Journal of Intelligent Information Systems (2020) 54:129–156
https://doi.org/10.1007/s10844-018-0521-8

Topic and sentiment aware microblog summarization for twitter

Syed Muhammad Ali[1] · Zeinab Noorian[1] · Ebrahim Bagheri[1] · Chen Ding[2] · Feras Al-Obeidat[3]

Received: 1 October 2017 / Revised: 26 July 2018 / Accepted: 27 July 2018 /
Published online: 8 August 2018

Abstract
Recent advances in microblog content summarization has primarily viewed this task in the context of traditional multi-document summarization techniques where a microblog post or their collection form one document. While these techniques already facilitate information aggregation, categorization and visualization of microblog posts, they fall short in two aspects: *i*) when summarizing a certain topic from microblog content, not all existing techniques take topic polarity into account. This is an important consideration in that the summarization of a topic should cover all aspects of the topic and hence taking polarity into account (sentiment) can lead to the inclusion of the less popular polarity in the summarization process. *ii*) Some summarization techniques produce summaries at the *topic* level. However, it is possible that a given topic can have more than one important *aspect* that need to have representation in the summarization process. Our work in this paper addresses these two challenges by considering both topic sentiments and topic aspects in tandem. We compare our work with the state of the art Twitter summarization techniques and show that our method is able to outperform existing methods on standard metrics such as ROUGE-1.

Keywords Microblogging · Twitter · Summarization · Topic Modeling

1 Introduction

Microblogging services have become one of the prominent platforms for sharing, disseminating and consuming user generated content. In order to understand and exploit this information, microblogging services enable users to search for posts that contain topic

✉ Ebrahim Bagheri
bagheri@ryerson.ca

[1] Laboratory for Systems, Software and Semantics (LS3), Ryerson University, Toronto, Canada

[2] Department of Computer Science, Ryerson University, Toronto, ON, Canada

[3] College of Technological Innovation, Zayed University, Dubai, United Arab Emirates

phrases, and return the results sorted by metrics that consider recency and relevancy (Sharifi et al. 2013). To get a snapshot of what users are primarily saying about a particular topic, it is necessary to acquire a summary or gist of these posts as opposed to returning all of the posts that match the search criterion. Without effective data reduction or summarization mechanisms, users are often confronted with an overwhelming amount of replicated information, which makes it difficult for them to understand the essence of the topics and therefore, possibly miss valuable information. Applying summarization methods on microblogging services, e.g., Twitter, facilitates the generation of condensed summaries about a certain topic discussed by the users in real-time with less time and effort, which can be specifically advantageous for individuals, companies, agencies or institutions seeking public opinion. Therefore, it would be of great benefit if effective mechanisms can be developed for summarizing various aspects of a topic of interest on microblogs (Bian et al. 2015).

Microblog summarization can be viewed and formulated as a multi-document summarization (MDS) task (Jones 2007), which has been widely studied in information retrieval. MDS allows users to quickly capture the essential information contained in a large cluster of documents by producing a summary about a particular topic. In recent years, several MDS approaches ranging from cluster-based (Wan and Yang 2008) and graph-based (Mihalcea and Tarau 2005) to semantic-based approaches (Wang et al. 2008; Hennig and Labor 2009) have been proposed to analyze the information contained in a document set and extract highly salient sentences for generating a summary. While existing works in MDS are designed for well-organized documents, they could still be applicable to microblogs since they are based on simple frequency based methods (Erkan and Radev 2004a; Mihalcea and Tarau 2004). However, research shows that techniques for noise removal and extensive preprocessing on microblog content is required due to the informal, short and noisy nature of the content posted on microblogging services (Bian et al. 2015). Recently, a few efforts have been undertaken for microblog summarization (Sharifi et al. 2013; Bian et al. 2015; Inouye and Kalita 2011; Nichols et al. 2012; Lin et al. 2012). In this paper, we propose a novel summarization method based on sentiment and topical aspect analysis to generate a holistic summary for trending topics in microblogs. Specifically, the proposed method comprises of three stages. First, after pre-processing and semantically enriching microblog posts, we extract the topics and sentiments expressed by each post. Then, we build a sentiment-based Word Graph for each topic and cluster the graph to extract different aspects of the topic. Finally, we apply state-of-the-art summarization methods to summarize each topical aspect individually; and aggregate all aspect-level summaries to generate the holistic summary for each topic.

To the best of our knowledge, there have been limited work that consider sentiments, topics and aspects detection algorithm and semantic enrichment process of tweets in tandem for microblog summarization. As such the core contributions of our work can be enumerated as follows:

- We propose a sentiment-based approach in our summarization method to automatically consider sentiments (positive or negative) of semantically enriched tweets when generating summaries. This is useful because it allows positive and negative feedback to be aggregated equally into the final generated summary.
- In addition to sentiments, we propose to construct a word graph, inspired by KeyGraphs (Ohsawa et al. 1998), to automatically extract *aspects* of a topic to be used in the summarization process. Therefore, our work benefits from topic aspects and sentiments simultaneously when generating summaries.

– We conduct experiments on an already available Twitter dataset presented by Abel et al. (2011) consisting of approximately 3M tweets. By comparing with several state of the art summarization methods, we show improved summarization performance by our proposed approach.

The rest of the paper is organized as follows: in the next section we review the related literature, followed by the presentation of the overview of our proposed approach. In Section 4, the details of our work are presented and Section 5 covers the experimental setup. We present our findings in Section 6 and the paper is finally concluded in Section 7.

2 Related work

Work in automated multi-document summarization has drawn much attention in the recent years. A number of algorithms have been developed to improve summarization as well as to perform summarization on new forms of documents such as Web pages (Sun et al. 2005), discussion forums and blogs (Zhou and Hovy 2006; Ku et al. 2006). General MDS methods can be separated into two categories: extractive and abstractive (Knight and Marcu 2002; Jing and McKeown 2000). Extractive summarization involves assigning saliency scores to sentences and paragraphs of the documents calculated by a set of predefined features such as term-frequency-inverse document frequency (TF-IDF), sentence or term position (Lin and Hovy 2002; Yih et al. 2007), and the number of keywords (Yih et al. 2007), and extracting those with the highest scores. Notable extractive MDS methods include SumBasic (Vanderwende et al. 2007) and centroid-based methods such as MEAD (Radev et al. 2004). The underlying premise behind SumBasic is that words which occur more frequently across documents have a higher probability of being selected for human created multi-document summaries over those words that occur less frequently. SumBasic tends to favour longer sentences, as they are more likely to have higher average probabilities leading to increased recall, as noted in Sharifi et al. (2013). On the other hand, MEAD (Radev et al. 2004), which is a centroid-based method, scores sentences based on sentence-level and inter-sentence features, including cluster centroids, position, TF-IDF, and other features. NeATS (Lin and Hovy 2002) uses sentence position, term frequency, topic signature and term clustering to select important content. Unlike MEAD and SumBasic, the Maximal Marginal Relevance (MMR) (Goldstein et al. 1999) method has been developed as an extractive method that decides which sentences should be removed to reduce the original document to a summary as opposed to selecting sentences to be included in the summary.

Abstractive summarization employs techniques from information fusion (Knight and Marcu 2002; Jing and McKeown 2000), rule based approach (Genest and Lapalme 2012); sentence compression (Knight and Marcu 2002); sentence merging based on semantics (Liu et al. 2015), and graph-based algorithms (Ganesan et al. 2010; Bhargava et al. 2016). The abstractive summarization framework proposed by Liu et al. (2015) parses input sentences to build Abstract Meaning Representation (AMR) graphs based on semantic representation of text. AMR graphs are then converted to a summary graph using a perceptron model prediction algorithm to select the subgraphs, which can be further used for summary generation. Ganesan et al. (2010) describes an approach that uses the original sentence word order to build directed graphs in order to generate abstractive summaries. Their technique leverages the graphical form of the input text to reduce redundancy. In Lloret and Palomar (2011), the authors propose a summarization approach which builds a directed weighted word

graph in which each word is represented by a node in the graph and the edge indicates the adjacency relation between the words. The weight of the edge is calculated by a combination of their pagerank value and the frequency of the words. Important sentences are determined by selecting the first n words with the highest TF-IDF score. Sentence correctness is also ensured using basic rules of grammars.

Most recently, graph-based ranking methods have been proposed for MDS which rank sentences based on votes or recommendations. TextRank (Mihalcea and Tarau 2004) and LexPageRank (Erkan and Radev 2004b) use algorithms such as PageRank and HITS to rank important keywords of n-grams in a corpus. Thus, their summaries appear to be short snippets of the corpus. These methods first construct a graph representing the relationship between sentences and then evaluate the importance of each sentence based on the topology of the graph. LexRank uses Erkan and Radev (2004a), a modified cosine similarity function to construct an adjacency matrix with similarity values of two sentences. This matrix is treated as a markov chain, and an iterative algorithm is used to compute the stationary distribution. Each value of the stationary distribution represents a weight for the corresponding document, and the one with the highest weight is chosen to represent the summary. Mihalcea and Tarau (2005) also propose an algorithm based on PageRank, which exploits a meta-summarization process to summarize the meta-document generated by assembling all the single summaries of each document.

Some other methods have been designed that identify semantically important sentences for summary generation. Gong and Liu (2001) propose a method that uses Latent Semantic Analysis (LSA) to select highly ranked sentences for summarization; while the work in Haghighi and Vanderwende (2009) exploits a hierarchical LDA-style model to represent content specificity as a hierarchy of topic vocabulary distributions, based on which sentences are selected according to these distributions. Wang et al. (2008) have proposed a framework based on Sentence Level Semantic Analysis (SLSS) and Symmetric Nonnegative Matrix Factorization (SNMF) to capture relationships between sentences in a semantic manner and factorize the similarity matrix to obtain meaningful groups of sentences. Other methods include NMF-based topic-specific summarization (Lin and Hovy 2003); Conditional Random Fields (CRF) based summarization (Lin and Hovy 2002); and Hidden Markov Model (HMM) based methods (Mani 2001).

