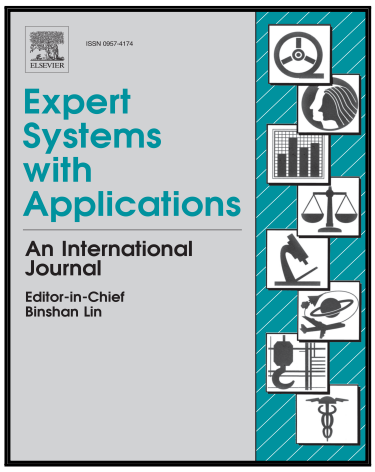

Graph Coloring and ACO based Summarization for Social Networks

Mohamed Atef Mosa , Alaa Hamouda , Mahmoud Marei

PII: S0957-4174(17)30011-8
DOI: 10.1016/j.eswa.2017.01.010
Reference: ESWA 11058

To appear in: *Expert Systems With Applications*

Received date: 24 July 2016
Revised date: 26 December 2016
Accepted date: 9 January 2017

Please cite this article as: Mohamed Atef Mosa , Alaa Hamouda , Mahmoud Marei , Graph Coloring and ACO based Summarization for Social Networks, *Expert Systems With Applications* (2017), doi: 10.1016/j.eswa.2017.01.010

Highlights

- An unconventional GR-ACO-LS-STS for short text summarization it is proposed.
- The mechanism relies on graph coloring mixed with Ant colony optimization and local search.
- The mechanism was found more efficient than the other traditional algorithms.

Graph Coloring and ACO based Summarization for Social Networks

Mohamed Atef Mosa[1], Alaa Hamouda[2], Mahmoud Marei[3]

[1]*National Authority for Remote Sensing and Space Sciences, Cairo, Egypt*

[2, 3]*Faculty of Computer Engineering, Al-Azhar University, Cairo, Egypt*

[1]*m.atef@narss.sci.eg,* [2]*alaa.hamouda@azhar.edu.eg,* [3]*marie@azhar.edu.eg*

Abstract*—Due to the increasing popularity of contents of social media platforms, the number of posts and messages is steadily increasing. A huge amount of data is generated daily as an outcome of the interactions between fans of the networking platforms. It becomes extremely troublesome to find the most relevant, interactive information for the subscribers. The aim of this work is to enable the users to get a powerful brief of comments without reading the entire list. This paper opens up a new field of short text summarization (STS) predicated on a hybrid ant colony optimization coming with a mechanism of local search, called ACO-LS-STS, to produce an optimal or near-optimal summary. Initially, the graph coloring algorithm, called GC-ISTS, was employed before to shrink the solution area of ants to small sets. Evidently, the main purpose of using the GC algorithm is to make the search process more facilitated, faster and prevents the ants from falling into the local optimum. First, the dissimilar comments are assembled together into the same color, at the same time preserving the information ratio as for an original list of comment. Subsequently, activating the ACO-LS-STS algorithm, which is a novel technique concerning the extraction of the most interactive comments from each color in a parallel form. At the end, the best summary is picked from the best color. This problem is formalized as an optimization problem utilizing GC and ACO-LS to generate the optimal solution. Eventually, the proposed algorithm was evaluated and tested over a collection of Facebook messages with their associated comments. Indeed, it was found that the proposed algorithm has an ability to capture a good solution that is guaranteed to be near optimal and had realized notable performance in comparison with traditional document summarization algorithms.

***Keywords*—** *short text summarization, social networks, ant colony optimization, graph coloring, text mining, local search.*

I. INTRODUCTION

In recent years, social media services (SMS) are widespread and have become a significant communication recent years. In accordance with many statistics by the greatest social networking site Facebook and Twitter, there are so many billion active users have numerous daily interactions. Due to the platform's popularity, many celebrities, companies, and organizations also create social pages to interact with their fans. Note that, users are able to express their opinions for each message by forwarding, giving a like, share, and leaving a comment on it. It is possible to note that a number of comments are so vast and the generation rate is dreadfully high. This short text can then be used by fans to improve the understanding of current issues such as, what topics or arguments are significant? For a product, what aspects of this product do fans enjoy? Ultimately, the collective opinions of fans could lead to change purchasing decisions, political decisions, and influences an organization's view of the world. By all means, users needlessly go over the whole comments of each post. With this motivation, and to get over these challenges, an innovative technique has been developed to summarize the user-contributed aggregation comments in SMS. Moreover, user-contributed comments

summarization faces many unique challenges: e.g., length of comments are high variability. The comments have numerous competing opinions, abundant colloquial spelling mostly, implicit references to earlier comments, and so forth.

Abundant studies and systems have proposed mechanisms and algorithms to generate different types of summaries on comment streams. One major objective aims to extract important and representative comments from the messy discussion. Like Facebook (Facebook site) and YouTube (YouTube site), these popular services allow users to find whether a comment is more recommendable or attractiveness of the list. But such this methodology relies on fans' contributions solely. On the other hand, some studies are modeled in this issue as classification (Sriram B. et al., 2010) (Rosa K. D. et al., 2011) (Perez-Tellez F.et al., 2010) or recommended (Khabiri E. et al., 2011) (Chen J. et al., 2012) (Takamura H.et al., 2011) (Chakrabarti D .et al., 2011) (Becker H .et al., 2010) tasks based on machine learning mechanisms. Moreover, to discover the hidden emotions in the comments, a sentiment analysis has been applied (N. Diakopoulos A. and Shamma D. A., 2010) (Li CT .et al., 2011) (Bollen J .et al., 2011) (Tumasjan A. et al., 2010) to do that as well.

In this study, we do not concentrate on conventional comment streams that usually reveal a complete information, such as the extended discussion on movies or products. Our main goal is in the short text style of the comment streams in SMS that are with casual/colloquial language usage. In addition, the user-contributed comment's summarization problem is a process that should cover the manifest aspects and topics of the interested comments subject to certain constraints: (1) ranging of comment's lengths is varied from one/two/three words to many paragraphs, (2) multiple competing opinions arise, or high similarity between comments, and (3) an implicit reference to earlier comments is shown. It can be perceived that this problem can be modeled as an optimization problem. To satisfy this requirement, this problem is formalized mainly as an ant colony optimization ACO task. However, ACO as a well-performed algorithm in solving intricate combinatorial optimization quandaries has not been applied in these types of problems. Besides, the ACO is an optimization problem in which the space of possible solutions is discrete and finite. ACO typically, the feasible solutions are determined by their ability to satisfy certain constraints. Therefore, ACO is one of the many appropriate algorithms for such a problem that's guaranteed to be the near-optimal solution.

In our opinion, considering ACO during solving the short text summarization STS problem, at least, has the following advantages. Noticeably, ACO method is affiliated with the graph coloring GC (Zufferey N. et al., 2008) which guarantees a summary could capture the essence of the whole input text speedily and efficiently. It is therefore, an ACO mixed with graph coloring GC and local search LS to address the problem of short text summarization. Unfortunately, according to our knowledge, no study has taken into account the retention ratio of information (RR) or the length of the summary. Where the RR is a percentage of retaining the information in the final summary as for the primary text; namely, how to retain the gist of the original text. Actually, this work aims to fill this gap. Indeed, the GC algorithm is employed in a modern way to protect the ACO of likely cycling. Initially, the massive amount of data is divided into a number of colors, each is able to generate the final summary standalone. Another motivation behind this work is to apply ACO mixed with LS to solve a short text summarization problem. In the future, except GC, ACO, and LS plenty of heuristic algorithms may be employed to solve this conundrum, such as Bee Colony Optimization, Genetic Algorithm, Local Search, Simulated Annealing, and Taboo Search. In fact, GC and ACO mixed with LS have been demonstrated to be very effective in solving

complex combinatorial optimization problems. In this study, we want to discover the most interactive, representative group of inconsistent opinions that make users easily and rapidly understand. It follows, an innovative technique will be introduced when addressing the STS problem, aiming to enhance the summary efficiency.

The rest of the paper is organized as follows. In the next section, background and related work are introduced. In section III, our proposed algorithm for summarization problem is presented. Section IV exhibits different methods of evaluation and performance analyses of the proposed algorithms and other alternative algorithms. Section V provides the conclusions.

II. BACKGROUND AND RELATED WORKS

Multi-document summarization (MDS) is a similar work to the short text summarization. MDS is the process of filtering a significant information from the set of documents to produce a summary version for particular applications and users. In other words, it can be said that the MDS is an extension of a single document summarization. It is the type of news aggregation is used to reduce the huge corpus of documents into an abbreviated summary. Which represents the most remarkable key topics within the overall corpus. MDS is typically applied to a corpus of lengthy documents, rather than short comments. These documents are standard and coherent texts. Comparing the majority of comments, it was found that they contain many colloquial/slang words. That's why we have to deal professionally with many challenges that could appear. A number of algorithms have been developed for various aspects of MDS during recent years (Marujo et al., 2015) (Dos Santos and Luıs, 2015) (Bing et al., 2015) (AL-Dhelaan and Mohammed, 2015) (Canhasi and Ercan, 2014) (Marujo et al., 2016) (Glavas et al., 2014).

In recent years, social media (SM) is widespread and has the largest platforms in our daily life. With the tremendous increasing amount of online user contributed comments on social media, users unnecessarily go over the whole comments to discover the useful one. In addition to the user contributed comments data discussed in this book, preceding works also target at various types of social data. In this section, we can particularly classify all of these works into four categories: (1) summarization technique, (2) sentiment analysis (3) topic and event detection, and (4) filtering and rating.

