Performance of Opinion Summarization towards Extractive Summarization

H.Iboi, S.Chua, B.Ranaivo-Malançon, and N.Kulathuramaiyer
Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.
16020133@siswa.unimas.my

***Abstract*—Opinion summarization summarizes opinion in texts while extractive summarization summarizes texts without considering opinion in the texts. Can opinion summarization be used to produce a better extractive summary? This paper proposes to determine the effectiveness of opinion summarization generation against extractive text summarization. Sentiment that includes emotion which indicates whether a sentence may be positive, negative or neutral is considered. Sentences that have strong sentiment, either positive or negative are deemed important in text summarization to capture the sentiments in a story text. Thus, a comparative study is conducted on two types of summarizations; opinion summarization using the proposed method, which uses two different sentiment lexicons: VADER and SentiWordNet against extractive summarization using established methods: Luhn, Latent Semantic Analysis (LSA) and LexRank. An experiment was performed on 20 news stories, comparing summaries generated by the proposed opinion summarization method against the summaries generated by established extractive summarization methods. From the experiment, the VADER sentiment analyzer produced the best score of 0.51 when evaluated against the LSA method using ROUGE-1 metric. This implies that opinion summarization converges with extractive summarization.**

***Index Terms*—Extractive summarization; opinion summarization; LexRank method; LSA method; Luhn method.**

I. INTRODUCTION

The abundance of opinions on the Web has inspired the research of opinion summarization in the last few years. Opinion summary is the outcome of sentiment analysis which summarizes opinions in texts. The objective of opinion summary is to assist the reader to understand the huge collection of opinions in an efficient way [1]. This summarization approach involves text clustering, sentiment analysis, text mining and natural language processing (NLP). Nevertheless, it is unlike common text summarization because opinion summarization emphasizes on the opinionated parts while the common extractive summarization emphasizes on extracting informative parts and redundancy removal.

Sentiment analysis is part of opinion summarization. It has been a popular platform in gauging sentiments on the Web and social media. Sentiment analysis distinguishes and extracts subjective or emotion information in texts by using NLP, text analysis and computational linguistics [2]. It focuses on the expressed opinion of a text, disregarding the topic of the text itself. There are three levels in sentiment analysis; document level, sentence level and phrase level. Document level sentiment analysis determines whether the whole document gives a positive, negative or neutral sentiment. The advantage of this level of analysis is the ability to determine the overall text sentiment classification. As for sentence level sentiment analysis, it classifies whether each sentence indicates a positive, negative or neutral opinion [3]. Phrase level is also known as feature based sentiment analysis in which sentiment is directly assigned to the features.

With the growth in the number of digital documents, there is an important need for text summarization. When reading a text, a reader usually tends to skim through the text for the first time to grab the general idea of the text. Text summarization can generally be described as the process of forming a summary out of the textual elements of a text narrative. A summary is defined as a text that is generated from one or more texts, that delivers important information in the original text, and that is no longer than half of the original text [2]. The original text can be very long and this may put the casual reader off. Thus, automatic text summarization (ATS) can aid the reader to understand the gist of the text in just a fraction of time by providing a concise summary. ATS is helpful when a useful summary is needed from a very lengthy text.

The question that remains to be answered is how does opinion summarization correlate with extractive summarization? This study was undertaken to compare the result of the proposed opinion summarization method against the result of established text summarization methods: Luhn, LSA and LexRank. The metric used for evaluation is ROUGE-N, looking for overlapping fragments of text.

II. RELATED WORKS

The scene of text summarization research had evolved over the years. The earliest works on summarization largely made use of statistical-based techniques based on word frequency [4, 5] and sentence position [5]. These techniques form the foundation of feature extraction in text summarization and are still largely adopted in most text summarization approaches. Subsequently, machine learning and NLP techniques for text summarization followed. Machine learning techniques are used for selecting the best feature to extract in text summarization [6-8] while NLP techniques allow elements of the natural language such as text structure, concepts in documents [6] and lexical chains [7] to be exploited for text summarization. The major approaches to text summarization are also summarized in [8], highlighting the literature for summarization through extraction and abstraction.

More recent approaches to text summarization looks at sentence ordering [9, 10], extracting salient sentences in given document(s) by modeling text summarization as an optimization problem [11], constraint-driven models [12], correlation of sentences, removal of redundant sentences and using fuzzy logic extraction and latent semantic analysis. The drawback of all these methods for text summarization is that they focus mainly on textual content and not on how a human understands a text. Current extraction techniques were limited by their inability to convey implicit information, the author's intention, the reader's intention, the context of influence and the general world knowledge as well as sentiments embedded within a text. In general, text summarization techniques extract sentences from text based on word frequency, sentence position, text structure, concepts and lexical chains to name the least. These sentences are then put together into a summary. At best, the summary is understandable and acceptable.