These methods are specifically designed for formal texts and documents, thus applying them on a microblog dataset may not produce the best results (Bian et al. 2015). With the increasing interest in using microblog services to disseminate information, a few works have shifted their focus to process microblog data. Summarizing microblogs can be considered as an instance of extractive MDS which deals with informal documents derived from a wide range of users and topics. Most of the prior work on Twitter data summarization are about topic-level summarization. In Zhou et al. (2016), the authors present CMiner, an opinion mining system for Chinese microblogs. CMiner adopts an unsupervised label propagation algorithm to extract target opinions based on the assumption that similar messages are focused on similar opinion targets. On this basis, a summarization framework is proposed to generate opinion summaries for different opinion targets. CMiner clusters the extracted opinion targets based on different similarity measures and ranks both the opinion targets and microblog sentences based on the proposed co-ranking algorithm.

Chakrabarti and Punera (2011) have formalized the problem of tweet summarization for highly structured and recurring events. The authors discuss how events can be subdivided into smaller events using Hidden Markov Models (HMMs). They assert that HMMs are useful in detecting bursty events and are able to learn differences in language models of sub-events automatically. This method is useful when a training set can describe changes

in events; otherwise, it is difficult to use this method effectively. Lin et al. (2012) adopt a graph optimization method to generate event storyline of an ongoing event from microblogs. Temporal information is utilized for event representation. This framework is only suitable for relatively long-term events, which makes it less effective for our task given the fact that the hot period of most social events to be summarized is usually very short. TwitInfo (Marcus et al. 2011) has been designed to aggregate tweets about a topic into visual summaries on peaks of high tweet activity and display the summaries on users' timelines. Given a search query related to an event, the streaming algorithm identifies and labels event peaks; highlights important terms and tweets in the conversation; and provides an aggregative view of users' sentiments. These visualizations must be interpreted by users and do not include sentence-level textual summaries. Carrillo-de Albornoz et al. (2016) propose a novel methodology for evaluating the performance of three representative summarization algorithms such as LexRank, Follower and Single Voting on generating summaries in the context of online reputation reports from Twitter. This work exploits the RepLab Dataset(Amigó et al. 2013) in which the tweets have been manually annotated by experts for relevancy, polarity, topic and priority. The empirical results indicate that incorporating priority signals improve the summarization task.

One of the significant contributions in microblog summarization is the work presented by Sharifi et al. (2010, 2013). In this paper, the authors have developed two multi-document summarization algorithms, which include: 1) a clustering based algorithm; and 2) a Hybrid TF-IDF algorithm, which is a direct extension of TF-IDF used in single-document summarization algorithms. They also propose a Phrase Reinforcement algorithm which uses a graph to represent overlapping phrases in a set of related microblog sentences and summarizes Twitter hot topics through finding the most commonly used phrases that encompass the topic phrase. Sharifi et al. evaluated the performance of their proposed algorithms in comparison with other notable summarizers, including MEAD (Radev et al. 2004), LexRank (Erkan and Radev 2004a), TextRank (Mihalcea and Tarau 2004) , and SumBasic (Vanderwende et al. 2007).

Our proposed method is a topic-level microblog summarization approach, which first semantically enriches the microblogs and constructs a Word Graph for each class of sentiment in each topic and then applies clustering algorithms on each graph for the purpose of topical aspect extraction. Different cluster-based multi-document summarization methods are explored and exploited to produce a final summary.

The idea of incorporating aspect-relevance and sentiment intensity of documents for the summarization purpose has already been considered in the literature. For instance, in (Xu et al. 2011), the authors propose an aspect-based summarization method which considers representativeness and diversity of online reviews in order to generate summaries with maximum coverage and minimum redundancy in terms of sub-aspects and their opinions. Furthermore, in Wu et al. (2016), the authors adopt a direction different from prior work on aspect extraction by directly mapping each review sentence into pre-defined aspect categories, assuming that users already know what aspects each product can have. This paper proposes two convolutional neural network-based approaches to 1) extract implicit mentions of an aspect in a review and assign them to a pre-defined set of aspects, and 2) predict the sentiment polarity of the input sentences. In Lin and Hovy (2003), the authors present a framework for identifying and extracting important pieces of information along with their sentiments in the document to form a summary. This approach categorizes the textual content in two subjective and objective categories and then aggregates them using specific rules. In Piryani et al. (2018), the authors present a summarization framework to generate aspect-wise extractive sentiment summary for textual reviews. However, this approach focuses on

certain domains, e.g., laptop reviews, and identifies the set of aspects manually. In Hu et al. (2017), the authors present a model to classify the reviews from the Trip Advisor website into predefined aspects and then apply a topic modelling technique along with sentiment analysis on the classified reviews to identify hidden information which would be further exploited to generate summaries.

It should be noted that most existing approaches for aspect extraction and polarity classification for the summarization purpose are designed for formal texts and documents, which might not be precise enough or suitable for short texts such as microblog posts (Ling et al. 2008; Titov and McDonald 2008; Zhuang et al. 2006). Besides, many existing approaches have been built on the assumption of the existence of a pre-defined set of aspects and did not propose any aspect detection algorithm to automatically derive aspects from free-form texts.

We have chosen Sharifi's proposed algorithms presented in Sharifi et al. (2013) as the baselines since they have shown strong performance in comparison to the state of the art summarization approaches for Twitter.

3 Approach overview

The objective of our work is to automatically generate a textual summary for any given topic present on Twitter. The flowchart of our proposed microblog summarization work is illustrated in Fig. 1.

As seen in the figure, there are three main stages in our proposed work. In the first stage, after preprocessing the textual content of microblog posts, we discover the active topics that are present in microblog posts using the LDA topic modelling approach (Blei et al. 2003) and then further employ sentiment analysis to derive the sentiment of each microblog post, accordingly. Once the topics are identified and the sentiment of each microblog post is determined, we group the microblog posts into several *topic-sentiment* clusters that contain posts related to similar topics and with similar sentiments. In the second stage, we create a Word Graph (WG) for each cluster of microblogs that share the same topic and sentiment in order to model the co-occurrence of the words within that cluster. Once WG is constructed, graph clustering is applied to identify the different topical aspects of the specific topic-sentiment cluster. The final stage re-assigns microblogs to the identified topical aspects in each WG and further exploits document clustering algorithms to summarize each topical aspect, accordingly. Finally, we aggregate all topical aspect-level summaries to derive a holistic summary for each topic .

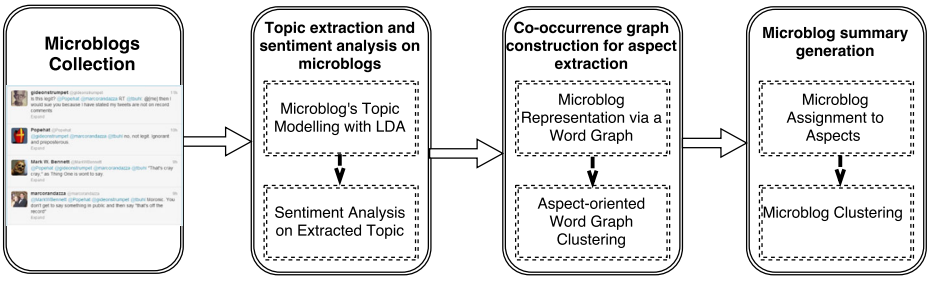

Fig. 1 Flowchart of our proposed microblog summarization method

4 Our proposed approach

In this section, we elaborate on the details of the proposed microblog summarization method including topic extraction, sentiment analysis, word graph construction and topical aspect extraction, and aspect level summary extraction as shown in Fig. 1.

4.1 Microblog topic and sentiment extraction

Given the core of our work is to summarize topics on Twitter, we assume that a collection of microblog posts denoted as $M = \{M_1, ., M_i, .., M_{|n|}\}$ are present that collectively form k topics represented by $TP = \{tp_1, .., tp_k\}$. By considering M, the set of microblog posts as a textual corpus, it is possible to extract a set of active topics TP using topic modelling techniques such as LDA. As proposed in Varga et al. (2014) and Saif et al. (2012), to obtain better topics without modifying the standard topic modelling methods, we enrich each microblog post M_i using an existing semantic annotator, i.e., TagMe (Ferragina and Scaiella 2010) and employ the extracted entities within the topic modeling process. This has shown to result in the reduction of noisy content within the topic detection process (Zarrinkalam et al. 2015; Zarrinkalam et al. 2016). Therefore, in our work, each microblog post is represented as a set of one or more semantic entities that collectively denote the underlying semantics of the microblog post. We view a topic, defined in Definition 1, as a distribution over entities.

Definition 1 (Topic): Let M be a microblog collection and $E = \{e_1, e_2, ..., e_{|E|}\}$ be a vocabulary of entities. A topic, tp, is defined to be a vector of weights, i.e., $(g_{tp(e_1)}, ..., g_{tp(e_{|E|})})$, where $g_{tp(e_i)}$ shows the participation score of term $e_i \in E$ in forming topic tp. Collectively, $TP = \{tp_1, ...tp_K\}$ denotes a set of K topics extracted from M.

To extract topics from microblogs using LDA, documents should naturally correspond to microblog posts. We treat each microblog post that has been enriched with entities as a single document and train LDA on all microblog posts M. LDA has two parameters to be inferred from the corpus of documents: document-topic distribution θ, and the K topic-term distribution ω. Given that each document corresponds to the microblog entities, by applying LDA over all microblog posts, K topic-entity distributions will be produced, where each topic entity distribution associated with a topic $tp \in TP$ represents one active topic in M.