Regarding the field of short text summarization, numerous researchers have been concentrated on modeling this problem as a classification and recommended tasks in recent years. And they were focused on micro-blogging (Marcus A. et al., 2011) (Sankaranarayanan J. et al., 2009) (Weng et al., 2011) (Sharifi B. et al., 2010) (Harabagiu SM. et al., 2011). A variety of mechanisms have been developed to satisfy the inconsistent needs of summarization. In (Sharifi B. et al., 2010), with collections of short messages on a particular topic, the short summary sentences were created to describe and pick the primary gist of what fans were saying. In addition, a descriptive system (Marcus A. et al., 2011) is presented to enable the appropriate browsing of the large collection of tweets by detecting the peaks of highly-discussed activity. On the other hand, another system Twitter Stand (Sankaranarayanan J. et al., 2009) is developed to obtain the latest breaking news automatically based on the geographic location of tweets. Besides, in (Weng et al., 2011) the posts were classified into four pre-defined categories, after that, a different strategy has been utilized to pick the final summary. With similar intention, the micro-blogging messages have been collected on the same topic into a prose description.

On the other hand, the study topic has also attracted much attention is analyzing product reviews (Baccianella S. al., 2009). Generally, the first main point obtains the most attractive

aspects of production from review texts. Following, to achieve various summarization needs, the traditional techniques of data mining, natural language processing, and machine learning were incorporated.

Regarding sentiment analysis, many researchers investigate the sentiment analysis (Pang B. et al., 2008) to find out an emotion hidden in social comments. In general, to classify comments into predefined sentiment categories, the sentiment classification model is employed to capture the positive and negative sentiment. In (N. Diakopoulos A.and Shamma D. A., 2010), the authors suggested evaluating the presidential debate performance using the Twitter platform depending on the negative and positive messages. Aside from, over six sentiment labels in (Li CT. et al., 2008) and (Bollen et al., 2011) have evaluated each tweet by aggregating the distribution of sentiment.

On the other hand, some studies endeavor to highlight the most representative information by electing the comments that reveal the different opinions of the group or contain important information. The assessment mechanisms (Khabiri et al., 2011) (Becker H .et al., 2010) (Takamura H.et al., 2011) (Chen J. et al., 2012) (Chakrabarti D .et al., 2011) are excessively developed to determine the significant messages. In addition, to keep meaningful messages and eliminate redundant ones, several types of filtering techniques (Sriram B. et al., 2010) (Perez-Tellez F.et al., 2010) (Daly E. et al., 2011) have also been originated. The work of (Khabiri et al., 2011) aims to select the best top-k representative and informative comments from a list of comments for a specific video. Initially, an improved model of Latent Dirichlet Allocation (LDA) is applied to group the similar comments with each other. After that, select informative comments for each group.

Concerning the micro-blogs selection via Tweeter (Takamura H.et al., 2011) (Chakrabarti D .et al., 2011) (Becker H .et al., 2010), the authors in (Becker H .et al., 2010) propose a new strategy to pick the highest-quality messages. A new selection method is proposed to select the closest nearer messages from the center of the cluster. These messages reflect mainly key aspects of the event. In addition, in (Takamura H.et al., 2011) model of tweets selection as the p-median problem is presented. In (Chakrabarti D .et al., 2011), a modified Model is utilized to divide the event timeline, and then in each segment, the system will pick the closest messages to all other ones. Other researchers (Rosa K. D. et al., 2011) (Sriram B. et al., 2010) aim to classify the messages into a number of topics using derived bag-of-words features. That's why the users can easily find the desired information. In (Perez-Tellez F.et al., 2010) the K-Means clustering algorithm is employed to develop a term expansion methodology.

On the other hand, the main motive of the topic detection on SMS is to assist users in understanding the social stream (Michelson M. and Macskassy S. A., 2010) (O'Connor B., 2010). In (Michelson M. and Macskassy S. A., 2010), an entity-based topic profile for each user is generated by checking if the user's entities mentioned previously. Topics in (O'Connor B., 2010) will be clustered for more directed examination, after that extracting a set of themes and grouped for faceted navigation. Moreover, in (Liu F. et al., 2011), to strengthen the influence of topic summarization on Twitter, multiple text sources are merged with each other.

III. PROPOSED ALGORITHM FOR SHORT TEXT SUMMARIZATION

Extraction of the summary of the social media is one of the biggest problems, especially when the amount of fan-contributed comments are so huge. Thus the main purpose is finding the best minimal subset that has the highest RR of information. Meanwhile, it should have multiple

salient topics without any redundancy in the summary. Therefore, there is no concept of ordering the path/solution. The size of the solution is predefined. Moreover, to select the most appropriate group of comments, the candidate comment should be influenced by the gathered comments into the solution. Furthermore, the produced solutions are necessary to have the same size.

In this section, the detail of the algorithm is expressively presented. The overall process of the algorithm unravels to illustrative components as shown in Fig 1. First, Natural Language Processing (NLP) module is employed to transform the comments to a set of n-terms. Secondly, to protect the ants of likely cycling and to reduce the massive amount of data into a number of groups, each is able to generate the final summary standalone. Thereby, a graph coloring GC-ISTS module is exploited to gather the dissimilar or less similar comments together into the same color. Expressively, the use of the graph coloring module in this problem to constitute a minimum number of groups, where it may be formed an initial set of the summaries. These relatively small groups will prevent the ants from falling into a likely local optimum, improve the efficiency and increase the speed of the algorithm. Meanwhile, the generated solutions should preserve on the RR of information in each color as much as possible. Thirdly, a cyclic semi-graph is constructed within each color by taking into account the extremely lengthy comments are isolated from the graph. Finally, the hybrid ant colony optimization mixed with local search ACO-LS-STS algorithm is applied in all colors in parallel form to discover which subset has the smallest comments, interactive, and a representative that has multiple opinions and remains the highest RR of information as for the original comments.

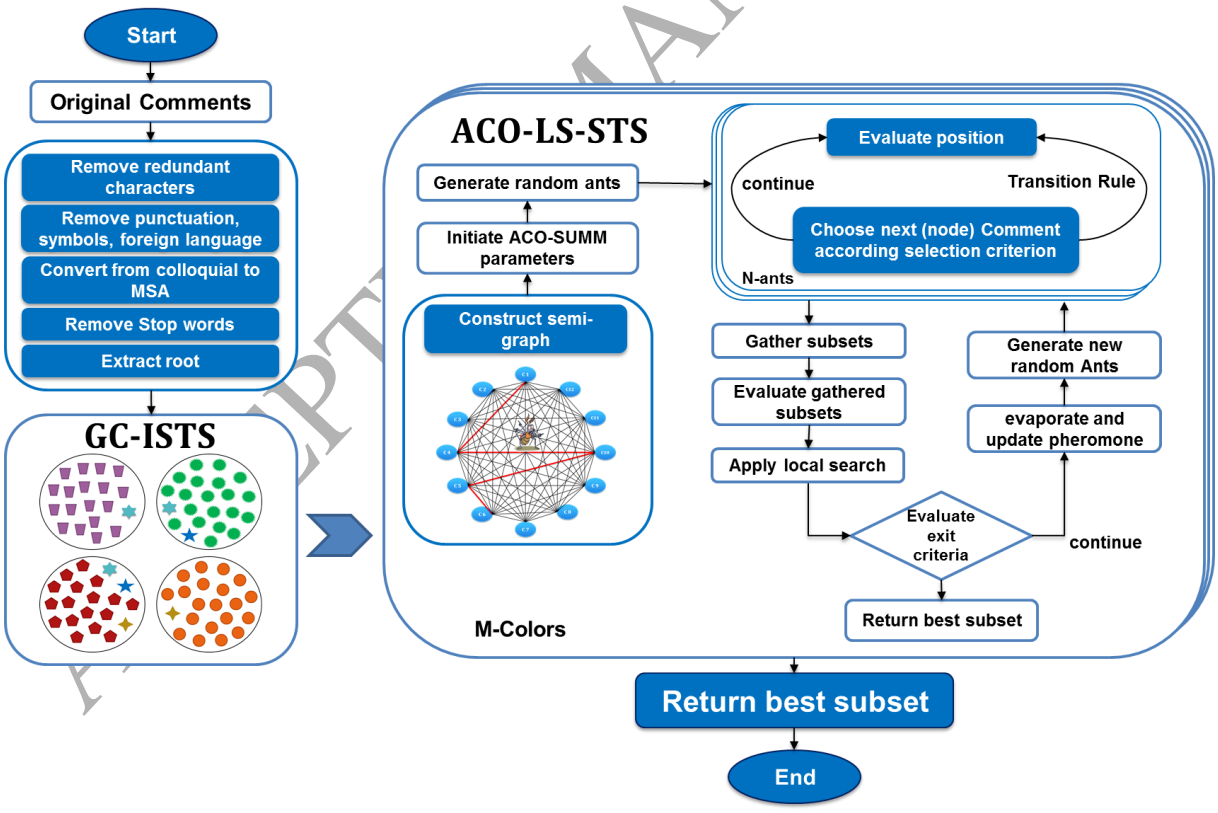

Fig.1. System model of the proposed framework

A. NLP representation for short text summarization

In this section, elaborating the natural language processing (NLP) module that transforms an Arabic comment into a set of terms. The complexity of Arabic language arises in both structure and morphology, and highly derivation, the inflectional language with many term forms. The same three-letter can give many different meanings of the words predicated on different suffixes, affixes, and prefixes. A short texts on SM is unstructured, have many stop words, and many informal syntaxes (El-Fishawy et al., 2014). NLP is a very important practicality employed to reveal the significant words and unify the terms that appear in many forms to compute the term frequency accurately. Fig. 2 shows the procedure flows with an illustrative example.