Sentiment classification distinguishes the semantic orientation of words, sentences and documents [1]. Sentiment classification is a significant step in opinion summarization. Opinion summarization involves a holistic method to generate summaries from the raw opinionated text. The objective of summarizing opinions is different from summarizing general texts. Thus, opinion summarization has different characteristics from the common extractive summarization. Opinion summarization focuses on the sentiment polarities of the sentences. Nevertheless, extractive summarization techniques can still be applied in opinion summarization for sentence selection and summary generation [1].

Opinion summarization techniques consist of aspect-based and non-aspect-based summarization [1]. Aspect-based summarization classifies input texts into aspects which are known as subtopics and features. Then, a summary is generated for each aspect. Non-aspect-based summarization generates the summary without considering the aspects.

Balahur et al. proposed a method of summarizing positive and negative opinions in blog threads [13]. They employed a sentiment classification system and a text summarizer in their approach. They classified the sentences into three groups: positive, negative and neutral or objective sentences. The positive and negative sentences were processed by a text summarizer to produce the summary of each group but the group of neutral or objective sentences is not considered to be in the summary. Thus, they generated two summaries, positive and negative summaries for each blog thread. They ran a sentiment analysis system and delivered the result to a standard LSA-based text summarization system. They applied WordNet Affect [14], SentiWordNet [15] and MicroWNOp [16] as their lexicons to classify the sentiment polarity for the opinionated sentences. For the evaluation metrics, they used the ROUGE metric: ROUGE-N, (where, N=1 and 2, ROUGE_{SU4} and ROUGE_{L}. The results for sentiment analysis were presented as follows: negative sentences scored 0.98 for precision, 0.54 for recall and 0.69 for F-score, whereas positive sentences scored 0.07 for precision, 0.69 for recall and 0.12 for F-score. Other than that, they evaluated the summarization performance on LSA summarizer on each negative and positive posts and the performance of LSA summarizer using the 2008 Text Analysis Conference Summarization track (TAC08).

Yadav et al. proposed an extraction-based summarization that included sentiment [17]. Their approach consisted of three main stages; sentence scoring, redundancy removal and summary evaluation. For sentence scoring, they proposed two techniques; statistical technique and sentiment technique. The scoring of statistical technique was based on four features; location, aggregation similarity, frequency and centroid. As for the sentiment technique, the entities that appeared in the sentences were identified and given sentiment scores. The total sentiment scores of all the entities in a sentence was the score of the sentence. The sentences were arranged in descending order based on the total score. In the second stage, the top most scored sentences would be put together as the summary if the length of the summary was less than the desired length and the similarity between summary and sentence is lower than the predetermined threshold. The last stage is the evaluation of the summary. The authors used the ROUGE evaluation package and they could obtain high precision most of the time. The highest score was when evaluated against MEAD-10 model summary in which the summary length is limited to 10%. The evaluation measure is ROUGE-1 and the results of 0.46 for precision, 0.71 for recall and 0.56 for F-score were obtained.

III. Proposed Method

Sentiment analysis can be employed in different tasks such as determining text subjective or objective polarity, positive or negative polarity and determining the strength of the text polarity (weak, mild or strong). The focus of this work is to apply sentiment analysis on sentence-level positive or negative polarity. The opinion summarization method is based on strong sentiment sentence extraction, either positive or negative. The framework for a comparative study of opinion summarization and extractive summarization is illustrated in Figure 1.

Briefly, this study is conducted by comparing the results of opinion summary and extractive summary. In opinion summarization, the methods are divided into two stages: sentiment classification and summary generation. In Stage 1, the news stories are processed through the sentiment analyzers for both SentiWordNet and VADER lexicons. In this stage, the words in each sentence will be assigned their sentiment scores and polarity automatically from the sentiment analyzers. In Stage 2, the sentences with assigned sentiment scores are ranked in descending order based on the total sentiment scores, taking both the positive and the negative polarity and then considering only the magnitude of the scores. Then, the top N sentences are selected to be an opinion summary.

For extractive summarization, the same set of newspaper stories are processed through three established extractive summarization methods: Luhn, LSA and LexRank. These methods will each generated their respective summaries. The generated opinion and extractive summaries are then evaluated using the ROUGE 2.0 toolkit.

A. Proposed Opinion Summarization Method

The proposed opinion summarization method is based on strong sentiment sentence extraction, either positive or negative. There are two stages in this method, which are i) sentiment classification and ii) summary generation. The generated summary from this method consists of sentences with positive and negative polarity. In the first stage, the raw sentences are assigned to positive or negative sentiment polarity by using two different sentiment analyzers. The two sentiment analyzers are respectively using two different lexical resources, which are SentiWordNet [15] and VADER [18].