In the next step, given the set of enriched microblog posts with entities, we incorporate the semantic features of the microblogs into a Naive Bayes (NB) classifier using the semantic augmentation method (Saif et al. 2012; Go et al. 2009) to identify the sentiment of each microblog post in M. The assignment of a sentiment class c to a given microblog post M_i in a NB classifier can be computed as:

$$\begin{aligned} \hat{c} &= \arg\max_{c \in C} P(c|e) \\ &= \arg\max_{c \in C} P(c) \prod_{1 \le i \le N_e} P(c|e) \end{aligned} \tag{1}$$

where N_e is the total number of entities in microblog M_i , $P(c)$ is the prior probability of a post appearing in class c, $P(e|c)$ is the conditional probability of entity e occurring in a microblog post of class c .

In multinomial NB, $P(c)$ can be estimated by $P(c) = N_c/N$ where N_c is the number of microblog posts in class c and N is the total number of microblog posts. $P(e|c)$ can be estimated using maximum likelihood with Laplace smoothing:

$$P(e|c) = \frac{N(e,c)+1}{\sum_{e' \in V} N(e'|c)) + |V|} \tag{2}$$

where $N(e, c)$ is the occurrence frequency of word e in all training microblogs of class c and $|V|$ is the number of entities in the vocabulary. In our work, we only consider polarity when determining sentiment and therefore, $|c| = 3$; consisting of positive, neutral and negative classes.

4.2 Co-occurrence graph model for topical aspect extraction

In this section, we present the notion of *Word Graph (WG)*, a co-occurrence graph model, to build a graph of entities for microblog posts (hereafter, tweets) for each class of sentiments derived for a particular topic. Our goal is to apply graph clustering algorithms on WG to discover different topical aspects of a topic, which will be further exploited to generate summaries for that topic. The idea of applying clustering algorithms on long texts (multi-documents) as well as short texts (microblogs) has been well-studied in the literature. Various approaches have been proposed to cluster similar tweets ranging from using hierarchical clustering methods (Abdullah and Hamdan 2015; Jashki et al. 2009), term-frequency analysis (Atefeh and Khreich 2015) to tweet attribute analysis such as favourite and re-tweet counts (Bild et al. 2015). These approaches apply clustering algorithms directly on tweets to induce different attributes and topical aspects of the topic for the summarization task. However, we argue that building a graph of entity co-occurrence for each class of sentiments of a particular topic prior to clustering would significantly enhance the summarization task for two reasons:

1. a graph representation captures different relations pertaining to node attributes (e.g., favorite counts in tweets) between directly connected nodes as well as hidden relations between indirectly connected nodes; and
2. as the number of tweets with positive and negative sentiments are imbalanced (i.e., consider the topic that is dominated by highly positive tweets and few negative ones), directly applying clustering algorithm on tweets may result in the domination of only positive tweets, and negative tweets most likely would be overlooked in the generated summary.

Let Mtp^1 be a set of tweets related to topic $tp_1 \in TP = \{tp_1, ..., tp_k\}$, denoted as $Mtp^1 = \{Mtp^1_{sen_p} \cup Mtp^1_{sen_n} \cup Mtp^1_{sen_o}\}$ where $Mtp^1_{sen_p}$ indicates the set of microblog posts with a positive sentiment; $Mtp^1_{sen_n}$ indicates a set of microblogs with the negative sentiment; and $Mtp^1_{sen_o}$ indicates the set of microblogs with a neutral sentiment. We construct a Word Graph (WG) for each sentiment of a particular topic based on Definition 2, as follows:

Definition 2 (Word Graph Representation): Given a collection of tweets such as Mtp, a co-occurrence word graph can be defined as an undirected, weighted graph $G = (V, E)$, where the set of vertices V correspond to the entities of the tweets and E is the set of edges that represent relationships among these entities. The weight of the edges is calculated based

on the total number of times the two entities have co-occurred in different tweets and the number of times those tweets have been favorited (i.e., liked, re-tweeted) as follows,

$$Weight_{E(e_i,e_j)} = \sum_{\forall M_k \in Mtp} Occurred(e_i, e_j, M_k) \times \sum_{\forall M_k \in Mtp} Count(e_i, e_j, M_k) \quad (3)$$

such that:

$$Occurred(e_i, e_j, M_k) = \begin{cases} 1 & \text{if } e_i, e_j \in M_k \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

and,

$$Count\left(e_i, e_j, M_k\right) = \begin{cases} 1 & \text{if } e_i, e_j \in M_k \text{ and } M_k \text{ isFavourited} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

This process will generate three word graphs per topic, i.e., one word graph per positive, neutral and negative aspect of each topic, respectively.

For example, consider the following two tweets for the topic, "Costco":

– All the **retailers** are **closed** now, except **Costco**. [liked: 0]
– **Costco reports** 2 percent **profit** after **stock market closed**. [Liked =1]

4.2.1 Word graph clustering

A natural form of graph clustering is to partition the vertex set into disjoint subsets called clusters. One of the common requirements for graph clustering is *modularity* (Blondel et al. 2008; Newman 2006), which formalizes that the connections within graph clusters should be dense, and the connections between different graph clusters should be sparse.

In this paper, we have explored various well-known graph clustering algorithms for partitioning sparsely connected dense word subgraphs from each other. As indicated in Table 1, there is no single superior algorithm among existing work and their quality is dependent on the characteristic of the specific graph under study. Therefore, in our paper, we adopt three clustering algorithms that consider *edge weights* in their process: 1) InfoMap (Rosvall and Bergstrom 2008); 2) Newman's Eigenvector Method (Newman 2006); and 3) Blondel's Multi-level Clustering (Blondel et al. 2008). The overview of these algorithms is presented in Table 1.

We compare these three algorithms using two clustering metrics to measure the quality of the produced clusters, namely: 1) Variance of Information (VI) (Meila 2003), which measures the amount of information loss when changing from one clustering to another; and 2) Split-Joint distance (Dongen 2000), which is a non-commutative metric that measures the overlap between two different clusters. The reason why we exploit such metrics is because they make no assumptions about how the clusterings were generated and apply to both soft and hard clusterings (Meila 2003). The algorithm with the highest values for VI and Split-joint distance was selected as the representative graph clustering technique to induce topical aspects (e.g. See Table 2) in our experiments.

4.3 Microblog summary generation

In this subsection, we explore how to utilize the discovered topical aspects of the constructed word graphs to facilitate the generation of the summary for a specific topic. We propose

Table 1 Summary of the graph clustering algorithms

Clustering Technique	Overview	Output	Edge Weights
Edge betweeness	Removes the most commonly used edges to connect. shortest path between every vertex pair in the graph	Remaining clusters	No
Bicomponent	Runs a depth-first search to find the biconnected components of the graph.	All components of a graph with a property that at least two vertices must be removed in order to disconnect the graph.	No
Weak component	Runs a breadth-first search to find maximal subgraph in which all pairs of vertices in the subgraph are reachable from one another.	A set of weakly components.	No
Voltage clustering	Algorithm by Wu and Huberman (2004) combined with k-means for determining cluster membership.	Number of clusters not higher than requested amount.	No
InfoMap	Random walks to reveal community structure.	Arbitrary amount of clusters. Returns high amount as demanded.	Yes
Newman's Eigenvector Method	Uses eigenvectors of matrices to find community structures.	Produced similar amount of clusters as Blondel's algorithm.	Yes
Blondel's Multi-level Clustering	Optimize modularity at the local level and at the community level.	Arbitrary number of clusters, usually lower than Newman's method.	Yes

an approach for text summarization based on cluster-based multi-document summarization algorithms such that different sentiments regarding a single topic would be equally weighted and aggregated in the final summary. The key idea is to apply cluster-based multi-document summarization algorithms on every topical aspect of the constructed word graphs in order to group tweets into clusters; and then select the representative tweets from each cluster to generate the final summary for a specific topic. For this purpose, we first need to re-assign the set of tweets with the given topic to one or more of the topical aspects that we have extracted from the word graphs.

Table 2 Sample Aspects for the *Tsunami* Topic

	Aspect 1 (Words in Aspect: 223)	Aspect 2 (Words in Aspect: 119)
Nouns	Earthquake, Depth, Epicenter, December, November	Volcano, Mount, Eruption, Bromo, Indonesias, Alert, Ash, Toll, Death
Adjectives	Southern	Beautiful, Hot, Safer, Highest, Volcanic

Definition 3 (Microblog and topical aspect similarity): Given a topical aspect ($\mathcal{A}$) and the entities in a microblog post m, we calculate the similarity of the microblog post to the topical aspect as follows:

$$Sim(\mathcal{A}, m) = \frac{\sum_{m_i \in m} \sum_{t_j \in \mathcal{A}} sim(m_i, t_j)}{|\mathcal{A}| \times |m|}. \tag{6}$$

where sim is a semantic relatedness measure that calculates the similarity of two semantic entities (Feng et al. 2017). We employ the similarity measure proposed in Ferragina and Scaiella (2010) for this purpose. A microblog m will be assigned to a topical aspect $\mathcal{A}$ if their similarity score is larger than a threshold, λ.