Initially, remove an additional punctuation, symbols, and any foreign language except Arabic to eliminate inessential punctuation marks. In fig. 2, the "#media" word is eliminated. Furthermore, the heuristic procedure for redundant character removal is developed for recovering the words to the original shape. Thus, the word "humbugsssssssssss", "بيزيييييييييط" is transformed into "humbugs", "بيزيط", especially, the redundant character in a word is oftentimes used to emphasize the great sentiment. The succeeding step is converting the colloquial words to a modern standard Arabic (MSA) (SHAALAN et al., 2007). For instance, the word "humbugs", "الحقايق" is transformed into "fabricate", "يزيف". Later, remove stop words. These words are common words that appear within the text but are meaningless. They only have a syntactic function. Eliminating stop words from the original text helps in identifying the most significant terms accurately. The list of stop words is supplied in (Li et al., 2011). So, the word "the, but", "أحيانا, لكن" will be removed. Finally, Stemming stage is employed. It is one of the most significant factors which use in many natural language processing applications to enhance their performance. Stemming is the process of extracting the root of the word after pruning prefixes and suffixes (Al Hajjar et al., 2010). The main purpose of employing Stemming procedure is to unify the word style that may have many shapes. Subsequently, the word "Media", "الاعلام" will be transformed into "Medium", "علم".

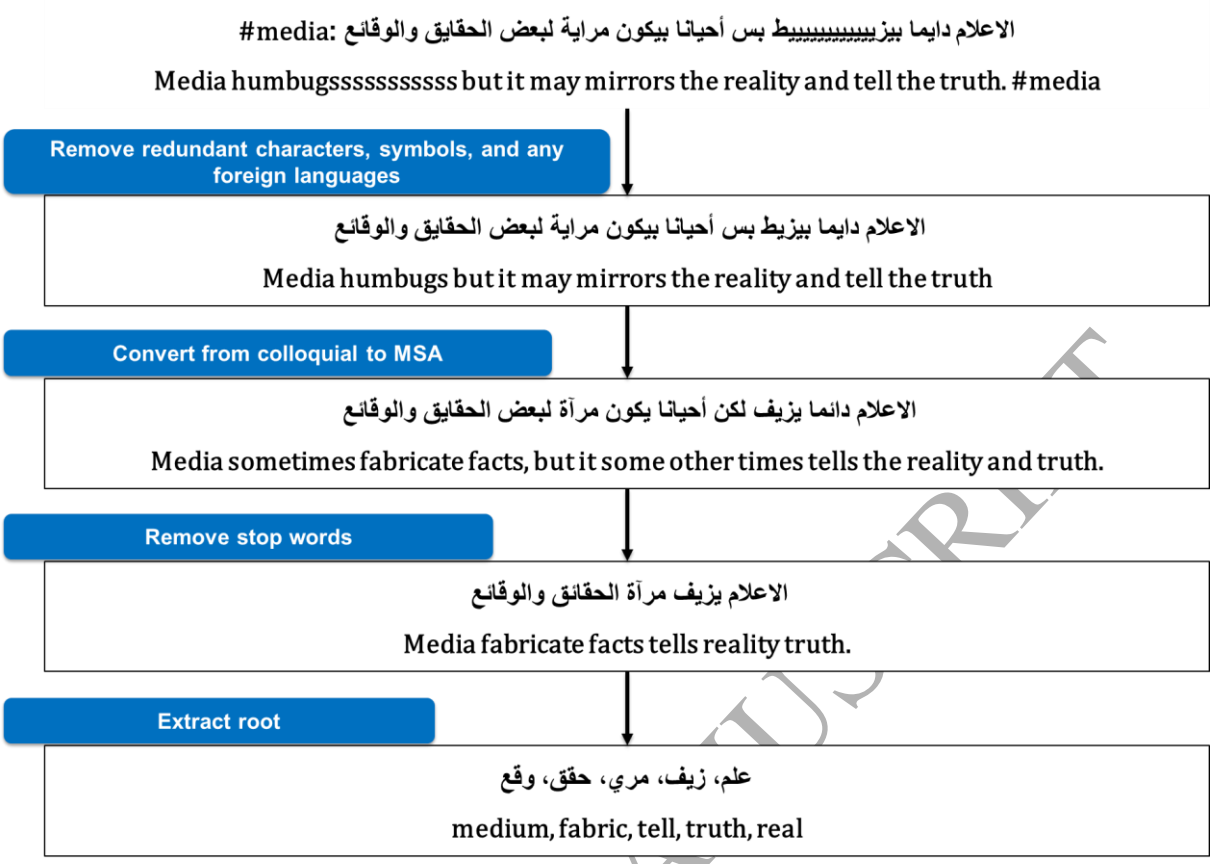

Fig. 2: NLP of the proposed framework

Algorithm GC-ISTS

Input: L: the comments list
θ_s: the similarity threshold
∂_{RR}: the retention rate threshold
Output: minimum number of colors C which have the quite dissimilar comments and highest Retention Rate percentage

1. Initialize $C = \emptyset$;
2. **For** each element m_i of L **Do**
3. **For** each color C_k **Do**
4. **For** each element m_j of color C_k where $i \neq j$ **Do**
5. **If** distance $(m_i, m_j) > \theta_s$
6. Add m_j to $Conflict\_comments$;
7. **If** distance between m_i and all comments into color $C_k \leq \theta_s$
8. Add m_i to C_k;
9. **Else If** $Conflict\_comments \neq \emptyset$
10. Calculate RR of C_k as RR_{ck} according Eq. 10;
11. Replace $Conflict\_comments$ by m_i into C_k and re-calculate the RR of C_k as RR_{new};
12. **If** $RR_{new} < RR_{ck}$
13. Replace $Conflicted\_comments$ by m_i;
14. Create a new color C_{new} with the comment m_i;

15. $C = C + C_{new}$;
16. **If** m_i hasn't assigned to any color
17. Create a new color C_{new} with the comment m_i;
18. $C = C + C_{new}$;
19. **For** each color C_i of C **Do**
20. **If** any color C_i has the RR percentage < ∂_{RR}
21. **For** each element m_i of list C_i **Do**
22. **If** m_i of C_i didn't assigned to any color
23. **For** each element m_j of other colors C_g where $C_g \neq C_i$ **Do**
24. Get the maximum distance between m_i of C_i and m_j of the C_g;
25. Calculate Retention rate percentage of C_g;
26. Append m_i of C_i to the color that has the highest distance and RR;
27. Output minimum number of colors C which have dissimilar and highest RR percentage;
28. **End**

Algorithm 1. Algorithmic form of GC-ISTS

B. GC-ISTS for short text summarization

STS problem is an intelligent method seeks to constitute the best subset from a huge amount of various comments. Some pages on YouTube or Facebook may have a tremendous interaction. The attached comments on posts/videos in some cases may exceed 10,000 comment. Which makes the dealing with all this huge number of the comments so difficult, take a very long time, and the final created summary may not be satisfactory. That's why, a novel technique is employed to facilitate dealing with that huge amount of short messages without significant waste of time, at the same time maintain the goodness of the solutions. The best mechanism to overcome these obstacles is graph coloring GC. The GC module can be efficiently represented for solving the STS problem in tentative form. A given a graph is G = (V, E) with vertex set V and edge set E, where vertices are the comments and edges are the relation between them and given an integer C, where C is the number of colors. In STS problem, vertices represent the comments should be assigned to a color class, while two adjacent vertices (comments) do not have the same color. If two adjacent vertices (comments) x and y have the same color, vertices x and y are said to be conflicting. The vertices x and y aren't assigned to the same color. The C-graph coloring optimization problem is to determine M-coloring of G that minimizes the number of conflicting vertices. It is exploited adeptly to generate more one group, each is able to generate a summary. Eventually, the final summary could be extracted from the best one, vice versa the concept of clustering. With the clustering concept, the comments that have the smallest distance are gathered with each other in the same cluster, after that to realize the summarization process, a number of the comments are picked from each cluster depending on the significance and the length of the required summary. While in this matter, the GC module will constitute many colors, whereas each has a collection of considerably various comments, and all different topics that have been addressed by users. Later, the final summary will be extracted from just the best one. It should be noted that the main purpose of employing the GC module is to reduce the number of comments in each color, subsequently to protect the ants of falling in the likely local optimum.

In this study, the main purpose of employing the graph coloring module is to divide the comments into a minimum set of C colors, each can generate the final summary standalone. The algorithmic form of GC-ISTS in outlined in Algorithm 1. GC-ISTS takes the whole comment set L as the input. The second input is the similarity threshold θ_s used for determining how dissimilar the comments are in a color. The third input is the RR threshold ∂_{RR} used for measurement the RR of information in each color. There are two main steps in GC-ISTS. The aim of the first step, shown in lines 2-18 of Algorithm 1, is to find all disconnected comments into the set L. It can be noticed that if there is a link between two comments as their distance, these comments are assigned to different colors. In addition, the same comment can be added to more one color. In line 6, gathering the conflict comments that prevent m_i to be added to the same color. In lines 9-15, the retention rate of information RR_{ck} of color C_k will be calculated according to Eq. 10 twice if the C_k has the conflict comments with the m_i element. The first one, without the m_i element, and the second one, when the conflict comments are replaced by the m_i. Later, the algorithm remains on the comments that achieve the highest RR information. The new color is created for the other one thereafter. Subsequently, in lines 19-26, the objective of the second main step is to remain the colors that have the RR of information greater than the threshold ∂_{RR}. To meet this requirement, for each color C_i that has a bad percentage of information, we exclude it's comments m_i. Meanwhile, in line 21, it will be checked whether m_i is assigned to any color. If not, the comment m_i will be appended to the color that has the minimum similarity and the largest RR when m_i is added. After this step, all colors will satisfy the RR of information restriction and similarity threshold as much as possible. Finally, GC-ISTS outputs the minimum colors which have the quite dissimilar comments and highest RR.

