# Tweet-Biased Summarization

Evi Yulianti[*], Sharin Huspi[**] and Mark Sanderson

School of Computer Science and Information Technology, RMIT University, Melbourne, Australia.

[*] Faculty of Computer Science, Universitas Indonesia, Kampus UI Depok 16424, Jawa Barat, Indonesia.

[**]Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia.

Email: evi.y@cs.ui.ac.id, sharin@utm.my, mark.sanderson@rmit.edu.au

## Abstract

We examined if the micro-blog comments given by people after reading a web document could be exploited to improve the accuracy of a web document summarization system. We examined the effect of social information (i.e. tweets) on the accuracy of the generated summaries by comparing the user preference of TBS (Tweet-Biased Summary) with GS (Generic Summary) in a crowdsourcing-based evaluation. Comparing TBS with two different GS baselines, we found that the user preference for TBS was significantly higher than GS. We also took random samples of the documents to see the performance of summaries in a traditional evaluation using ROUGE, which in general TBS was also shown to be better than GS. We further analysed the influence of the number of tweets pointed to a web document on summarization accuracy, finding a positive moderate correlation between the number of tweets pointed to a web document and the performance of generated TBS as measured by user preference. The results show that incorporating social information into summary generation process can improve the accuracy of summary. The reason of people choosing one summary over another in a crowdsourcing-based evaluation also presented in this paper.

## 1 Introduction

Summarization focuses on generating a condensed version of a document that covers the document's main topic. This benefits people to get an understanding of the document content quickly. As a result, people can make a decision faster whether a certain document is relevant for them. For example: in search engine results, people usually read the result snippet first before deciding to read the full document. A summary is useful in many other areas, such as in customer review of products (Minqing Hu & Liu, 2004), movie reviews (Zhuang, Jing, & Zhu, 2006), medical documents (Afantenos, Karkaletsis, & Stamatopoulos, 2005) and legal documents (Galgani, Compton, & Hoffmann, 2012). Summarization techniques can be applied not only to text documents, but also to speech (Maskey & Hirschberg, 2005) and video (Ma, Hua, Lu, & Zhang, 2005).

There are different approaches to summarization. *Extractive summaries* only contain sentences or sentence fragments from a document to be summarized. In *abstractive summaries*, some further modifications to the sentences can be performed such as: revision, fusion, and compression (Nenkova & McKeown, 2011). A *generic summary* is produced with respect to the content of a document without any additional clues while a *biased summary* is generated with respect to an additional source of information: for example in information retrieval, query-biased summaries are used in search result snippets.

In social media, people often give their comments about a web document that they have read. For example, in a microblog (e.g. Twitter), people post the URL of a web document and comment on it. In general, such comments reflect certain parts of the document that are considered important or interesting. We assumed that this information can give additional clues to choose better sentences in generating a document summary. Most summarization approaches, however, did not consider this social information although some of them also used

additional information obtained from hyperlinks (Delort, Bouchon-Meunier, & Rifqi, 2003), click-through data (Sun et al., 2005), related documents (Wan & Xiao, 2010), etc.

In this work, we used the approach in (Parapar, López-Castro, & Barreiro, 2010) to generate a *tweet-biased summary* (TBS) which adopts the concept of extractive summarization and query-biased summarization. The TBS was then compared with two different *generic summary* (GS) that is only generated using the information from document content. In Parapar et al's work, the approach was applied to the blog domain, while in this work we applied it to the Twitter microblog domain. Blog post comments and tweets have different characteristics in terms of length and location that make this work different with the work performed by Parapar et al (2010), who also evaluated their summaries differently. In this work, we also analysed the influence of number of social media information used to generate a summary on summary accuracy, which has not been done in the previous summarization researches that exploited social media (Meishan Hu, Sun, & Lim, 2008; Yang et al., 2011). We compared the performance of TBS and GS by doing pairwise comparison approach using crowdsourcing service (i.e. CrowdFlower) in which the approach was also used in (Glaser & Schütze, 2012) to evaluate single sentence summaries of product review. In this work, however, we also did traditional ROUGE evaluation using small subset of the data and looked at the agreement between user preference and ROUGE.

This research was conducted to answer the following questions:

Q1. Can social information be used to select more important sentences in generating an extractive summary of a web document?

Q2. How much would the TBS be better or even worse than GS?

Q3. Does the number of tweets pointed to a web document influence the accuracy of TBS (i.e. the higher number of related tweets impacts on the better quality of TBS)?

Q4. What are the aspects that people consider when choosing one summary over another?

This article is organized as follows. In the next section, we review related work followed by a detailing of the research methodology. The results and analysis of the evaluation are presented next, before the work concludes.

# 2 Related Work

## 2.1 The Use of Social Media in IR Applications

Social media was defined by (Kaplan & Haenlein, 2010) as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content. The use of social media has exploded in recent years. The high availability of such information is then used by many researchers. One of the most popular social media platforms is Twitter, a microblogging site created in 2006. According to the Twitter Statistics measured on April 19th, 2013, the number of active Twitter users reached 554,750,000 and there are 58 billion of tweets posted every day ("Twitter Statistics," 2013).

Phelan, McCarthy, and Smyth ( 2009) identified emerging topics of interest in Twitter and used them to recommend news stories taken from an RSS feed. The researchers found that the Twitter-based strategy had higher average click-throughs per user than a content-based strategy. Diakopoulos, De Choudhury, and Naaman (2012) used the information shared by eyewitnesses of certain events in Twitter to find interesting and trustworthy sources for their reports. They performed an exploratory study by asking seven journalists to use the system in finding information that could add the coverage of the event. They found that the journalists' responses were positive towards the system helping them find and rapidly assess the information sources. Asur and Huberman (2010) exploited the rate of related tweets to generate box-office revenue predictions for new release movies. Then, they analyzed the sentiment discussed in tweets after the movie was released to improve the predictions. They showed that their predictions were better than the existing state of the art approach.

Meishan Hu et al. (2008); Parapar et al. (2010) used comments attached to blog posts with the aim of generating an effective summary. Lee and Croft (2013) used blog and forum comments that pointed to web documents to improve retrieval of those documents. They built *query dependent* and *query independent* features from the comments showing such features improved effectiveness.

## 2.2 Generic Summarization

Commonly, a summary is generated by considering only the document content without taking into account additional information. The earliest work of automatic summarization was done by Luhn (1958) using a sentence extraction approach. He argued that the important sentences cover many descriptive words close to each other. The descriptive words were identified based on their frequency of occurrences. Later work was done by Edmundson (1969) using machine learning approach. He defined some features extracted from a document and combined them using a linear combination approach to weight the sentences.

