EXTRACTIVE SUMMARIZATION ON FOOD REVIEWS

(Date received: 12.01.2022/Date accepted: 10.04.2022)

Yuen Kei Khor[1]*, Chi Wee Tan[1], Tong Ming Lim[2]

[1] *Faculty of Computing and Information Technology, Tunku Abdul Rahman University College, Malaysia.*

[2] *Centre for Business Incubation and Entrepreneurial Ventures, Tunku Abdul Rahman University College, Malaysia.*

**Corresponding author: khory0k-wm17@student.tarc.edu.my*

ABSTRACT

Text summarization is a technique to summarize the content of a sizeable text but meanwhile, it keeps the key information. Extractive summarization and abstractive summarization are the main techniques for text summarization. TextRank algorithm, an extractive summarization technique is applied to perform automatic text summarization in this study. Furthermore, GloVe pre-trained word embedding model is used to map each word from the reviews to a vector representation. In the end, the PageRank algorithm is applied to rank the sentences based on their sentence ranking scores. The more important and relevant sentences which can be the representatives of a summary will be placed in a higher rank. The objective of our study is to extract the top five reviews with the highest sentence ranking scores which can form a summary to provide a conspectus of a cookies brand in Amazon food reviews. Besides, a detailed description of the implementation is discussed to provide an overview on using TextRank to create a summary. An analysis of the customer perception based on the summary generated is conducted to understand their needs and level of satisfaction. The final summary demonstrates that Amazon customer reviews for certain cookies brand are generally positive.

Keywords: Text Summarization, Extractive Summarization, TextRank

List of notations

n is the number of individual sentences

1.0 INTRODUCTION

With the advancement of technology, social media platforms and websites have evolved into a place for the public to freely share their opinions, experiences and thoughts about products, services, and breaking news. This vast volume of text contains essential information but reading it all and creating a summary is inefficient for humans so text summarization comes in handy.

Maybury (1999) defined text summarization as the process of distilling the most important information from one or more sources to produce an abridged version for one or more users and one or more tasks. To be more specific, text summarization produces a summary from one or multiple plain texts while retaining important information. Abstractive summarization and extractive summarization are the two main techniques of automatic text summarization. Abstractive summarization made use of advanced natural language techniques such as the deep learning approach to create a completely new and shorter text which consists of the key information from original source. On the other hand, extractive summarization is the extraction of a subset of important sentences from the original source.

This study is motivated by the desire to replace human power in the task of summarizing a lengthy text into a few sentences in a short period of time. Aside from time-consuming issues, human knowledge and language ability level also greatly affect the quality of summaries. Humans may occasionally misinterpret the meaning of text documents. We choose to apply the extractive summarization approach in this study is because it always outperforms abstractive summarization. This is due to the reason that abstractive summarization needs to address issues such as natural language generation, semantic representation, and inference which is difficult for sentence extraction (Allahyari *et al.*, 2017). TextRank is applied due to the research result of Mihalcea and Tarau (2004) proved that TextRank is competitive or better in some cases when compared to previously proposed algorithm that using supervised system. Besides, TextRank also adaptable to different languages and domains as training corpus is not required.

The objective of this study is to extract the top five reviews with the highest sentence ranking scores which can summarize the overall reviews of a cookies brand from Amazon fine food reviews. The top sentence represents the highest chances of the topic discussed by customers in overall product reviews of the product. Therefore, we can have an overview of the customers' perceptions toward certain food products based on the summary.

2.0 LITERATURE REVIEW

This section mainly discusses the main approaches of extractive summarization which are widely used in research works. Basically, there are three independent tasks to perform extractive summarization: create an intermediate representation of the document, score sentences based on the representation, and create a summary by selecting the few most important sentences based on their scoring. There are a few common approaches widely used in text summarization that will be further discussed below.

In the early research on extractive summarization, researchers use features from the sentences such as their position in the text, word frequency, or key phrases indicating the importance of the sentences (Erkan & Radey, 2004). Term Frequency-Inverse Document Frequency (TF-IDF) method is used to determine how important a word is to a collection of documents. TF-IDF scores increase when the number of times that a word appears is increased in a document. This method works in the weighted term-frequency and inverse sentence frequency. Sentence frequency refers to the total number of sentences containing a specific term in the document. The sentence vectors will be scored by similarity and the sentences with the greatest similarity scores are chosen as a part of the summary (Saranyamol & Sindhu, 2014). Christian *et al.* (2017) proposed an automatic text summarizer that uses TF-IDF to extract three to five sentences with the highest TF-IDF scores to be the final summary, where the number of sentences is decided by users. When compared to another online automatic summarizer, their proposed text summarizer yields a 67% accuracy.

