A Network Based Stratification Approach for Summarizing relevant Comment Tweets of news articles

Roshni Chakraborty, Maitry Bhavsar, Sourav Dandapat, and Joydeep Chandra

Indian Institute of Technology, Patna, India
{roshni.pcs15,bhavsar.mtcs15,sourav,joydeep}@iitp.ac.in

Abstract. Social media platforms like Twitter have become extremely popular for exchanging information and opinions. The opinions expressed through Twitter can be exploited by news media sources to obtain user reactions centered around different news articles. A comprehensive summary of the user reactions with respect to a news article can be crucial due to various reasons like: i) obtaining insights about the diverse opinions of the readers with respect to the news and ii) understanding the key aspects that draw the interest of the readers. However extracting the relevant opinions from tweets is a challenging task due to the enormous volume of contents generated and difference in vocabulary of social media contents from the published article. Existing supervised learning based techniques yield poor accuracy due to unavailability of sufficient training data and large heterogeneity in the features of various news articles, while the unsupervised techniques fail to handle the noise and diversity of the tweets.

In this paper, we propose a network community based unsupervised approach that effectively handles the problem of noise and diversity in tweet feeds to capture the relevant and the diverse opinions with respect to a news article. Using a combined metric that considers both relevance and diversity, we show that our proposed approach produces $16 - 25\%$ improvement over existing schemes. Results based on human annotations also validate the effectiveness of the extracted summary tweets with respect to specific news articles.

Keywords: Tweet Summarization, Summarization, Twitter, News Articles, Relevance, Diversity

1 Introduction

The popularity of Twitter as a news media platform encourages a large fraction of the news readers to post tweets on news articles so that the same can be widely discussed by a large user base [1]. A comprehensive summary of the users' opinions extracted from these tweets would help the readers to obtain an insight of the larger debate centered around a news published in the article

and at the same time it would also help the news agencies to understand the key aspects that are likely to draw the interest of the readers. However, the summary must be relevant to the news article and at the same time it should represent a holistic view (diverse in nature) of the users' opinions. Ensuring above phenomenon in summary is a fundamentally challenging job. These challenges need to be collectively addressed, otherwise, would yield poor accuracy due to propagation of errors occurring at each step of the process.

Extracting relevant tweets based on keyword similarity is not effective because of the large difference in vocabularies of the news articles and the corresponding tweets [2]. There are few supervised learning approaches that mainly dealt with the problem of extracting relevant tweets for specific news articles. Like recent works in [2–4] have proposed mechanisms for mapping tweets to articles using supervised learning approaches based on language as well as topic models. Further, certain tweet specific features like presence of hashtags and user mentions, language independent features like non-lexical markers as well as lexical features like presence of question words and sentiment words have also been used in classification. However, the major drawback of such supervised classification approaches is the requirement of large manually annotated dataset to train the classifiers. Moreover, annotation is also error-prone as inconsistencies may arise depending upon the perception and knowledge of the annotators.

On the other hand, unsupervised summarization techniques like LexRank [5] and Latent Semantic Analysis (LSA) [6] that are based on keyword frequency works well for multi-document summarization but are inefficient for summarizing opinion tweets with respect to specific news articles. This is primarily because, in contrast to text documents, tweets are inherently noisy with low word count and high variance in word frequencies. Further they also feature high redundancy (high number of similar tweets) as well as large diversity in the vocabulary. These existing techniques are not suitable to handle these features and hence fail to capture the diverse opinions in the tweet summary. Hence to produce an effective tweet summary, techniques must be developed for extracting tweets with respect to news articles that satisfy the following objectives:

1. *Relevance*: The tweets generated in the summary must reflect relevant opinions or feelings of the readers with respect to the news article. They should not be mere facts or quotes from the news article. Metric related to relevance is introduced shortly.
2. *Diversity*: The tweets generated should capture the diverse opinions of the readers and should not be similar in nature. Metric related to diversity is introduced shortly.