Now, given a collection of tweets that are assigned to the different topical aspects of a certain topic, we exploit various cluster-based algorithms to group microblog posts into clusters given each specific topical aspect of the word graphs. In this work, we have adopted two clustering algorithms that are widely-used for document summarization, namely, Agglomerative Clustering (Ackermann et al. 2014); and Bisect K-Means++ Clustering (Steinbach et al. 2000). These algorithms are described in more detail later in the paper. After grouping the microblogs into clusters within each topical aspect, in each cluster, we compute the score of a tweet to measure how important the tweet is to be included in the summary. We rank the tweets based on a tweet score calculated as follows:

$$Score(m_i, C_k) = \frac{1}{N_{C_k} - 1} \sum_{m_j \in C_k / m_i} sim(m_i, m_j) \tag{7}$$

where $Score(m_i, C_k)$ measures the average similarity score between tweet m and all the other tweets in the cluster C_k, and N_{C_k} is the number of tweets in C_k.

Finally, given the set of word graphs built based on different classes of sentiments of a certain topic, we select k tweets with the highest score, calculated by (7), from each topical aspect of the word graphs to form the final summary.

5 Experimental setup

5.1 Dataset and pre-processing

Our experiments were conducted on the available Twitter dataset released by Abel et al. (2011). It consists of approximately 3M tweets sampled between November 1 and December 31, 2010. Given the fact that tweets are an informal way of communicating, they were preprocessed to remove spam and other noise features. We adopted the Datumbox Framework package and followed the steps proposed in Sharifi et al. (2013) to preprocess the tweets, as indicated in Table 3.

5.2 Topic modeling and sentiment analysis on tweets

We have enriched the processed tweets with Wikipedia entities using the TagMe (Ferragina and Scaiella 2010) semantic annotator. To derive topics and their associated sentiments from the tweets, we adopted the Stanford Topic Modeling Toolbox (Ramage

Table 3 Tweet Preprocessing Steps

Step	Description
1	Converted any HTML-4 and HTML-3 encoded characters into ASCII.
2	Removed any Unicode characters (e.g. '\x000')
3	Removed any embedded URLs (e.g. http://), HTML tags (e.g.), other tags (e.g. <>), tokenize any smileys, remove any accents, and user mentions
4	Discarded the document if it is not English. We used Language Detection tool by Shuyo3 5. Removed duplicate posts by the same user.
5	Removed any terms that are equal or larger than 20 characters. This is to ensure that any long hash-tags or other obscure terms are removed.
6	Removed any consecutive question marks that are 6 characters or longer (e.g. ??????)
7	Removed the stopwords.
8	For Phrase Reinforcement algorithm: The documents were broken into sentences. Most tweets have only one sentence.
9	For Phrase Reinforcement algorithm: The longest sentence was detected that contains the topic phrase and used it to represent the tweet.
10	For Word Graph construction, punctuation marks were removed.

and Rosen 2011) to run LDA on tweets, and the Naive Bayes implementation provided in Datumbox official website http://www.datumbox.com/ to discover tweet sentiments in our work.

5.3 Word graph construction

In order to construct Word Graphs for tweets, we exploited Jung (Java Universal Network/Graph Framework)[1] API and extended it to meet our purposes. The graphs were ported to Pajek[2] format and used in Python iGraph library for clustering. Each Word Graph was uniquely identified by the topic and the overall sentiment of positive, negative, or neutral. Thus, each topic had a Word Graph for each sentiment. Furthermore, we also investigated if a tweet had any replies associated with it. On Twitter, users can reply to tweets, favorite them, or re-tweet them. In many instances, replies may not have the same semantic annotation as the original tweet and, hence, will not be assigned to the same topic as the original tweet. In order to ensure that replies are also assigned to the same topic, we looked at the *replyToId* attribute of the tweet and linked the original tweet to it. So if the original tweet is assigned to a topic, automatically, all of its replies would be assigned to that topic as well. Each edge weight between the original tweet and the reply would be equal to the number of times the reply is favourited plus 1, consistent with the methodology of weights applied to co-occurrence of words within a tweet itself.

[1] http://jung.sourceforge.net/
[2] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

6 Evaluation

6.1 Outline

We have proposed a method for extracting summaries from a collection of tweets. The proposed method builds a sentiment-based word graph for each topic and clusters the word graph into k clusters each of which represents a specific aspect of the topic. Then, each topical aspect is summarized individually through document clustering. In the next subsection, we perform testing of various graph clustering approaches to select the best method for WG clustering. We then evaluate how the performance of existing document clustering algorithms are improved when they are used in our proposed method for summarization. More specifically, our goal is to determine if having word graphs to induce topical aspects of the topic, before summarization, will improve the quality of the generated summaries or not. An articulated example on the end-to-end process of our proposed mechanism in automatic summary generation is provided in the Appendix.

6.2 Selection of the clustering method

As indicated in Table 1, we have considered different clustering techniques to cluster our Word Graphs, namely, InfoMap, Newman and Blondel clustering algorithms. We compare these methods using two clustering metrics in order to identify the most suitable one for our experiments. Table 4 shows the comparative result of the three algorithms with respect to the Variance of Information (VI) metric. As is indicated, VI, which measures the amount of information loss when changing from one type of clustering to another, shows smaller values for Blondel compared to Newman and InfoMap, meaning that Blondel shows better clustering performance, given the current graph structure, compared to the other clustering approaches.

We also measure the split-join-distance for these three clustering algorithms. The split-join-distance shown in Table 5 are higher for Blondel compared to Newman and InfoMap, meaning that the amount of overlap would be low if we change our clustering from Blondel to one of the other two clustering algorithms. Newman also showed a comparatively high score; however, since the VI metric is more stable for Blondel, we choose Blondel to cluster our Word Graphs in the rest of the experiments.

6.3 Summarization algorithms

We compare the results of our proposed microblog summarization method with baseline algorithms and well-known multi-document clustering algorithms. The detailed overview of the comparative algorithms is presented in Table 6. We have chosen the work presented in Sharifi et al. (2013) as our baseline, due to that fact that it has already shown better performance compared with other state of the art approaches such as Random Summarizer

Table 4 Variance of information for the three clustering algorithms

Clustering algorithms	Variance of information (VI)
Blondel-Newman	1.720
Blondel-InfoMap	1.435
NewMan-InfoMap	2.091

Table 5 Split-join-distance for the three clustering techniques

From/To	Blondel	Newman	InfoMap
Blondel	–	221	377
Newman	261	–	423
InfoMap	42	112	–

(Sharifi et al. 2013), Most recent summarizer (Sharifi et al. 2013), SumBasic (Vanderwende et al. 2007), MEAD (Radev et al. 2004), LexRank (Erkan and Radev 2004a) and TextRank (Mihalcea and Tarau 2004). Thus, our baselines include Bisect K-Means++ with Hybrid TF-IDF; Hybrid TF-IDF; as well as the Phrase Reinforcement algorithm.

6.4 Evaluation process

To evaluate the efficiency of our proposed method in generating the summary for a tweet collection, we compare the performance of different multi-document summarization algorithms under two different conditions: 1) when adopting our proposed microblog summarization method; and 2) without adopting our proposed method. There are two types

Table 6 Description of the baseline algorithms

Comparative summarization algorithms	Description
Agglomerative clustering (Ackermann et al. 2014)	A hierarchical bottom-up clustering process in which each document starts in its own cluster, and pairs of closest clusters (given different strategies for measuring document similarity), will be merged as one moves up the hierarchy until only one cluster remains.
Bisect K-Means++ clustering (Steinbach et al. 2000)	A top-down variant of hierarchical clustering technique, where all documents are merged to form one mega-cluster, and subsequently sub-clusters are formed using a k-meansalgorithm based on document similarities.
Hybrid TF-IDF (Sharifi et al. 2010, 2013)	The authors redefine TF-IDF in terms of a hybrid document, where the term frequencies are calculated across all microblogs but the IDF component treats each document as a separate microblog. Each document weight is divided by a normalization factor, which is the maximum of minimum threshold or the number of words in the sentence, in order to reduce the effect of bias towards longer sentences.
Phrase reinforcement algorithm (Sharifi et al. 2010, 2013)	Specifically designed for single document summarization, the main idea of the PR algorithm is to find the most heavily overlapping phrase centered around the topic phrase. The algorithm creates a directed graph where each vertex represents a word and is weighted based on their counts and their unique position with respect to the topic phrase. The final summary is obtained by finding the directed path with the highest total weights which includes the topic phrase.

of workflows for evaluation. In the first workflow, we directly cluster the documents in k groups and find a representative tweet from each cluster to come up with a k sentence summary. In the second workflow, we first create Word Graphs to induce aspects of the topic and then cluster each topical aspect into k groups, and pick a representative tweet from each cluster to come up with a topical aspect summary. As described in Table 6, we exploit Agglomerative clustering, Bisect K-Means++ clustering; Hybrid TF-IDF algorithm, and Phrase Reinforcement algorithm (Sharifi et al. 2013) as our baseline comparative approaches for document summarization.

In the first workflow, the first two comparative approaches involve directly applying Agglomerative and Bisect K-Means++ algorithms on tweets without building any word graphs. These *k* clusters are then passed to a 1-means algorithm to determine the centroid which would be used as the representative tweet for that cluster. The other comparative approaches involve applying the above-mentioned algorithms with the exception of exploiting Hybrid TF-IDF approach instead of a 1-means algorithm. Sharifi et al concluded that Bisect K-Means++ combined with Hybrid TF-IDF is the best methodology for tweet summarization. The next comparative method is to directly apply Hybrid TF-IDF on tweets, which has shown the highest summarization performance with respect to F-measure according to Sharifi et al. In order to keep the documents from being too similar in content, Sharifi et al. (2013) conducted preliminary tests to determine the best cosine similarity threshold, which was reported to be 0.77.

In the the second workflow, the comparative approaches are aimed at evaluating the effectiveness of building word graphs in improving the quality of summarization. Thus, after building word graphs for each class of sentiment, we apply the five above-mentioned methods on the microblog posts related to each topical aspect of the word graph. The next baseline method is dedicated to exploiting the PR algorithm proposed by Sharifi et al. (2013) to obtain a one sentence summary phrase for each topical aspect, without applying any clustering algorithm on that topical aspect.