Machine learning can be applied for text summarization if the dataset or documents consists of a summary for each observation. This is because machine learning required a large amount of labelled data to train the model. Machine learning models will learn the patterns by identifying those relevant features values that are correlated with the labelled data. Feature extraction take an important role to improve the accuracy of the summarization result. Having more training data leads to better accuracy of the model as they can learn more different patterns. As a result, an extractive summary can be produced for each document when new documents are given to the model. Neto *et al.* (2002) present a text summarizer using two well-known algorithms, Naive Bayes and C4.5 decision tree algorithm with a set of features that are classified into two categories: statistics-oriented and linguistic-oriented. The performance of these two algorithms is compared with two baseline methods with two sets of experiments by employing automatically-produced extractive summaries and manually-produced summaries. Results show that Naïve Bayes outperforms all the summarizers. Besides, the deep learning approach is also quite common in automatic text summarization (PadmaPriya, 2014; Day & Chen, 2018; *Patel et al.*, 2018).

Another well-known approach of extractive summarization is a graph-based approach. The graph-based approach consists of two elements which are nodes and edges. Nodes refer to the sentences while edges are the similarity between sentences. If two sentences share certain common words, they are connected with an edge. When a node has a large number of edges connected to it, then it is considered as an important sentence that should be included in the summary. TextRank is a well-known graph-based approach for text summarization and it is inspired by PageRank (Brin & Page, 1998) which is implemented by Google. Simply, TextRank is used to rank sentences while PageRank rank web pages in Google search engine results. Important pages will have a higher PageRank score and rank higher in the search engine results. Actually, PageRank can be used in text summarization to select the most important sentences from the original text document. In the study of Mallick *et al.* (2019), they proposed a modified PageRank algorithm that assumes that the important sentences are linked (similar) to other important sentences in the text document. Li and Zhao (2016) also proposed a TextRank algorithm by exploiting Wikipedia for short keywords extraction. Their findings show TextRank model constructed based on Wikipedia as external knowledge works better than traditional TextRank which uses TF-IDF.

3.0 METHODOLOGY AND FRAMEWORK

This section shows the framework of the TextRank algorithm in Figure 1 and a detailed description for each step is discussed.

