Design Your Own NLP Project

BSc Computer Science

Declaration

I certify that the material contained in this dissertation is my own work and does not contain unreferenced or unacknowledged material. I also warrant that the above statement applies to the implementation of the project and all associated documentation. Regarding the electronically submitted work, I consent to this being stored electronically and copied for assessment purposes, including the School's use of plagiarism detection systems in order to check the integrity of assessed work. I agree to my dissertation being placed in the public domain, with my name explicitly included as the author of the work.

Name: Vanessa Chew Thong Xin
Date: 22/3/2024

Contents

Abstract

The evolution of technology created an explosion of information that is accessible to the public, hence creating a need for automatic summarization to simplify and upgrade the quality of life when going through articles or webpages online. The advent of Artificial Intelligence (AI) released various inventions such as Chatbots and AI related products, that gradually became a part of human daily life experience, improving aspects of daily life tasks. This project aims to examines the emerging role of ChatGPT 3.5 in the context of automatic text summarization.

1. Introduction:

1.1 Aims and Objectives:

This project aims to investigate the capabilities of ChatGPT 3.5 when undertaking extractive text summarization tasks. In this project, despite ChatGPT 3.5 leaning towards providing abstractive outputs, an extractive approach was chosen as the method for the summarization task. The dataset used is based on restaurant reviews to allow food seekers to make decisions swiftly, according to the restaurant reviewer's opinion.

2. Background

This section reviews the literature related to automatic text summarization, ChatGPT 3.5 and methods used in automatic extractive text summarization. The background work will be This is done to understand the development of automatic text summarization, the possible methods within the field and the role of ChatGPT 3.5 in automatic text summarization.

As technology evolved, especially in recent times, with the exponential increase in accessible information/ the information people have access to has increased exponentially, which increases emphasis on having a method to reduce time for information gathering in an accurate and rapid manner to produce a concise and comprehensible summary (Allahyari et. al., 2017; Abualigah, Laith, Bashabsheh, Mohammad, 2020; El-Kassas et. al., 2021). Manual summarization was insufficient to deal with the volume of text generated by this change, in terms of cost and time (El-Kassas et. al., 2021).

For the reason stated above, automatic text summarization has been a major area of interest within the field of Natural Language Processing (NLP) since the 1950s (Allahyari et al., 2017).

Extensive research has made many discoveries in recent decades. Mitkov (2022) defined a summary as “a text that is produced from one or more texts that contains a significant portion of the information in the original text(s) and is no longer than half of the original text(s)”. ‘Text’ can be composed of media such as “speech, document, multimedia, hypertext, etc” (Sukmandhani et al., 2022). ‘Text’ in document format is also not limited to only a single document, it is also inclusive of multiple documents. Summaries can be classified as shown in the table below:

Types of summaries	**Definitions**
Indicative Summaries	an idea of what the text is about without giving much content
Informative Summaries	a shortened version of the content
Extracts	created by reusing portions (words, sentences, etc.) of the input text verbatim
Abstracts	created by regenerating the extracted content using new phrasing

Table 1: Definitions of types of summaries (Mitkov, 2022)

Text summarisation can be dealt with using multiple approaches. Thus far, previous studies have demonstrated that single document summarization and multiple document summarization could be done (Haque, Pervin & Begum, Widyassari et al., 2020). In single document summarisation, summaries are derived from one source document, revolving around the same topic. Multi-document summarization takes various sources that have the same interests and generates a concise summary from those documents. Both input types are equally important and have various functionalities for different scenarios (Widyassari et al., 2020). Previous research has established that single documents can produce summaries from sources such as lectures, scientific articles, shows, news articles (Nenkova, Mckeown, 2011). Initial stages of research on multi-document summarization demonstrated that summarising news articles clusters of an event to create a compilation of that event or for search engine results is possible, which gradually led to summarising search engine articles that are like the user's search result can be produced based on being of interest (Nenkova, Mckeown, 2011).

There are two fundamental approaches to produce summary outputs that are currently being adopted in research into automated text summarization (Abualigah et al., 2020). One is extractive text summarizατιον ανδ τηε other abstractive text summarization. Extractive summarization differs from abstractive summarization in the important way that they are created differently, with different methods, even if they both provide outputs (Madhuri, Kumar, Abualigah et al., 2020). Extractive summaries are produced by finding vital information within the text or document, such as sentences or expressions, based on linguistic and statistical features of the text. On the other hand, abstractive summarization paraphrases and restructures the content. That is done by conceptualising the key concepts and implications of the text or documents, which is completed by making use of linguistic methods.