In this paper we propose an unsupervised approach for summarizing diverse opinion tweets related to specific news articles that satisfy both the above stated criteria. To handle word diversity and vocabulary gap of the tweets (with respect to the news article), the relevant tweets are captured by considering both the keyword similarity with respect to the news article as well as contextual similarity captured by features like presence of relevant co-occurring hashtags. Subsequently, a weighted tweet network is formed by linking relevant tweets based

on their relative keyword similarity as well as relevance of those words with the published article. Finally, to capture the diverse opinions, we group tweets by applying a community identification algorithm on the weighted tweet network. Communities and the representative tweets from those communities are selected for summary based on certain relevance and diversity metrics. Validation on empirical data set of around 800 articles shows that tweets extracted using our proposed approach are not only relevant with respect to the news article but maintains high diversity as compared to tweets extracted by using existing approaches like LexRank and LSA.

The organization of the paper is as follows: We discuss the related works in section 2. In section 3, we present a formal definition of the problem, a brief description of certain preliminary approaches that we use along with an overview of the dataset that we use for our experiments. In section 4, we detail our proposed approach. We discuss the experiments and observations in section 5 and finally conclude our study in section 6.

2 Related Works

There is a plethora of available research related to the various aspects of our proposed work like i) extracting relevant posts ii) summarization and iii) opinion mining. A few of these works are highlighted next.

Extracting Relevant Posts: Although features like word similarity [7, 8] can extract blogs relevant to news articles, it fails in micro-blogs like Twitter due to the vocabulary difference between tweets and the original news article [2].

Krestel et al. [4] extracted relevant and diverse tweets corresponding to news articles by applying language modeling, topic modeling and logistic regression on a set of tweets with either specific seed hashtags like *obama*, *snowden*, *merkel* or hashtags co-occurring with these hashtags. However, our experiments indicate that topic models or language models fail to derive relevant information when each tweet is treated as a document. Certain works attempt to extract relevant tweets through a search process starting from the tweets containing the URL of the article [9] or the hashtags specific to the news [10] but these systems fail to capture many tweets that are relevant to the news article. In [11] Cao et al. generated relevant comments for Chinese news by mining micro-blog posts. However, the authors did not consider diversity of the comments while selecting them.

Unsupervised topic modeling techniques like Latent Dirichlet Allocation [12] fail to extract topics from tweets due to inherent noise present in tweets [13]. Mehrotra et al. [14] pools all tweets of a hashtag in a document which improves the performance over the existing topic modeling techniques in Twitter. TweetMotif [15] follows an unsupervised approach to cluster messages by considering hashtags as an approximate indicator of tweet topics. However, our observations indicate that neither keywords (as used in topic model based techniques) nor hashtags can independently represent a news item uniquely. Tweets related to a specific news article may include several hashtags of different news events and

several news articles covering different news may be mapped to same hashtag set. Thus, extraction of specific and appropriate hashtag set for a particular news article requires both content similarity with the specific news article and contextual similarity with the news tweets(i.e. mostly quotes from the news article.)

Summarization: Public opinion to a news article covers multiple facets; thus the extracted relevant comment tweets should be summarized in a way to provide a holistic view of the opinions.

Document summarization techniques like LexRank and LSA have also been applied for summarizing tweets. In LexRank, sentences are linked based on their keyword similarities, thus forming a weighted network. Subsequently, a sentence is selected into the summary based on the importance of it's constituent keywords and it's connections to other important sentences. However, this algorithm fails in summarizing opinion tweets that are highly redundant and have large diversity in the vocabulary. LSA uses singular value decomposition on the term sentence matrix to derive a set of concepts and further select representative sentences from different concepts to maintain *diversity* in the summary. However, the major problem in applying LSA for summarizing news-specific tweets is the large variance in the frequency of the terms in the term-sentence matrix.

For multi-tweet summarization, researchers have proposed using both contextual features of the tweets as well as social influence of the users to extract predetermined length summary tweets of certain events [16, 17]. Entity based summarization [18] algorithms use *affinity propagation* algorithm to group relevant hashtags into coherent topics and further summarize tweets from these groups into most relevant, insightful and diverse opinions of an entity. However, unlike the single dimensional search space of an entity the search space of news specific tweets is inherently multi-dimensional that is difficult to explore.

Opinion Mining: Opinion mining is a process of extracting and summarizing important opinions about a particular product or entity [19]. Previous works accurately identify the semantic orientations [20], syntactic relations [21] and sentiments [22–24] between opinion words and target to provide a holistic view of the existent expressions. Recent studies have also been made to extract relevant comment tweets (expressing opinions on news articles) specific to an article using supervised learning approaches [3]. However, creating such a universal dataset for all kind of news articles is difficult. Further, a common set of features might not be able to model the variance in characteristics of the comment tweets which differs across the entities involved and the region.