Current research in generic summarization used various approaches, such as: LexRank (Erkan & Radev, 2004), CRF (Shen, Sun, Li, Yang, & Chen, 2007), affinity graph (Wan & Xiao, 2010), and similarity with novelty detection algorithm (Parapar et al., 2010). Erkan and Radev (2004) proposed three different degree-based methods (i.e. degree centrality, LexRank and continuous LexRank) that determine the importance of sentences based on its centrality in a graph representation of sentences. Their methods were then compared with centroid-based methods that measured the importance of the sentence according to how close the sentence was to the centroid of a cluster. They showed that all of their proposed methods outperformed the centroid-based method. In addition, LexRank method with a threshold also outperformed two other degree-based methods. Shen et al. (2007) proposed a method using CRFs (Conditional Random Field) by labeling the sentences in sequence with 1 (if it is included in the summary) and 0 (if it is excluded). The label given to one sentence will influence the label given to its nearby sentences. They showed that their approach could improve accuracy over the best-supervised baseline (i.e. HMM) and unsupervised baseline (i.e. HITS). Wan and Xiao (2010) produced generic summaries by building a within-document affinity graph that represents the relationship of sentences within a document.

## 2.3 Summarization by Using Social Media

Some past researches have studied the use of social media information to generate the summary of a web document. Meishan Hu et al. (2008) summarized blog posts by considering not only its content but also comments left by readers. They built three kinds of graph (topic, quotation, and mention) to model the relationship among comments, which was then combined into a multi-relation graph. They determined the importance of each comment using graph-based and tensor-based scoring. The results showed that the summaries generated by utilizing comments give significant improvements over summaries that do not use comments. Yang et al. (2011) considered how informative sentences were and the interests of social users to summarize the web document. They proposed DWFG (Dual Wing Factor Graph) which uses mutual reinforcement between web documents and their social context information. They found that their approach showed significant improvements over all baseline methods.

Gao et al. (2012) summarized both the news documents and tweets by jointly discovering the representative and complementary information from them. They identified and measured the complementary sentence-tweet pairs using a topic modeling approach. According to the experimental results, the news summary as well as the tweet summary significantly outperformed all baseline summaries in terms of ROUGE score. Parapar et al. (2010) exploited blog comments to guide the sentence selection process when summarizing. They proposed a query-biased summarization approach that only relied on a similarity and novelty detection algorithm to produce the summary. Their experimental approach was to compare the effectiveness (according to MAP score) of retrieval systems searching on the different versions of the summaries: the biased summaries were found to be better.

While the above works described the summarization of web documents using social media information, Liu, Liu, and Weng, (2011) performed the summarization on the social media itself. They employed a concept-based optimization framework and generated a summary of the twitter topic by using multiple text sources, such as: (1). the original tweet, (2). a normalized tweet, (3). webpages pointed by tweets, and (4). combination of tweets and webpages pointed by them. The result showed that the web content can provide extra information to the summary and the best performance was achieved by using the combination of normalized tweets and webpages.

# 3 Research Methodology

## 3.1 Dataset

We used the Twitter dataset from TREC 2011 microblog track that was collected between January 23rd and February 8th, 2011. We used a set of twitter-corpus-tools[1] to download 15,167,481 tweets in total. We extracted some information from each tweet in HTML format, such as: status (i.e. content of the tweet); URL of any web document that is pointed to by the tweet, author, and timestamp. A URL would be used to get the web document that would be summarized, and the related tweets for this web document were obtained from all tweets in the dataset that pointed to the URL of this web document. We removed duplicate tweets by assuming that two tweets written by the same author at the same timestamp are similar tweets, and obtained 14,940,758 unique tweets.

We used the language identifier tool Langdetect[2] to filter tweets written in English. This tool uses a Naïve Bayesian classifier and is claimed to have over 99% precision for 53 languages. We converted the tweets to lowercase; removed stop words using the default English stop words list[3]; removed punctuation and some Twitter symbols, such as: username (@...), hashtag keyword (#...), and retweet (RT); and applied Porter stemming (Porter, 1980). We obtained 5,487,411 English tweets.

We filtered those that contained a URL and obtained 1,244,360 tweets. We implemented a program to convert the URLs in short format (e.g. http://tinyurl.com/46hlnq7) into long format. We found there are some different URLs that are similar in some initial parts, have similar contents. For example:

http://online.wsj.com/article/SB10001424052748703439504576116083514534672.html

http://online.wsj.com/article/SB10001424052748703439504576116083514534672.html?mod=wsj_share_twitter

http://online.wsj.com/article/SB10001424052748703439504576116083514534672.html#mod=djempersonal

According to this condition, we grouped together the URLs that fell into this case (using character "?" and "#" as an indicator) and sorted the groups in descending order based on the total number of tweets pointed to them. We scanned 115,716 URLs in the sorted list to get the URLs pointed to by a minimum of ten tweets. We then examined manually the target page of the URLs to determine valid articles to be summarized. We used the following heuristics in that determination: the URL had the title or ID of article; if the URL was a homepage, only a picture, advertisement, or pointed to a page with a small amount of text; then it was excluded. This left us with 493 URLs. The distribution of the number of tweets pointed to is plotted in logarithmic scale in Figure 1.

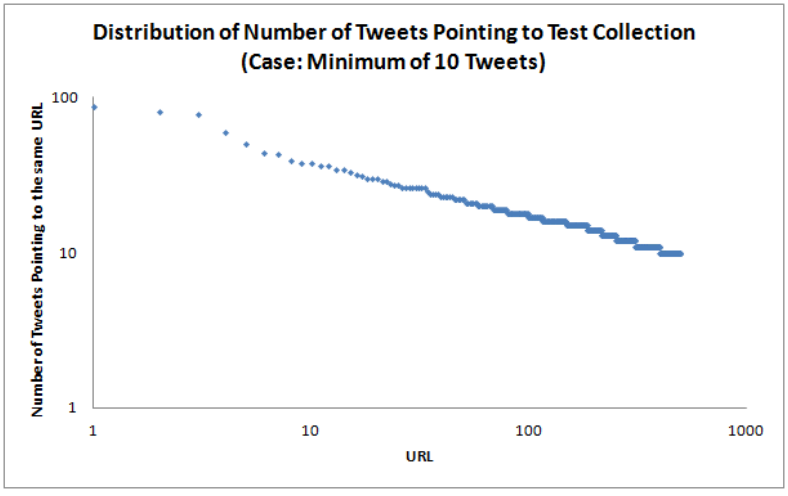


***Figure 1.*** **Distribution of Number of Tweets pointed to the URLs in the Test Collection.**

[1] https://github.com/lintool/twitter-tools
[2] http://code.google.com/p/language-detection/
[3] http://www.ranks.nl/resources/stopwords.html

We also used web documents pointed by less than ten tweets. Because there are many URLs that are pointed by a small number of tweets, we did not conduct the manual process like before. We instead learned the domains of URLs that appear frequently in the result of 493 URLs. If the URLs in the list matched with our defined domains, then we included them in our results. This process generated 24,410 URLs. We performed random sampling to choose fifty documents from each tweet number (1-9) and obtained 450 URLs. Combining this result with the previous result, we had 943 URLs in our test collection.