Extractive summaries are more rigid in terms of producing summaries. Abstractive summarization is a significantly harder task as it requires a larger amount of natural language processing (Gothankar et al., 2022). However, abstractive summarization is much flexible and can provide increasingly diverse outputs, however it may not be able to contain content that follows the original text in a more grammatical way and may generate nonfactual content (Abualigah et al., 2020, Zhang, Liu & Zhang,). The issue in abstractive summarisation is a nonissue in the case of extractive summarization as sentences are directly chosen to be included in the summary, hence the grammar and context remains intact, following the original text (Zhang, Liu & Zhang,).

ChatGPT is a recent addition due to the advancement of AI, a large-scale language model developed and trained by OpenAI ((Zohery, 2023). It can interact with users with a humanlike way, conversing with users (OpenAI, 2022). According to OpenAI (2022) on a showcase of ChatGPT's abilities, it can carry out a multitude of general tasks, including answering questions, refusing to comply and provide a response towards potentially illegal activities, understanding and performing anaphora resolution throughout the conversation it holds with a user, and generating text content and responding to continuing questions. The goal of creating and develop a powerful tool like ChatGPT was to design an advanced and adaptable AI language model to carry out NLP tasks.

ChatGPT's versatile nature is due to it being a language model with an architecture style of Generative Pre-Trained Transformer (GPT), specifically the GPT-3.5 architecture ((Ray, 2023). This architecture style means that ChatGPT was developed to undertake NLP tasks, and ChatGPT has the potential to carry out much more advanced tasks.

2.1 Word Frequency Algorithm:

Word frequency algorithm was used as an attempt to create summaries from the dataset provided. Promising output was generated from using the weighted frequency method; however, it was insufficient as an algorithm to fulfil the project aims due to the nature of the algorithm.

Extensive research on weighted frequency have demonstrated a method to calculate weighted frequency. The weighted frequency of a term is calculated by dividing the occurrence of the term with the highest occurring word in the text (Abinayan,).

The formula for Weighted Frequency:

$$Wf = \frac{F_w}{F_{mow}}$$

Wf represents the weighted frequency of a term, Fw is the frequency of the current term and Fmow refers to the frequency of the most occurring term in the text.

An example sentence from the dataset provided as shown:

"Look no further than Akoko. To the credit of Akoko each dish was meticulously explained to us by our servers, who were more than happy to engage in conversation and respond to the geeky queries of your reviewer and his dining comrade."

Word	Word Count	Weighted Frequency
Look	1	1/3
no	1	1/3
further	1	1/3
than	1	1/3
Akoko	2	2/3
to	1	1/3
the	3	3/3
credit	1	1/3
of	1	1/3
each	1	1/3
dish	1	1/3
was	1	1/3
meticulously	1	1/3
explained	1	1/3
to	1	1/3
us	1	1/3
by	1	1/3
our	1	1/3

Table 2.1: Visualising Calculating Weighted Frequency

Word	Word Count	Weighted Frequency
servers	1	1/3
who	1	1/3
were	1	1/3
more	1	1/3
than	1	1/3
happy	1	1/3
engage	1	1/3
in	1	1/3
conversation	1	1/3
and	1	1/3
respond	1	1/3
geeky	1	1/3
queries	1	1/3
your	1	1/3
reviewer	1	1/3
his	1	1/3
dining	1	1/3
comrade	1	1/3

Table 2.1: Visualising Calculating Weighted Frequency

The sentence score is a sum of all the weighted frequencies of each term in a sentence, which is derived from the weighted frequency formula (Abinayan,).

The formula for Sentence Scores:

$$S_s = \sum_{1}^{w} Wf$$

Ss is the sentence score, and it is a summation of the total number of words w within that sentence.

Generating summaries with the word frequency method will require the average scores for all sentences within the text. This average is the threshold value, and it filters out sentences that are lesser than its own score from the summary and only allows sentences with higher scores. This value can be adjusted based on the requirements needed. The average sentence score can be found by dividing the total sentence scores (Abinayan,).

The formula for Threshold Value:

$$Threshold = \frac{\sum S_s}{S_t}$$

St is the total number of sentences, while Threshold is the average score of all sentences.

2.2 Term Frequency - Inverse Document Frequency (TF-IDF):

Term Frequency - Inverse Document Frequency, TF-IDF for short, was utilised as an alternative method used to summarise the text (M. S M et al., 2020). There are two parts to this method to summarise text.

Firstly, Term Frequency measures the occurrence of a word appearing in a document, text, paragraph or sentence ((M. S M et al., 2020). In this section's case, since the example provided has a length of two sentences, TF will be explained in terms of terms and sentence. In short, TF calculates the occurrence a word appears in the text. The frequent occurrence of a term might mean the term is significant, however if it appears too frequently in other texts or document, it is not a unique identifier, and a lower score will be provided to it.