We next provide a formal definition of the problem and outline our approach with respect to the drawbacks in existing mechanisms.

3 Problem Definition and Preliminaries

In this section, we initially present a formal description of the problem. Subsequently, we provide a brief overview of the proposed approach to address the problem.

3.1 Problem Definition

Given a news article $\mathcal{N}$, let $\mathcal{T} = \{t_1, t_2, \cdots\}$ be the set of opinion tweets relevant to $\mathcal{N}$. The objective is to select at most n tweets in our final summary ($\mathcal{O}_n$) in such a way that it maximizes both the pairwise diversity of the tweets selected and relevance of the selected tweets with the news article. Further, we can impose additional constraints on the minimum value of the total relevance score (say R_{min}) of the selected tweets in $\mathcal{O}_n$. We can thus formally define the problem as:

$$\begin{aligned} f = \arg\max(&\alpha Div(O) + \\ &(1-\alpha)Sim(O, \mathcal{T} - O)) \quad O \subseteq \mathcal{T}, \alpha \in [0,1] \\ \text{subject to} \quad &Rel(\mathcal{N}, O) \geq R_{min} \\ &||O|| \leq n, \end{aligned} \tag{1}$$

where $Div(O)$ is a function representing the pairwise diversity of the tweets in subset O and $Sim(O, \mathcal{T} - O))$ represents the similarity score of the tweets in O with respect to the tweets not selected in O (i.e. $\mathcal{T} - O$). The similarity function ensures that the representative tweets in the summary captures the holistic opinions expressed through the tweets. The objective is to select a subset O such that f is maximized subject to the relevance of selected tweets with news article (represented as $Rel(\mathcal{N}, O)$) and size constraints. $\mathcal{O}_n$ is that subset O for which f is maximized. Since the problem defined is a 0-1 Integer Programming Problem which is known to be NP-complete [25], existing methods of summarization are mainly approximation to the optimality problem. However, the major challenge for any summarization technique is to handle the redundancy of the tweets. The large diversity of the tweet vocabulary as well as the noise further adds to the complexity in tweet summarization.

3.2 Outline of the Proposed Approach

We address these issues in the proposed approach by initially finding the set of relevant opinion tweets for a news article based on their keyword similarity as well as their contextual similarity. Subsequently, we create a weighted network of the relevant tweets, where the weight is determined based on the keyword similarity in tweets and relevance of those overlapping keywords with the news article. The tweet network is analyzed further to identify the different closely connected tweets (communities), while each community represents a set of similar opinions. We then sample representative tweets from each of these communities using a *maximum marginal relevance* measure that satisfy the relevance constraint and also ensures relative diversity among the selected tweets, thus satisfying equation 1. The steps of our proposed approach are highlighted in figure 1. Steps are described in more detail in section 4, where we describe our proposed approach.

3.3 Dataset

The dataset consists of both news articles as well as tweet sets.

News Article Dataset: We have considered only major (front page) political news from *New York Post*.We have crawled 800 articles for 2 months duration starting from 1^{st} *July*, 2016 to 31^{st} *August*, 2016.
Twitter Dataset: Twitter allows free access to approximately 1% random tweets using the Streaming API. We have used this API to crawl the tweets for 2 months duration starting from 1^{st} *July*, 2016 to 31^{st} *August*, 2016. We collected around 2.1 billion tweets with 77 million hashtags and 36 million unique users.

3.4 Preprocessing

We next highlight the major steps of preprocessing followed for both the articles as well as the tweets.

Preprocessing of Articles We initially extract the nouns and verbs (that are important to differentiate articles with similar entities) from the headline and the body of every article, using the POS tagger of the NLTK toolkit. These keywords are further ranked by their *tf-idf* score computed based on the documents published in the same day. We then select the top 10 keywords of every article as its representative keywords [26].

Preprocessing of Tweets The raw tweets collected are inherently noisy and hence needs to be cleaned for further analysis. We briefly outline the preprocessing steps that we perform on the tweets and then describe the extraction of related tweets of an article.
Types of Tweet: Based on the nature of tweet contents with respect to a specific news article, we categorize tweets into 3 different types. We provide one example of each type using tweets related to the news article entitled *Newly released Clinton emails show favors for foundation donors* published in New York Post on 22^{nd} August, 2016.

