } b = (w_i, w_j) \in \mathcal{B}^*, \\ 0 & \text{otherwise.} \end{cases}$$

Identifying Tweets Relevant to the Topic

We have used a graph-based ranking algorithm on our constructed lexical graph to identify key points discussed in the tweet collection. One of the most famous graph-based ranking algorithm, the PageRank [15], is based on intuitive notion of endorsement. In PageRank, a page can have high ranking if many pages point to it or some other high–ranking pages point to it. Similarly, in our tweet collection, if a word co-occurs with many different words in the identified statistically significant co-occurrences, we can interpret it as an important event-related word that has been used by many users to report some important happenings in the event. Moreover, the words in the lexical graph endorse each other in proportion to their strength of association.

$$\mathbf{PR}(v_j) = (1 - d) + d \sum_{(v_i, v_j) \in E} \frac{\mathbf{W}_{ij} * \mathbf{PR}(v_i)}{\sum_{(v_i, v_k) \in E} \mathbf{W}_{ik}} \quad (1)$$

We have applied the weighted, undirected version of PageRank algorithm defined in eq. 1 on the constructed lexical graph for each topic cluster. Parameter $d \in (0, 1)$, called the *damping factor*, is the probability at each page that the random surfer will get bored and request another random page. We have used $d = 0.85$ following [15]. Initially, $\mathbf{PR}(w)$ is set to 1 for all $w \in \mathcal{U}^b$. Initial scores of the nodes can be set to any unique value [15]. Eq. 1 is a recursive equation which iterates until convergence. Following [18], we have used a convergence threshold of $\zeta = 0.0001$; i.e., the algorithm converges if in two successive iterations, the rank of any of the nodes does not change yb more than ζ. Upon completion, each unigram $w \in \mathcal{U}^b$ receives its score in $\mathbf{PR}(w)$. Let ρ is a vector in $\mathbb{R}^l$ such that $l = |\mathcal{U}^b|$ and $\rho = \{\mathbf{PR}(w) : w \in \mathcal{U}^b\}$. Let, $y : t \rightarrow \{0, 1\}^l$ be a function representing the set of words

[1] http://www.textfixer.com/resources/common-english-words.txt

TABLE I. TRACK KEYWORDS USED FOR TWITTER SEARCH API

#debate, #debates, #Denverdebate, #election2012, "#obama", "#romney", #barakobama, #mittromney, #obama2012, #romneyryan2012, #presidentialdebate

$w \in \mathcal{U}^b$ present in tweet t. Then the score associated with any $t \in \mathcal{T}$ is $s(t) = y(t) \cdot \rho$. The score of a tweet is a relative measure indicating how many of the popular co-occurrences or terms the tweet contains in comparison to the other tweets in the collection. After removing the near duplicate tweets from the collection, the tweets are sorted in descending order of their scores. Top-K tweets from each topic cluster are recommended to the users. The collection of recommended tweets from all clusters forms the "**recommended set**" for the event.

Many of the tweets convey the same information in slightly different forms which often frustrates search users looking for new content [2]. We removed duplicate tweets during pre-processing step. Tao et al. [2] did a comprehensive study on the effectiveness of various similarity measurement methods for identifying tweets demonstrating different levels of similarity. For simplicity, we use a variation of Jaccard distance, sometimes referred to as Simpson or Overlap distance [19], to remove near-duplicate tweets from the set of tweets recommended to the users. Let, the function $S(t)$ returns the set of words in tweet t. Then, Simpson distance between two tweets is defined as $simp(t_1, t_2) = 1 - \frac{|S(t_1) \cap S(t_2)|}{min(|S(t_1)|, |S(t_2)|)}$. The set of tweets, $\mathcal{R}_l$, recommended for each cluster $\mathcal{C}_l \in \mathcal{C}$ is selected by maximizing the objective function in eq. (2).