The formula for Term Frequency:

$$Tf = \frac{count\ of\ t\ in\ s}{number\ of\ words\ in\ s}$$

Tf represents the term frequency of term t in sentence s, “count of t in s” is the number of occurrences of term t in sentence s, and “number of words in s” denotes the sum of the words in sentence s.

We will be reusing the example in 3.2.1.1:

Term	Frequency
Akoko	2
Look	1
further	1
To	1
credit	1
each	1
dish	1
meticulously	1
explained	1
us	1
servers	1
happy	1
engage	1
conversation	1
respond	1
geeky	1
queries	1
reviewer	1
dining	1
comrade	1

Table 3.1: Term Frequency Visualizer

IDF is Inverse Document Frequency, and it rectifies the overscoring of non-important yet frequent words such as stop words

The formula for Inverse Document Frequency:

$$IDf = \log\left(\frac{number\ of\ s\ in\ the\ corpus}{number\ of\ s\ in\ the\ corpus\ with\ t}\right)$$

IDf represents the inverse document frequency, which allocates higher values to rarer words and lower values to words that are more common in the sentences s.

Tf and IDf are paired together as Tf would miscount words such as stop words and provides higher scores to it, while IDf identifies words which are rare occurrences allocates higher scores to it whereas stop words are scored lower due to IDf.

Term	IDF
Akoko	log (10/1) = 1
Look	log (10/1000) = -2
further	log (10/1000) = -2
To	log (10/1000) = -2
credit	log (10/1000) = -2
each	log (10/1000) = -2
dish	log (10/1000) = -2
meticulously	log (10/1000) = -2
explained	log (10/1000) = -2
us	log (10/1000) = -2
servers	log (10/1000) = -2
happy	log (10/1000) = -2
engage	log (10/1000) = -2
conversation	log (10/1000) = -2
respond	log (10/1000) = -2
geeky	log (10/1000) = -2
queries	log (10/1000) = -2
reviewer	log (10/1000) = -2
dining	log (10/1000) = -2
comrade	log (10/1000) = -2

Table 3.2: Inverse Document Frequency Visualiser

Term Frequency- Inverse Document Frequency (TF-IDF) can be calculated using the scores from TF and IDF.

$$Tf - IDf = Tf \cdot IDf$$

Term	IDF	TF-IDF
Akoko	2	1
Look	1	-2
further	1	-2
To	1	-2
credit	1	-2
each	1	-2
dish	1	-2
meticulously	1	-2
explained	1	-2
us	1	-2
servers	1	-2
happy	1	-2
engage	1	-2
conversation	1	-2
respond	1	-2
geeky	1	-2
queries	1	-2
reviewer	1	-2
dining	1	-2
comrade	1	-2

Table 3.3: TF - IDF

2.3 Rapid Algorithmic Keyword Extraction (RAKE):

Rapid Algorithmic Keyword Extraction (RAKE) is an algorithm that extract keywords from individual documents, and it is an “unsupervised, domain-independent and language-independent method for extracting keywords from individual documents” (Rose et al.,). RAKE starts keyword extraction by getting a set of candidate keywords, which are words that are split into individual words using specific word delimiters

Steps to compute RAKE include:

Calculate Word Co-occurrence

Word	Co-occurrences
Look	[no, further, than]
no	[Look, further, than]
further	[Look, no, than]
than	[Look, no, further, Akoko]
Akoko	[than, credit, each]
To	[the, credit, explained]
the	[To, credit, explained]
credit	[To, the, Akoko, each]

Table 4.1: RAKE

Calculate Word Degree (Number of co-occurrences)

Word	Degree
Look	3
no	3
further	3
than	4
Akoko	3
To	3
the	3
credit	4

Table 4.2: RAKE

Calculate Keyword score

Word	Frequency	Degree	RAKE Score
Look	1	3	3
no	1	3	3
further	1	3	3
than	1	4	4
Akoko	2	3	1.5
To	1	3	3
the	1	3	3
credit	1	4	4

Table 4.3: RAKE

Rank Keywords

Word	RAKE Score
than	4
credit	4
Look	3
no	3
further	3
To	3
the	3
Akoko	1.5

Table 4.4: RAKE

3. Methodology

The next chapter describes the procedures and methods used in this investigation. This section will be split into three subsections: Dataset, ChatGPT 3.5 and Implementation. Dataset will describe the type of dataset used for this study, and the selection and editing process. ChatGPT 3.5 is included as a subsection as it is an integral part of this study. Implementation will introduce the basic libraries and software used for this project and RAKE will be

3.1 Dataset:

The dataset is collected and edited from various restaurant review blogs or websites, where the reviewers recorded their experience dining in the restaurants with words and some even attached images of the food and beverage they were served while dining at the restaurants.