Definition 1. *News Tweet* is a tweet with no sentiment polarity that states a fact specific to the news article or to a related news. Most often these are direct quotes, for example, the tweet*, new emails show Clinton foundation sought access to state department on donors behalf* is a news tweet based on the above mentioned article.

Definition 2. *Opinion Tweet* is a tweet that represents an opinion regarding a news article and has some associated sentiment polarity. For example, the tweet *Hillary Clinton is guilty of sin, is a traitor to our country, and is unfit to be president* indicates a negative sentiment related to the same article.

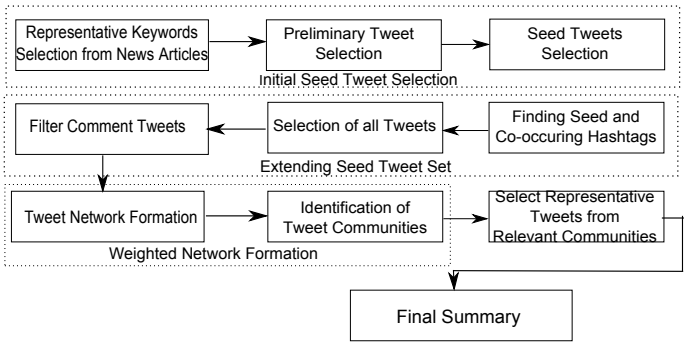
Definition 3. *Irrelevant Tweet* is a tweet that does not provide any relevant information with respect to the news article. For example, the tweet *rt twitter users call out pbs for using stock footage of dc fireworks show* is not related to this article.

As the definition of the different types of tweets suggest, each type of tweet highlights different kind of information in relation to an article. We are only interested in the *opinion tweets* with respect to a news article because it captures

the reaction of readers. We identify an opinion tweet by it's sentiment score that we calculate using Vader Sentiment Analyzer[27]. The steps of further preprocessing used to filter out the opinion tweets with respect to a news article is described next.

1. *Keyword Extraction* We remove the duplicate tweets and subsequently remove user mentions, retweet tags and URLs as well as the stop words from the tweet text. From the remaining words, we extract nouns and verbs using a POS tagger that forms the representative keywords of the tweet.
2. *Seed Tweet Set Creation* We create an initial *seed tweet set* of an article that includes tweets having maximum keyword overlap with the article's representative keywords. The *seed tweets* are used to extract a set of most frequent hashtags(*seed hashtags*) that are present in the *seed tweets set* and the *co-occurring hashtags*, that appear in more than a threshold percentage of tweets of any *seed hashtag*. The collective set of seed as well as co-occurring hashtags is termed as a *related hashtag set* of the article.
3. *Extended Tweet Set* We extend the seed tweet set by including relevant opinion tweets that have at least one of the related hashtags or one level of indirection in word overlap. Due to short length of tweets and vocabulary gap, we might not find direct word overlap of tweets with news article. However, with high probability we can find word overlap of such tweets with other tweets which does have direct word overlap with news. This step helps us to extend our relevant tweet set as well as to get rid off irrelevant tweets.

We next discuss our proposed approach that is applied on the filtered tweets.



FF

Fig. 1. Block Diagram Representing the Proposed Approach

4 Proposed Approach

In this section, we detail our methodology to extract relevant and diverse opinion tweets specific to the targeted article. We first describe the steps for creating a tweet network from the extended opinion tweet set and identifying the different communities in the network. Subsequently, we highlight the method of selecting the summary of opinion tweets from these communities.

4.1 Tweet Network Formation and Community Identification

We consider extended tweet pool for forming network and community analysis.

Tweet Network Formation: To form the tweet network, we consider every tweet as a node. Two nodes are connected by an edge if the edge weight between them exceeds a predefined threshold. Edge weight is computed based on a parameter that we term as *significance score*. *Significance score* of a keyword represents the importance (based on tf-idf) of a keyword with respect to the news article. If any keyword, say w, belongs to the representative keywords of the article then its significance score is the relative importance of w with respect to the least important keywords in representative keyword set, otherwise it is 1.

$$\begin{aligned} \mathcal{S}(w) &= \frac{\text{tf-idf}(w)}{min(\text{tf-idf}(u) : \forall u \in \mathcal{K})} \quad \text{if } w \in \mathcal{K} \\ &= 1 \quad \text{otherwise} \end{aligned} \tag{2}$$