We used the FortiGuard[4] website to assign a category to each web page. This was based on the dominant content of web page in the URL[5]. Most URLs in our test collection were found to come from the "News and Media" category. Such website claimed that its rating database contains over 26 million websites and over several billion rated web pages that has been assigned to 78 categories (Fortinet, 2007). Table 1 describes Top 5 category of URLs in our test collection.

**Table 1. Top 5 URL Categories in the Test Collection**

| Category | #URL | Example of URL |
|---|---|---|
| News and Media | 730 | http://www.bbc.co.uk/news/uk-northern-ireland-12377862 |
| Information Technology | 108 | http://searchengineland.com/google-bing-is-cheating-copying-our-search-results-62914 |
| Reference | 27 | http://www.helium.com/items/1815603-bgp-border-gateway-protocol |
| Personal Websites and Blogs | 20 | http://googleblog.blogspot.com.au/2011/02/microsofts-bing-uses-google-search.html |
| Sports | 15 | http://sports.espn.go.com/chicago/nfl/news/story?id=6056888 |

This result was matched with our results after grouping each URL in the test collection according to its domain. The top five domains are described in Table 2. We can see that all domains in the following table are news websites.

**Table 2. Top 5 URL Domains in the Test Collection**

| Domain | #URL |
|---|---|
| mashable.com | 259 |
| cnn.com | 104 |
| bbc.co.uk | 77 |
| techcrunch.com | 54 |
| huffingtonpost.com | 53 |

We crawled the web documents according to the URLs in our test collection by using HTML parser JSoup[6]. Each web document was split into sentences using a Java library from MorphAdorner[7]. Finally, we gathered the tweets pointing to the same URL in test collection as one group and made this as the related tweets for the web document located in this URL. After doing all the above steps to preprocess the dataset, our test collection consisted of 943 web documents and their related tweets (i.e. 9,658 tweets in total). Those data would be inputted to our summarization systems.

[4] http://www.fortiguard.com/ip_rep.php
[5] http://www.fortiguard.com/static/webfiltering.html
[6] http://jsoup.org/
[7] http://morphadorner.northwestern.edu/morphadorner/download/

## 3.2 Tweet-Biased Summarization System

In this work, we generated TBS (Tweet-Biased Summary) using the summarization approach described in (Parapar et al., 2010). In general, there are two main processes in this approach that are based on a query-biased summarization concept: (1) **Generate the query** that would bias the summary and (2) **Run the query** to generate the summary. In original approach, Parapar et al. (2010) defined the length of summary to be 30% of the size of blog post. However, in this work we do not use that definition by considering that using a percentage as a length constraint will make a length of summary depends on a length of original document. So, we referred to the summary length imposed in TAC (Text Analysis Conference) on 2008 until 2011 and also the length used by (Wan & Xiao, 2010) to limit the summary in 100 words.

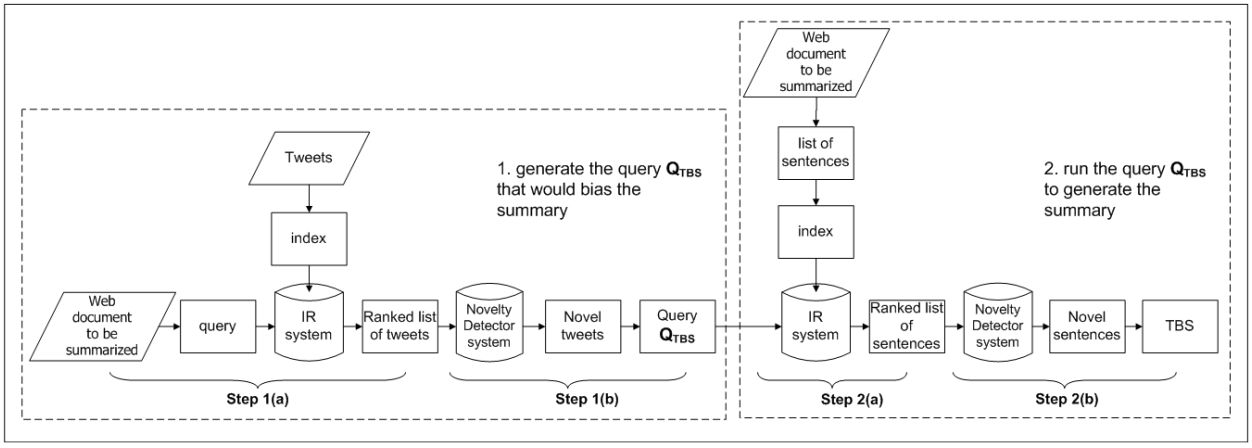


***Figure 2*. Framework of Tweet-Biased Summarization System.**

Figure 2 illustrates the framework of tweet-biased summarization system that can be explained as the following:

1. Generate the query $Q_{TBS}$ that would bias the summary:
   a) **Rank the tweets according to the relevance with the entire web document.** Index the tweets linking to the web document. Tokenize the web document, remove stop words & symbols, apply Porter stemming, and combine the results as a query. Rank the tweets based on their similarity to the query.
   b) **Select the most novel tweets to form a query $Q_{TBS}$.** Input the ranked list of tweets to the novelty detector system to select the 30% most novel tweets. This process was performed to remove redundant tweets. The resulting novel tweets were combined to form a query $Q_{TBS}$.
2. Run the query $Q_{TBS}$ to generate a summary:
   a) **Sort the sentences of the web document according the relevance with a query $Q_{TBS}$.** Split the web document into sentences, remove stop words & symbols, apply Porter stemming and index the sentences into an inverted index. The title of the document was not indexed because the system only choose the sentences from the body of the document. Then run the query $Q_{TBS}$ to generate a ranked list of sentences.
   b) **Select the most novel sentences to form a TBS containing 100 words.** Input the ranked list of sentences to the novelty detector system to re-rank the sentences according to the novelty score. Select the most novel sentences until 100 words are produced. The sentences are then sorted in document position order. If the last sentence contains more words, it was truncated so that the length of the summary was exactly 100 words.

We can see from the above steps that step 1 and step 2 used the same process: sort according to the similarity score and re-rank according to the novelty score. Note that the novelty detection applied in the step 1 and step 2 has different purposes. In step 1, it was applied to the ranked list of tweets; in step 2, it was applied to the ranked list of sentences.