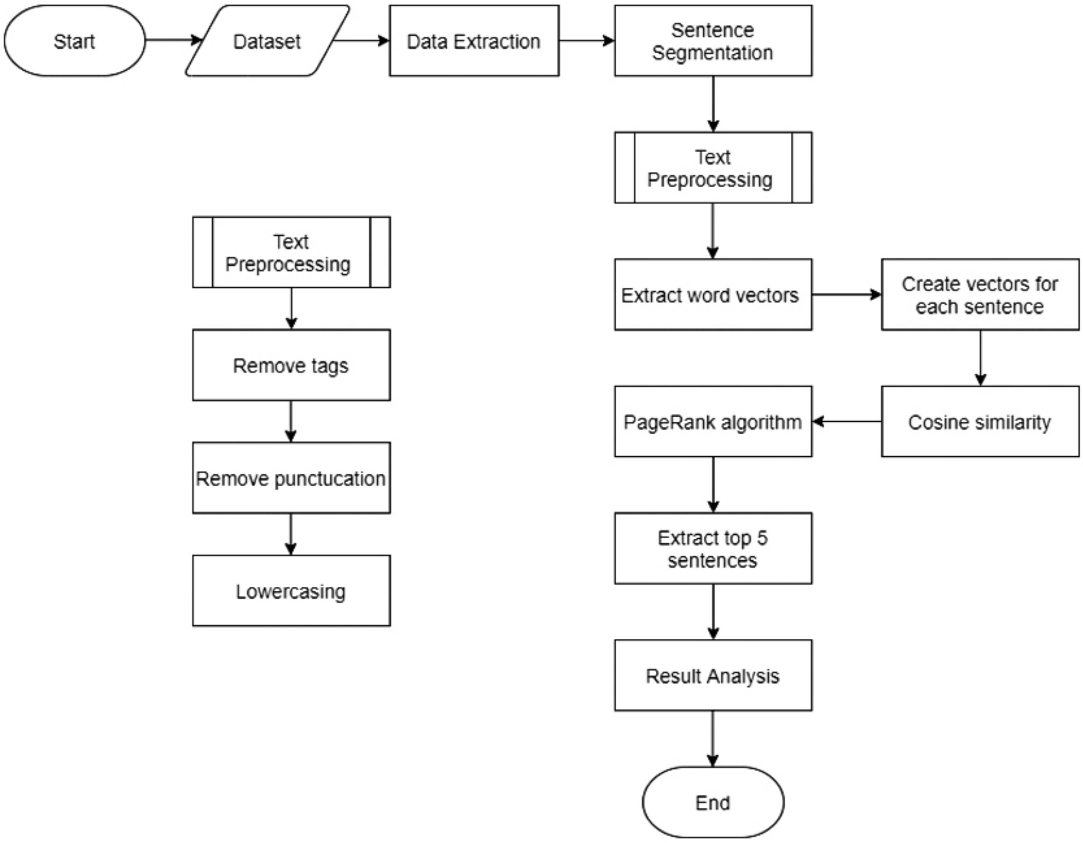


Figure 1: TextRank flowchart of Amazon fine food reviews summarization

3.1 Experiment

This section discusses how to generate a summary of food reviews from Amazon fine food dataset using TextRank algorithm. A total of eight steps that includes dataset, data extraction, sentence segmentation, text preprocessing, extract word vectors, cosine similarity scores, PageRank algorithm and extract top-ranked sentences are explained as follows:

3.1.1 Dataset

The dataset used in our study is Amazon fine food reviews which can be accessed from Kaggle. This dataset contains 568,454 reviews to 74,258 products which are collected from October 1999 to October 2012.

3.1.2 Data Extraction

A summary is generated on the product with the highest number of reviews in the dataset which is a cookie brand but we will not disclose the brand name due to privacy reasons. 910 reviews left after extracting the related reviews of this cookies brand and dropping duplicates.

3.1.3 Sentence Segmentation

zof a few sentences, so it is necessary to segment it into individual sentences for further processing. Segmentation occurred when the sentence ends at the segmentation point such as full stop, question marks, and exclamation marks. Therefore, a total of 910 reviews are segmented to 3661 individual sentences. Figure 2 shows a few sentences from 3661 individual sentences and ranking score is given to each of them to indicate their importance (refer to 3.1.7 PageRank Algorithm). 5 out of 3661 sentences will be selected for inclusion of summary according to the ranking score.

3.1.4 Text Preprocessing

Amazon reviews are usually in the form of unstructured which consists of noises and affect the performance of text summarizing if noise removal is not done perfectly. Noises such as HTML tags, punctuations and stop words are removed and converted all letters to lowercase.

3.1.5 Extract Word Vectors

Techniques are applied to map each word to a real-valued vector which is called word embedding because machines are not able to recognize the semantic and syntactic similarity between words in a text document. Word embedding is typically in the form of a real-valued vector that is used for the representation of words in a vector space. Figure 3 shows an example graph of word embedding, each word represented as a real-valued vector in a vector space. Words that are close to each other in the vector space tend to have associated meanings (McDonald & Ramscar, 2001). Based on the graph, ‘cookie’ and ‘biscuit’ are close to each other so they are expected to have a similar meaning. Therefore, Global Vectors (GloVe) (Pennington *et al.*, 2014) which is an unsupervised learning algorithm for obtaining word vector representations are used to convert each word in the sentences to word vectors. It is trained on the global word-to-word co-occurrence statistics by estimating the frequency of words co-occurs with one another in a given corpus. Pre-trained word vectors with 100 dimensional of 400k words computed on 2014 dump of English Wikipedia is used to create vectors for sentences and are available at (https://nlp.stanford.edu/projects/glove/). Each word will have 100 vectors in the 100 dimensional pre-trained word vectors. Next, 100 vectors of each word will be fetched, and calculate the total vectors of each word in the sentence. The final vector is computed by taking the sum of vectors and dividing by the total number of words in a sentence.