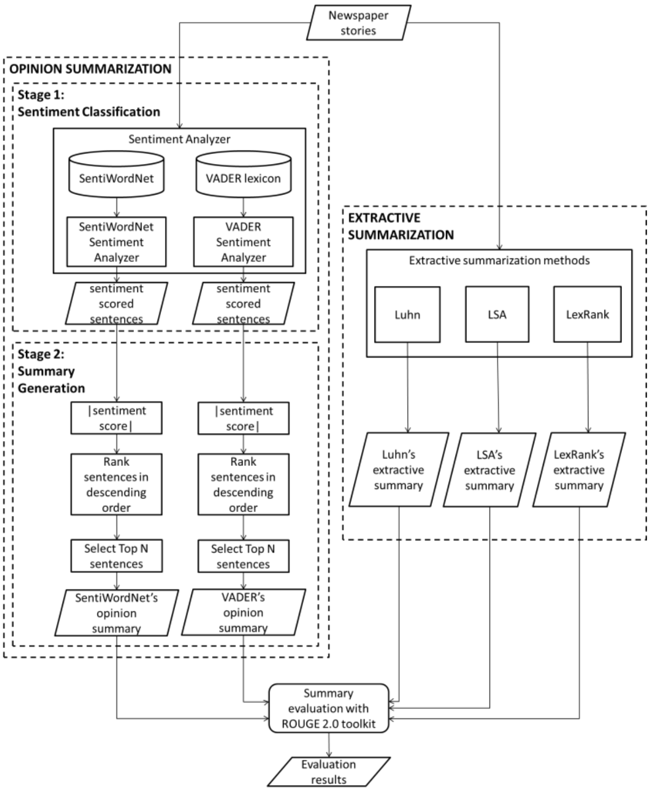

Newspaper stories
OPINION SUMMARIZATION
Stage 1: Sentiment Classification
Sentiment Analyzer
SentiWordNet
VADER lexicon
SentiWordNet Sentiment Analyzer
VADER Sentiment Analyzer
sentiment scored sentences
sentiment scored sentences
Stage 2: Summary Generation
|sentiment score|
|sentiment score|
Rank sentences in descending order
Rank sentences in descending order
Select Top N sentences
Select Top N sentences
SentiWordNet's opinion summary
VADER's opinion summary
EXTRACTIVE SUMMARIZATION
Extractive summarization methods
Luhn
LSA
LexRank
Luhn's extractive summary
LSA's extractive summary
LexRank's extractive summary
Summary evaluation with ROUGE 2.0 toolkit
Evaluation results

Figure 1: Framework for a comparative study of opinion summarization and extractive summarization

In the summary generation stage, the sentences with assigned sentiment are ranked based on the sentiment scores, taking both the positive and the negative polarity and then considering only the magnitude of the scores. The top N scored sentences are selected to form a summary. The selection of the summary length is based on compression ratio set for each summary. The length of the summaries for both opinion and extractive summaries is predefined before summary generation. The length of the summaries is measured by the number of sentences. The Compression Ratio (CR) method is used to determine the length. CR is calculated using the following formula:

$$CR = \frac{length\ of\ summary}{length\ of\ full\ text} \times 100\% \qquad (1)$$

As mentioned by Morris et al., the best summary is around 20% to 30% from their original texts [19]. In the experiment conducted, the value 30% was used as the threshold for CR. If the case where the length is not a whole number, the value will be rounded down.

This method is illustrated in Algorithm 1 and 2.

Algorithm 1: Computing sentiment scores

```
input: An array A of n sentences
output: Sentence-score matrix
    A = {1, 2, 3, …$n$}
    for i←1 to $n$ do
            Score each sentence using sentiment analyzers
    end for
    return sentence-score matrix
```

Algorithm 2: Opinion Summarization Method

```
input: Sentence-score matrix
output: Opinion Summary
    for i←1 to n do
        scores = |score|
    end for

    /*Sort the sentences in descending order according to their scores:*/
    for i←1 to n-1 do
        j = i;
        do while (j>0) and (A(j) > A(j-1)
                temp = A(j)
                A(j) = A(j-1)
                A(j-1) = temp
                j = j-1
        end do
    end for

    /*Select top N sentence*/
    set CR = 30
    N = (CR*n)/100
    if N is in decimal value then
                round down the value

    /*Create the summary with top N sentences*/
    for i = 1 to N
                if (summary, i$^{th}$ sentence) ≤N then
                        summary = summary, i$^{th}$ sentence
                end if
    end for
    return summary
```

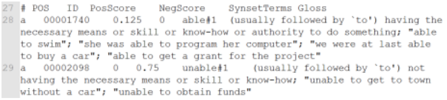
The selected sentiment lexical resources are SentiWordNet and VADER. Both lexicons are easily available and are capable of providing sentence polarity scores. Both of the lexicons are given the sentiment scores in between the range of -1.0 (most negative) to +1.0 (most positive).