$$\mathcal{R}_l = \{\arg\max_{t_i \in \mathcal{C}_l \setminus \mathcal{R}_l} s(t_i) : \forall t_j \in \mathcal{R}_l \; simp(t_i, t_j) < \tau \text{ and } |\mathcal{R}_l| = K\} \quad (2)$$

IV. EVALUATION AND RESULTS

The objective of the proposed method is to select a reasonably small set of tweets that can summarize the happenings of a real-time event. We call this set the "recommended set". We have compared the performance of the proposed method against two other baseline methods. We have used precision and recall as the performance measures. We have used the first presidential debate held on October 3, 2012 between President Obama and Governor Romney as the real-time event to test the performance of the competing models. We have used Twitter's Streaming API with the track keywords used in table I to collect a total of 270,337 English tweets posted by 212,308 different users during the course of the event. For performance evaluation, we have divided the 90-minute event into fifteen 6-minute segments and generated recommended set for each segment independently. Here we present the evaluation process in detail.

A. Focal Points

To measure the performance of the models, we need ground truth data describing what exactly happened in the event. Four annotators went through the video of the debate, its transcript, the highlights presented by major news media such as BBC, CNN and Fox News and tried to evaluate independently which points should be included in a summary of the event. Of course, a search user would not be interested in every single detail of the proceeding. Hence, the annotators were asked to pick up the important moments from each debate segment. An important moment might be a key discussion point or a rhetoric made by a candidate. Finally, we selected only those moments from each segment on which majority of the annotators could reach in accord. This set contained an average of eighteen important moments per segment. We call them the *focal points* of the segment. For segment $i \in [1, 15]$, the set of focal points are denoted as $\mathcal{S}_i$. Table II presents a sample of identified focal points from some selected debate segments.

TABLE II. SOME EXAMPLE FOCAL POINTS

Segment	Focal Point
1	*The question here tonight is not where we've been, but where we're going.*
5	*Two wars were paid for on a credit card, two tax cuts that were not paid for and then a massive economic crisis.*
8	*Governor Romney's plan would turn medicare into a voucher program.*
9	*Does anybody out there thinks that the big problem we had is that thre was too much oversight and regulations on Wall Street?*
10	*In Obamacare, an unelected board will tell people what kind of treatment they can have.*
13	*We're a nation that belives that we're all children of the same God.*

B. Recommended Set

The tweet collection was also segmented according to the time boundary of the debate segments; i.e. any tweet generated within the beginning and ending time boundary of a debate segment was put in the corresponding tweet segment. Let, $\mathcal{T}_i$ denote the tweet segment corresponding to debate segment i. We applied the competing models on each tweet segment separately to produce a recommended set for the debate segment.

We first divided the tweets in $\mathcal{T}_i$ into topical clusters. Each of the competing models were applied to each topic cluster to select a set of K tweets that they identified as suitable for putting in the recommended set. Hence, if $\mathcal{T}_i$ had k topical clusters, then the recommended set for the debate segment i would have $k * K$ tweets in it. In this evaluation we have used $K = 10$. We observed that, by increasing the value of K recall increases and precision decreases for all three models. This has also been reported in [3].

C. Evaluation Measures

We have used the standard measures of *precision* and *recall* for comparing the performance of the proposed model against the baseline models. *Recall* for a segment i is the fraction of focal points in $\mathcal{S}_i$ that has been referred or mentioned in any of the tweets in the recommended sets for segments i to the last segment, 15. It should be noted here that, though a tweet t might have been generated within the boundary of a particular debate segment, say j, it can refer to a focal points in $\mathcal{S}_p$ where $p \leq j$.

Precision for segment i is the fraction of tweets in its recommended set, which satisfies either of the following criteria:

- It refers to some focal point in $\mathcal{S}_p$ where $p \leq i$, or
- It can be categorized as a "narration" tweet.