6.5 Manual summarization model (Gold Standard)

For building the gold standard, we adopted the approach proposed in Sharifi et al. (2013). Our manual multi-document summaries were created by four volunteers for six topics. Each topic had at least 80 positive tweets and 80 negative tweets. The choice for our sentiment analysis method was motivated by the findings reported in Saif et al. (2012). Each topic was analyzed twice by two separate individuals. So, every annotator analyzed three topics and provided one manual summary for each aspect. Annotators worked independently and did not share any information. An example of the selected topics and the number of tweets per topic is provided in the Appendix.

Volunteers had to perform summarization tasks for the first workflow (prior to Word Graph construction) as well as the second workflow (post-Word Graph construction). For evaluating summaries prior to Word Graph construction (first workflow), volunteers were given three sets of tweets for each topic: positive, negative, and neutral. For each set of tweets, they were required to group the tweets into clusters and then pick a representative tweet from each cluster to obtain a summary for that sentiment. Afterwards, the volunteers were asked to provide content scores for different algorithms to evaluate the automated summaries against the manual summaries in order to understand whether or not we are truly achieving human-comparable summaries.

For evaluating summaries after Word Graph construction (second workflow), volunteers were given three sets of tweets for each topic: positive, negative, and neutral. Each of these

Table 7 Content scores for the algorithms after the application of the word graph

Techniques	Content score
Agglomerative algorithm	3.15
Bisect K-Means++ / Hybrid TF-IDF	2.93
Agglomerative + Hybrid TF-IDF	2.91
Bisect K-Means++	2.75
Hybrid TF-IDF	2.72
Phrase reinforcement	1.08

sets contained subsets of tweets corresponding to the topical aspects. For each set of tweets, volunteers were required to group the tweets into clusters and then pick a representative tweet from each cluster to obtain a summary for that sentiment. Afterwards, as suggested by Sharifi et al. (2013), the volunteers were asked to provide content scores for different algorithms by comparing their manual summary to the summary generated by the algorithms. Table 7 presents the average results of content score for algorithms after the word graph construction. As seen in the figure, the human evaluators believed that the proposed Agglomerative method produced the best summaries for the topics after the word graph was constructed. While the Agglomerative method was the best method from the perspective of the human evaluators, our proposed Agglomerative + Hybrid TF-IDF method showed competitive performance to the baseline Bisect K-Means++ / Hybrid TF-IDF method. We further show how the human evaluators' perception of the quality of the summaries changed before and after the application of the word graph. As seen in Fig. 2, regardless of whether the method was proposed in this paper or by the baseline, the application of the word graph leads to higher quality topic summaries. We conclude that (1) employment of word graphs

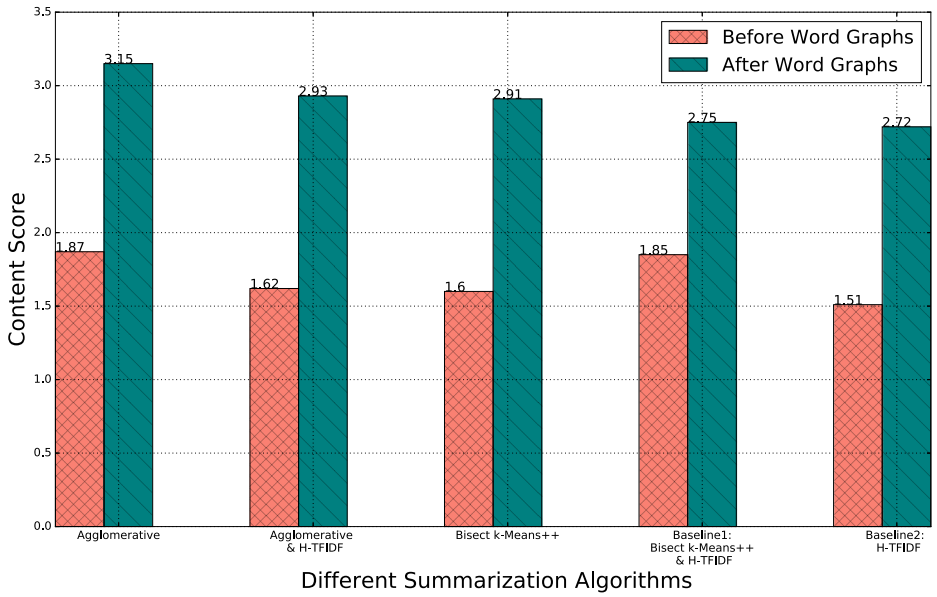

Fig. 2 Content score comparison after word graphs construction

leads to higher quality summaries regardless of the method that is used and (2) our proposed Agglomerative method produces the most desirable summaries from the perspective of the human evaluators compared to both the baselines and the other variations of our proposed methods.

6.6 Evaluation metrics

In Saggion et al. (2010) and Louis and Nenkova (2009), the authors have mentioned different complex methods to evaluate automatically generated summaries. In this paper, we employ ROUGE, a widely-use summarization evaluation method (Lin and Hovy 2003) which automatically determines the similarity of a summary compared to human generated gold standards.

ROUGE-N is an N-gram recall metric computed as follows:

$$ROUGE - N = \frac{\sum_{s \in MS} \sum_{n-gram \in s} match(n - gram)}{\sum_{s \in MS} \sum_{n-gram \in s} count(n - gram)} \tag{8}$$

where MS is the set of manual summaries; n is the length of n-grams, $count(n - gram)$ is the number of n-grams in the manual summary; and $match(n - grams)$ is the number of co-occurrences where an n-gram was found in both the manual summary and automated summary. Since the baseline (Sharifi et al. 2013) used the $ROUGE - 1$ metric, we will

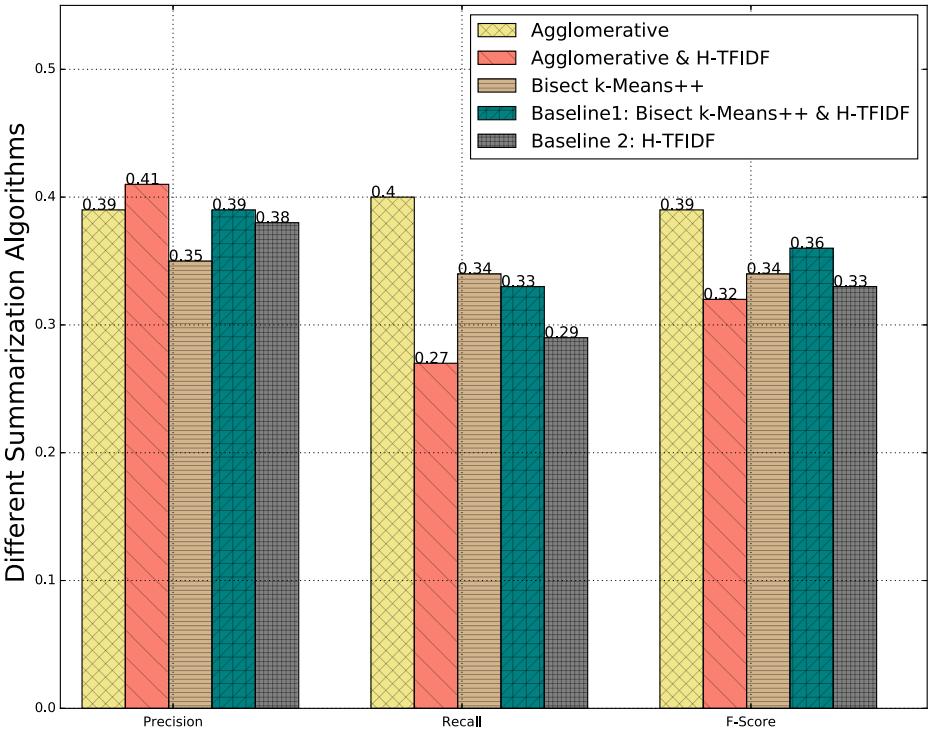

Fig. 3 Performance of document clustering techniques prior to word graphs construction

adopt the same for comparison to the baselines. The $ROUGE-1$ metric can be modified to obtain precision of automatically generated summaries as follows:

$$p = ROUGE-1' = \frac{\sum_{m \in MS} \sum_{u \in m} match(u)}{|MS| \times \sum_{u \in a} count(u)} \tag{9}$$

where $|MS|$ is the number of manual summaries; a is the automatically generated summary. The recall of the automated summaries can be also computed using related formulation of the ROUGE metric, as follows:

$$r = ROUGE-1 = \frac{\sum_{m \in MS} \sum_{u \in m} match(u)}{\sum_{m \in MS} \times \sum_{u \in m} count(u)} \tag{10}$$

where u is the set of unigrams in a particular manual summary. Finally, $F-measure$ which is a harmonic average of precision and recall is computed as follows:

$$F-measure = \frac{2 \times p \times r}{p + r} \tag{11}$$

6.7 Evaluation results

In this subsection, we evaluate the effectiveness of our proposed work compared to several summarization approaches. Figure 3 shows the results of ROUGE based on clustering techniques prior to Word Graphs. We observe that Sharifi et al.'s baseline of Bisect K-Means++ with Hybrid TF-IDF was outperformed by Agglomerative clustering with 1-means pass.