### 3.2.1 IR System

This is one of main component in our summarization system to index each related tweets for a web document (in step 1(a)) and each sentences of a web document (in step 2(a)), then sort them according to a given query. We used the IR toolkit Indri (v5.4[8]) developed at University of Massachusetts. Indri provides full search engine functionality that is based on a combination of language modeling and inference network retrieval framework (Strohman, Metzler, Turtle, & Croft, 2005). To index the tweets in step 1(a) of the TBS system, we put all tweets pointing to the same web documents into one corpus file, adopting a "trectext" format.

### 3.2.2 Novelty Detector System

We implemented a cosine distance algorithm that was also used in (Allan, Wade, & Bolivar, 2003; Parapar et al., 2010) to filter redundant content. Each sentence/tweet was represented an *m*-dimensional vector where *m* was the number of unique terms in the collection. The novelty score was defined by computing the cosine similarity between a current vector and the top 10% of retrieved sentences/tweets; following work from Allan et al (Allan et al., 2003). In the summarization system framework, this component re-ranked the result given by IR system according to the novelty score. In the case that two sentences have a similar novelty score, then they would be re-ranked according to the relevancy score.

$$N_{cd}(s_i|s_1,\dots,s_{i-1}) = \min_{1\le j\le i-1} N_{cd}(s_i|s_j)$$

$$N_{cd}(s_i|s_j) = -\frac{\sum_{k=1}^{m} w_k(s_i)\ w_k(s_j)}{\sqrt{\sum_{k=1}^{m} w_k(s_i)^2\ \sum_{k=1}^{m} w_k(s_j)^2}}$$

In the above equation, $w_k(s_i)$ is the weight of word $w_k$ in sentence $s_i$ that was computed using the following TF.IDF formula in (Allan et al., 2003):

$$w_k(s_i) = \frac{tf_{w_k,s_i}}{tf_{w_k,s_i} + 0.5 + (1.5 * \frac{len(s_i)}{asl})} \cdot \frac{\log\frac{n+0.5}{sf_{w_k}}}{\log(n+1)}$$

*n* is the number of presumed relevant sentences, $tf_{wk,si}$ is the number of word $w_k$ in sentence $s_i$, $len(s_i)$ is the number of words in sentence $s_i$, *asl* is the average number of words in presumed relevant sentences, and $sf_{wk}$ is the number of presumed relevant sentences that contain word $w_k$.

## 3.3 Generic Summarization System

We have implemented two different generic summarization systems to generate baseline summaries (i.e. $GS_{sn}$ and $GS_{ag}$) to be compared with TBS. These systems only use the information from document content in the summary generation process. Both GS were generated in the same length with TBS.

### 3.3.1 $GS_{sn}$ System

This system was built using a similar approach explained in the section 3.2 to generate TBS, except the query that would bias the summary was obtained from the document content. As our TBS are formed using Parapar et al's approach, we also use their approach for generating generic summaries. Parapar et al. (2010) state that to build their blog post summarization system they "*…followed exactly the same steps [as the TBS] but the post text itself was used … to guide the sentence selection process*". Therefore in our system, the only difference between the GS system framework and TBS framework is in the step 1(a) of the query generation process, where instead of tweets being indexed, sentences from the document to be summarized are indexed. In the first step, the ranking of the sentences was aimed to get the subset of sentences that best represent the document, which further they were combined to form a query $Q_{GS}$ (that will bias the summary). In the second step, the sentences were again ranked to get the sentences that are most relevant with $Q_{GS}$. The summary generated from

[8] http://www.lemurproject.org/indri/

this system was called $GS_{sn}$ ('sn' is an abbreviation for Similarity and Novelty detection as the basic building block of this system). Note, by following Parapar et al's methodology, we are aware that the novelty detection on sentences was run twice. It was decided that keeping to Parapar et al's methodology was important so as to replicate the past results in the new context. Applying novelty detection twice does not harm the accuracy of the summarizer.

### 3.3.2 $GS_{ag}$ system

This system was built using the Affinity Graph algorithm (Wan & Xiao, 2010) which represents each sentence in the document as a single node and the similarity between two sentences as a link between nodes. This method was also used by Wan and Xiao (2010) as their baseline for their proposed approach that considered the additional knowledge from neighbor documents. We created a sentence-sentence pair matrix for all sentences in the document that represents the importance of each sentence in the within-document affinity graph. The matrix is then normalized and used to calculate the informativeness score (*IF_score*) of each sentence of the document. The sentences with the high informativeness score were then chosen to produce a generic summary. The summary generated from this system was called $GS_{ag}$ ('*ag*' is an abbreviation for Affinity Graph).

## 3.4 Experiment Results and Analysis

### 3.4.1 Experiment Design

We applied pairwise comparison using a crowdsourcing platform to ask people to choose which summary was the best between TBS and GS. Some previous researches also used a crowdsourcing service to help them collecting the data for evaluation purposes. Sanderson et al. (2010) performed pairwise comparison by asking people to choose better search results from two different search engines. Yang et al. (2011) asked people to select some important sentences from document and tweets in order to build gold standard summary. Diakopoulos et al. (2012) asked people to assign a label (i.e. "eyewitness" or not) into each tweet in their dataset. Glaser & Schütze (2012) applied comparative approach using crowdsourcing service to evaluate their summaries (i.e. single sentence) of product review.

Based on the success of Glaser & Schütze (2012), we asked users to make a binary selection on which summary they preferred. We chose binary selection because it would be easier for people to "choose A or B" rather than to "give score to each summary in a pair" or "create gold-standard summaries". Scoring the summary is likely to produce bias as two different people might have different standard in giving the score. Creating gold-standard summary using crowdsourcing service was considered as difficult and time-consuming job, implies the quality of result negatively (Lloret et al. 2013). Lloret et al (2013) obtained that the fast and easy to perform task can help to get a better result.

In this work, we used the crowdsourcing platform CrowdFlower[9]. For each question, we presented the original document and a pair of summaries. We instructed the workers read the original document carefully and select the best summary that represents it and give the reasons for their selections. The screenshot of question in CrowdFlower is illustrated in Figure 3.