The weight W_{ij} of the edge connecting nodes i and j is calculated as the sum of the significance score of intersecting keywords of the two tweet nodes, normalized by the total significance score of all the unique keywords. Thus if K_i and K_j represent the keywords in nodes i and j respectively, then

$$W_{ij} = \frac{\sum_k \mathcal{S}(k)}{\sum_l \mathcal{S}(l)} \quad k \in K_i \cap K_j, \quad l \in K_i \cup K_j \tag{3}$$

The edge weight between two tweet nodes will be high if they have significant keyword overlap; however if the non-overlapping keywords have high significance score (implying tweets carry different opinions and should be part of different communities), then the edge weights will be significantly lower.

Extracting Tweet Communities: The communities in a weighted network are identified by logically partitioning the network into groups of nodes where the nodes within a partition are highly interconnected as compared to interconnections across the partitions. The quality of partitioning is measured by a term called *modularity* that compares the actual weight of the edge between two nodes within a community to the possible weight in a scenario if the connections between them were purely random.

By identifying the communities in the network, the group of tweets representing a similar class of opinions are identified, whereas each group represents a different opinion class. We show later in section 5, partitioning of tweets into communities tends to make the frequency of keywords uniform with less diversity. This eases the process of selecting representative tweets from the community for summarization. We next discuss the steps taken to extract the final set of representative summary tweets.

4.2 Extracting Final Summary of Relevant and Diverse Tweets

We describe the steps followed to extract the final summary of relevant and diverse tweets based on the identified communities in the tweet network.

Ranking Communities Based on Relevance: For each community detected using the Louvain algorithm, we initially compute a community relevance score (*CommRel*) based on the cosine similarity of the top ten keywords of a community (the highest frequency keywords in all tweets of the community) and that of the top ten keywords of the seed tweets. We consider a community as relevant to the news article if the *CommRel* value is above a predefined threshold. The irrelevant communities are discarded and not used for further consideration.

Selecting a Diverse Tweet Set from Relevant Communities: After filtering out the irrelevant communities, we select a diverse set of tweets from relevant communities.The final summary tweets are selected from each relevant community C_j such that *maximum marginal relevance* [28] is achieved. *Maximum marginal relevance* ensures selection of such tweets which are highly relevant with seed tweet set, however, having minimum similarity with already selected tweet set.

$$MMR_S = max_{T_i \in C_j}[\beta Sim(K_{tweet}, K_{seed}) - (1-\beta)Sim(K_{tweet}, K_{selected})] \quad (4)$$

where, $0 < \beta < 1$, T_i are the tweets of the relevant community C_j and $Sim(K_a, K_b)$ denotes the similarity of two keyword sets K_a and K_b. Finally, we merge the summaries from all the communities to get the aggregated summary. We next highlight the evaluation technique to observe the efficiency of the proposed method and the observed experimental results.

5 Results and Discussion

In this section, we initially describe the performance metrics and the existing summarization approaches and then compare the performance of the proposed approach with the existing summarization algorithms like LSA and LexRank. Further, we also perform a manual annotation of the extracted tweets to validate the accuracy of the proposed relevance measure. Finally, the specificity of the extracted tweets with respect to the news article is observed using a case study.

5.1 Performance Metrics

We use four different performance metrics to evaluate the quality of the summary (tweet sets).

Mean Relevance Score: The relevance score of a tweet with respect to an article is defined by the cosine similarity of the tweet keywords with the representative set of words of the news article. If a_i and b_i denotes the frequency of i^{th} term in two documents a and b respectively, then the cosine similarity S_{ab} of the two documents will be represented as

$$S_{ab} = \frac{\sum_i a_i b_i}{\sqrt{\sum_j a_j^2}\sqrt{\sum_j b_j^2}} \quad (5)$$

A value near to 1 indicates high similarity between a and b. If T represents a tweet set and S_{id} denotes the relevance score (cosine similarity) of the i^{th}

tweet and the news article d then the mean relevance score, $\bar{R}_T$, is given as $\bar{R}_T = \frac{\sum_{i=1}^{k} S_{id}}{k}$, where k denotes the total number of tweets in T.

Mean Diversity Score: We calculate diversity score of tweets based on two different similarity measures.