Note that the bottleneck of GC-ISTS is the first step that finds out all the distances between all comments in set L. The worst case scenario is when the distances between all pairs of comments computed, indicating that the complexity of GC-ISTS is $O(n^2)$, where n is the total number of comments in L. To further finalize the summarization process and made it more efficient, a fully ACO-LS-STS algorithm is introduced in section D.

On the other hand, one commonly adopted metric is used to measure the distance between comments: Jaccard method that is defined as:

$$\text{Jaccard}_{(C_i,C_j)} = \frac{C_i \cdot C_j}{C_i \cup C_j} \tag{1}$$

The Jaccard coefficient measures the distance between finite sample sets and is defined as the size of the intersection between two term list divided by the size of the union of the sample sets: The Jaccard similarity of two comments ranges from 0 to 1. Note that the comments are not equivalent in length. That's why, in this situation, the number of common terms is accounted additionally.

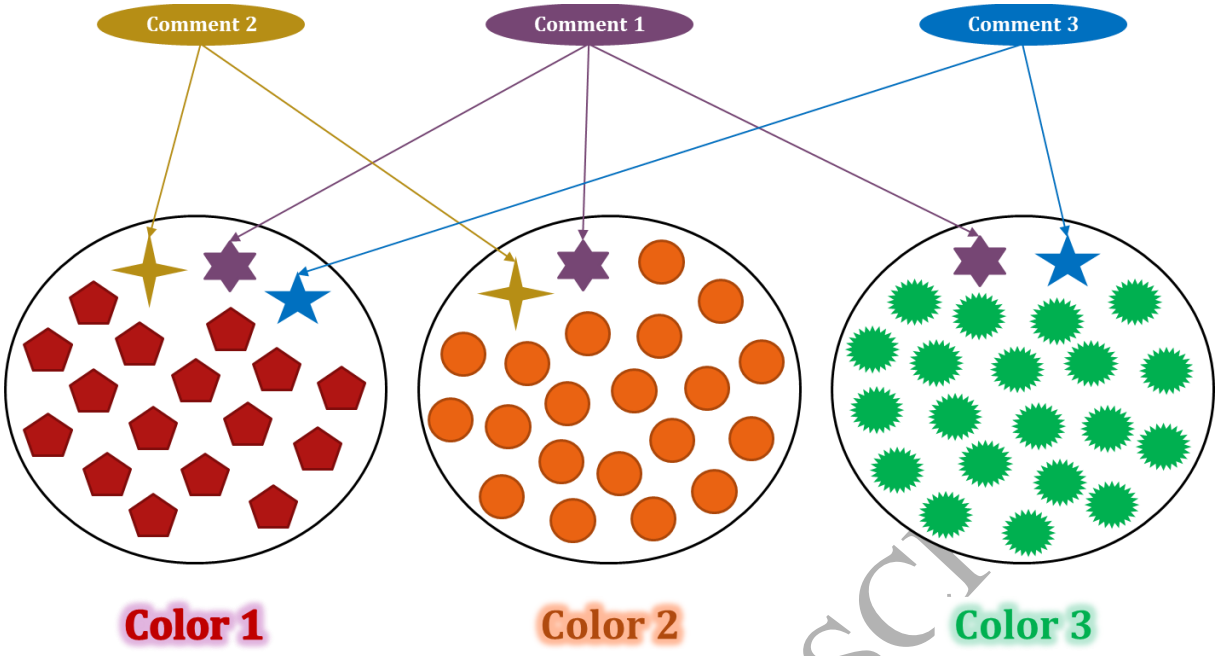

Fig.3: Graph coloring representation for STS

C. Graph Representation of comments in colors

STS problem can be reformulated into an ACO problem explicitly. ACO requires a represented problem in a graph. In this problem, nodes represent the comments and directed edges designate the possible candidate comments. First, inside each color, a semi-graph is constructed. It is worth mentioning that, as shown in fig. 4, to avoid noise and keep our graph less impure, the graph isolates the extremely lengthy comments. It is noticed that the very lengthy comments don't carry any meaning, even if they may be not belong to the put forward topic (El-Fishawy et al., 2014) (Alaa El-Dine and Fatma, 2012).

The following step in each color, is the random ant's travels through the graph where a few comments are visited that satisfies the traversal exit criterion. It can be observed in fig 4, that the ant is currently at the comment C_1. The ant has to choose an adequate comment depending on the heuristic information and transition rule to add to the path. Comment C_4 will be chosen up to reach comment C_5. Later, the generated path $\{C_1, C_4, C_{10}, C_5\}$ is determined after the stop criterion is satisfied. Eventually, after the path is created, the local search LS algorithm is applied to ensure that the collected comments didn't neglect the most interactive comments.

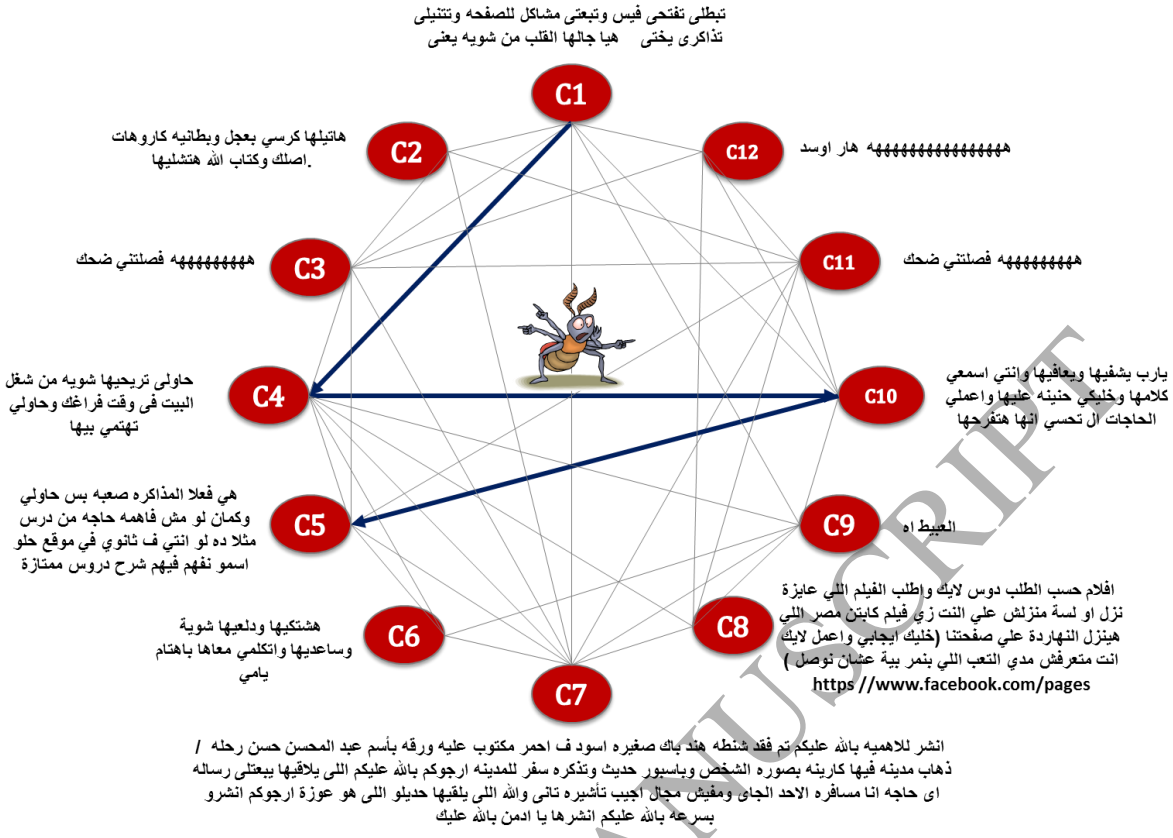

Fig.4: ACO problem representation for STS

Algorithm ACO-LS-STS

Input: C colors

Output: S^N: the most interactive, express subset of comments

1. Define pheromone and heuristic information, set parameters, $Ant_s, \alpha, \beta, \gamma, \theta, \lambda, \omega, \nu, \tau_0, \tau_{max}, \tau_{min}, \rho, Max_i, Min_i$.
2. **For** each color **Do** */*in parallel form*/*
3. **While** $NC \leq Max_i$ **Do**
4. Randomly select number of Ant_n (comments) ∈ S, where Ant_n== Ant_s;
5. start to search a path in graph;
6. **For** each ant **Do** */*in parallel form*/*
7. Construct a candidate list $j_k(i)$ according to Eq. 2;
8. **If** $j_k(i) \neq$ ∅ **Then**
9. Select a comment C_l in a $j_k(i)$ has the highest $\rho_{ij}^k(t)$;
10. Add the selected comment to the partial solution;
11. Delete the comment C_i from candidate list;
12. Record the solutions were generated by the colony in this generation;
13. Get the unselected comments that have the highest priority into $\boldsymbol{CU_i}$ according to Eq. 6;
14. **While** $CU_i \neq \emptyset$ **Do**
15. **For** each comment C_j into solution S_j **Do**
16. **If** C_i priority of $CU_i \geq C_j$ priority of solution S_j
17. Get the Max (SIMI) with C_i of solution S_j into $SIMI_i$;
18. Get the Max (SIMI) with C_j in solution into $SIMI_j$;
19. Calculate the RR of the solution associated with C_i into RR_i;

20. Calculate the RR of the solution associated with C_j into RR_j;
21. **If** $SIMI_i \leq SIMI_j$ & $RR_i \geq RR_j$
22. Insert $C_i$into the solution;
23. Remove C_j from the solution;
24. Evaporate, and update pheromone on the visited edges according to Eq. 7, 8;
25. Record the score of the best path according in this generation as S^N;
26. Update additional pheromone on the visited edges of the best path;
27. **If** $NC \geq Min_i$ **Do**
28. Catch the best solution S^N according to fitness function is shown in Eq. 10;
29. **If** $S^N \geq S^O$ /*where S^O *the old solution* */
30. $S^O = S^N$;
31. **Else If** there is no improvement for 10 times consecutively
32. $NC = Max_i$; /* *the algorithm is terminated* */
33. $NC = NC + 1$;
34. Output the best S^N from all colors;
35. **End**

Algorithm 2. Algorithmic form of ACO-LS-STS

D. ACO-LS-STS for short text summarization

The primary concept of ACO-LS-STS is how to create the optimal solution in each color, where the best one who has the best feature without violation of certain conditions. Algorithm 2 outlines the algorithmic form of ACO-LS-STS. Three main steps are involved in ACO-LS-STS are described below in detail.