The dataset consists of a maximum of 50 sentences and has at least 20 sentences. Some examples of websites and blogs include Major Foodie, GourmanGunno, The Foodaholic, The Graphic Foodie, Stefan's Gourmet Blog, Andy Hayler, A Girl has to Eat, etc. These websites were chosen for the subject and information included within as the dataset is limited for restaurant reviews. As the project focuses on text summarisation, the images and additional text labels on the images were removed. A total of 50 pieces of text were collected and recorded into an excel sheet, which was then converted to a .csv file.

3.2 ChatGPT 3.5

For this project, ChatGPT 3.5 (ChatGPT) was utilised as it is Open-Source and has minimal requirements for usage, requiring either email logging or Google Account logging. Both logging methods were used as ChatGPT has a limit of 30 prompts per hour. For this study, ChatGPT was used to provide extracts for further experimenting. As mentioned in the introduction section, this project aims to investigate the capabilities of ChatGPT when undertaking extractive summarization tasks.

To generate extractive summaries, the following prompt was inputted to ChatGPT: 'Please extract word-by-word, the 15 most important sentences from the text provided'. Generating 15 sentences from ChatGPT was crucial. These sentences are deemed the most important sentences in the text by ChatGPT, however, to check the scoring, more sentences were required to be inputted for further processing.

3.3 Implementation:

This is an NLP text summarization project, hence, to implement the program, Python 3, and libraries such as Natural Language Toolkit (NLTK) was used (NLTK Team, 2023). The NLTK website boasts that the library is at the forefront of the platform to develop python programs for human language data. NLTK covers over "50 corpora and lexical resources", alongside libraries for text processing which are essential in NLP tasks. The main reason Python 3 and NLTK were both used is because both are free and open source, as it is maintained by a community.

3.4 Rake-based Extractive Summarization (RAKE)

As mentioned in the background section, RAKE can extract key phrases with scores. This algorithm is much suitable due to its ability to extarct ke.

For this project, the dataset was stored in a .csv file. To generate coherent sentences, stop word preprocessing was not done to the dataset, instead text was split into words and rejoined. Preprocessing of data such as original text and the summary from ChatGPT was merely done to separate data into different columns.

RAKE was used to identify keywords or phrases of variable lengths from the original text and ChatGPT summaries. Scores of the phrases were then stored into dictionaries alongside the keywords. This was done to ensure the keyword or phrases with the highest score appears in the final summary.

After extracting the keyword-score pairs, the top ten sentences with the highest scores will be chosen as summary. The sentences are sorted based on the sum of each word in the sentence, hence sentences containing words or phrases with higher scores have a higher chance of appearing in the summary.

To sum it all up, the general steps taken in the program is listed below as follows:

1. Separating data into different columns.
2. Extract and score keywords using RAKE from both the original text and ChatGPT summary provided.
3. Store the extracted keywords and scores.
4. Sort sentences according to the weightage of the sentence, which is the total of all the words in the sentence.
5. Print out the summary.

3.5 Evaluation Methods:

In this section, methods used to carry out the evaluation process and findings from the results will be presented. As this project revolves around text summarisation, human evaluation was

used to gather results. Krippendorff's alpha (α), an evaluation method, was carried out to test for inter-annotator agreement between participants. Krippendorff's alpha was used to ensure the evaluation of results were robust and accurate, by having multiple annotators to evaluate the summaries generated by the program (Krippendorff, 2011).

3.5.1 Human Evaluation: Questionnaire

A questionnaire with ten questions was created to investigate participant's responses. Using a random number generator, ten texts were chosen from the dataset, and alongside the text, the two summaries generated from the program would be blindly set and labelled as 'Summary 1' and 'Summary 2'. For the questionnaire, the four metrics used were informativeness, readability, grammatically and coherence. Definitions for the tested metrics were included in the start of the questionnaire to allow participants to have a clearer understanding of the metrics used and criteria used for scoring.

Definitions provided in the questionnaire for the metrics are as such:

- Informativeness: Extent to which the summary relays important information.
- Readability: How easy it is to read and understand the summary.
- Grammatically: Extent to which the sentences in the summary follow the rules of English Grammar.
- Coherence: The extent to which the structure of the summary is organized.