A. SentiWordNet Lexicon

SentiWordNet is an open source resource and has a web-based graphical user interface. SentiWordNet is a lexical resource that is constructed from WordNet [15]. SentiWordNet is grouped into adjectives, nouns, adverbs and verbs in synonym sets (synset). Each set is assigned to three numerical scores Obj(s), Pos(s) and Neg(s) to distinguish between objective, positive and negative terms in the synset [15]. The value of positive and negative scores are assigned in SentiWordNet by adapting synset classification to decide the PN-polarity (positive negative) and SO-polarity (subjective objective) polarity of terms [15]. This method depends on training a set of ternary classifiers, which are able to determine positive, negative or objective polarity of a synset. Then, the objective score is calculated by the following formula:

$$ObjScore = 1 - (PosScore + NegScore) \qquad (2)$$

The lexicon is arranged by part-of-speech (POS) tags, term ID, positive scores, negative scores and the glossary of synset terms. Each part is separated only by spaces. Figure 2 shows an example of the lexicon representation.



```
# POS   ID  PosScore    NegScore    SynsetTerms Gloss
a   00001740    0.125   0   able#1  (usually followed by `to') having the
necessary means or skill or know-how or authority to do something; "able
to swim"; "she was able to program her computer"; "we were at last able
to buy a car"; "able to get a grant for the project"
a   00002098    0   0.75    unable#1    (usually followed by `to') not
having the necessary means or skill or know-how; "unable to get to town
without a car"; "unable to obtain funds"
```

Figure 2: SentiWordNet's lexicon arrangement for 'able' and 'unable' terms

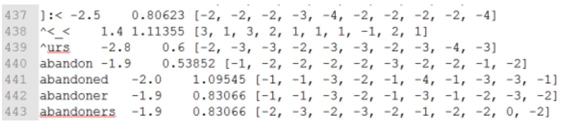
SentiWordNet has 117659 entries or synsets. Each synset has three numerical scores ranging from 0.0 to 1.0 for Obj(s), Pos(s) and Neg(s) and the total score for a synset is equal to 1.0. The scores represent the magnitude for each word in the synset. A synset may have nonzero scores for the three terms categories because each sense has a certain degree of polarity. For example, a term may be positive in some sense and negative in another sense.

B. VADER Lexicon

Valence Aware Dictionary for sEntiment Reasoning (VADER) is a lexicon and rule-based sentiment analysis tool written in Python. It is specifically used to identify sentiments conveyed in social media but it operates well on other general texts [18].

VADER lexicon is developed by analyzing existing well-established sentiment word-banks which are Linguistic Inquiry Word Count (LIWC), Affective Norms for English Words (ANEW) and General Inquirer (GI). Then, they merged common sentiment expression in social media which are the emoticons, sentiment-related acronyms and initialisms. There are currently more than 9000 lexical feature candidates. These candidates are evaluated based on their applicability to express sentiment. This results in the VADER lexicon to have only 7517 lexical features with validated valence scores that determine sentiment polarity and intensity. Sentiment polarity assigns positive and negative polarity while sentiment intensity is ranged from -4 to +4.

The implementation of VADER focuses on sentence-level sentiment analysis method. It classifies the sentences to determine their positivity or negativity. VADER is an open source tool and gives a good performance observed in various experiments conducted in the works of Ribeiro et al. [20]. Figure 3 shows the arrangement of each lexical feature in VADER lexicon.



```
]:< -2.5    0.80623 [-2, -2, -2, -3, -4, -2, -2, -2, -2, -4]
^<_<    1.4 1.11355 [3, 1, 3, 2, 1, 1, 1, -1, 2, 1]
^urs    -2.8    0.6 [-2, -3, -3, -2, -3, -3, -2, -3, -4, -3]
abandon -1.9    0.53852 [-1, -2, -2, -2, -2, -3, -2, -2, -1, -2]
abandoned   -2.0    1.09545 [-1, -1, -3, -2, -1, -4, -1, -3, -3, -1]
abandoner   -1.9    0.83066 [-1, -1, -3, -2, -1, -3, -1, -2, -3, -2]
abandoners  -1.9    0.83066 [-2, -3, -2, -3, -2, -1, -2, -2, 0, -2]
```

Figure 3: VADER lexical features

VADER sentiment analyzer produces four different types of score; positive (*pos*), neutral (*neu*), negative (*neg*) and compound [18]. The *pos*, *neu* and *neg* scores are ratio for proportions of the text that fit in each category. These metrics are beneficial for multidimensional measures of sentiment for a given sentence. The compound score is calculated by adding the valence score of each word in the lexicon by following its parsimonious rule-based modeling and the score is normalized between -1 (the most negative) and +1 (the most positive). This metric is a normalized, weighted composite score [18]. It is suitable when analyzing a sentence's sentiment for a single unidimensional measure.

C. Reference Extractive Summarization Methods

There are three established extractive summarization methods that are adopted for comparison in this research. They are Luhn [4], Latent Semantic Analysis (LSA) [21] and LexRank [22]. These methods are used to generate benchmark summaries to compare with the summary generated by the proposed method.

D. Luhn method

This method uses two features to identify the important sentences in a text. The two features are (i) the presence of significant words and (ii) the distance between these significant words. A word's significance is based on the occurrence of the word in the whole text. The distance is computed from the number of non-significant words between two significant words. If the distance is more than a pre-determined threshold value, then the significant and non-significant words within the count of the threshold value will be grouped into a cluster. The score of each sentence is given based on the following formula [3].

$$sentence\ score = \frac{(significant\ words\ in\ cluster)^2}{total\ words\ in\ cluster} \qquad (3)$$

E. Latent Semantic Analysis (LSA) method

This method is used to identify the important sentences by considering the semantic features [23]. LSA extracts and makes up semantic knowledge of the text from the observation of the term frequency [24]. It constructs a semantic space with a massive dimension from the statistical analysis of term frequency for the whole text. This method is implemented by performing latent semantic indexing which uses singular value decomposition (SVD) to generic text summarization [25]. SVD is used to reflect an important topic or concept of the document and the value shows the importance level of the topic or concept.