[9] https://crowdower.com/

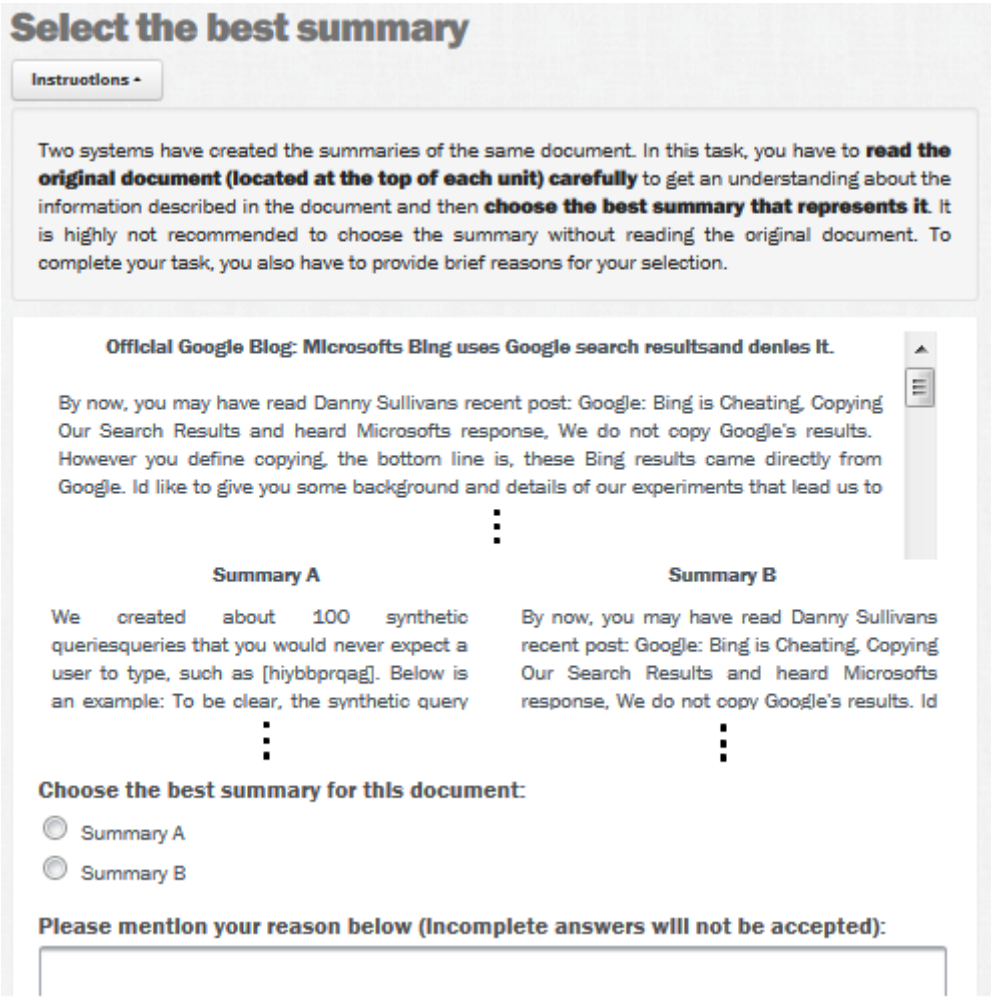


***Figure 3*. Screenshot of Question Presented to the Contributors Containing the Instruction, Original Document, Pair of Summaries, Answer Buttons, and Text Box for Their Reason of Selection.**

We applied a quality control by creating gold questions that displayed a pair of summaries in which one of them was generated from a different document. If people did not understand the task well, they would not get a high accuracy in answering question. If the accuracy of contributors on gold questions dropped below 70%, their answers were not included in the results. CrowdFlower presented gold questions in each page of five questions and randomized their presentation. Each question was answered by minimum of five contributors. For each question, we put TBS and GS in a random order to avoid any bias to the left or right position. We piloted the test to see how the experiments would perform and made sure that the contributors can understand the task. Contributors were paid 15¢ per page; total cost spent in our evaluation was $479. Average gold question accuracy of trusted contributor was 94%.

### 3.4.2 User Preference of TBS vs GS

As we have two different baselines, then we run the experiment using CrowdFlower in two different settings: *TBS vs* $GS_{sn}$ and *TBS vs* $GS_{ag}$. This experiment was held using the summaries of web documents pointed by minimum of 10 tweets in the test collection, which were 493 summaries respectively for TBS, $GS_{sn}$, and $GS_{ag}$.

Firstly, we compared TBS with $GS_{sn}$. We removed some pairs of summaries because the length of the summary is less than our defined summary length or summaries in a pair is exactly similar. Examining 465 pairs of summaries, TBS was preferred over $GS_{sn}$, in percentage terms 169.05%. Secondly, we compared TBS with

$GS_{ag}$. Using the same exclusion criteria, this experiment was held using 440 pairs of summaries, resulted that TBS was also preferred over $GS_{ag}$, in percentage terms 163.64%. We found a $\chi^2$ test for both settings resulted in $p<0.0001$: the difference was statistically significant. This result was summarized in Table 3.

**Table 3. User Preference of the Best Summary**

| | TBS vs $GS_{sn}$ | TBS vs $GS_{ag}$ |
|---|---|---|
| **#TBS** | 339 | 319 |
| **#GS** | 126 | 121 |
| **Total** | 465 | 440 |

The result of both experiment settings shows that in the majority of summaries, people preferred TBS over GS, means that the social media information (i.e. tweet) could benefit the sentence selection process in generating a summary. The result also showed that the outperforming level of TBS over $GS_{ag}$ is getting lower compared with $GS_{sn}$ which gives a clue that $GS_{ag}$ was better than $GS_{sn}$. This presumption was supported by the number of user agreement for TBS in each summary pair described in Table 4. This table shows the result for overlapped summaries examined in both settings: 433 results. When TBS was compared with $GS_{ag}$, the number of high agreement for TBS is getting lower and the number of low agreement for TBS is getting higher than when it was compared with $GS_{sn}$. The 100% agreement means that all people who performed the job in CrowdFlower chose TBS as the best summary, on the contrary the 0% agreement means that all people chose GS as the best summary.

**Table 4. User Agreement for TBS**

| Agreement for TBS | TBS vs $GS_{sn}$ | TBS vs $GS_{ag}$ |
|---|---|---|
| 100% | 84 | 64 |
| 80% - < 100% | 138 | 112 |
| 60% - < 80% | 97 | 134 |
| 40% - < 60% | 60 | 71 |
| 20% - < 40% | 39 | 32 |
| 0% - < 20% | 15 | 20 |
| **Total** | 433 | 433 |

We then split the results according to some web categories and web domains in order to see the user preference for TBS and GS across different subsets. In general, TBS is performing better than GS for each subset of documents. This is illustrated in Table 5.