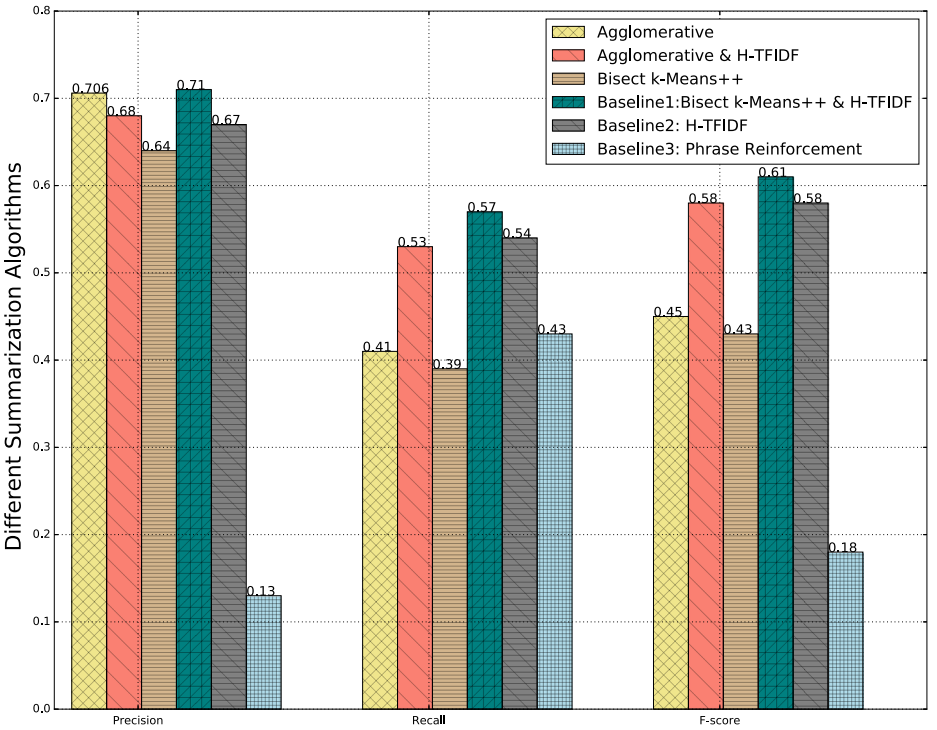

Fig. 4 Performance of document clustering techniques after word graphs construction

The standalone Bisect K-Means++ method, which picks out the centroids from each cluster with a 1-means pass also performed competitively. Agglomerative clustering technique with 1-means pass, which has not been evaluated in any previous work for microblog clustering and is proposed in this paper, outperforms all the other algorithms.

Figure 4 shows the performance of the methods after word graph construction, with the addition of Phrase Reinforcement. We wanted to bring in the PR method in this experiment to observe the effectiveness of choosing the most weighted phrases. In many instances, tweets did not contain the topic phrases that existed in the concept titles of the topic. Hence, the performance of the PR algorithm was low.

We notice that all algorithms performed better after Word Graph construction. The best overall summarizer was Bisect K-Means++ with Hybrid TF-IDF. The Hybrid TF-IDF algorithm, whether with or without Word Graphs, did not perform as well as expected. We also evaluate which summarization algorithm benefits the most from the word graphs. Figure 5 shows the results of the performance deltas. We can observe that Agglomerative clustering with Hybrid TF-IDF (F-Measure delta: +0.255) benefits the most, followed by Bisect K-Means++ with Hybrid TF-IDF (F-Measure delta: +0.251), and followed by Hybrid TF-IDF only ((F-Measure delta: +0.249). We also noted that all techniques had positive deltas, which was desirable and points to the fact that when word graphs are used as proposed in this paper, a more focused set of tweets are considered when generating summaries.

Finally, we conclude that Agglomerative clustering technique with 1-means pass has been the better summarizer without the construction of Word Graphs. With the construction of Word Graphs, all summarization algorithms that we experimented had improved F-Measures compared to the baselines. In particular, the Agglomerative clustering technique with 1-means pass reported the highest performance improvement.

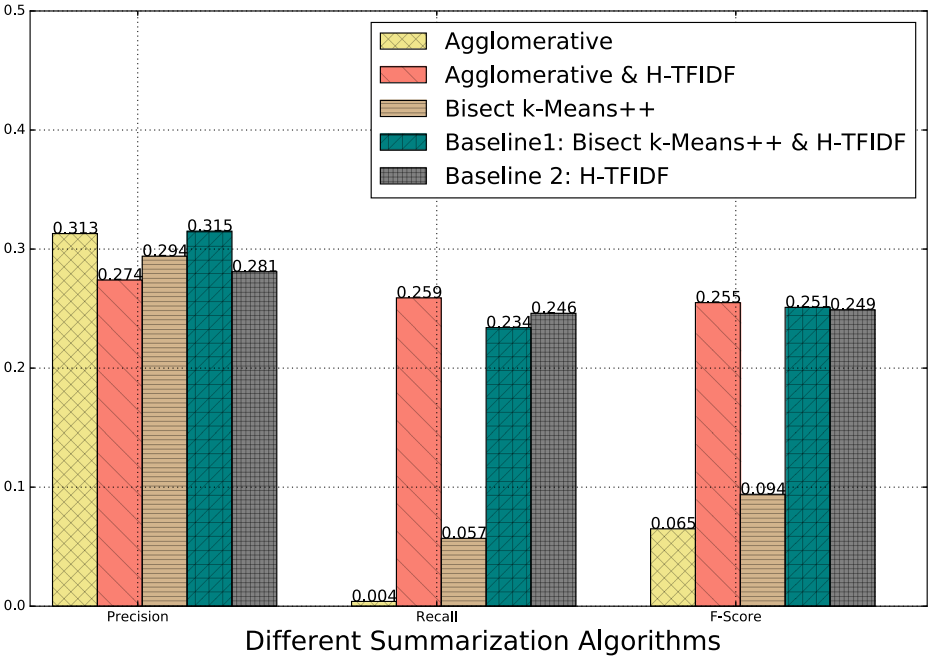

Fig. 5 Delta of the performance of document clustering techniques

Our research findings can be summarized as follows:

1. The content score of Agglomerative clustering is higher *prior* and *post* Word Graph construction compared with other summarization algorithms, meaning that the automatically generated summaries by our proposed Agglomerative clustering based approach are better matched with the human perception of summarization quality.
2. The quality of summaries generated by Agglomerative clustering technique with 1-means pass outperforms other state-of-the-art approaches without the construction of Word Graphs.
3. Building a sentiment-based Word Graph to extract different aspects of the topic, as proposed in this paper, improves the performance of the summarization task. The F-Measures of all representative summarization algorithms are consistently higher *after* the construction of the Word Graphs.
4. Agglomerative clustering with Hybrid TF-IDF has the highest improvement with the construction of Word Graphs, with an F-Measure delta of +0.255.

7 Concluding remarks

In this paper, we presented a summarization method based on sentiment topical aspect analysis to automatically generate a holistic textual summary for the topics present on Twitter. The proposed approach features the exploration of the intrinsic sentiments associated with the microblog posts as well as the analysis of the topical aspects for enhancing the summarization performance. In particular, we propose three major stages to accomplish summarization. First, we employed an approach for semantically enriching the microblog post and extracting the sentiment and the topic associated with each microblog post. Then, we create a Word Graph for each topic-sentiment cluster of microblog posts to identify different topical aspects of each specific cluster. Finally, we generate a holistic summary for each topic by applying different state-of-the-art summarization algorithms on each topical aspect of the Word Graph.

For the purpose of evaluation, we conducted a series of experiments on a Twitter dataset (Abel et al. 2011) to comparatively evaluate the performance of our proposed work against existing document clustering algorithms. Experimental results showed that inducing topical aspects with the word graphs will improve the quality of the generated summaries compared to the state of the art. We also comparatively evaluated the performance of our work against leading document clustering algorithms without the construction of the Word Graph. We found that our proposed Agglomerative clustering based approach with 1-means pass, which to the best of our knowledge has not been employed for the summarization task before and is one of the contributions of our paper, outperforms the others.

There are areas where our evaluations could be strengthened. There have been work in the document summarization literature that argue ROUGE-2 is a more reliable measure of summary quality compared to the ROUGE-1 metric. As such, we intend to perform additional evaluation in our future work to evaluate our work based on the ROUGE-2 metric. The primary reason for selecting ROUGE-1 in this paper has been motivated by this choice in the baseline paper. Additionally, performing more in-depth evaluation of the performance of each of the steps of our proposed method, in addition to the overall performance evaluation of the summarization technique, can provide insight as to which steps are causing the issues and are in fact the primary causes for errors

in summarization. We also intend to undertake such evaluation as a part of our future work.

In addition and as a part of our future work, we intend to improve the proposed approach in the following ways. First, our summarization process does not currently consider the time of occurrence of the tweets. Motivated by Time-Aware Knowledge Extraction (TAKE) methodology presented in De Maio et al. (2016), we plan to extend our summarization process to incorporate the temporal evolution of tweets by identifying temporal peaks of tweets frequency through analyzing their timestamps. Second, it would be interesting to evaluate the performance of the proposed summarization process in the context of question-answering systems. Previous work on query-oriented summarization (Miao and Li 2010; Torres-Moreno et al. 2009; Biryukov et al. 2005) mostly aim to automatically extract information from documents. We intend to study how our proposed approach can be adopted in question-answering systems such that relevant and useful answers are generated from the corpus of tweets for a given query. Finally, in the first stage of our approach, we have performed tweet content wikification (Feng et al. 2018) by enriching each tweet using an existing semantic annotator (i.e., TagMe) which links entities to the pertinent Wikipedia articles. However, our work could benefit from various annotation systems such as Carrillo-de Albornoz et al. (2016), De Maio et al. (2014), and Hu et al. (2009) which use Wikipedia as a resource to support accurate algorithms for keyword extraction and word sense disambiguation. One of the possible venues for future work will be to conduct different experiments to compare the efficacy of each annotation algorithm in the context of the proposed summarization process.