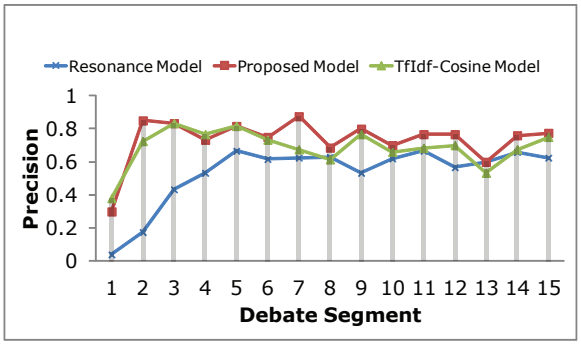

Fig. 1. Performance comparison between the proposed model and the baseline models at $K = 10$ in terms of Precision

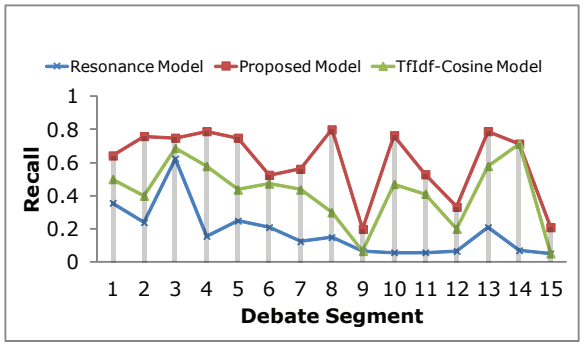

Fig. 2. Performance comparison between the proposed model and the baseline models at $K = 10$ in terms of Recall

A "*narration*" tweet does not refer to any focal point of the debate. However, it conveys useful insight about the proceedings of the debate. For example, "*Jim Lehrer needs to fake a heart attack to get some moderating done. It is a complete chaos here ...#debates*". This tweet certainly does not refer to any points discussed in the debate. However, it gives a clear picture of the debate environment.

D. Baseline Models

1) tf-idf-cosine Model: This model has been used in earlier research [3] for determining the relevance of tweets to be recommended to the users. The objective of this model is to select those tweets which are closest to all other tweets in the tweet collection. Hence, the objective function is quite similar to the objective function we laid out in the introduction part of this paper. Hence, this model is a good candidate as a baseline model for evaluating the performance of the proposed model. In this model, each tweet is represented as a length $|\mathcal{V}|$ vector of *tf-idf* of its constituent words, where $\mathcal{V}$ is the set of vocabulary. Let, $tf_{w,t}$ be the normalized term frequency of term w in tweet t. The inverse document frequency of a term in the tweet collection $\mathcal{T}$ is represented as $idf_{w,\mathcal{T}} = \log \frac{\mathcal{T}}{|t \in \mathcal{T} : w \in t|}$. tf-idf$_{w,t} = tf_{w,t} * idf_{w,\mathcal{T}}$. Cosine similarity between two vectors u and v is defined as $cosine(u, v) = \frac{u.v}{\|u\|\|v\|}$. The ranking score of a tweet t is determined as: $score(t) = \sum_{t' \in \mathcal{T}} cosine(t, t')$. For the sake of fairness among competing models, we removed duplicate and near duplicate tweets from the model's *recommended set* using *Simpson distance* method described earlier.

2) Resonance Model: Twitter uses a specialized ranking function which among other indicators also considers the resonance signal to compute a relevance score for each tweet [1]. Resonance signal includes the users' interactions with a tweet, e.g. number of times the tweet has been replied or retweeted. Hence, a relevant tweet, which is retweeted many times, enjoys higher score. The "Resonance Model" uses the *retweet–count* (number of times a tweet has been retweeted) of a tweet to emulate the resonance signal. In this model, the retweet–count of a tweet is considered as its relevance score. To avoid duplicate tweets from appearing in the tweets recommended from each cluster, only the tweet with the highest retweet count among a set of *peer retweets* (set of retweets whose source tweet is the same) is considered. Simpson distance is used to get rid of near duplicate tweets.