1. Heuristic Desirability

Initially, in each color, a solution establishment commences with an empty partial path. The number of ants is set on the number of comments randomly. Each ant applies an arbitrary probability to make an intelligent decision concerning which comment should be visited later. Afterward, the feasible comment which has the highest probability will be a candidate to be selected in accordance with the amount of pheromone and heuristic importance. In line 2 the ACO is applied in parallel form inner each color standalone. First, in lines 3-5, the algorithm selects the random ants. In line 6 the ants search in the solution area so create its path in parallel form. Subsequently, in line 7, the probability of selecting the candidate comments $\rho_{ij}^k(t)$ is shown in eq. 2:

$$\boldsymbol{\rho_{ij}^k(t)} = \begin{Bmatrix} \dfrac{[\boldsymbol{\tau_{ij}(t)}]^{\alpha}.[\boldsymbol{\eta_{ij}}]^{\beta}}{\sum_{\boldsymbol{l \in \mathcal{N}_i^k}} [\boldsymbol{\tau_{il}(t)}]^{\alpha}.[\boldsymbol{\eta_{il}}]^{\beta}} & \textbf{if } \boldsymbol{l \in j_k(i)} \\ \mathbf{0} & \boldsymbol{otherwise} \end{Bmatrix} \quad (2)$$

Where k is an ant appearing at the comment τ_i at iteration t to determine the successor comment τ_j from its neighborhoods of unvisited comments $\mathcal{N}_i^k$.where τ_{ij} is a pheromone trail associated with the edge from C_i to C_j. η_{ij} represents the greedy heuristic information. $\propto$ and β have represented the weights of pheromone and heuristic information. $j_k(i)$ represents the accumulation of possible candidate successor comments of τ_i. Furthermore, In order to invest an effective, numerous heuristic information to this problem, some of the pivotal features are

elected according to (El-Fishawy et al., 2014) (Khabiri et al., 2011) (Weng et al., 2011). The desirable heuristic information η_{ij} is defined as:

$$\eta_{ij} = \frac{\theta . MI(C_i) + \lambda . PR(C_i)}{\upsilon . Max\left[simi(C_i, C_{ps})\right] + \omega . (\sum_{w_i \epsilon C_i} w_i)} \quad (3)$$

Where $MI(C_i)$ is inspired with that comments containing more significant terms are themselves appear plenty. Note that there are numerous ways to select salient terms. An effective approach MI (Bollen et al., 2011) is considered to select the salient terms from the colors. To find the MI of the term t of a resource which is grouped in color C, the MI is defined as follows:

$$MI = \frac{N_{11}}{N} log_2 \frac{N_{11}N}{N_{1.}N_{.1}} + \frac{N_{10}}{N} log_2 \frac{N_{10}N}{N_{1.}N_{.0}} + \frac{N_{01}}{N} log_2 \frac{N_{01}N}{N_{0.}N_{.1}} + \frac{N_{00}}{N} log_2 \frac{N_{00}N}{N_{0.}N_{.0}} \quad (4)$$

Where N shows the number of comments, N_{11} shows the number of comments that contain term t and are in color C, N_{10} shows the number of comments in which contain term t and are not in color C, N_{01} shows the number of comments that don't contain term t and are in color C, and N_{00} shows the number of comments which don't contain term t and aren't in color C.

Regarding the second feature $PR(C_i)$ (Khabiri et al., 2011) (El-Fishawy et al., 2014), is the precedence-based ranking algorithm for C_i. The main target behind this hypothesis is that earlier comments that insert new ideas that are repeated by others after that should be rewarded. These comments may be the source of inspiration for later comments. In contrast, comments that are not never cited are less significant. The PageRank technique is a random walk style across a constructed graph. The PageRank has formulated as:

$$PR(C_i) = \mathrm{d} \sum_{C_j \epsilon N_{C_i}} PR(C_j) \left(W_{out}^{ij} \Big/ \sum_{C_k \epsilon N_{C_j}} W_{out}^{jk} \right) + (1\text{-d}) \quad (5)$$

Where N_{C_i} is the set of neighbors for the comment C_i, and W_{out}^{ji} is the weight of out-link edges from C_i to C_j. d is a damping factor. On the other hand, to make the solution more diverse and shorter, the previous two feature are reduced by the more similar and lengthy comments. The $Max\left[simi\ (C_i\ ,\ C_{ps})\right]$ is the largest similarity between the candidate comment and comments are previously appended into the partial solution. $\sum_{w_i \epsilon C_i} w_i$ is the number of terms of the candidate comment. In addition, the λ, θ, ω and υ represent the weights of different kinds of heuristic information respectively. The heuristic information η_{ij} indicates that ants consider a desirable way to select the comment that has the highest significance and more citation. Meanwhile, has a dissimilar comments and the minimum number of terms as much as possible.

2. Local search

The ACO and LS can supplement with each other. Namely, GC and ACO can provide a promising initial summary for LS, which successively can further optimize the summary obtained by ants effectively. The LS mechanism employing with ACO is straightforward but efficient. In each solution, the high priority unscheduled comments are ordered and try to insert them into the existing solution. If the insertion increases the profit of the summary, then the

summary will be updated. Otherwise, the insertion is invalid and the solution will not be updated. This LS is based on the classical add and move comments.

Our hypothesis on how to determine the priority of the comments was introduced by the fans themselves. Meaning that users give an importance for their comments by like or reply to those comments or mentioning in other comments (#LRM). # LRM is the most important trait which gives a real opportunity for the straightforward outstanding interaction with the fans. Whereas, it is an important feedback evidence that contributes to the advancement the algorithm proficiently, where the priority expresses the number of replies, likes and mentions of a comment. When a comment has numerous likes, replies, and mentions, it is probably more attractor and expresses the view most of the fans. The algorithm exploits these social media feature in a skilled way to count the real number of like, mention, and replay for each comment. Actually, not all granted like, mention and reply have the same importance and equality. Most users are interested in what is written by public, celebrated users who have numerous followers. They have a massive interaction through their messages and comments. Apparently, counting the number of followers is an ingenious way to quantify the importance that fans and what they write have. Thereby, to get an effective exploitation of these features, the factual significance number of likes, replays and mentions are accounted according to the number of followers that the user has as shown below.

$$\boldsymbol{priority}(\boldsymbol{C_i}) = \sum_{\boldsymbol{i}\ \boldsymbol{\epsilon}\ \boldsymbol{\#LRM}} \left[\frac{\boldsymbol{C^i_{\#Followers}}}{\boldsymbol{\#MaxFollowers}}\right]^{\boldsymbol{1}}_{\boldsymbol{0.05}} \qquad (6)$$

Where $C^i_{\#followers}$ is the number of followers of fan i, $\#Maxfollowers$ is the maximum number of followers among all fans associated with a specific post. To normalize the value, the result is restrained within a value range [0.05, 1].

This LS is based on the classical add and move comments. Let CU_i denote the collection of unscheduled comments associated with solution S_i. The procedure of LS is described in lines 13-23, the algorithm initially picks the high priority unselected comments. Later, if the comment C_j conflict with the low priority comment C_i, the maximum similarity will be calculated with the gathered comments into the final solution, and the RR as well twice. One, when C_j is associated with the solution instead of C_i, second with C_i. Eventually, if the comment C_j achieve the higher RR and the lower similarity, the C_i is replaced with C_j. Else, the solution will not be updated.

3. Pheromone evaporation and Update Rules

It is worth mentioning that the seeking behavior of ant algorithms can be characterized by two main traits: exploitation, and exploration. Exploration is the ability of the algorithm to seek out through the solution space lengthily. Whereas exploitation is the ability of the algorithm to search exhaustively in the local neighborhood where good solutions have previously been found. Higher exploitation is indicated in the fast convergence of the algorithm to a suboptimal solution. Whereas higher exploration results in better solutions at a higher computational cost due to the slow convergence of the method. In ant colony algorithm, an adequate trade-off between exploration and exploitation has been adeptly developed (Ezzat, 2013).

Eventually, when all ants have found their paths, the pheromone trails are evaporated. In line 24, to help the ants to forget previous decisions expeditiously, especially bad ones, all visited edges are reduced by a constant factor. Additionally, each ant k deposits a quantity of pheromone $\Delta\tau_{ij}(t)$ on the edges have been visited as shown in eq. 7. The localized pheromone

update rule can be visual as a diversification mechanism which can avoid getting into a local optimum too early.

$$\boldsymbol{\tau_{ij}(t+1)} = \left[(\mathbf{1}-\boldsymbol{\rho}) \,.\, \boldsymbol{\tau_{ij}(t)} + \sum_{k \,\epsilon\, m} \Delta\boldsymbol{\tau_{ij}(t)^k} \right]_{\tau_{min}}^{\tau_{max}} \quad (7)$$

Where ρ ($0 < \rho < 1$)is the pheromone evaporation coefficient. $\Delta\tau_{ij}(t)$ is the amount of pheromone to be deposited by ant k on the edge (i,j). Furthermore, in line 26, a supplemental pheromone deposit on the edges of the best solutions, to enhance the opportunity of picking in the later iterations. Expressly, the purpose of the global pheromone updates is to reinforce pheromone of better solutions in order to enhance the search process more cleverly.

$$\Delta\boldsymbol{\tau_{ij}(t)} = \begin{Bmatrix} \frac{\sum_{j \,\epsilon\, S^k(t)} \boldsymbol{\eta_{ij}}}{\boldsymbol{Q}} & \textbf{if } \boldsymbol{ij \,\epsilon\, S^k(t)} \\ \mathbf{0} & \boldsymbol{otherwise} \end{Bmatrix} \quad (8)$$

Where $S^k(t)$ is the best selected subset by ant k at iteration$t$, j represents the collected comments intos the optimum solution, and Q is the positive constant value. Besides, to prevent pheromone value from accumulating or evaporating too much, the pheromone is restrained within a value range $[\tau_{max},\tau_{min}]$, where τ_{max} and τ_{min} are 1, 0.001 the boundaries of the pheromone trail.