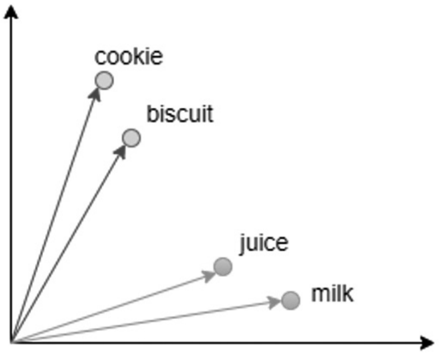
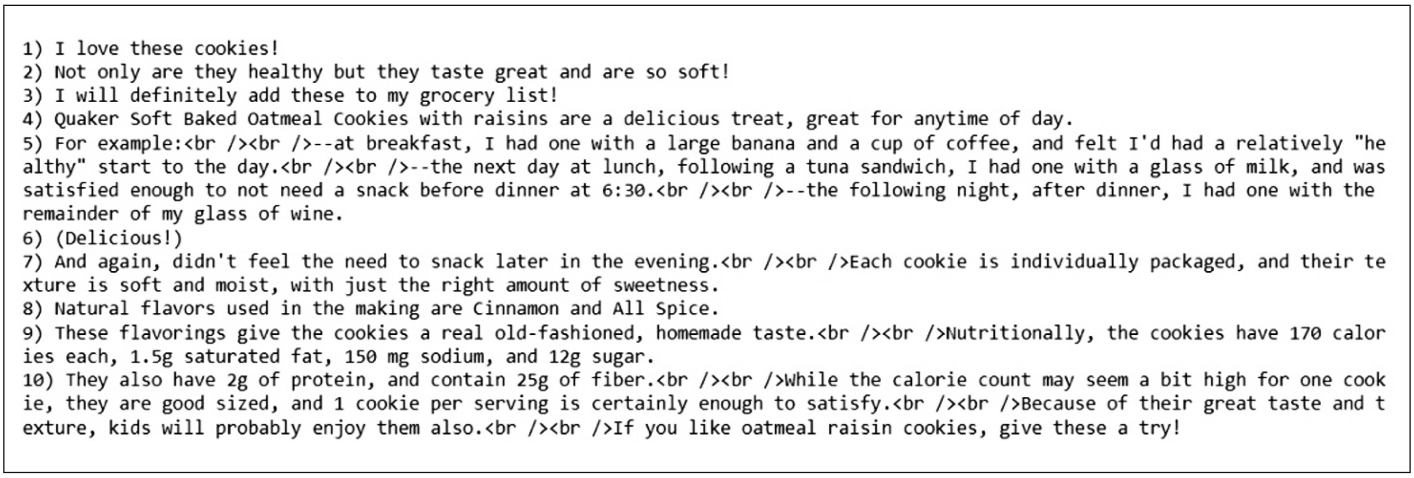


Figure 3: Word Embedding



```
1) I love these cookies!
2) Not only are they healthy but they taste great and are so soft!
3) I will definitely add these to my grocery list!
4) Quaker Soft Baked Oatmeal Cookies with raisins are a delicious treat, great for anytime of day.
5) For example:<br /><br />--at breakfast, I had one with a large banana and a cup of coffee, and felt I'd had a relatively "he
althy" start to the day.<br /><br />--the next day at lunch, following a tuna sandwich, I had one with a glass of milk, and was
satisfied enough to not need a snack before dinner at 6:30.<br /><br />--the following night, after dinner, I had one with the
remainder of my glass of wine.
6) (Delicious!)
7) And again, didn't feel the need to snack later in the evening.<br /><br />Each cookie is individually packaged, and their te
xture is soft and moist, with just the right amount of sweetness.
8) Natural flavors used in the making are Cinnamon and All Spice.
9) These flavorings give the cookies a real old-fashioned, homemade taste.<br /><br />Nutritionally, the cookies have 170 calor
ies each, 1.5g saturated fat, 150 mg sodium, and 12g sugar.
10) They also have 2g of protein, and contain 25g of fiber.<br /><br />While the calorie count may seem a bit high for one cook
ie, they are good sized, and 1 cookie per serving is certainly enough to satisfy.<br /><br />Because of their great taste and t
exture, kids will probably enjoy them also.<br /><br />If you like oatmeal raisin cookies, give these a try!
```