**Table 5. User Preference across Top 3 Web Categories and Web Domains**

| | | #TBS | #$GS_{sn}$ | #TBS | #$GS_{ag}$ |
|---|---|---|---|---|---|
| **Web Categories** | **News and Media** | 222 | 79 | 215 | 86 |
| | **Information Technology** | 60 | 19 | 60 | 19 |
| | **Reference** | 7 | 2 | 8 | 1 |
| **Web Domains** | **Mashable.com** | 106 | 46 | 114 | 38 |
| | **Cnn.com** | 31 | 6 | 26 | 11 |
| | **Bbc.co.uk** | 13 | 4 | 13 | 4 |

### 3.4.3 ROUGE evaluation

Measurements using preference only determine if one summary is better than another. Such an evaluation provides no information on whether the summaries are of any value. Therefore, we also evaluated our summaries using ROUGE[10]. To compute this measure, we chose randomly from our dataset a number of web documents that were pointed by minimum of 10 tweets. Only web documents with a higher number of related tweets were considered because we were interested to see the agreement between user preference and traditional evaluation measure for documents having stronger effects of social information, rather than documents that were only pointed by one or two tweets. For this purpose, there were 55 manual summaries created by 21 postgraduate students from different discipline in Melbourne (i.e. RMIT University, Melbourne University, and Victoria University). They were asked to select *n* sentences from a web document to generate a summary and the total length of the summary could not be greater than 100 words. We assigned each person to summarize 5 documents, where each of the documents has two summaries created by different people. We obtained the average term-level Kappa ratio between two different human summaries is 0.33. This agreement is comparable with previous studies, which reported a Kappa is about 0.35 and 0.39 respectively for manual summaries of news reports and columns (Hori, Hirao, & Isozaki, 2004) and 0.38 for manually annotated answer passages (Keikha, Park, & Croft, 2014).

Measuring the summaries using ROUGE, (see Table 6), it can be seen that the summary accuracy is similar to scores obtained in other summarization exercises, such as DUC. We also tried to compare the summarization systems using ROUGE. Here, TBS was found to be more accurate than both $GS_{sn}$ and $GS_{ag}$. The outperforming level of TBS over $GS_{sn}$ was better than $GS_{ag}$, indicates that $GS_{ag}$ is more powerful baseline than $GS_{sn}$. This result agrees with the result measured by user preference. The paired t-test calculation, however, shows that only the difference between TBS and $GS_{sn}$ was significant (for all ROUGE scores).

**Table 6. ROUGE score for TBS and GS**

| | **TBS** | **$GS_{sn}$** | **$GS_{ag}$** |
|---|---|---|---|
| **ROUGE-1** | 0.55544 | 0.47984 | 0.53173 |
| **ROUGE-2** | 0.37790 | 0.27248 | 0.33734 |
| **ROUGE-L** | 0.52696 | 0.44796 | 0.49877 |
| **ROUGE-SU4** | 0.39625 | 0.30260 | 0.35344 |

Because the difference of all ROUGE scores between TBS and $GS_{ag}$ was not significant, we conducted a post-hoc power analysis (Faul, Erdfelder, Lang, & Buchner, 2007) to see how likely it was that a Type II (false negative) error occurred. The greater the power, the lower the chance of Type II error to occur. We calculated the power values for our ROUGE results comparing TBS and $GS_{ag}$ using a sample size $n$=55, $\alpha$-*error rate*=0.05 and effect size $d$=0.26 (considered as small effect), where the Effect Size calculation for the t-test is based on (Cohen, 1988). We obtained the power is 0.47, which is less than the suggested power (i.e. 0.8) to detect a statistical effect defined in (Cohen, 1988). As the power is low, then the probability of finding a difference between TBS and $GS_{ag}$ (when it actually exists) is also small as it is influenced by the high probability of Type II error. The post-hoc power analysis showed that with effect size $d$=0.26, we would need approximately 119 sample sizes to achieve the suggested statistical power of 0.8.

We also counted the number of summary pairs where both user preference and ROUGE score agreed, displayed in Table 7. The user preference was said to agree with the ROUGE score if the summary that has higher user preference also has higher ROUGE score. Referring to this definition, we did not count the agreement for summary pairs where the ROUGE result is similar and where the summary does not have user preference result (because of exclusion criteria specified in section 3.4.2). For both settings, we found that the number of agreements between user preference and ROUGE was higher than the number of disagreements.

[10] The parameter setting used to run ROUGE: -a -n2 -m -2 4 -u -c95 -r1000 -fA -p0.5 -t0 -l100

**Table 7. The Agreement of User Preference with the ROUGE Evaluation Measure**

| | TBS vs $GS_{sn}$ | | TBS vs $GS_{ag}$ | |
|---|---|---|---|---|
| | **Agree** | **Disagree** | **Agree** | **Disagree** |
| **ROUGE-1** | 38 | 12 | 25 | 21 |
| **ROUGE-2** | 41 | 9 | 29 | 18 |
| **ROUGE-L** | 37 | 13 | 28 | 20 |
| **ROUGE-SU4** | 39 | 11 | 29 | 19 |

We contend that ROUGE and preference measures can be used together to assess the accuracy of summaries. ROUGE computed over a small number of summaries can establish an overall accuracy of a summary system and the crowdsourced user preference scheme can determine fine grained differences between pairs of summarizers.

### 3.4.4 Number of tweet vs User Preference

We performed another experiment using CrowdFlower for the summaries of web documents with fewer than ten tweets pointing to them, to examine accurately the correlation between number of tweet and user preference. We run this experiment using first setting (i.e. TBS vs $GS_{sn}$) because it was better agree with ROUGE compared with the second setting. The similar exclusion criteria specified in section 3.4.2 was applied. The results showed that TBS (289) still outperformed $GS_{sn}$ (130) eventhough the level of improvement decreased: TBS was preferred over $GS_{sn}$ by 122.31%. A $\chi^2$ test resulted in $p<0.0001$. Combining this result with the result given in section 3.4.2, we obtained the user preference for TBS (628) is clearly higher than GS (256) in which TBS outperforms GS by 145.31%. Note that regarding this result, we did not include in the evaluation 15 pairs of summaries in which the content of TBS and GS is exactly the same. Assuming we included them in the evaluation, meaning that TBS and GS respectively has similar user preference for those summary pairs, this level would decrease become 137.27%.