This questionnaire utilises Likert scales, with a five-point scale and labels to inform and instruct participants to fill out and complete the form appropriately. These are the labels provided in the questionnaire: 'Excellent', 'Good', 'Fair', 'Poor' and 'Very Poor.'

A total of six participants were recruited: three males and three females. The participants were recruited from social media and word of mouth, and they have varying ages.

3.5.2 Krippendorff's Alpha (α): Inter-annotator agreement

Krippendorff's alpha (α) also known as reliability coefficient, as explained by Krippendorff (2011), is a method to calculate the extent to which individuals or instruments agree with each other when evaluating or assigning values to unstructured phenomena; in short, it is an inter-rater reliability scorer. Individuals can have roles such as "observers, coders, judges, raters, and annotators." This reliability coefficient was initially used in content analysis, and it has been used in other fields as it is applicable to collect data of various data types and can expand to an arbitrary number of raters ((Krippendorff, 2011). Krippendorff's alpha is widely used because the calculated information is useful to analyse. The range of the reliability coefficient is between -1 and 1, where one represents total agreement among annotators and -1 showing perfect disagreement between annotators ((Krippendorff, 2011). This metric will be used to calculate the inter-annotator agreement to observe any agreement within participants in the latter section.

Krippendorff's Alpha is usually calculated using the ratio between the observed disagreement among raters (disagreeo) and the disagreement expected by chance (disagreee).

The formula of Krippendorff's Alpha:

$$\alpha = 1 - \frac{disagree_o}{disagree_e}$$

α represents the reliability coefficient and it is computed by the ratio of weighted percent agreement (disagreeo) and weighted percent chance agreement (disagreee) deducted from 1.

4. Evaluation and Results

In the previous chapter, the methodology for the code and for the project was introduced the process and methods used to attain the results for this project. The section that follows establishes the framework for discussion and analysis of the results.

4.1. Code

The output for the program presents a pair of summaries for each piece of data inserted into the program. Each summary contains 10 sentences of variable length, and each row has 2 summaries. As demonstrated in the examples of the list of sentences below, two different rows have drastically different types of output, and the summaries generated from the original text and ChatGPT also have individual differences.

```
Top 10 Sentences with the Highest Scores for Row 10:
RAKE Summary:
['Although trained in imperial Chinese cuisine, his restaurant served Sichuan food, and he popularised this style of
cooking in Japan via appearances in various TV shows.', 'This restaurant is part of a group, the Singapore branch being
located on the 35th floor of the Mandarin Orchard hotel.', 'The Singapore branch, opened in 2014, has a third generation of
the family in charge of the kitchen, Chen Kentaro.', 'The cavernous dining room has a high ceiling and an impressive array
of chandeliers.', 'By contrast, the group’s restaurants in Japan, such as the Fukuoka branch and the Tokyo branch in
Akasaka, remain a Michelin free zone.', 'The restaurant was given two Michelin stars in the inaugural Michelin Guide to
Singapore in 2016.', 'The other thirteen branches (at the time of writing) are in Japan, spread across Fukuoka, Matsuyama,
Nagoya, Takamatsu, Tokyo and Yokohama, where the restaurant empire began.', 'Tables are large and well spaced.', 'The group
was founded in 1958 by Chen Kenmin, a Chinese immigrant.', 'His son Chen Kenuchi appeared regularly in the famous TV series
Iron Chef.']


ChatGPT Summary:
['Bak choy was lightly cooked and tender, while a spicy dry noodle dish with minced meat had excellent texture and
flavor.', 'The menu is vast, offering Peking duck, xiao long bao, and a variety of Cantonese dishes, with the Sichuan
influence hidden in the menu descriptions.', 'This restaurant is part of a group, with branches in Japan and Singapore,
where the restaurant empire began.', 'Spicy prawns were well-cooked and flavorful, served with cashew nuts and dried red
peppers with a hint of Sichuan peppercorn.', 'The xiao long bao and siew mai were pleasant, with reasonable texture and
good fillings.', 'The cavernous dining room has a high ceiling and an impressive array of chandeliers, with large and well-
spaced tables.', 'The group was founded in 1958 by Chen Kenmin, a Chinese immigrant known for popularizing Sichuan food in
Japan.', 'Hokkaido scallops were disappointing, overcooked and lacking in flavor due to clunky cooking and spicy topping.',
'The Singapore branch, opened in 2014, was given two Michelin stars in the inaugural Michelin Guide to Singapore in 2016.',
'Crab and sweet corn soup was a superior version, with a generous amount of crab and a thick, broth-like texture.']
```