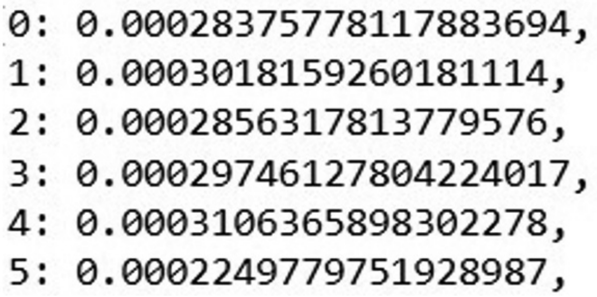
Figure 2: Part of Sentences from the Segmented Individual Sentences

3.1.6 Cosine Similarity Score

Cosine similarity is used to find the similarity between sentences even though TF-IDF is commonly used in text summarization to calculate the relevance and importance of sentences. TF-IDF is too long and sparse because sentences may not share the same words. Despite the fact that no common words appear in two sentences but this does not imply that the sentences have no associated meaning (Han *et al.* 2012). In contrast, cosine similarity measures only focus on common words between sentences and measure their similarity. Therefore, a zero similarity matrix *(n*n)* is created where the size of the matrix is equal to the number of individual sentences. Cosine similarity is used to compute similarity scores between sentences vectors and assigned to the matrix. There will be no relationship between two sentences if the score is 0.

3.1.7 PageRank Algorithm

The similarity matrix is then converted into a graph that has two elements: nodes and edges. Nodes represent the sentence whereas edges reflect the similarity scores between sentences. With the aid of a graph, the PageRank algorithm is used to compute the sentence rankings scores. The scores are used to determine the importance and relevance of sentences in generating a summary. Figure 4 shows the sentence ranking scores for each individual sentences. The first sentence (0.0002837) is less important compared to second sentence (0.0003018) so the ranking of first sentence must lower than second sentence.



```
0: 0.00028375778117883694,
1: 0.00030181592601811114,
2: 0.00028563178137779576,
3: 0.00029746127804224017,
4: 0.00031063658983022278,
5: 0.00022497797519289887,
```

Figure 4: Sentence Ranking Scores of Individual Sentences

3.1.8 Extract Top-Ranked Sentences

Sentences are sorted in descending order based on their sentence ranking scores to generate a summary. The higher the scores, the more relevant the sentences to be extracted for being a part of the summary. In our study, the top five reviews with the highest sentence ranking scores are extracted to form the summary of cookies reviews. This is because five reviews are often the ideal length of a summary, three is too short while ten might be too long for a summary.

4.0 RESULT

Table 1 shows the top five sentences with the highest sentence ranking scores which can be used to form a summary of cookie reviews. Customers who commented on the first and third sentences got a cookie sample from Influenster, product discovery and review platform, and they really liked it because of the softness or freshness. Unlike the second and fifth comments, customers dislike the cookies as they are not fresh, crumbled, or dry. Meanwhile, the third commenter enjoys the taste and softness as well. Based on the summary, we can infer that customers enjoy the oatmeal flavour of this cookies brand and it could be the most popular flavour among customers. In an overall view, customers praised the taste, softness, and freshness of this cookies brand but nevertheless, it is also disliked by customers because the cookies were too dry, not crumble, and not fresh enough.

Table 1: Top 5 highest-ranked sentences

Ranking	Summary
1	I GOT TO TRY THIS QUAKER SOFT BAKED OATMEAL COOKIE THROUGH THE GOOD FOLKS FROM INFLUENSTER AFTER RECEIVING THEIR 2012 MOMVOX BOX, AND I MUST SAY I LOVE IT, FIRST OF ALL OATMEAL COOKIES ARE MY FAVORITE, SO THERE WASNT ANY DISAPPOINTMENT THERE, THE COOKIE RETAIN ITS SOFTNESS/FRESHNESS OFTER BEING OPENED BY ME FOR A WEEK NOW, AND THAT WAS GOOD, PLUS IT TASTE GREAT SO THUMBS UP
2	Maybe it was the baking process? These cookies, although individually packed (so good for school lunches), came out a bit dry and crumbly. Sure, maybe I am just a messy eater but a soft baked cookie just not crumble as much as the cookies I got crumbled. Maybe if you get them at the supermarket they would be less dry. Maybe if is just a general problem with the way they are produced.
3	i received a free sample from Influenster and let me tell you it was so good and soft it crumbles up right in your mouth and its a big cookie my daughter also loved it i would definitely recommend buying it if you like oatmeal and raisins
4	yummy great cookie just like my momma makes this is definitely a second best of course after my mom’s cooking love how soft and chewy they are a must buy
5	I love soft baked cookies, but I find that whenever i try to buy ones that are already made, they don't taste fresh.