The method begins by creating a term by sentences matrix A = $[A_1, A_2, \ldots, A_n]$ with each column vector A_i, indicating the weighted term-frequency vector of sentence *i* in the document [25]. SVD of A is formulated as:

$$A = U\Sigma V^T \qquad (4)$$

where,
U= $[u_{ij}]$ is a $m \times n$ column-orthonormal matrix
Σ = diag(σ_1, σ_2, …, σ_n) is an $n \times n$ diagonal matrix
V= $[v_{ij}]$ is an $n \times n$ orthonormal matrix

The formula can be illustrated as shown in Figure 2.

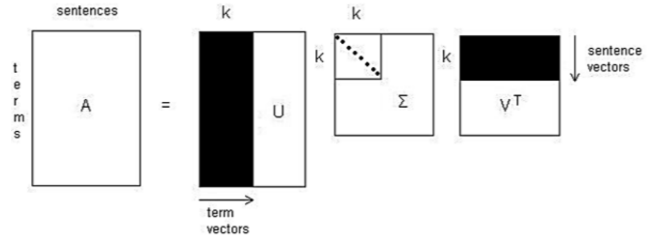


Figure 4: Singular Value Decomposition [25]

F. LexRank method

Lexical PageRank or LexRank is a method that constructs the text into a graph that consists of nodes which represent the sentences and edges which represent the similarity relation between sentences [22]. LexRank calculates similarities among the sentences by applying cosine similarity function. The formula is shown as follows [22]:

$$idf - modified - cosine(x, y) = \frac{\Sigma_{w \in x,y} tf_{w,x}\, tf_{w,y} (idf_w)^2}{\sqrt{\Sigma_{x_i \in x} (tf_{x_{i,}x} idf_{x_i})^2} \times \sqrt{\Sigma_{y_i \in y} (tf_{y_{i,}y} idf_{y_i})^2}} \qquad (5)$$

where,
tf = term frequency
idf = inverse document frequency
$tf_{w,s}$ = number of occurrences of the word w in the sentence *s*

A sentence is ranked higher if it is cited by other highly ranked sentences as inspired from the idea of the PageRank algorithm [26]. The summary is generated by taking the top ranked sentences using a pre-determined threshold value.

G. Evaluation Metric

In this comparative study, the generated opinion summary is evaluated by using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) toolkit by calculating the overlapping of words between the opinion summaries and the extractive summaries from Luhn, LSA and LexRank. The ROUGE metric is used in this study as it is the commonly used metric for evaluating summaries. It is able to measure the quality of a summary by comparing it against the ideal summary [27]. ROUGE is a recall-based metric which is based on n-gram co-occurrence for constant-length summaries [22]. This is known as ROUGE-N, which is available in the ROUGE 2.0 evaluation toolkit. ROUGE-N is a recall-related measure. The denominator of the equation is the total sum of the number of n-grams occurring at the reference summary side [27].

$$ROUGE - N = \frac{\sum_{S \epsilon \{ReferemceSummaries\} gram_n \epsilon S} \Sigma\, Count_{match}(gram_n)}{\sum_{S \epsilon \{ReferemceSummaries\} gram_n \epsilon S} \Sigma Count\, (gram_n)} \qquad (6)$$

ROUGE-1 had been proven to be a good measure for a short summary of a single document [22]. ROUGE-1 searches for the overlapping of unigram in the whole text against the model summary. To evaluate the generated summaries using ROUGE in a fair manner, the length of the summaries needs to be fixed. ROUGE gives three values of measurement: recall, precision, and F-score. Some past results reported in the literature are as follows. The ROUGE-1 F-score for the summarization of clinical text notes is around the value of 0.28 to 0.48 by using different summarization methods such as Random and Oracle methods [28]. The ROUGE-1 F-score for different variations of LexRank summarization algorithm is around the value of 0.36 to 0.44 [22]. The ROUGE-1 F-score of LSA-based text summarization, when utilized on blog posts, is 0.22 on negative posts and 0.21 on positive posts [13]. These results in the literature served as an overview of the range of results obtainable from the methods used.

F-score measure is used to compare the performance of the summaries as F-score represents the combination of recall and precision. The following formula describe the context of the evaluation metrics in summary evaluation [29].

$$\text{Precision } = \frac{\text{correct}}{correct + wrong} \qquad (7)$$

$$\text{Recall } = \frac{\text{correct}}{correct + missed} \qquad (8)$$

$$\text{F} - \text{score } = \frac{2\ x\ (Precision\ x\ Recall)}{Precision + Recall} \qquad (9)$$

where:
correct = the number of sentences in opinion summary that are correctly identified as important sentences and appear in extractive summary;
wrong = the number of sentences in opinion summary but not in extractive summary;
missed = the number of sentences that are not in opinion summary but appear in extractive summary

IV. EXPERIMENTAL RESULTS

A. Summary Generation

Two types of summaries were generated: (1) Opinion summary, which is the generated summary from the proposed opinion summarization method. (2) Extractive summary, the generated summary from three established extractive summarization methods: Luhn, LSA and LexRank. Two types of opinion summaries were generated, each using the SentiWordNet lexicon and the VADER lexicon respectively. Table 1 shows an overall description of the types of summaries generated.