1. *Cosine Similarity* : If S_{ij} denotes the cosine similarity (as mentioned in equation 5) of the i^{th} and j^{th} tweets in T, then the mean diversity score, D_T, of the set T is given as $D_T = \frac{\sum_{i,j}(1-S_{ij})}{m}$, where m denotes the total number of tweet pairs in T.
2. *Jensen Shannon Divergence* : *Jensen-Shannon* (JS) divergence [29] provides a measure of similarity between two probability distributions. It's based on *Kullback–Leibler* (KL) divergence. For two discrete probability distributions P and Q of the keywords in two tweets, the JS divergence of P and Q is given as

$$JSD(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M), \tag{6}$$

where $M = \frac{1}{2}(P+Q)$ and $D_{KL}(P||Q)$ is the KL divergence from Q to P is defined as

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}, \tag{7}$$

and represents the information gain achieved in using P instead of Q. If J_{ij} denotes the Jensen Shannon divergence of the i^{th} and j^{th} tweets in T, then the mean diversity score, D_T, of the set T is given as $D_T = \frac{\sum_{i,j}(J_{ij})}{m}$, where m denotes the total number of tweet pairs in T. Both the divergence scores vary between 0 (lowest diversity) and 1 (highest diversity).

As tweets are highly redundant and often contains duplicate information, summary tweets of a news article must ensure both *relevance (i.e., coherence to the main article)* as well as *diversity (i.e., different from each other)*. Thus, we devised the following metrics to measure the quality of summary of any article.

1. *Coverage* : *Coverage*, C_T, of the tweet set T with respect to a news article by a weighted sum of it's mean relevance score ($\bar{R}_T$) and mean diversity score, D_T, i.e., $C_T = \alpha R_T + (1-\alpha)D_T$, where $0 < \alpha < 1$. We select α as 0.5 to give equal importance to both relevance and diversity.
   Though high coverage indicates a good quality summary, however, which factor (relevance or divergence) is contributing how much is not clear from coverage.
2. *Diversity-Relevance Balance Factor* : It is to understand how balanced relevance and divergence are and computed as $S_T = \frac{min(R_T, D_T)}{max(R_T, D_T)}$. A higher score ensures uniformity in both *mean relevance score* and *mean diversity score* which is the objective of summarization.

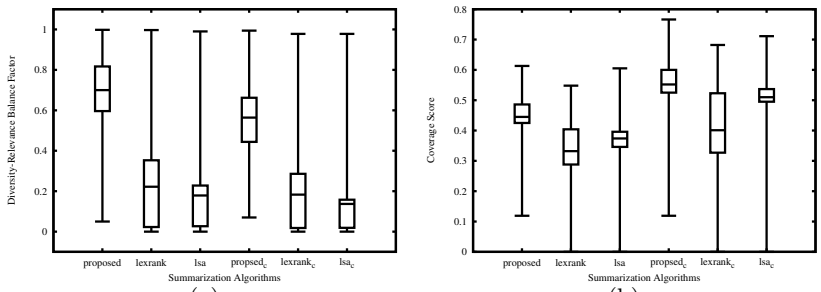
Comparison with Existing Techniques We compare our proposed method with the existing techniques of summarization. Each of these approaches are described as follows:

1. *LexRank:* LexRank is one of the most popular document summarization algorithms that selects *representative sentences* from a document by their *page rank values*. We apply *LexRank* to summarize the extracted opinion tweets of a news article.
2. *LSA:* Latent Semantic Allocation produces the underlying concepts from documents by analyzing the existent term relationship in that document and further select sentences from each concept thus representing the whole document. We apply *LSA* on the extracted opinion tweets of a news article to give a summary of the major opinions of the article.

5.2 Results

In this section, we compare the performance of the proposed approach with all existing methods described above.

Comparing Performance Metrics Initially, for every article, we compute the mean relevance score, mean diversity score, diversity-relevance balance factor and the coverage of the summary tweets extracted using the proposed approach and the existing methods. We then calculate the minimum, first quartile, average, third quartile and maximum of both the coverage score and diversity-relevance balance factor score of all these articles to compare the results of different algorithms.



(a) (b)

Fig. 2. Fig. 2(a) shows Diversity-Relevance Balance Factor and Fig. 2(b) shows coverage score.