4.1 Limitations

One of the drawbacks of this experiment is TextRank takes a long time to compute. The process of computing similarity matrix and sentence ranking scores for roughly 3600 sentences takes a few hours to complete. The computation time increases as the number of sentences extracted to perform summarization grows. This is because the increase in similarity matrix size required a longer time to compute the similarity scores between the sentences. Other than that, the summary is deemed lengthy to read even though it is made out of the top five original reviews with the highest-ranking score from the cookies reviews. Sometimes the reviews can be wordy and difficult to read at a single glance. Researchers may be dissatisfied with the summarising outcome because it is still not a thorough summary of the reviews. In addition, the accuracy of the cookies brand summary should be taken into account. This is owing to the fact that only 100 vectors are used for each word to compute the sentence vector representations.

4.2 Future Work

Computation time can be reduced by expanding the stopwords list from Natural Language Toolkit (NLTK) library. Expanding the stopwords list helps to remove more words from the sentences during text preprocessing which can greatly reduce the number of words to create vectors for each of them. As a result, the computation time needed to calculate the average vectors of each word and the sum of vectors for each sentence is decreasing. We will also study more techniques to summarize each review into a few words and apply our algorithm to obtain a shorter and more relevant summary. In addition, we can fetch more vectors for each word in the sentence to increase the accuracy of summarization. GloVe word embedding consists of pre-trained word vector models with 50, 100, 200 and 300 dimensions for each word. Since 50 dimensions pre-trained model is used to create sentence vector representations in this study, word vectors can be extracted from 200 dimensions or 300 dimensions pre-trained word vector model in our future study. However, there is a trade-off between accuracy and computation time. Increasing the dimensionality of word embedding shall improve the accuracy as it implies the ability to compute more accurate word representation but longer computation time is required. We will explore more in order to take account of the processing speed and accuracy.

5.0 CONCLUSIONS

A summary of a cookies brand's public reviews is created by using the TextRank algorithm to extract the top 5 reviews with the highest sentence ranking score. According to the summary, the majority enjoy this cookie brand because it is fresh, soft, and tastes well but some reviewers may think it is dry, not fresh, and crumble enough. Other than that, the oatmeal cookie could be the most popular product among customers as many compliments have been received by them. The overall summary tends to be positive however there is room for improvement in terms of the moistness and crumble. We can now understand customers' perception of the food product without any human power to read and produce a summary. This study has shown that summarizing reviews required long computation time and unacceptable summary length produced which are needed to be improved in the future.

6.0 ACKNOWLEDGMENTS

The authors would like to express our deepest appreciation to Tunku Abdul Rahman University College for providing financial and resource supports for the implementation of this study. ■

REFERENCES

[1] Allahyari M, Pouriyeh S, Assefi M *et al.* (2017) Text summarization techniques: a brief survey. International Journal of Advanced Computer Science and Applications (IJACSA) 8(10), http://dx.doi.org/10.14569/IJACSA.2017.081052.

[2] Brin S and Page L (1998) The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems 30(1-7): 107-117, http://dx.doi.org/10.3844/jcssp.2014.1.9.

[3] Christian H, Agus MP and Suhartono D (2016) Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). ComTech: Computer, Mathematics and Engineering Applications 7(4): 285-294, http://dx.doi.org/10.21512/comtech.v7i4.3746.

[4] Day MY and Chen CY (2018) Artificial intelligence for automatic text summarization. In 2018 IEEE International Conference on Information Reuse and Integration (IRI), pp. 478-484, http://dx.doi.org/10.1109/IRI.2018.00076.

[5] Erkan G and Radev DR (2004) Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research 22(1): 457-479.

[6] Han JW, Kamber M and Pei J (2012) Getting to know your data. In Data mining (Third Edition), pp. 39-82, https://doi.org/10.1016/B978-0-12-381479-1.00002-2.