Image 1: Example output for a row (Row 10)

```
Top 10 Sentences with the Highest Scores for Row 40:
RAKE Summary:
['The £181 for one (with sake) dinner price tag was more digestible than most high-end Kaseiki restaurants in Japan and
although it was a good experience to explore Osaka, there were not enough high-end moments to making me want to leap back
for a return visit here.', 'Frequent readers of mine will know that I have no qualms in saying whether I believe Michelin
are pitching too high or low on a verdict, but in this case I very much agree that this experience did not merit the
ultimate accolade, certainly nowhere near the comparison to Kitcho or Mizai and sadly, I actually questioned whether the 2
stars was generous.', 'At the time of my visit this restaurant was graded as a 3 Michelin starred restaurant but, in a
follow-up, it lost a star in the 2020 guide two months later.', 'This restaurant is run by head chef Shintaro Matsuo who
was friendly but could not engage too much with diners owing to the language barrier.', 'What was a nice touch was his
staff members running around the restaurant showing pictures of the fish and produce all were eating on an iPad and
explaining where possible.', 'Bonito with aubergine and roe sauce was our first bite (the same menu given to all diners at
the same time) and this was a pleasurable snack.', 'You could do a lot worse of course and this would serve as a useful
‘beginner’ Kaiseki venue.', 'In 2020 the Michelin guide pronounced three 3 Michelin starred restaurants within Osaka
(Hajime, Taian and Kashiwaya).', 'Koryu now slips into the 2 Michelin starred family of which, there are an impressive 15
restaurants of this category in Osaka alone.', 'Before all diners received their appetisers, several large shrimps heads
were placed on a grill to gently cook the brains.']


ChatGPT Summary:
['The £181 for one (with sake) dinner price tag was more digestible than most high-end Kaiseiki restaurants in Japan and
although it was a good experience to explore Osaka, there were not enough high-end moments to making me want to leap back
for a return visit here.', 'At the time of my visit this restaurant was graded as a 3 Michelin starred restaurant but, in a
follow-up, it lost a star in the 2020 guide two months later.', 'Crab with vinegar jelly, shiitake mushroom and berries was
wonderful – a light sweetness to the jelly with beautifully smoked mushroom and very well done.', 'The tuna with slow-
cooked egg yolk, wasabi and soy was one of the rare stunning moments of this though.', 'The sashimi platter was beautifully
presented and included yellowfin with egg yolk, herring, squid with plum sauce, beef noodle, snapper and shrimp and
purified saltwater.', 'The squid was lovely with plum sauce but texturally was quite hard and not as massaged as it was in
Mizai (bit hard in comparison) by a long shot.', 'In 2020 the Michelin guide pronounced three 3 Michelin starred
restaurants within Osaka (Hajime, Taian and Kashiwaya).', 'Koryu now slips into the 2 Michelin starred family of which,
there are an impressive 15 restaurants of this category in Osaka alone.', 'Before all diners received their appetisers,
several large shrimps heads were placed on a grill to gently cook the brains.', 'The soup contained grouper and matsutake
mushrooms (very chewy but with a strong mushroom aroma).']
```

Image 2: Example output for a row (Row 40)

The RAKE summary for Row 10 focuses mainly on the restaurant and its history and accomplishments throughout the years while the ChatGPT summary pivots on the dishes and provided multiple examples of the types of food served. The similarity between these two summaries was about the building structure, furnishings, restaurant’s history and Michelin star rating.

The RAKE summary for Row 40 focuses on the detailed experiences of the writer, such as noting down the star rating for the restaurant and the opinions of the writer regarding the food while the ChatGPT summary heavily emphasises on food and atmosphere while dining.

4.1.2 Human Evaluation: Questionnaire

As mentioned previously, six annotators were recruited to evaluate the summaries. Their responses were recorded, as shown in the tables below. There were an equal number of male and female participants, to differentiate between the genders, the first three response slots were allocated to male participants, while the last three were allocated to female participants.

According to the male annotators’ responses, the ratings that was chosen more frequently was ‘Good’ and ‘Fair’, following up with equal number of ‘Excellent’ ratings in both summaries. None chose ‘Very Poor’ as a rating, and ChatGPT summaries were given five more ‘Poor’ ratings. This is seen in the Informativeness metric, where P1 provided more frequent scores in the ChatGPT summaries.

Summary Types	Excellent(M)	Good(M)	Fair(M)	Poor(M)	Very Poor(M)	Total Scores

ChatGPT summary	21	44	37	18	0	120
RAKE summary	21	39	47	13	0	120
Total Scores	42	83	84	31	0	240

Table 5.1: Table containing the accumulated scores of male participants

Summary Types	**Excellent(M)**	**Good(M)**	**Fair(M)**	**Poor(M)**	**Very Poor(M)**	**Total Percentage**
ChatGPT summary	17.5	36.67	30.83	15	0	100
RAKE summary	17.5	32.5	39.17	10.83	0	100
Total Percentage	35	69.17	70	25.83	0	200

Table 5.2: Table containing the average scores between male participants

Scores were provided for each rating level to calculate the aggregation scores to check for the differences between responses of the annotators for both summaries.