Appendix

Table 8 shows the topics and their associated number of tweets that are used in our experiments. Note that the internal cohesion of all the topics is 1. Table 9 shows samples of summaries generated by different clustering algorithms along with the manual summary generated by our volunteers based on the topics from (Table 10). The set of extracted aspects are reported in Table 11, which are then assigned to respective aspects as reported in Table 12. Finally, we pick one representative tweet for each sentiment-aspect pair in order to generate a summary shown in Table 13.

Table 8 Topics and their associated tweets in our experiments

Topics	Keywords	Positve tweets	Negative tweets	Neutral tweets
Topic 1	Vehicle,Accident Anchorage, Alaska Snow	80	80	80
Topic 2	Upgrade U, IPhone 3G, Apple Inc., Lawsuit	80	80	80
Topic 3	Privacy, Facebook	172	162	387
Topic 4	China, Inflation	80	80	80
Topic 5	HIV/AIDS, Malaria, World Pneumonia Day	89	237	266
Topic 6	Bloomberg Businessweek, Economy of the United States, Retail	95	247	262

Table 9 Sample generated summary for different clustering algorithms

Topic & Sentiment	Agglomerative	Bisect K-Means++ w/H-TFIDF	Hybrid TF-IDF	Manual
Topic1 (Negative Tweets) UnprocessedTweets	RT @wxchannel: NASA Modis satellite imagery showing snow cover and the storm over the Northeast Monday: http://bit.ly/hfBdLa#eastsnow rt sharp gradient in snow on western edge contrast forecast snowfall from washington dc to boston, rt colder air is filtering inblue canyon is 29 degreessnow at tahoe should all be snow above 5000 feet now,rt it now appears heavy snow is not likely in indiana christmas eve tilt a few snow showers are possible north east	Brrrr...Today in 1947, over 26 inches of snow fell on New York City; it was the city's heaviest snowfall on record., RT @13News: #Snow for the record books: NORFOLK – Sunday's snowfall was the 3rd heaviest on record for Norfolk.... http://bit.ly/fevHrz,At least 46 states had snow this #Christmas. More than 50% of Lower 48 had #snow cover Christmas morning: http://ow.ly/3uWAU,9:00am Snow Update: LATEST: Slow and steady would describe the snowfall here in the viewing area. Here is the l... http://bit.ly/fhLvqv	Brrrr...Today in 1947, over 26 inches of snow fell on New York City; it was the city's heaviest snowfall on record.,9:00am Snow Update: LATEST: Slow and steady would describe the snowfall here in the viewing area. Here is the l... http://bit.ly/fhLvqv, RT @13News: #Snow for the record books: NORFOLK – Sunday's snowfall was the 3rd heaviest on record for Norfolk.... http://bit.ly/fevHrz,' Tis the season to be... snowy! Snowfall has begun in the northern areas of the U.S. Share your experiences here:,http://on.cnn.com/fuxyVc	RT @colbertema: Winter Weather Advisory until 6pm today. Moisture is bringing moderate snow showers. New snow accumulation of 1" likely.,'Tis the season to be... snowy! Snowfall has begun in the northern areas of the U.S. Share your experiences here: http://on.cnn.com/fuxyVc, I hate snow.. It's so.. Snowy,Crap!! Looking at forecast for next 10 days Rain and Snow showers are in for 12/11 #SantaCon! PLEASE, PLEASE, PLEASE DON'T MESS UP MY DAY!
Topic2 (Positive Tweets) Processed Tweets	apple sued over privacy,apples black friday shopping event starts in the us, apples ipad helps israeli hospital treat patients reuters, rt apples new energy efficient devices arent so great for the environment	rt apples latest ipad ad is magically amazing ,rt mashable news apples black friday shopping event starts in the us,apples new energy efficient devices arent so great for the environment,rt apples ipad helps israeli hospital treat patients	rt obama praises the success of apples steve jobs rt apples black friday shopping event starts in the us,rt apples latest ipad ad is magically amazing, apples new energy efficient devices arent so great for the environment	rt apples new energy efficient devices arent so great for the environment apples patent may unlock 3d technology,googles new android music player,how did the apple logo come to be

Table 10 Tweet corpus for the *snowfall* topic with associated sentiments

Tweet	Sentiment
rt powerful western storm dumps inches of rainfeet of snow for ca today 12ft wasatch and rockies too storm warms u	Positive
rt mornin all woke up to snow falling outside my house today first snowfall of the season for us beautiful	Positive
wow 13 ft of fresh snow in our mountainsguess there is an upside to 7 days of rain have fun so cal skiers	Positive
hey tweeties hope you had a blessed day we had snow here today just a little but still saw tons if accidents	Positive
wow crazy weather around the world high elevations of ca could get 15 ft of snow	Negative
mountains news epic storm could drop 8 feet of snow on colorado high country	Negative
rt powerful western storm dumps inches of rainfeet of snow for ca today 12ft wasatch and rockies too storm warms u	Negative
rt my goodness its snowing really hard here and its only 500 ft elevation	Negative
how to keep airports open even at 2 ft of snow in helsinki which hasnt been closed since cont	Negative
rt the uk continues to reel from a few inches of snow but im trying to think of a way to get to the 2 ft of powder that hit	Negative
i know the snow is bad but an ice storm is really bad i wondered if it would be heavy wet snow instead of the powder kind	Negative
powerful western storm dumps inches of rainfeet of snow for ca today 12ft wasatch and rockies too storm warms up the east late wk	Negative
bkken i have the best tree ever its like 20 tall the snow just made it amazing preprocess-docem1 smh its 10 ft	Negative
snow showers continue today in indy area	Negative
snow showers and squalls will increase today some will be heavy at times leading to quick accumulations and snow covered roads	Negative

Table 11 Aspects extracted from the word graph based on the tweets and their sentiments

	Positive tweets	Negative tweets
Aspect 1	ft, Snow, Beautiful	ft, hard, elevation
Aspect 2	Accidents	snow, shower

Table 12 Selected tweets for two different aspects

	Positive tweets	Negative tweets
Aspect 1	1. rt powerful western storm dumps inches of rainfeet of snow for ca today 12ft wasatch and rockies too storm warms up.	1. wow crazy weather around the world high elevations of ca could get 15 ft of snow
	2. rt mornin all woke up to snow falling outside my house today first snowfall of the season for us beautiful	2. rt powerful western storm dumps inches of rainfeet of snow for ca today 12ft wasatch and rockies too storm warms up
	3. wow 13 ft of fresh snow in our mountains guess there is an upside to 7 days of rain have fun so cal skiers	3. rt my goodness its snowing really hard here and its only 500 ft elevation
		4. how to keep airports open even at 2 ft of snow in helsinki which hasnt been closed since cont
		5. rt the uk continues to reel from a few inches of snowbut im trying to think of a way to get to the 2 ft of powder that hit
		6. powerful western storm dumps inches of rainfeet of snow for ca today 12ft wasatch and rockies too storm warms up the east late wk
		7. bkken i have the best tree ever its like 20 tall the snow just made it amazing preprocess docem1 smh its 10 ft
Aspect 2	1. hey tweeties hope you had a blessed day we had snow here today just a little but still saw tons if accidents	1. snow showers continue today in indy area, snow showers and squalls will increase today some will be heavy at times leading to quick accumulation and snow covered roads.
		2. mountains news epic storm could drop 8 feet of snow on colorado high country
		3. i know the snow is bad but an ice storm is really bad i wondered if it would be heavy wet snow instead of the powder kind

Table 13 The set of summary tweets for the two aspects

	Representative Tweet (Positive)	Representative Tweet (Negative)
Aspect 1	rt mornin all woke up to snow falling outside my house today first snowfall of the season for us beautiful	powerful western storm dumps inches of rainfeet of snow for ca today 12ft wasatch and rockies too storm warms up the east late wk
Aspect 2	hey tweeties hope you had a blessed day we had snow here today just a little but still saw tons if accidents	snow showers and squalls will increase today some will be heavy at times leading to quick accumulations and snow covered roads

References

Abel, F., Gao, Q., Houben, G.-J., Tao, K. (2011). Analyzing user modeling on twitter for personalized news recommendations, User Modeling, Adaption and Personalization, pp. 1–12.

Abdullah, Z., & Hamdan, A. (2015). Hierarchical clustering algorithms in data mining.

Ackermann, M.R., Blömer, J., Kuntze, D., Sohler, C. (2014). Analysis of agglomerative clustering. *Algorithmica*, *69*(1), 184–215.

Amigó, E., De Albornoz, J.C., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., De Rijke, M., Spina, D. (2013). Overview of replab 2013: Evaluating online reputation monitoring systems. In *International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 333–352 Springer*.

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence*, *31*(1), 132–164.

Bhargava, R., Sharma, Y., Sharma, G. (2016). Atssi: Abstractive text summarization using sentiment infusion. *Procedia Computer Science*, *89*, 404–411.

Bian, J., Yang, Y., Zhang, H., Chua, T.-S. (2015). Multimedia summarization for social events in microblog stream. *IEEE Transactions on Multimedia*, *17*(2), 216–228.

Bild, D.R., Liu, Y., Dick, R.P., Mao, Z.M., Wallach, D.S. (2015). Aggregate characterization of user behavior in twitter and analysis of the retweet graph. *ACM Transactions on Internet Technology (TOIT)*, *15*(1), 4.

Biryukov, M., Angheluta, R., Moens, M.-F. (2005). Multidocument question answering text summarization using topic signatures. *JDIM*, *3*(1), 27–33.

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Carrillo-de Albornoz, J., Amigó, E., Plaza, L., Gonzalo, J. (2016). Tweet stream summarization for online reputation management. In *European Conference on Information Retrieval, pp. 378–389 Springer*.