[7] Li W and Zhao J (2016) TextRank algorithm by exploiting Wikipedia for short text keywords extraction. In 2016 3rd International Conference on Information Science and Control Engineering (ICISCE), pp. 683-686, https://doi.org/10.1109/ICISCE.2016.151.

[8] Maybury M (1999) Advances in automatic text summarization. MIT press.

[9] McDonald S and Ramscar M (2001) Testing the distributioanl hypothesis: The influence of context on judgements of semantic similarity. In Proceedings of the Annual Meeting of the Cognitive Science Society 23(23).

[10] Mihalcea R and Tarau P (2004) Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing, pp. 404-411.

[11] Neto JL, Freitas A and Kaestner CA (2002) Automatic text summarization using a machine learning approach. In Brazilian symposium on artificial intelligence, Berlin, Heidelberg, vol. 2507, pp. 205-215, http://dx.doi.org/10.1007/3-540-36127-8_20.

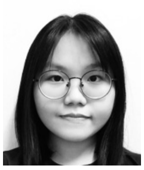
[12] PadmaPriya G (2014) An approach for text summarization using deep learning algorithm. International journal of trends in computer science 10(1): 1-9, http://dx.doi.org/10.3844/jcssp.2014.1.9.

[13] Patel M, Chokshi A, Vyas S and Maurya K (2018) Machine Learning Approach for Automatic Text Summarization Using Neural Networks. International Journal of Advanced Research in Computer and Communication Engineering 7(1), http://dx.doi.org/10.17148/IJARCCE.2018.7133.

[14] Saranyamol CS and Sindhu L (2014) A survey on automatic text summarization. International Journal of Computer Science and Information Technologies, 5(6): 7889-7893.

[15] Pennington J, Socher R and Manning CD (2014) Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543.

PROFILES



KHOR YUEN KEI received her BCS (Hons) in Data Science in the year of 2021 from Tunku Abdul Rahman University College (TAR UC) and is currently pursuing her master's degree in Computer Science at TAR UC. Her research interest focuses on Natural Language Processing (NLP), particularly in code-mixed multiword expressions identification, sentiment and emotion analysis. She is also a data scientist in Work At Cloud Sdn. Bhd. and majorly involved in sentiment analysis tasks and turning data into valuable insights to customers.
Email address: khoryk-wm17@student.tarc.edu.my

DR TAN CHI WEE received BCompSc(Hons) and PhD degrees in year 2013 and 2019 respectively in Universiti Teknologi Malaysia. Currently, he is a Senior Lecturer cum Programme Leader at Tunku Abdul Rahman University College and actively involved in the Centre of Excellence for Big Data and Artificial Intelligent (CoE) and become the research group leader for Audio, Image and Video Analytics Group under Centre for Data Science and Analytics (CDSA). Dr Tan's main research areas are Computer Vision (CV), Image Processing (IP) and Natural Language Processing (NLP) and Artificial Intelligence (AI). He is an enthusiastic researcher experienced in conducting and supporting research into Image Processing. Being a meticulous and analytical researcher with Train-The-Trainer certificate of many years of educational and hands-on experience, he was invited to Université d'Artois (France) under Marie Skłodowska-Curie Research and Innovation Staff Exchange (RISE) programme for collaborative research between European countries with Southeast Asian countries on motion detection and computer vision and being involved in industry project as professional consultant.
Email address: chiwee@tarc.edu.my

PROFESSOR LIM has about 10 years of industry experiences in the design, development, implementation and maintenance of commercial software from 1989 to 1999 after departing from TARC where he spent his early days with TARC as an IT lecturer from 1987 to 1989 after returning from Mississippi State University USA with a Master of Computer Science degree. He is currently the Director for CBIEV at TAR UC, Professor at FOCS at TAR UC and Head for Big Data Analytics Centre. His research interest involving Natural Language Processing, Sentiment Analysis and Code-Mixed language analysis. In the last 15 years, his work has consistently focused on organizational knowledge sharing and technology acceptance, social media analytics and social influence maximization in Sunway University and Tunku Abdul Rahman University College (TAR UC). Professor Lim has graduated more than 20 master and 2 PhD students while he was with Monash, UTAR and Sunway University.
Email address: limtm@tarc.edu.my