Rating level of ‘Excellent’ was given a score of FIVE, ‘Good’ was allocated a score of FOUR, ‘Fair’ was provided a score of THREE, ‘Poor’ was given a score of TWO, ‘Very poor’ was allocated a score of ONE.

Although the there was a ten-count difference between the ‘Fair’ rating between ChatGPT and RAKE, the final aggregation score for ChatGPT and RAKE was the same due to the five-count difference.

Aggregation Scores for ChatGPT summary = (21x5) + (44x4) + (37x3) + (18x2) + (0x1) = 428
Aggregation Scores for RAKE summary = (21x5) + (39x4) + (47x3) + (13x2) + (0x1) = 428

Summary Types	**Aggregation Scores**
ChatGPT summary	428
RAKE summary	428

Table 5.3: Table containing aggregation scores for both summaries

While male annotators had no difference in aggregation scores for the summaries, there was a difference in the female category. There was a significant larger number of 'Excellent' ratings for ChatGPT summaries and an increase in the numbers for the 'Very Poor' rating. P5 had awarded the four of the ratings towards the summaries of ChatGPT. According to the overall results, most women rated 'Excellent' towards ChatGPT's summaries while RAKE summaries often received rating of 'Good'.

Summary Types	**Excellent(F)**	**Good(F)**	**Fair(F)**	**Poor(F)**	**Very Poor(F)**	**Total Scores**
ChatGPT summary	47	27	21	21	4	120
RAKE summary	33	40	30	15	2	120
Total Scores	80	67	51	36	6	240

Table 6.1: Table containing the accumulated scores of female participants

Summary Types	**Excellent(F)**	**Good(F)**	**Fair(F)**	**Poor(F)**	**Very Poor(F)**	**Total Percentage**
ChatGPT summary	39.17	22.5	17.5	17.5	3.33	100
RAKE summary	27.5	33.33	25	12.5	1.67	100
Total Percentage	66.67	55.83	42.5	30	5	200

Table 6.2: Table containing the average scores between female participants

Rating level of 'Excellent' was given a score of FIVE, 'Good' was allocated a score of FOUR, 'Fair' was provided a score of THREE, 'Poor' was given a score of TWO, 'Very poor' was allocated a score of ONE.

Female annotators scored ChatGPT a lot higher due to a larger number of 'Excellent' scores.

Aggregation Scores for ChatGPT summary = (47x5) + (27x4) + (21x3) + (21x2) + (4x1) = 452
Aggregation Scores for RAKE summary = (33x5) + (40x4) + (30x3) + (15x2) + (2x1) = 447

Summary Types	Aggregation Scores
ChatGPT summary	452
RAKE summary	447

Table 6.3: Table containing aggregation scores for both summaries

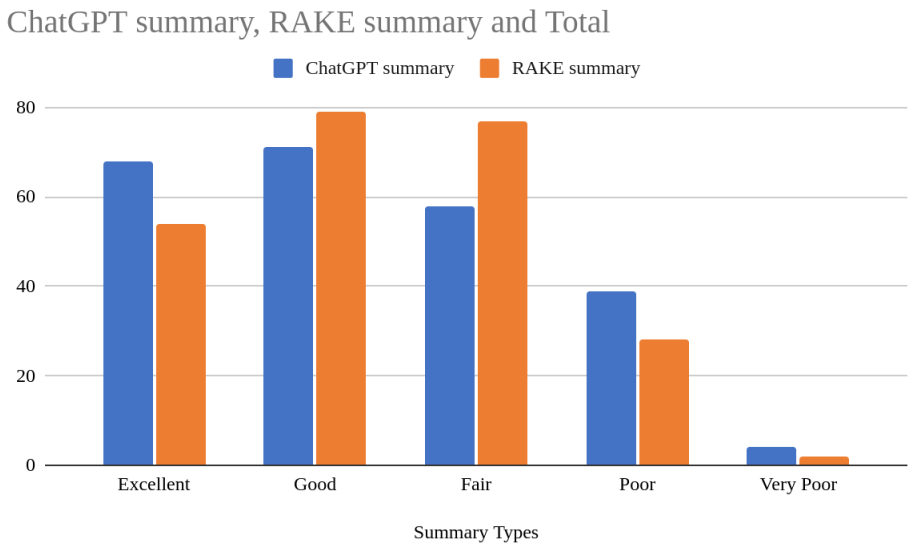

Figure 1: Bar Chart. This chart shows the number of times a score was chosen for all four metrics for all genders.