E. Parameter Setting

Our model has several parameters as described in the "Proposed Method" section. Some of these parameter values were chosen based on the results reported by earlier research works and we mentioned them where the parameters have been introduced. For the parameter τ in eq. (2), we determined its value using a development dataset while checking over a range of values, [0,1]. The development dataset was different from that used for the performance evaluation of the proposed method. Based on the experimental results we finally set the values $\tau = 0.6$.

F. Results

Figure 1 and figure 2 show the achieved precision and recall by the three competing models. The proposed model performs better than both the baseline models in terms of both precision and recall. We shall discuss the results in detail in the discussion section.

To evaluate the impact of topical clustering, we performed an experiment where the proposed tweet retrieval algorithm was applied on each tweet segment without performing topical clustering. Figure 3 shows the results of the experiment. To evaluate the impact of duplicate removal we applied the *tf-idf-cosine* model on each tweet segment $\mathcal{T}_i$ and generated the recommended set without removing duplicate or near duplicate tweets. The results are shown in figure 4.

V. Discussion

We refer to the tweets pointing to some focal points as "*citation*" tweets. Tweets that are neither *"citation"* nor *"narration"* tweets are referred to as *"distant"* tweets. From figure 1, it is evident that the precision of the proposed model outperforms that of both the baseline models for almost all debate segments. The precision of the resonance model is substantially lower than the other two models for the first six segments. We found that many distant tweets, which were generated even days before the debate commenced, already had high retweet-count and were thus included in the resonance models recommended set. However, as the debate progressed, citation and narration tweets acquired sufficient retweet-count to override the distant tweets. Hence, the precision of resonance model ameliorated in later part of the debate. Though the *tf-idf-cosine* model closely contends with the proposed model in terms of precision, it fails to demonstrate similar

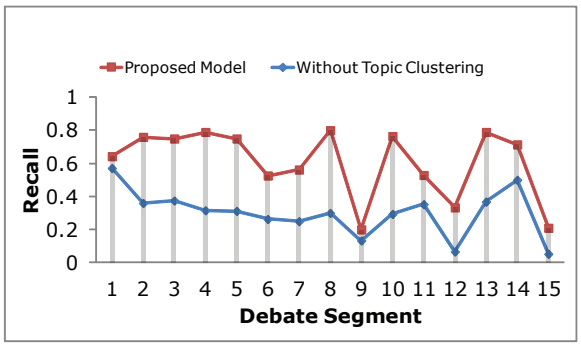

Fig. 3. Impact of topical clustering in performance at $K = 10$ in terms of Recall

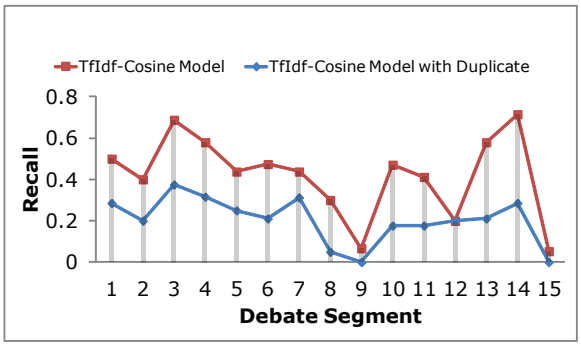

Fig. 4. Impact of duplicate and near–duplicate removal from recommended set at $K = 10$ in terms of Recall

competence in terms of recall for most debate segments (figure 2).

Though vanilla LDA does not offer the best topical clustering of tweets [20], it still helps to improve the recall as suggested by figure 3. Hence, replacing vanilla LDA with another topical model, which is customized for short text, would further ameliorate the recall. Removal of duplicate tweets also helps to achieve better recall (figure 4).