Table 1
Types of summaries generated

Summary types		Description
Opinion summary	SentiWordNet	Summary generated from the proposed method using SentiWordNet lexicon
	VADER	Summary generated from the proposed method using VADER lexicon
Extractive summary	Luhn	Summary generated from benchmarked extractive summarization using the Luhn method
	LSA	Summary generated from benchmarked extractive summarization using the LSA method
	LexRank	Summary generated from benchmarked extractive summarization using the LexRank method

The dataset used in this work comprised a collection of online newspaper articles taken from the Borneo Post, New Straits Times, The Independent and USA Today. 20 newspaper articles were used as our full texts. The texts were preprocessed first to eliminate irrelevant features such as images and their captions. The maximum number of sentences is 36 while the minimum number of sentences is 13. The number of sentences for each summary is calculated using the CR formula as detailed in Section III (A). The statistics for the dataset are shown in Table 2.

B. Summary Evaluation

The summaries were evaluated using the ROUGE 2.0 toolkit with different ROUGE-N score (N=1 to 10). The measures of F-score for both the system summary and model summary were obtained. Here, system summary refers to the opinion summary while model summary refers to the extractive summary. Both the system summaries and model summaries generated have the same number of sentences in each set.

The results for the comparison of the system summary against the model summary are shown in Figure 5 and Figure 6.

Table 2
Statistics for the dataset

Text ID	Number of sentences		CR (%)	Source
	Full text	Summary		
1	36	10	27.78	Borneo Post
2	23	6	26.09	Borneo Post
3	21	6	28.57	Borneo Post
4	29	8	27.59	Borneo Post
5	20	6	30.00	Borneo Post
6	33	9	27.27	Borneo Post
7	20	6	30.00	Borneo Post
8	21	6	28.57	Borneo Post
9	17	5	29.41	Borneo Post
10	37	11	29.73	Borneo Post
11	36	10	27.78	The Independent
12	25	7	28.00	The Independent
13	26	7	26.92	New Straits Times
14	24	7	29.17	New Straits Times
15	34	10	29.41	New Straits Times
16	25	7	28.00	New Straits Times
17	39	11	28.21	New Straits Times
18	19	5	26.32	New Straits Times
19	26	7	26.92	USA Today
20	38	11	28.95	USA Today

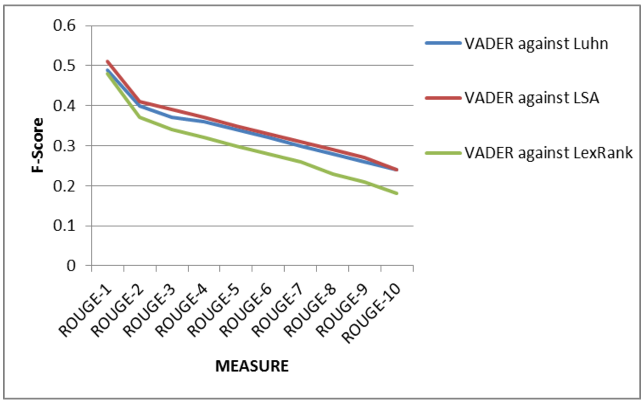


Figure 5: VADER against the three established methods

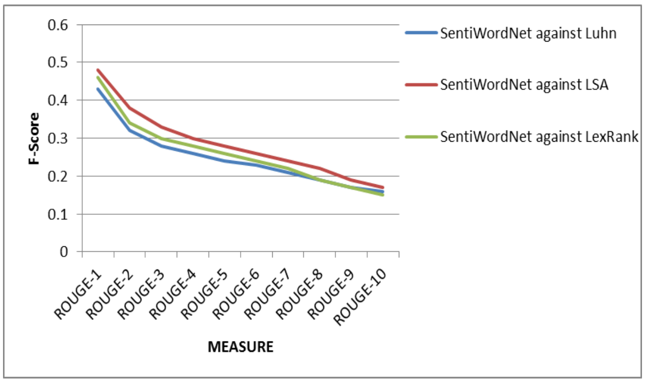


Figure 6: SentiWordNet against the three established methods

The performance of the proposed opinion summarization method is measured in terms of F-score with respect to the ROUGE-1 metrics. Table 3 summarizes the best F-score of the opinion summaries using VADER lexicon and SentiWordNet lexicon against the established extractive summarization methods.