Fig. 2 highlights the performance of the proposed and existing approaches in terms of the different metrics and summarization techniques. Fig.2(a) and 2(b) shows the performance when Jensen-Shannon diversity score (first three) and cosine diversity score (last three) is used for summarization. Results shown in Fig. 2(a) indicate that the *Proposed* approach ensures better balance factor compared to existing methodologies. While Fig.(2(b) ensures that the proposed method also maintains better coverage score. The final coverage score of the

proposed approach shows improvement up to 25% and 16% with respect to $LexRank$ and LSA respectively.

		Y	
		Relevant	Irrelevant
X	Relevant	0.798	0.028
	Irrelevant	0.09	0.082
	Summary	**Match** = 89%	**Mismatch** = 11%

Table 1. Fraction of articles for which final extracted tweet sets are marked as relevant or irrelevant based on manual annotation as well as mean relevance score. **X**=Manual Annotation Based and **Y**= Mean Relevance Score Based

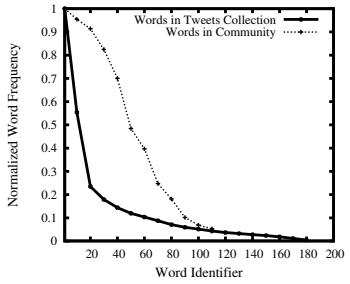


Fig. 3. Comparing the normalized word frequency (**Y** axis) of all the words (**X** axis) in relevant tweets with respect to six articles and the corresponding communities. Number of articles considered is 200.

Comparison using manual annotation To evaluate the accuracy of the proposed relevance measure, we also verified the relevance of the final extracted tweets manually. To validate our results, we generated the final extracted tweets of a random 200 news articles. We provided the set of final tweets and the corresponding articles to the three manual annotators (who does not have knowledge about the details of our algorithm). Each of them annotated all the articles. An annotator marked the tweet relevant or irrelevant based on his knowledge of the article. We strongly encouraged them to try their best to understand the comment before labeling (as tweets are sometimes confusing with informal expressions and sarcasms). To compute inter-annotator agreement, we took the majority of the three for the set of tweets for an article. The final extracted tweet set was classified as relevant to the news article, if a threshold percentage (90%) of its total tweet set was marked as relevant by the majority of the manual annotator. For comparison, we also classify the final extracted tweet set as relevant based on the mean relevance score, if the value is greater than threshold (0.26). We summarize in table 1, the fraction of articles for which the final extracted tweets are classified as relevant or irrelevant using both the techniques. We had repeated the same experiment thrice for different random set of 200 news articles. It can be observed that for 89% of the total articles, the classification made using both the approaches match, thus highlighting the accuracy of the proposed relevance measure.

Effectiveness of communities If we rank words in selected relevant tweet set in terms of importance computed based on word frequency (normalized respect to the highest frequency), then we find very small number of words are very important. These words may be related to main theme, however, a summarization based on this importance would lose diversity. However, when we group similar tweets using community analysis and do the same experiment (normalized frequency separately in each community) there would be important keywords in every community and this would help to select a diverse set of tweets in final summary. Effects of normalization is shown in Fig. 5.2.

Specificity of the Tweets to news article To establish that proposed approach is capable of extracting tweets specific to news article we do the following case study. We considered *four* related articles published in the same day. We show that tweets extracted using the proposed approach are relevant to the corresponding article while it is irrelevant when compared with other articles. In Table 2 we show the mean relevance score of extracted tweets of different news articles related to the same news event, *Dallas Shooting*. As can be observed, the mean relevance scores at the diagonals are much higher than the rest of the values indicating high specificity of the extracted tweets with respect to the corresponding news article.

Extracted Tweet Set for Articles / Article Titles	1	2	3	4
1. *Dallas cop killed in attack survived three tours in Iraq*	0.28	0.04	0.1	0.1
2. *Trump, Clinton postpone campaign events after Dallas attacks*	0	0.28	0	0
3. *Dallas PD first to use a robot to kill suspect*	0.175	0.04	0.43	0.09
4. *Arsenal found in Dallas sniper's suburban home*	0.04	0.04	0.2	0.608

Table 2. Mean relevance score of extracted tweet set against different articles related to Dallas Shooting. A value in cell ij indicates the mean relevance score of article number i with the tweets extracted for article number j