Recalls for all three models plummet for the ninth debate segment (figure 2). A closer analysis revealed that the segments preceding and following segment nine, were related to medicare and Obamacare respectively. Segment nine was mostly about regulations (table II). Common people are much more concerned about healthcare and social security than Wall Street regulations. Hence, the discussion points of the tweets in these three segments were dominated by healthcare and social security related issues.

VI. Conclusion

In this paper, we have presented a method for summarizing the important occurrences of a real-time event using a limited number of relevant-tweets. The proposed method is completely unsupervised and hence can be leveraged for any public event that causes upsurge of traffic in Twitter and occurs within a limited span of time. The method tries to incorporate in the recommended set, the most informative event-relevant tweets that cover various topical aspects of the event while minimizing repetition of information. Evaluation performed on real-world dataset shows that the method can summarize real-time events with high precision and recall. In our future work we plan to use online versions of LDA as proposed in [21], [22]. This would help the proposed method to generate real-time dynamic summaries of ongoing events.

References

[1] M. Busch, K. Gade, B. Larson, P. Lok, S. Luckenbill, and J. Lin, "Earlybird: Real-time search at twitter," in *Proc. IEEE Data Engineering (ICDE)*. IEEE, 2012, pp. 1360–1369.

[2] K. Tao, F. Abel, C. Hauff, G.-J. Houben, and U. Gadiraju, "Groundhog day: near-duplicate detection on twitter," in *Proc. of WWW*, 2013, pp. 1273–1284.

[3] D. Chakrabarti and K. Punera, "Event summarization using tweets," in *Proc. of ICWSM*. AAAI, 2011, pp. 66–73.

[4] B. Sharifi, M.-A. Hutton, and J. K. Kalita, "Experiments in microblog summarization," in *Proc. of IEEE Second International Conference on Social Computing*, 2010.

[5] J. Nichols, J. Mahmud, and C. Drews, "Summarizing sporting events using twitter," in *Proc. of IUI*. ACM, 2012, pp. 189–198.

[6] Y. Hu, A. John, D. D. Seligmann, and F. Wang, "What were the tweets about? topical associations between public events and twitter feeds," in *Proc. ICWSM*. AAAI, 2012.

[7] B. OConnor, M. Krieger, and D. Ahn, "Tweetmotif: Exploratory search and topic summarization for twitter," in *Proc. of ICWSM*. AAAI, 2010, pp. 2–3.

[8] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, "Twitinfo: aggregating and visualizing microblogs for event exploration," in *Proc. of CHI*. ACM, 2011, pp. 227–236.

[9] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[10] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[11] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *Proc. of ICWSM*, vol. 5, no. 4. AAAI, 2010, pp. 130–137.

[12] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews," in *NAACL-HTL*. ACL, 2010, pp. 804–812.

[13] Z. Niu, D. Ji, and C. Tan, "I2r: Three systems for word sense discrimination, chinese word sense disambiguation, and english word sense disambiguation," in *Proc. of the 4th International Workshop on Semantic Evaluations*. ACL, 2007, pp. 177–182.

[14] E. Levine and E. Domany, "Resampling method for unsupervised estimation of cluster validity," *Neural computation*, vol. 13, no. 11, pp. 2573–2593, 2001.

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

[16] K. Toutanova, D. Klein, C. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proc. of NAACL-HLT-Volume 1*. ACL, 2003, pp. 173–180.

[17] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational linguistics*, vol. 19, no. 1, pp. 61–74, 1993.

[18] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts," in *Proc. of EMNLP*, vol. 4. Barcelona, Spain, 2004, pp. 404–411.

[19] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring semantic similarity between words using web search engines," in *Proc. of WWW*, vol. 7, 2007, pp. 757–786.

[20] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proc. of WWW*, 2013, pp. 1445–1456.

[21] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent dirichlet allocation," in *advances in neural information processing systems*, 2010, pp. 856–864.

[22] L. Yao, D. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," in *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 937–946.