Table 3
ROUGE-1 metric value for opinion summary against extractive summary

Summary type	VADER	SentiWordNet
Luhn	0.49	0.43
LSA	0.51	0.48
LexRank	0.48	0.46

V. Discussion

From the experiment conducted, it could be seen that the score obtained from using the VADER lexicon produced higher scores than using the SentiWordNet lexicon. This indicated that the VADER lexicon is more suitable for in the context of our experiment [20]. However, SentiWordNet still performed well in this experiment by having the F-score results of 0.43, 0.48 and 0.46 for Luhn, LSA and LexRank methods respectively. When comparing the three model summaries; Luhn, LexRank and LSA methods, the proposed method using the VADER lexicon was shown to work more similarly to the LSA method as it gives the highest score among the three model summaries with the score of 0.51 for F-score. The findings from the experiment conducted are as follows:

A. Extractive summarization includes opinion summarization

As highlighted by Kim et al. opinion summarization is different from general text summarization from several perspectives [1]. While the polarities of input opinions are very important in opinion summarization, they have no importance in general text summarization. While the summaries outputted by opinion summarization are more structured as they are divided by topics and polarities, the summaries generated by common text summarization remain texts, and thus unstructured. Nevertheless, the same authors brought the attention on the usefulness of text summarization techniques for opinion summarization: “After separating input data by polarities and topics, classic text summarization can be used to find/generate the most representative text snippet from each category.” [1]. From the point of view of this work, opinion summarization can be useful for extractive text summarization. Conceptually, extractive summarization selects significant sentences without any constraint on whether the sentences convey polarities. It means that the sentence space selection is larger in extractive summarization than in opinion summarization. In addition, the ROUGE-1 recall when comparing VADER opinion summarization against LSA-based extractive summarization indicates that there are around 51% overlaps between the two generated summaries. Thus, this indicates that the contents in opinion summary appear in extractive summary as well.

B. LSA-based extractive summarization shows good correlation with opinion summarization

The good performance of LSA-based extractive summarization may not be surprising. LSA makes use of semantic features and opinion summarization depends usually on semantic classification, which is “determined by the semantic orientation of words, sentences, and documents” [1]. Thus, when an extractive summarization injects some semantic features in its process, it can capture opinionated sentences.

C. Opinion summarization relies on the quality and size of its opinion lexicon

LexRank extractive summarization has a recall value below LSA but above Luhn method when the evaluated opinion summarization is VADER. VADER lexicon is more for microblog/social network-type texts; still work for newspapers; can be considered as gold standard.

D. How far can opinion summarization or extractive summarization perform on news articles?

The dataset used for our experiments corresponds to news articles and thus, they contain certainly less expressed opinions. The main task of a journalist is to report events and not to communicate his or her opinion even though today many journalists go beyond their main task. And because of such attitude, some news articles convey opinions. When VADER was evaluated against human rater annotations and run on opinion news articles (“5,190 sentence-level snippets from 500 New York Times opinion editorials” [16]), its overall F-score was 0.55 (recall = 0.49 and precision = 0.69) [16], which is the lowest value since VADER can reach 0.96 F-score on tweets to go down to 0.63 on product reviews and 0.61 on movie reviews. In our experiments, VADER is compared to automatic text extractive summarization. The F-score value of LSA-based extractive summarization is not far from 0.55 as we obtained 0.52 on general news articles. One can conclude that whatever the content of news articles, with or without opinions, automatic summarization is limited to an F-score of below 0.60.

VI. Conclusion

The experiment conducted had successfully identified that the proposed opinion summarization method can produce acceptable summaries when compared against the established extractive summarization methods. The main contribution is the proposed opinion summarization method. The best results were produced when evaluated using ROUGE-1 metric. ROUGE-1 searches for overlapping of unigram in the opinion summary against the extractive summary. The use of the VADER lexicon in the proposed method produced the highest score when evaluated against the LSA extractive summarization method with the score of 0.51 for F-score. The summary generated by the proposed method using the SentiWordNet lexicon also produced the best result when evaluated against LSA with the value of 0.48 for F-score. The results of this comparative study imply that the proposed opinion summarization method is promising in generating summaries similar to the established extractive summarization methods.

Acknowledgment

This research is funded by the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme F08/FRGS/1489/2016. We would also like to thank Universiti Malaysia Sarawak for the research opportunity and support.

References

[1] H. D. Kim, K. Ganesan, P. Sondhi, and C. Zhai, “Comprehensive review of opinion summarization,” *Illinois Environ. Access to Learn. Sch. Tech. Rep,* pp. 1–30, 2011.

[2] A. V Gundla and M. S. Otari, “A Review on Sentiment Analysis and Visualization of Customer Reviews,” vol. 4, no. 9, pp. 2062–2067, 2015.

[3] S. Sun, C. Luo, and J. Chen, “A Review of Natural Language Processing Techniques for Opinion Mining System,” *Inf. Fusion*, vol. 36, pp. 10–25, 2017.

[4] H. P. Luhn, “The Automatic Creation of Literature Abstracts,” *IBM J. Res. Dev.*, vol. 2, no. 2, pp. 159–165, 1958.

[5] H. P. Edmundson, “New methods in automatic extracting,” *J. Assoc. Comput. Mach.*, vol. 16, no. 2, pp. 264–285, 1969.