E. Fitness function and exit criteria

The main purpose of exit criteria is to obtain a desirable balance between time consumption and solution fineness, where it avails to guidance the entire optimization procedure. Noteworthy, the maximum and minimum iteration number are adopted here. In line 28 in algorithm 2, the algorithm relies on Trivergence of probability distributions (TPD) as a beneficial fitness function (Cabrera-Diego et al., 2016) (Torres-Moreno, 2015). TPD one of the most attractive methods to evaluate the selected contents automatically. Obviously, this theoretically allows us to calculate the similarity among triplets of objects. TPD is a statistical measure that compares simultaneously three different probability distributions for three probability distributions P, Q, and R.

When the system realizes the number of minimal iterations, the optimal solution is evaluated using TPD to ensure it gets the best subset. The composite trivergence of Kullback-Leibler is defined as in Eq. 9:

$$\boldsymbol{T_c(P||Q||R)} = \sum_{\sigma \in P} \boldsymbol{p_\sigma} \log \frac{\boldsymbol{p_\sigma}}{\left(\sum_{\omega \in Q} \boldsymbol{q_\omega} \log \frac{\boldsymbol{q_\omega}}{\boldsymbol{r_\omega}} / \boldsymbol{N}\right)} \quad (9)$$

Where Q is the source of the original document, R is a summary to evaluate, and P is all the summaries of the set excepting R. ω is the words belonging to Q; q_ω and r_ω are the probabilities of ω to occur in Q and R respectively, σ is the words belonging to P; p_σ represents the probabilities of σ to occur in P. N is the normalization parameter is used to make the inner divergence to make the value smaller and more similar to a probability figure. The N is defined as the sum of the 3 distributions size (|P| + |Q| + |R|). On the other hand, it is possible to have in some distributions unseen elements. For instance, not all the elements in the distribution Q exist in distribution R. Therefore, a smoothing criterion is utilized based on Eq. 10:

$$\boldsymbol{smothing} = \frac{\mathbf{1}}{|\boldsymbol{P}| + |\boldsymbol{Q}| + |\boldsymbol{R}|} \quad (10)$$

Where |P|, |Q| and |R| corresponds to the respective distribution size of P, Q and R.
For instance, as shown in Fig. 5. The distribution of different words in all the comments relevant to an original text and two summaries. The good (green) selected summary has roughly the same distribution of words as for an original text. On the other hand, the red one shows a high trivergence from the original input text and is not considered as a good summary.

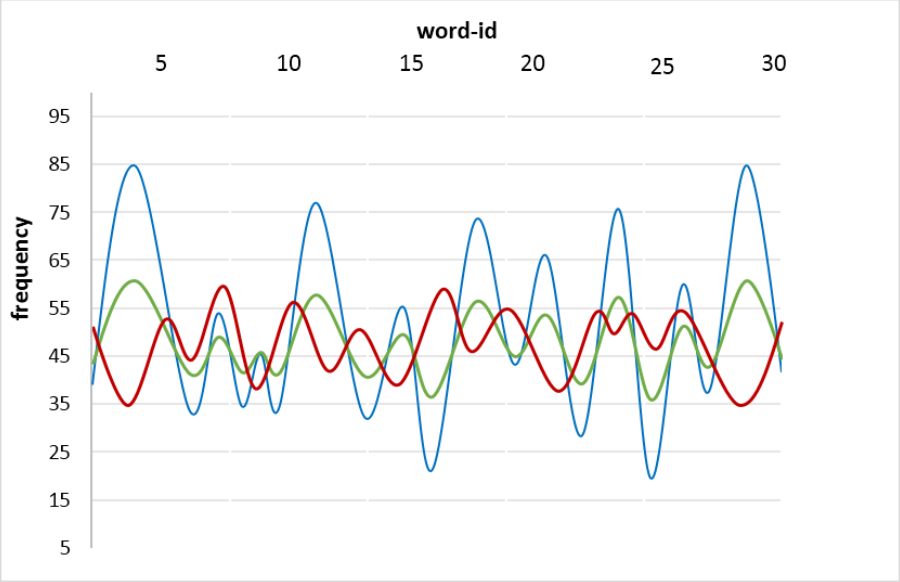

Fig.5: Example of a good and a bad summary for an input text using TPD measure.

Ultimately, when the system convey minimum iteration number and there is no any rising to the best for ten times, the algorithm is terminated to get the optimal solution that has satisfied the lowest TPD.

IV. EXPERIMENT

In this section, the exhaustive experiments are conducted with the short text streams data collected from Facebook to assess the performance of the proposed algorithm and other alternative algorithms. Our evaluation concentrates on answering the following issues. 1) Can GC-ISTS and ACO-LS-STS meet a need of STS on social media? 2) How do the GC and ACO affect the effectiveness of the system overall? First, the ACO-STS algorithm is evaluated. After that, the LS is mixed with ACO. Finally, the combination of these two methods is mixed with the GC that show the promising results. This happens for different perspectives that each method proposes to focus more on the term usage of the comments that have attracted additional attention for many fans.

A. Dataset

We relied on two data sets to be our benchmarks. One is developed by (Alaa El-Dine and Fatma, 2012). It involves an Egyptian Arabic corpus have been collected from a celebrated news Facebook pages; benchmark #1 BM#1. The total corpus size is 27 posts associated with 60 comments for each post. Additionally, the second dataset is developed by ourselves as benchmark #2 BM#2. Among the top 25 Arabic Facebook pages (with the highest number of likes, replies, and mentions) in November 2014, a hundred Facebook messages and their associated comments are collected. The candidate messages have from 500 to 3000 comments. Besides, the average numbers of likes for each message is over 3000, indicating each message is

exceedingly attractive. Finally, the data set is divided into two parts. 40% comments are exploited as a training data. The rest is utilized for testing.

B. Evaluation

Generally, it is so hard task to compare which of the output summaries is optimal among numerous choices of summarization. Predominately, there is not a total agreement among the human judges. So, an automatic assessment of summaries which is based on some heuristics of the fineness of a summary. Although automatic summarization is not as perfect as human assessment, it is just as a complementary method to an emphasis on human judges.

1. Human Evaluation

Next, to answer the later issues, five volunteers are demanded to evaluate the goodness and quality of summary for all algorithms in agreement with some of the standards. To make evaluation easy and possible, the list of comments are partitioned by the number of volunteers out of each message and showed them to our volunteers. We asked them to select the comments which they found interesting, more informative, retains the fresh news about the original comments, dissimilarity, and short enough. Ultimately, thirty comment streams are used for human evaluation.

The quality of algorithms is evaluated by one popular evaluation metric ROUGE that has been adopted since 2004. It is a suite of metrics that measures the similarity between a set of manual summaries and automated one automatically (C.-Y. Lin and E. Hovy., 2003). One of the simplest ROUGE metrics is the ROUGE-N metric:

$$\boldsymbol{ROUGE-N} = \sum_{\boldsymbol{s\in MS}} \sum_{\boldsymbol{n-gram\in S}} \boldsymbol{Match}\,(\boldsymbol{n-gram}) \Big/ \sum_{\boldsymbol{s\in MS}} \sum_{\boldsymbol{n-gram\in S}} \boldsymbol{Count}\,(\boldsymbol{n-gram}) \qquad (11)$$

Where n is the length of the $n-gram$, $Count\,(n-gram)$ is the number of the $n-gram$ in the manual summary, and $Match\,(n-gram)$ is the number of co-occurring $n-gram$ between the automated and manual summaries.

2. Automatic evaluation

On the other hand, the automatic evaluation is employed to measure how well a final summary could capture the essence of the original comments, shorter, and not similar. Three methods are employed to measure the quality of the extracted summary. The first method is retention ratio of information RR inspired by (Cabrera-Diego et al., 2016) (Torres-Moreno, 2015), it is defined as in Eq. 9: where RR shows the distribution of different meaningful terms in all the summary relevant to an original list of comments and other produced summaries.