Chakrabarti, D., & Punera, K. (2011). Event summarization using tweets. *ICWSM*, *11*, 66–73.

De Maio, C., Fenza, G., Gallo, M., Loia, V., Senatore, S. (2014). Formal and relational concept analysis for fuzzy-based automatic semantic annotation. *Applied intelligence*, *40*(1), 154–177.

De Maio, C., Fenza, G., Loia, V., Parente, M. (2016). Time aware knowledge extraction for microblog summarization on twitter. *Information Fusion*, *28*, 60–74.

Dongen, S. (2000). Performance criteria for graph clustering and markov cluster experiments.

Erkan, G., & Radev, D.R. (2004a). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479.

Erkan, G., & Radev, D.R. (2004b). Lexpagerank: Prestige in multi-document text summarization. In *EMNLP*, (Vol. 4 pp. 365–371).

Feng, Y., Bagheri, E., Ensan, F., Jovanovic, J. (2017). The state of the art in semantic relatedness: A framework for comparison, The Knowledge Engineering Review.

Feng, Y., Zarrinkalam, F., Bagheri, E., Fani, H., Al-Obeidat, F. (2018). Entity linking of tweets based on dominant entity candiyears. *Social Network Analysis and Mining*, *8*(1), 46.

Ferragina, P., & Scaiella, U. (2010). Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 1625–1628 ACM*.

Ganesan, K., Zhai, C., Han, J. (2010). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics, pp. 340–348 Association for Computational Linguistics*.

Genest, P.-E., & Lapalme, G. (2012). Fully abstractive approach to guided summarization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, pp. 354–358 Association for Computational Linguistics*.

Go, A., Bhayani, R., Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(2009), 12.

Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J. (1999). 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, pp. 121–128 ACM*.

Gong, Y., & Liu, X. (2001). 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, pp. 19–25 ACM*.

Haghighi, A., & Vanderwende, L. (2009). Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 362–370 Association for Computational Linguistics*.

Hennig, L., & Labor, D. (2009). Topic-based multi-document summarization with probabilistic latent semantic analysis. In *Ranlp* (pp. 144–149).

Hu, X., Zhang, X., Lu, C., Park, E.K., Zhou, X. (2009). Exploiting wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 389–396 ACM*.

Hu, Y.-H., Chen, Y.-L., Chou, H.-L. (2017). Opinion mining from online hotel reviews–a text summarization approach. *Information Processing & Management*, *53*(2), 436–449.

Inouye, D., & Kalita, J.K. (2011). Comparing twitter summarization algorithms for multiple post summaries. In *2011 IEEE 3rd international conference on privacy, security, risk and trust (PASSAT) and 2011 IEEE 3rd Inernational conference on social computing (SocialCom), pp. 298–306 IEEE*.

Jashki, M.-A., Makki, M., Bagheri, E., Ghorbani, A.A. (2009). An iterative hybrid filter-wrapper approach to feature selection for document clustering. In *Proceedings of the 22Nd Canadian Conference on Artificial Intelligence: Advances in Artificial Intelligence, Canadian AI '09* (pp. 74–85). Berlin: Springer.

Jing, H., & McKeown, K.R. (2000). Cut and paste based text summarization. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, pp. 178–185 Association for Computational Linguistics*.

Jones, K.S. (2007). Automatic summarising: The state of the art. *Information Processing & Management*, *43*(6), 1449–1481.

Knight, K., & Marcu, D. (2002). Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence*, *139*(1), 91–107.

Ku, L.-W., Liang, Y.-T., Chen, H.-H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI, pp. 100–107*.

Lin, C.-Y., & Hovy, E. (2002). From single to multi-document summarization: A prototype system and its evaluation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 457–464 Association for Computational Linguistics*.

Lin, C.-Y., & Hovy, E. (2003). 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-Volume 1, pp. 71–78 Association for Computational Linguistics*.

Lin, C., Li, J., Wang, D., Chen, Y., Li, T. (2012). Generating event storylines from microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 175–184 ACM*.

Ling, X., Mei, Q., Zhai, C., Schatz, B. (2008). Mining multi-faceted overviews of arbitrary topics in a text collection. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 497–505 ACM*.

Liu, F., Flanigan, J., Thomson, S., Sadeh, N., Smith, N.A. (2015). Toward abstractive summarization using semantic representations.

Lloret, E., & Palomar, M. (2011). Analyzing the use of word graphs for abstractive text summarization. In *Proceedings of the First International Conference on Advances in Information Mining and Management, Barcelona* (pp. 61–6).

Louis, A., & Nenkova, A. (2009). Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, pp. 306–314 Association for Computational Linguistics*.

Mani, I. (2001). Automatic summarization, Vol. 3, John Benjamins Publishing, Amsterdam.

Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S., Miller, R.C. (2011). Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 227–236 ACM*.

Meila, M. (2003). Comparing clusterings by the variation of information. In *Colt, vol. 3, pp. 173–187 Springer*.

Miao, Y., & Li, C. (2010). Enhancing query-oriented summarization based on sentence wikification. In *Workshop of the 33 rd Annual International* (p. 32).

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *EMNLP*, (Vol. 4 pp. 404–411).

Mihalcea, R., & Tarau, P. (2005). A language independent algorithm for single and multiple document summarization. In *Proceedings of IJCNLP*, Vol. 5.

Newman, M.E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, *74*(3), 036104.
Nichols, J., Mahmud, J., Drews, C. (2012). Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, pp. 189–198 ACM*.
Ohsawa, Y., Benson, N.E., Yachida, M. (1998). Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries, 1998. ADL 98. pp. 12–18 IEEE*.
Piryani, R., Gupta, V., Kumar Singh, V. (2018). Generating aspect-based extractive opinion summary: Drawing inferences from social media texts. *Computación y Sistemas*, *1*, 22.
Radev, D.R., Jing, H., Styś, M., Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*, *40*(6), 919–938.
Ramage, D., & Rosen, E. (2011). Stanford topic modeling toolbox.
Rosvall, M., & Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123.
Saggion, H., Torres-Moreno, J.-M., Cunha, I.d., SanJuan, E. (2010). Multilingual summarization evaluation without human models. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 1059–1067 Association for Computational Linguistics*.
Saif, H., He, Y., Alani, H. (2012). Semantic sentiment analysis of twitter. The Semantic Web–ISWC 2012, pp. 508–524.
Sharifi, B., Hutton, M.-A., Kalita, J.K. (2010). Experiments in microblog summarization. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on, pp. 49–56 IEEE*.
Sharifi, B.P., Inouye, D.I., Kalita, J.K. (2013). Summarization of twitter microblogs. *The Computer Journal*, *57*(3), 378–402.
Steinbach, M., Karypis, G., Kumar, V., et al (2000). A comparison of document clustering techniques. In *KDD workshop on text mining, vol. 400, pp. 525–526 Boston*.
Sun, J.-T., Shen, D., Zeng, H.-J., Yang, Q., Lu, Y., Chen, Z. (2005). Web-page summarization using clickthrough data. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 194–201 ACM*.
Titov, I., & McDonald, R. (2008). A joint model of text and aspect ratings for sentiment summarization, Proceedings of ACL-08: HLT, pp. 308–316.
Torres-Moreno, J.-M., St-Onge, P.-L., Gagnon, M., El-Beze, M., Bellot, P. (2009). Automatic summarization system coupled with a question-answering system (qaas), arXiv:0905.2990.
Vanderwende, L., Suzuki, H., Brockett, C., Nenkova, A. (2007). Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, *43*(6), 1606–1618.
Varga, A., Basave, A.E.C., Rowe, M., Ciravegna, F., He, Y. (2014). Linked knowledge sources for topic classification of microposts: a semantic graph-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, *26*, 36–57.
Wan, X., & Yang, J. (2008). Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 299–306 ACM*.
Wang, D., Li, T., Zhu, S., Ding, C. (2008). Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 307–314 ACM*.
Wu, F., & Huberman, B.A. (2004). Finding communities in linear time: a physics approach. *The European Physical Journal B-Condensed Matter and Complex Systems*, *38*(2), 331–338.
Wu, H., Gu, Y., Sun, S., Gu, X. (2016). Aspect-based opinion summarization with convolutional neural networks. In *Neural Networks (IJCNN), 2016 International Joint Conference on, pp. 3157–3163 IEEE*.
Xu, X., Meng, T., Cheng, X. (2011). Aspect-based extractive summarization of online reviews. In *Proceedings of the 2011 ACM Symposium on Applied Computing, pp. 968–975 ACM*.
Yih, W.-t., Goodman, J., Vanderwende, L., Suzuki, H. (2007). Multi-document summarization by maximizing informative content-words. In *IJCAI*, (Vol. 7 pp. 1776–1782).
Zarrinkalam, F., Fani, H., Bagheri, E., Kahani, M., Du, W. (2015). Semantics-enabled user interest detection from twitter. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, Singapore, December 6-9, 2015 - Volume I* (pp. 469–476).
Zarrinkalam, F., Fani, H., Bagheri, E., Kahani, M. (2016). Inferring implicit topical interests on twitter. In *European Conference on Information Retrieval, pp. 479–491 Springer*.

Zhou, L., & Hovy, E.H. (2006). On the summarization of dynamically introduced information: Online discussions and blogs. In *AAAI Spring symposium: Computational approaches to analyzing weblogs, p. 237*.

Zhou, X., Wan, X., Xiao, J. (2016). Cminer: opinion extraction and summarization for chinese microblogs. *IEEE Transactions on Knowledge and Data Engineering*, *28*(7), 1650–1663.

Zhuang, L., Jing, F., Zhu, X.-Y. (2006). Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 43–50 ACM*.