Now, we had a set of results $S = \{R_1, R_2, \ldots R_{884}\}$ where $R_i$ indicates the result for the $i^{th}$ web document. Each result ($R_i$) consists of number of tweets pointed to the original web document (#Tweets), number of people choosing TBS (#TBS), and number of people choosing GS (#GS). We sorted S according to #Tweets in ascending order, and then computed: (1) a rolling average of #Tweets; and (2) a rolling average of (#TBS - #GS) that indicates the user preference for TBS over GS. In such computation, we defined that each subset contains 200 elements. We plotted the result in the following scatterplot to illustrate the relationship between #Tweets and user preference for TBS over GS:

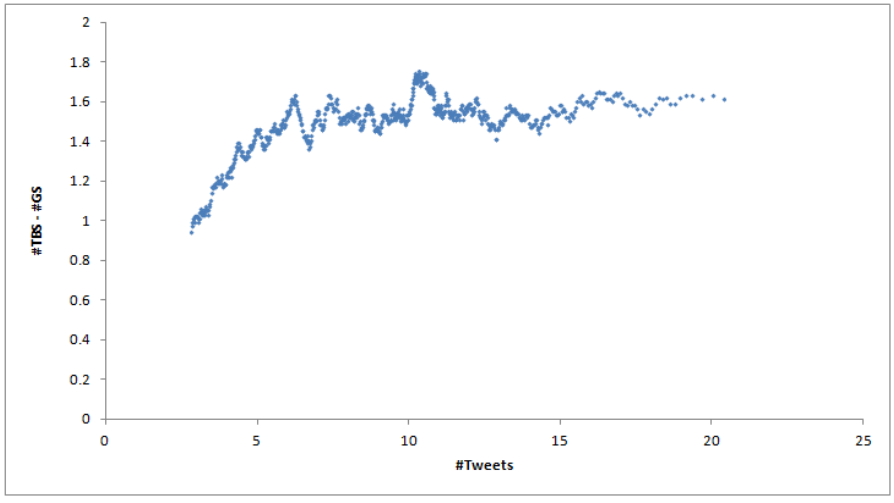


***Figure 4*. Relationship between the Number of Tweets pointed to Web Document and the User Preference for TBS over GS.**

We can see that the user preference for TBS over GS gets stronger as the number of tweets pointed to web documents increases. The highest average user preference is 1.75 that was obtained when the average tweets is 10.32. When the average tweets is 11 and higher, the user preference looks more stable but still shows slightly increasing trend. We computed a Pearson correlation between #Tweets and (#TBS - #GS) and obtained the correlation coefficient $r=0.64$, $p<0.01$ that shows there is a positive moderate correlation (Montcalm & Royse, 2002) between those two variables.

## 3.5 Analysis of Comments

We analyzed 4,403 comments from contributors in CrowdFlower that explains their reasons to choose one summary over another. We adopted an inductive approach (Thomas, 2006) to analyze the qualitative data which starts from building specific categories and ends up with producing general categories. We categorized these comments to get an overview of the aspects that influence contributors to prefer a particular summary. First, we read in detail the comments and labeled them to create a category. One comment can be assigned to more than one category. For example: comment "*stays on topic, more detailed*" was assigned to category "*more on topic*" and "*more detail*". There were some comments removed from our analysis as they did not contain a meaningful reason (e.g. sequence of random alphabets which does not have any meaning, copy of text from the original document / summary displayed). We obtained 28 categories where the statistic was displayed in the Table 8. Mostly, the reasons stated in the comment were related with the quality of summary content.

**Table 8. The Statistics of Initial Categories of Comment**

| Category | #Comment | Category | #Comment |
|---|---|---|---|
| 1. good quality | 1180 | 15. contains main point of document | 19 |
| 2. more detail | 528 | 16. gives better insight / overview | 16 |
| 3. more representative | 423 | 17. mentions name | 14 |
| 4. more related | 308 | 18. better flow | 11 |
| 5. contains more information | 263 | 19. easy to understand | 9 |
| 6. better written/presentation | 166 | 20. clearer | 9 |
| 7. more on topic | 135 | 21. more coherent | 9 |
| 8. more relevant | 99 | 22. more interesting | 8 |
| 9. more comprehensive | 94 | 23. more compact / tight | 8 |
| 10. almost the same | 58 | 24. gives example | 7 |
| 11. good introduction / backstory | 25 | 25. longer | 6 |
| 12. gives fact | 25 | 26. more sense / logical | 5 |
| 13. to the point | 22 | 27. contains information from main title | 4 |
| 14. less of useless information | 20 | 28. mentions actual quote | 3 |

In the next step, we reduced the redundancy among categories displayed in Table 8 by combining some overlap categories into single category and produced 14 categories. For example, category "*more on topic*", "*more representative*", "*give better insight*", and "*contains main point of document*" were merged into single category "*on topic*" because they mainly explained that the summary contains the document topic. We continued to group some specific categories into a broader category and produced three categories that are described in Table 9. Mainly category "*content*" was a merging of the following categories in Table 8: 1, 2, 3, 4, 5, 7, 8, 9, 12, 14, 15, 16, 17, 22, 24, 25, 27, 28; category "*writing/presentation*" was a merging of category 6 & 23; and category "*flow*" was a merging of category 11, 13, 18, 19, 20, 21, 26.

**Table 9. Final Categories of Comments**

| Category | Description | Sample of Quotation | #Comment |
|---|---|---|---|
| Content | This aspect related to having a good quality, on topic (discusses the point in original document), comprehensive, detail, relevant, factual, mentions some important text (e.g: quote, name, example, title, etc), interesting, and longer. | - "*Explains 3d better*"<br>- "*Summary A talks about adoption? Lol not relevant. Summary B is more on target.*"<br>- "*Summary A talks about tweetdeck, but summary B is better because it talks not only about Tweetdeck but tweetdeck.ly and the release of it.*"<br>- "*A is more on topic*"<br>- "*A gives more details and names of software, media, apps involved.*" | 3152 |
| Writing/ Presentation | This aspect is related to having a better writing style and compact. | - "*Gives the same info with less words*"<br>- "*Summary B is a more well-rounded summary of the article.*"<br>- "*A more compact summary is given in A. Summary B goes on at length and does not cover the basic facts simply*."<br>- "*It's presentation is better than B*" | 174 |
| Flow | This aspect is related to having a good introduction/backstory, good flow and to the point of story that makes the summary easy to understand. | - "*Summary A has a better structure.*"<br>- "*Summary A is the better description. Summary B is out of order and messy.*"<br>- "*I prefer the introduction as it helps the reader to understand what is happening*"<br>- "*A tells the back story while be jumps into the middle.*"<br>- "*It is more easy to comprehend and follow than A*" | 90 |

# 4 Conclusion

In this work, we utilized the information from social media to guide the sentence selection process of a summarization system in order to extract more important sentences from a web document. We adopted a query-biased summarization concept to generate a summary with respect to the information from tweets, called tweet-biased summary (TBS). It was then compared with two different summaries generated without using tweets, called generic summary (GS). We did pairwise comparison in a crowdsourcing-based evaluation to measure their performances. We also performed traditional ROUGE evaluation using small number of sampled documents to see the performance of summaries according to ROUGE score and analyzed the agreement between the ROUGE score and the user preference. Next, we also analyzed the influence of the number of tweets pointed to web documents on the performance of generated TBS.