The second is compression rate of information CR, how many numbers of terms are selected as for the entire list. CR is utilized to ensure the final summary is short enough, it is defined as below.

$$\boldsymbol{CR} = \frac{\boldsymbol{\#TermsInSummary}}{\boldsymbol{\#TermsInListOfComments}} \qquad (12)$$

Where $\#TermsInSummary$ is the number of words in the extracted summary, and $\#TermsInListOfComments$ is a number of words within the entire list of comments. On the other hand, a perfect summary should cover the whole new aspect of the topic lacking redundancy. Thereby, the third method is the measurement of maximum similarity SIMI among

selected comments in the final solution with each other. Namely, it should have the maximal distance between gathered comments with each other as follow:

$$\boldsymbol{SIMI} = \sum_{\boldsymbol{s}\,\boldsymbol{\epsilon}\,\boldsymbol{ps}} \sum_{\boldsymbol{d}\,\boldsymbol{\epsilon}\,\boldsymbol{ps}} \boldsymbol{Max}(\boldsymbol{SIMI_{sd}}) \qquad \boldsymbol{s} \geq \boldsymbol{d} \tag{13}$$

Where s, d are the numbers of comments on the partial path and $SIMI_{sd}$ is the similarity between comments on the path. Besides, there is a trade-off between CR, RR, and SIMI. A desirable summary will have a low CR, similarity and high RR. Thereby, the final accounted automatic formula is defined as shown:

$$\boldsymbol{Automatic - Evaluation} = \frac{\boldsymbol{RR}}{\boldsymbol{CR} + \boldsymbol{SIMI}} \tag{14}$$

3. Alternative Methods

To endorse the performance of the proposed algorithm, several alternative summarization algorithms are considered for comparison, including well-known traditional multi-document summarization algorithms as PageRank (Khabiri et al., 2011), MEAD (Radev et al., 2001) and LexRank (Erkan et al., 2004) graph-based ranking methods. Furthermore, Comparisons with the information theory-based method mutual information (MI), vector space based importance (TF-IDF) are undertaken. In addition, MMR in (Guo S. and Sanner S., 2010), Maximal Marginal Relevance MMR (Ma et al. 2012) and "Arabic summarization" method (Alaa El-Dine and Fatma, 2012). On the other hand, to enhance the task of summarizing for testing and comparison, tow celebrated representative clustering algorithms, hierarchical and K-Means are merged with all earlier methods. In addition, NLP module was utilizes as a preprocessing stage with the previously mentioned algorithms and our proposed algorithm.

C. Experiments with Human and automatic Evaluation

Table 1 presents the obtained results on the two datasets. Regarding the ROUGE-1 an importance of ACO-STS, ACO-LS-STS and GC-ACO-LS-STS in the table 1, it can be observed perceiving fineness about 89.8, 90.2% for ACO-STS in two benchmarks. Besides, when the algorithm inserts an unselected comments which have achieved higher RR and distance via the solution using LS, the performance was increased to 90.7%, 91%. This can be justified by these comments are considered as those of manual summaries. The algorithm ultimately has reached to the peak performance 94.8, 94.6 when the GC module is merged with ACO and LS. That's because the GC was able to shrink the solution area effectively to prevent the ants from a local optimum. The performance of our proposed algorithms had been exhibited higher than other algorithms. It was found that the methods, LexRank, PageRank, MEAD, MT, and TF-IDF are given a mediocre yield correspondingly. They may be not suitable in such a scenario alone. But they begin off improving when mixed with the clustering module, especially with a hierarchical algorithm. Even when the number of clusters is transmuted appropriately, the result is not changed concretely. Furthermore, when TF-IDF and MI are mixed with the clustering algorithms, the results show that MI has better performance than TF-IDF in selecting the farthest away representative comments. This can be justified by the MI focus on terms that contribute the furthermost to a particular cluster. On the other hand, when these algorithms had been merged with each other, the more improvement was achieved.

Algorithm	**ROUGE-1 (C.-Y. Lin and E. Hovy., 2003)**	
	Benchmark #1	**Benchmark #2**
LexRank	73.5%	72.5%

PageRank	72.7%	74.7%
Mead	77.5%	76.5%
TFIDF	67.8%	67.8%
MI	65.2%	65.2%
LexRank + K-means	81.90%	82.20%
PageRank + K-means	82.60%	81.30%
Mead + K-means	78.30%	80.00%
TFIDF + K-means	76.20%	76.20%
MI + K-means	79.40%	78.80%
MMR + K-means	80.70%	80.80%
LexRank + Hierarchical	85.80%	84.70%
PageRank + Hierarchical	84.80%	84.30%
Mead + Hierarchical	82.60%	82.90%
TFIDF + Hierarchical	75.10%	75.90%
MI + Hierarchical	80.20%	80.60%
MMR + Hierarchical	80.90%	81.80%
TFIDF+ Hierarchical+ LexRank	82.70%	84.20%
MI+ Hierarchical+ LexRank	86.00%	86.10%
TFIDF+ Hierarchical+ PageRank	82.80%	83.70%
MI+ Hierarchical+ PageRank	86.10%	87.10%
Arabic summarization	82.40%	81.00%
ACO-STS	89.80%	90.20%
ACO-LS-STS	90.70%	91.00%
GC-ACO-LS-STS	92.10%	92.30%

Table 1: Comparison of algorithms quality based on human Evaluation

On the other hand, when the automatic evaluation is employed according to the Eq. 14 to ensure that our algorithm is consistent and efficient. The automatic evaluation had confirmed that the previous results produced by humans are almost evenly matched. The results are illustrated in the table, 2 with benchmark #1, 2. Evidently, it's noticed that ACO-STS, ACO-LS-STS and GC-ACO-LS-STS in the table 2, show the highest performance in resolving STS problem. The ACO-STS, ACO-LS-STS and GC-ACO-LS-STS had exhibited the highest fineness about (92.80%, 93.30%, 94.80%) and (91.50%, 93.80%, 94.60%) in two benchmarks respectively. That emphasizes that our algorithm was able to get the best summary that had the lowest similarity, the highest retention ratio of information, and shorter enough. Moreover, there is no algorithm has the highest effectiveness constantly except our proposed algorithm. Moreover, the algorithms without combining with a clustering algorithm show the evil fulfilment. That's may justify by the similarity was high.

Algorithm	**TPD evaluation**	
	Benchmark #1	**Benchmark #2**
LexRank	66.30%	73.30%
PageRank	68.50%	74.70%
Mead	72.00%	75.00%
TFIDF	96.10%	70.90%
MI	70.90%	71.50%
LexRank + K-means	84.20%	85.00%
PageRank + K-means	81.50%	83.10%
Mead + K-means	83.00%	85.10%
TFIDF + K-means	84.10%	83.20%
MI + K-means	85.90%	84.70%
MMR + K-means	86.00%	86.00%
LexRank + Hierarchical	85.20%	85.60%
PageRank + Hierarchical	84.20%	84.80%

Mead + Hierarchical	85.10%	85.90%
TFIDF + Hierarchical	85.60%	85.60%
MI + Hierarchical	86.00%	85.70%
MMR + Hierarchical	86.20%	85.90%
TFIDF+ Hierarchical+ LexRank	87.00%	87.00%
MI+ Hierarchical+ LexRank	88.10%	88.60%
TFIDF+ Hierarchical+ PageRank	86.90%	86.70%
MI+ Hierarchical+ PageRank	88.20%	87.70%
Arabic summarization	84.20%	85.80%
ACO-STS	92.80%	91.50%
ACO-LS-STS	93.30%	93.80%
GC-ACO-LS-STS	94.80%	94.60%

Table 2: Comparison of algorithms quality based on automatic Evaluation

Determinately, ACO-STS, ACO-LS-STS and GC-ACO-LS-STS had exhibited the optimum performance in benchmark #1, 2. This can be justified by GC shrink the solution area to small dissimilar sets that protect the ants from falling local optimum. Later, the ACO-LS-STS can concentrate on the best path of the comments by availing the pheromone and heuristic information values, and how the LS is recommended the high priority comments.

D. Parameters effectiveness

In this section, the effects of efficiency caused by the parameters of GC-ISTS and ACO-LS-STS are assessed and analysed. To verify the desired threshold of similarity θ_s and RR ∂_{RR} attributes in which they have been utilized in Algorithm 1. GC-ISTS. The similarity threshold θ_s used for determining how dissimilar the comments are in a color using Jacard method. The link between two comments will be eliminated if it exceeds the θ_s threshold and it is possible to gather in the same color. To construct an effective color, first, the θ_s parameter is changed to obtain the best thresholds that are needed. The highest performance has been found when the value equals 0.1 or the common terms among two comments equal 2. Besides, the RR threshold attribute RR ∂_{RR} gave the best performance when it is 97%.

On the other hand, through the concentrated experimental study, the influence of changing the parameters has been studied empirically using different settings of ten independent executions of the algorithms. The maximum and the minimum number of tour constructions were 100, 20 respectively. Moreover, the number of ants Ant_Size, evaporation rate ρ, initial pheromone trail τ_0, the weights of different kinds of heuristic information ($\lambda, \theta, \omega, \upsilon$) and the weights of pheromone and heuristic information (α, β) are varied among candidate values except one parameter keeps unchanged. The ranges of these parameters are listed in table 3 associated with the maximal performance of them.

Parameter	Ranges	Optimal values
Ant_Size	{8, 10, 12, 15}	10
ρ	{0.03, 0.05, 0.07, 0.09}	0.05
τ_0	{0.01, 0.015, 0.02, 0.03}	0.02
$\lambda, \theta, \omega, \upsilon$	{(0.3, 0.2, 0.3, 0.2), (0.2, 0.3, 0.3, 0.2), (0.3, 0.2, 0.2, 0.3), (0.2, 0.3, 0.2, 0.3)}	(0.2, 0.3, 0.2, 0.3)
α, β	{(0.5, 1), (1, 0.5), (0.5, 1.5), (1.5, 0.5)}	(1.5, 0.5)

Table 3: Parameters ranges and values of ACO-LS-STS

V. CONCLUSION

In this paper, the STS problem was explored in a way that achieves performance above the state-of-the-art. The proposed algorithm STS was developed by reflecting diverse viewpoints of different fans while retaining the key of concepts with high text quality in the summary. The STS was formalized as an optimization problem for the first time. To satisfy the major requirements subject to specific constraints. The problem is formalized as a GC and ACO tasks. Where the list of comments is divided into the number of sets have dissimilar comments using GC, the final summary is generated based on ACO and LS from the best set. Moreover, the TPD mechanism had been employed to enhance the solutions initially obtained by ants. An innovative algorithm based on GC and ACO-LS has been performed and evaluated over a collection of Facebook messages. Eventually, comprehensive experiments via two benchmarks validated that the performance of the algorithm is satisfactory. During the comparisons and analysis of computational results, the algorithm had exhibited a high performance through several tests and had been found more efficient than the other traditional algorithms.