6 Conclusion

In this paper, we have proposed an unsupervised approach to summarize users' opinion on a specific news article. Proposed mechanism also ensures presence of diverse and relevant views in final summary. It can effectively handle the issues like large vocabulary gap and diversity of the tweets with respect to specific news articles while existing summarization techniques like LexRank and LSA fail to do so. We have introduced *coverage* and balance factor as metrics that capture both the relevance and diversity of the summary tweet sets with respect to an article. The summary tweets extracted using our proposed approach have very high coverage as compared to existing summarization approaches. Result shows that our proposed approach produces 16% - 25% improvement over existing

schemes. Further it is also observed that even in the presence of several related news articles and their corresponding tweets, our approach can easily identify the tweets that are relevant and specific to the targeted news article. Although the proposed method has been verified on political news stories, however, it can also be used for extracting summary tweets for other news categories as well. The accuracy of the proposed method can possibly be improved further by exploring techniques that establishes the contextual relationship among the keywords extracted from the articles and tweets.

References

1. H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600.
2. M. Tsagkias, M. De Rijke, and W. Weerkamp, "Linking online news and social media," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 565–574.
3. A. Kothari, W. Magdy, K. Darwish, A. Mourad, and A. Taei, "Detecting comments on news articles in microblogs." *ICWSM*, vol. 2013, 2013.
4. R. Krestel, T. Werkmeister, T. P. Wiradarma, and G. Kasneci, "Tweet-recommender: Finding relevant tweets for news articles," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 53–54.
5. G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Int. Res.*, vol. 22, no. 1, pp. 457–479, Dec. 2004.
6. Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '01. New York, NY, USA: ACM, 2001, pp. 19–25.
7. D. Ikeda, T. Fujiki, and M. Okumura, "Automatically linking news articles to blog entries." in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. AAAI, 2006, pp. 78–82.
8. Y. Takama, A. Matsumura, and T. Kajinami, "Visualization of news distribution in blog space," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, ser. WI-IATW '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 413–416.
9. T. Štajner, B. Thomee, A.-M. Popescu, M. Pennacchiotti, and A. Jaimes, "Automatic selection of social media responses to news," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 50–58.
10. B. Shi, G. Ifrim, and N. Hurley, "Be in the know: Connecting news articles to relevant twitter conversations," *arXiv preprint arXiv:1405.3117*, 2014.
11. X. Cao, K. Chen, R. Long, G. Zheng, and Y. Yu, "News comments generation via mining microblogs," in *Proceedings of the 21st International Conference on World Wide Web*. ACM, 2012, pp. 471–472.
12. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
13. W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *European Conference on Information Retrieval*. Springer, 2011, pp. 338–349.

14. R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving lda topic models for microblogs via tweet pooling and automatic labeling," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013, pp. 889–892.
15. B. O'Connor, M. Krieger, and D. Ahn, "Tweetmotif: Exploratory search and topic summarization for twitter." in *ICWSM*, 2010, pp. 384–385.
16. H. Becker, M. Naaman, and L. Gravano, "Selecting quality twitter content for events." in *Proceedings of International Conference on Weblogs and Social Media*, ser. ICWSM'11, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds. The AAAI Press, 2011.
17. Y. Chang, X. Wang, Q. Mei, and Y. Liu, "Towards twitter context summarization with user influence models," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, ser. WSDM '13. New York, NY, USA: ACM, 2013, pp. 527–536.
18. X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang, "Entity-centric topic-oriented opinion summarization in twitter," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 379–387.
19. M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.
20. X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the 2008 international conference on web search and data mining*. ACM, 2008, pp. 231–240.
21. G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," *Computational linguistics*, vol. 37, no. 1, pp. 9–27, 2011.
22. R. Ortega, A. Fonseca, and A. Montoyo, "Ssa-uo: unsupervised twitter sentiment analysis," in *Second joint conference on lexical and computational semantics (* SEM)*, vol. 2, 2013, pp. 501–507.
23. Z. Luo, M. Osborne, and T. Wang, "An effective approach to tweets opinion retrieval," *World Wide Web*, vol. 18, no. 3, pp. 545–566, 2015.
24. F. Bravo-Marquez, M. Mendoza, and B. Poblete, "Combining strengths, emotions and polarities for boosting twitter sentiment analysis," in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM, 2013, p. 2.
25. S. Sahni and T. Gonzalez, "P-complete approximation problems," *J. ACM*, vol. 23, no. 3, pp. 555–565, Jul. 1976.
26. J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, 2003.
27. C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.
28. J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 335–336.
29. J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information theory*, vol. 37, no. 1, pp. 145–151, 1991.