[6] I. Mani and E. Bloedorn, “Multi-document Summarization by Graph Search and Matching,” *Proc. fourteenth Natl. Conf. Artif. Intell. ninth Conf. Innov. Appl. Artif. Intell.*, vol. cmp-lg/971, pp. 622–628, 1997.

[7] R. Barzilay and M. Elhadad, “Using lexical chains for text summarization,” *Proc. ACL Work. Intell. Scalable Text Summ.*, vol. 17, no. 48, pp. 10–17, 1997.

[8] D. R.Radev, E. Hovy, and K. Mckeown, “Introduction to the Special Issue on Summarization,” *Comput. Linguist.*, vol. 28, no. 4, pp. 399–408, 2002.

[9] D. Bollegala, N. Okazaki, and M. Ishizuka, “A preference learning approach to sentence ordering for multi-document summarization,” *Inf. Sci. (Ny).*, vol. 217, no. 1, pp. 78–95, 2012.

[10] R. Zhang, “Sentence Ordering Driven by Local and Global Coherence for Summary Generation,” *Acl 2011*, no. June, pp. 6–11, 2011.

[11] R. M. Alguliev, R. M. Aliguliyev, M. S. Hajirahimova, and C. A. Mehdiyev, “MCMR: Maximum coverage and minimum redundant text summarization model,” *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14514–14522, 2011.

[12] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, “CDDS: Constraint-driven document summarization models,” *Expert Syst. Appl.*, vol. 40, no. 2, pp. 458–465, 2013.

[13] A. Balahur, M. Kabadjov, J. Steinberger, R. Steinberger, and A. Montoyoa, “Summarizing opinions in blog threads,” *PACLIC 23 - Proc. 23rd Pacific Asia Conf. Lang. Inf. Comput.*, vol. 2, pp. 606–613, 2009.

[14] C. Strapparava and A. Valitutti, “WordNet-Affect: an affective extension of WordNet,” *Proc. 4th Int. Conf. Lang. Resour. Eval.*, pp. 1083–1086, 2004.

[15] A. Esuli and F. Sebastiani, “SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining,” *Proc. 5th Conf. Lang. Resour. Eval.*, pp. 417–422, 2006.

[16] S. Cerini, V. Compagnoni, A. Demontis, M. Formentelli, and G. Gandini, “Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining,” in *Language resources and linguistic theory: Typology, second language acquisition, English linguistics*, 2007.

[17] C. S. Yadav, A. Sharan, R. Kumar, and P. Biswas, “A New Approach for Single Text Document Summarization,” *Adv. Intell. Syst. Comput.*, vol. 380, pp. 401–411, 2016.

[18] C. J. Hutto and E. Gilbert, “VADER: A parsimonious rule-based model for sentiment analysis of social media text,” in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014, pp. 216–225.

[19] A. H. Morris, G. M. Kasper, and D. A. Adams, “The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance,” *Inf. Syst. Res.*, vol. 3, no. 1, pp. 17–35, Mar. 1992.

[20] F. N. Ribeiro, M. Araújo, P. Gonçalves, and M. A. Gonçalves, “SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods,” *EPJ Data Sci.*, pp. 1–29, 2016.

[21] J. Steinberger and K. Ježek, “Using Latent Semantic Analysis in Text Summarization,” *Proc. ISIM 2004*, pp. 93--100, 2004.

[22] G. Erkan and D. R. Radev, “LexRank: Graph-based lexical centrality as salience in text summarization,” *J. Artif. Intell. Res.*, vol. 22, pp. 457–479, 2004.

[23] S. Xiong and Y. Luo, “A new approach for multi-document summarization based on latent semantic analysis,” in *Proceedings - 2014 7th International Symposium on Computational Intelligence and Design, ISCID 2014*, 2015, vol. 1, pp. 177–180.

[24] N. Alami, M. Meknassi, and N. Rais, “Automatic Texts Summarization: Current State of the Art,” *J. Asian Sci. Res.*, vol. 5, no. 1, pp. 1–15, 2015.

[25] O. M. Foong, S. P. Yong, and F. A. Jaid, “Text Summarization Using Latent Semantic Analysis Model in Mobile Android Platform,” *2015 9th Asia Model. Symp.*, pp. 35–39, 2015.

[26] B. V. Keong and P. Anthony, “PageRank: A modified random surfer model,” *2011 7th Int. Conf. Inf. Technol. Asia Emerg. Converg. Singul. Forms - Proc. CITA'11*, 2011.

[27] C. Y. Lin, “Rouge: A package for automatic evaluation of summaries,” *Proc. Work. text Summ. branches out (WAS 2004)*, no. 1, pp. 25–26, 2004.

[28] H. Moen *et al.*, “On evaluation of automatically generated clinical discharge summaries,” *CEUR Workshop Proc.*, vol. 1251, pp. 101–

114, 2014.

[29] N. Zamin and A. Ghani, “A Hybrid Approach for Malay Text Summarizer,” *Proc. Int. Multi-Conference Eng. Technol. Innov.*, 2010.