REFERENCES

Aji S, Kaimal R. "Document summarization using positive pointwise mutual information". International Journal of Computer Science & Information Technology. 4(2), 2012, 47.

Alaa El-Dine A, Fatma EL. "Automatic Summarization of Arabic post". the first International Conference for Faculty of Computers and Information (ICCI'12).

Al Hajjar AE, Hajjar M, Zreik K. "A system for evaluation of Arabic root extraction methods". In Internet and Web Applications and Services (ICIW), 2010 Fifth International Conference on IEEE. 2010, 506-512.

AL-DHELAAN, Mohammed. "StarSum: A Simple Star Graph for Multi-document Summarization". In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2015. p. 715-718.

Baccianella S, Esuli A, Sebastiani F. "Multi-facet rating of product reviews". Proc. of the 31st European Conference on IR Research on Advances in Information Retrieval (ECIR'09). 2009, 461–472.

Becker H, Mor N, and Luis G. "Selecting Quality Twitter Content for Events". Proc. of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11). 2011, 442–445.

Bing L, Li P, Liao Y, Lam W, Guo W, and Passonneau, R. "Abstractive multi-document summarization via phrase selection and merging". (2015). arXiv preprint arXiv: 1506.01597.

Bollen J, Mao H, Pepe A. "Modeling public mood and emotion: Twitter Sentiment and Socio-Economic Phenomena". Proc. Of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11). 11, 2011, 450–453.

Cabrera-Diego, L. A., Torres-Moreno, J. M., & Durette, B. (2016, June). Evaluating Multiple Summaries without Human Models: A First Experiment with a Trivergent Model. In International Conference on Applications of Natural Language to Information Systems (pp. 91-101). Springer International Publishing.

CANHASI and Ercan. "Graph-based models for multi-document summarization.Doktora Tezi". Ljubljana Universitesi, Slovenya, 2014.

Chakrabarti D, Punera K. "Event Summarization Using Tweets". Proc of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11). 2011, 66–73.

Chen J., Nairn R., Nelson L., Bernstein M., and Chi E. H. Short and Tweet: Experiments on Recommending Content from Information Streams. Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'10), pages 1185–1194, 2010.

GLAVAŠ, Goran, ŠNAJDER, Jan, "Event graphs for information retrieval and multi-document summarization". Expert systems with applications, 2014, 41.15: 6904-6916.

Daly E. M., Muller M., Millen D. R., and Social Lens L. Gou.: Personalization Around User Defined Collections for Filtering Enterprise Message Streams. Proc. of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11), pages 113– 120, 2011.

Dave K, Lawrence S, Pennock DM. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews". Proc. of the 12th International Conference on World Wide Web (WWW'03). 2003, 519–528.

Diakopoulos N. A. and Shamma D. A. Characterizing Debate Performance via Aggregated Twitter

Sentiment. Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'10), pages 1195–1198, 2010.

Dorigo M, Birattari M, Stutzle T. "Ant colony optimization". IEEE computational intelligence magazine", 1, 2006, 28-39.

Dos Santos M, Luıs C. "Event-based Multi-document Summarization". Diss. Universidade de Lisboa, 2015.

El-Fishawy N, Hamouda A, Attiya GM, Atef M. "Arabic summarization in twitter social network". Ain Shams Engineering Journal. 5(2) (2014): 411-420.

Erkan G, Radev DR. LexRank. "Graph-based lexical centrality as salience in text summarization". Journal of Artificial Intelligence Research. 22, 2004, 457-479.

Ezzat A. "Ant Colony Optimization Approaches for the Sequential Ordering Problem". (Doctoral dissertation, AMERICAN UNIVERSITY IN CAIRO). 2013.

Facebook. http://www.facebook.com/.

Guo S., Sanner S., Probabilistic latent maximal marginal relevance, Proceedings of 698 the 33rd International ACM SIGIR Conference on Research and Development in 699 Information Retrieval, ACM, 2010, pp. 833–834, doi:10.1145/1835449.1835639.

Harabagiu SM, Hickl A, "Relevance Modeling for Microblog Summarization", Proc. of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11). 2011, 514–517.

Hu M, Liu B. "Mining and summarizing customer reviews". Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04). 2004, 168–177.

Hu M, Sun A, Lim EP. "Comments-oriented blog summarization by sentence extraction". Proc. of the 16th ACM International Conference on Information and Knowledge Management (CIKM'07). 2007, 901–904.

Inouye D, Kalita JK. "Comparing twitter summarization algorithms for multiple post summaries". In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on 2011, 298-306.

Khabiri E, Caverlee J, Hsu CF. "Summarizing User-Contributed Comments". Proc. of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11). 2011, 534–537.

Li CT, Wang CY, Tseng CL, Lin SD. "Memetube: A sentiment-based audiovisual system for analyzing and displaying microblog messages". Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HCT'11), 2011, 32–37.

LIN, Chin-Yew; HOVY, Eduard. 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. Association for Computational Linguistics, 2003. p. 71-78.

Liu CY, Chen MS, Tseng CY. "Incrests: Towards real-time incremental short text summarization on comment streams from social network services". IEEE Transactions on Knowledge and Data Engineering. 27(11), 2015, 2986-3000.

Liu F., Liu Y., and Weng F. Why is "SXSW" trending? Exploring Multiple Text Sources for Twitter Topic Summarization. Proc. Of the ACL/HLT Workshop on Language in Social Media (ACL-LSM'11), pages 66–75, 2011.

Ma, Z.; Sun, A.; Yuan, Q.; and Cong, G. 2012. Topic-driven reader comments summarization. In Proceedings of the 21st ACM international conference on Information and knowledge management, 265–274. ACM.

Marcus A, Bernstein MS, Badar O, Karger DR, Madden S, Miller RC. "Twitinfo: aggregating and visualizing microblogs for event exploration". Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'11). 2011, 227–236.

Marujo L, Ling W, Ribeiro R, Gershman A, Carbonell J, de Matos D, and Neto J. "Exploring events and distributed representations of text in multi-document summarization". Knowledge-Based Systems, (2016). 94, 33-42.

Marujo L, Ribeiro R, de Matos D M, Neto J P, Gershman A, and Carbonell J. "Extending a single-document summarizer to multi-document: a hierarchical approach." arXiv preprint arXiv: (2015) 1507.02907.

Michelson M. and Macskassy S. A. Discovering Users' Topics of Interest on Twitter: A First Look. Proc. of the 4th Workshop on Analytics for Noisy Unstructured Text Data (AND'10), pages 73–79, 2010.

O'Connor B., Krieger M., and Ahn D. TweetMotif: Exploratory Search and Topic Summarization for Twitter. Proc. of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10), pages 384–385, 2010.

Pang B, Lee L. "Opinion mining and sentiment analysis". Foundations and trends in information retrieval. 2, 2008, 1–135.

Perez-Tellez F., Pinto D., Cardiff J., and Rosso P. On the Difficulty of Clustering Company Tweets. Proc. of

the 2nd International Workshop on Search and Mining User-Generated Contents (SMUC'10), pages 95–102, 2010.

Radev DR, Blair-Goldensohn S, Zhang Z. "Experiments in single and multi-document summarization using MEAD". Ann Arbor. 2001, 48-109.

Rosa D. K., Shah R., Lin B., Gershman A., and Frederking R. Topical Clustering of Tweets. Proc. of the ACM SIGIR's 3rd Workshop on Social Web Search and Mining (SWSM'11), 2011.

Sankaranarayanan J., Samet H., Teitler B. E., Lieberman M. D., and Sperling J. TwitterStand: News in Tweets. Proc. of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM-GIS'09), pages 42–51, 2009.

Shaalan Kh; BAKR H, ZIEDAN I. "Transferring Egyptian colloquial dialect into modern standard Arabic". In: International Conference on Recent Advances in Natural Language Processing (RANLP–2007), Borovets, Bulgaria. 2007. 525-529.

Sharifi B., Hutton MA, Kalita JK. "Experiments in Microblog Summarization". Proc. of the 2nd IEEE International Conference on Social Computing (SocialCom'10). 2010, 49–56.

Sriram B., Fuhry D., and Demirbas M. Short Text Classification in Twitter to Improve Information Filtering. Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10), pages 841–842, 2010.

Takamura H, Yokono H, Okumura M. "Summarizing a document stream". In European Conference on Information Retrieval. Springer Berlin Heidelberg. 2011, 177-188.

Torres-Moreno, J.M.: Trivergence of Probability Distributions, at Glance. Computing Research Repository (CoRR) abs/1506.06205 (2015).

Tumasjan A., Sprenger T. O., Sandner P. G., and Welpe I. M. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. Proc. of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10), pages 178–185, 2010.

Turney PD, Pantel P. "From frequency to meaning: Vector space models of semantics". Journal of artificial intelligence research. 37(1), 2010, 141-88.

YouTube. http://www.youtube.com/.

Wan X, Yang J. "Multi-document summarization using cluster-based link analysis". Proc. of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. 2008, 299-306.

Weng JY, Yang CL, Chen BN, Wang YK, Lin SD. "Imass: An intelligent microblog analysis and summarization system". Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACLHLT'11). 2011, 133–138.

Zhuang L, Jing F, Zhu XY. "Movie review mining and summarization". Proc. of the 15th ACM International Conference on Information and Knowledge Management (CIKM'06). 2006, 43–50.

Zufferey N., Amstutz P., Giaccari P., Graph coloring approaches for a satellite range scheduling problem, Journal of Scheduling 11 (4) (2008) 263–277.