Based on the result of our experiments, we obtained that TBS was significantly better than $GS_{sn}$ and $GS_{ag}$ according to the user preference, respectively in percentage of 169.05% and 163.64%. The ROUGE score also shows the same result that TBS is better than GS, but it is only significant for $GS_{sn}$ as a baseline. The user preference is reasonably agree with ROUGE score, however, the level of agreement is better when two summarization systems have quite different performance (according to ROUGE score). This finding answers the research question *Q2*. For the research question *Q3*, we showed that the number of related tweets influences the performance of TBS. We found that there is a positive moderate correlation between the number of tweets pointed to web documents and the accuracy of TBS as measured by user preference. Based on the above results, we can respond to research question *Q1* by concluding that social media can be used to select better sentences for generating a summary of web document, effects on improving the summary accuracy. Answering the last question *Q4*, after analyzing the comments given by people in the crowdsourcing-based evaluation, we obtained

three main aspects that people considered when choosing one summary over another: content, presentation and flow of the summary. Most of the comments discussed about the quality of summary content as a deciding factor in choosing the best summary.

In this work, we used the URL contained in a tweet as a means of gathering tweets related to a web document that would be summarized. From more than five million English tweets in the dataset, one-fifth of them contain a URL, in which then we found there is only a small number of URLs that were pointed by high number of tweets. However, the results of our experiments indicate that even a single tweet can improve the quality of the summary of a web page. Nevertheless, the technique as described can only be used for web pages pointed to by tweets. For future work, we need to implement a way for collecting tweets that relate to web pages as well as those that point directly to them. We might also try to add social information from different sources as a complement of tweets, such as: forums or blogs, by using the technique described in (Lee & Croft, 2013). Further work to compare user preference with ROUGE using higher number of data can also be done to get more accurate result.

## 5 Acknowledgments

This research is supported in part by the Directorate General of Higher Education of Indonesia (Ministry of National Education), the Indonesia Endowment Fund for Education (Ministry of Finance), the Ministry of Education (Malaysia), Universiti Teknologi Malaysia, and by the Australian Research Council (DP140102655).

## 6 References

Afantenos, S., Karkaletsis, V., & Stamatopoulos, P. (2005). Summarization from medical documents: a survey. *Artificial Intelligence in Medicine*, *33*(2), 157–77. doi:10.1016/j.artmed.2004.07.017

Allan, J., Wade, C., & Bolivar, A. (2003). Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 314–321).

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 492–499).

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd editio.). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Delort, J.-Y., Bouchon-Meunier, B., & Rifqi, M. (2003). Enhanced web document summarization using hyperlinks. *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia - HYPERTEXT '03*, 208. doi:10.1145/900095.900097

Diakopoulos, N., De Choudhury, M., & Naaman, M. (2012). Finding and assessing social media information sources in the context of journalism. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (pp. 2451–2460).

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, *16*(2), 264–285.

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.(JAIR)*, *22*(1), 457–479.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fortinet. (2007). *The Importance of Fortiguard Web Filtering as Part of a Multi-Threat Security System*.

Galgani, F., Compton, P., & Hoffmann, A. (2012). Combining different summarization techniques for legal text. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data* (pp. 115–123).

Gao, W., Li, P., & Darwish, K. (2012). Joint topic modeling for event summarization across news and social media streams. In *CIKM'12* (pp. 1173–1182).

Glaser, A., & Schütze, H. (2012). Automatic generation of short informative sentiment summaries. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 276–285).

Hori, C., Hirao, T., & Isozaki, H. (2004). Evaluation measures considering sentence concatenation for automatic summarization by sentence or word extraction. In *Proceedings of Workshop on Text Summarization Branches Out*.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168–177).

Hu, M., Sun, A., & Lim, E.-P. (2008). Comments-oriented document summarization: understanding documents with readers' feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 291–298).

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. doi:10.1016/j.bushor.2009.09.003

Keikha, M., Park, J. H., & Croft, W. B. (2014). Evaluating answer passages using summarization measures. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 963–966).

Lee, C.-J., & Croft, W. B. (2013). Incorporating social anchors for ad hoc retrieval. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval* (pp. 181–188).

Liu, F., Liu, Y., & Weng, F. (2011). Why is "sxsw" trending? exploring multiple text sources for twitter topic summarization. In *Proceedings of the ACL Workshop on Language in Social Media (LSM)* (pp. 66–75).

Lloret, E., Plaza, L., & Aker, A. (2013). Analyzing the capabilities of crowdsourcing services for text summarization. *Language Resources and Evaluation*, *47*(2), 337–369.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165.

Ma, Y.-F., Hua, X.-S., Lu, L., & Zhang, H.-J. (2005). A generic framework of user attention model and its application in video summarization. *Multimedia, IEEE Transactions on*, *7*(5), 907–919.

Maskey, S., & Hirschberg, J. (2005). Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *INTERSPEECH* (pp. 621–624).

Montcalm, D., & Royse, D. D. (2002). *Data analysis for social workers*. Allyn and Bacon Boston.

Nenkova, A., & McKeown, K. R. (2011). Automatic Summarization. *Foundations and Trends® in Information Retrieval*, *5*(2-3), 130. doi:10.1561/1500000015

Parapar, J., López-Castro, J., & Barreiro, Á. (2010). Blog snippets: a comments-biased approach. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 711–712).

Phelan, O., McCarthy, K., & Smyth, B. (2009). Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems* (pp. 385–388).

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, *14*(3), 130–137.

Sanderson, M., Paramita, M. L., Clough, P., & Kanoulas, E. (2010). Do user preferences and evaluation measures line up? In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 555–562).

Shen, D., Sun, J.-T., Li, H., Yang, Q., & Chen, Z. (2007). Document Summarization Using Conditional Random Fields. In *IJCAI* (Vol. 7, pp. 2862–2867).

Strohman, T., Metzler, D., Turtle, H., & Croft, W. B. (2005). Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis* (Vol. 2, pp. 2–6).

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).

Thomas, D. R. (2006). A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, *27*(2), 237–246. doi:10.1177/1098214005283748

Twitter Statistics. (2013).

Wan, X., & Xiao, J. (2010). Exploiting neighborhood knowledge for single document summarization and keyphrase extraction. *ACM Transactions on Information Systems (TOIS)*, *28*(2), 8.

Yang, Z., Cai, K., Tang, J., Zhang, L., Su, Z., & Li, J. (2011). Social context summarization. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 255–264).

Zhuang, L., Jing, F., & Zhu, X.-Y. (2006). Movie review mining and summarization. *Proceedings of the 15th ACM International Conference on Information and Knowledge Management - CIKM '06*, 43. doi:10.1145/1183614.1183625