Summary Types	Excellent	Good	Fair	Poor	Very Poor	Total
ChatGPT summary	68	71	58	39	4	240
RAKE summary	54	79	77	28	2	240
Total	122	150	135	67	6	480

Table 7: Compilation of the occurrence of scores in table format, exclusive of gender

As seen on the table above, annotators mainly pick ‘Good' and ‘Fair’ participants mainly chose the option of ‘Good’, and the least of ‘Very Poor’, and this is reflected in both types of summaries.

4.1.3 Krippendoff’s Alpha

As mentioned in the methodology section, Krippendorff's alpha would be used to calculate the reliability of the responses received from the participants for RAKE summaries and ChatGPT summaries. For this section of analysis, different approaches were taken to analyse the results. The first method was to check for agreement between genders. It was found that there was no agreement between all four metrics in both summaries for female annotators, while male participants mostly agreed with each other, except for Informativeness in ChatGPT summaries and a slight disagreement of 0.006 in Coherence for the RAKE summaries.

	Informativeness	Readability	Grammaticality	Coherence
Male	-0.125	0.006	0.141	0.112
Female	-0.222	-0.295	-0.294	-0.347

Table 7: Krippendorff’s alpha scores for ChatGPT summaries, between male and female annotators

	Informativeness	Readability	Grammaticality	Coherence
Male	0.191	0.023	0.168	-0.006
Female	-0.301	-0.197	-0.221	-0.11

Table 8: Krippendorff’s alpha scores for RAKE summaries, between male and female annotators

When gender was combined to check for inter-annotator agreements, there was mostly no agreement between annotators, which was due to the larger number of annotators involved. The only agreement between annotators were on the Readability metric, with a value of 0.023.

Informativeness	Readability	Grammaticality	Coherence
-0.074	-0.089	-0.047	-0.064

Table 9: Krippendorff’s alpha scores for ChatGPT summaries, between all annotators

Informativeness	Readability	Grammaticality	Coherence
-0.015	0.023	-0.024	-0.079

Table 10: Krippendorff’s alpha scores for RAKE summaries

5.1 Limitations

This project focuses on creating extractive summaries for Restaurant Reviews, hence the text used is unstructured, hence no summaries could be used to conduct further analysis for the dataset.

5.2 Future Work

Gender and annotator bias could be done more in further research.

Conclusions:

This project's overall aim is to investigate the capabilities of ChatGPT 3.5 when undertaking extractive text summarization tasks. This project partially fulfilled the requirements needed.

Appendices

References

Abinayan (a) *Extractive Text Summarization Using Word Frequency Algorithm for English Text*.

Abinayan (b) *Extractive Text Summarization Using Word Frequency Algorithm for English Text*.

Abualigah, L., Bashabsheh, M., Alabool, H., et al (2020) Text Summarization: A Brief Review.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K. & Trippe, E.B.D. (2017) *Text Summarization Techniques: A Brief Survey*.

El-Kassas, W.S., Salama, C.R., Rafea, A.A. & Mohamed, H.K. (2020) *Automatic text summarization: A comprehensive survey*, Elsevier BV.

Gothankar, A., Gupta, L., Bisht, N., Nehe, S. & Bansode, P.M. (2022) *Extractive Text and Video Summarization using TF-IDF Algorithm*, International Journal for Research in Applied Science and Engineering Technology (IJRASET).

Haque, M.M., Pervin, S. & Begum, Z. *Literature Review of Automatic Single Document Text Summarization Using NLP*.

Krippendorff, K. (2011) *Computing Krippendorff's Alpha-Reliability*.

M. S M, R. M P, A. R E & E. S. G SR (2020) *Text Summarization Using Text Frequency Ranking Sentence Prediction,* 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP).

Madhuri, J.N. & Kumar, G.R. *Extractive Text Summarization Using Sentence Ranking*.

Mitkov, R. (2022) *The Oxford Handbook of Computational Linguistics.* 2nd ed. ed.

Nenkova, A. & Mckeown, K. (2011) *Automatic Summarization*, Now Publishers.

NLTK Team (2023). *NLTK 3.8.1.* Available at: https://www.nltk.org/index.html .

OpenAI (2022). *Introducing ChatGPT.* Available at: https://openai.com/blog/chatgpt .

Qaiser, S. & Ali, R. (2018) Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications,* 181(1), 25-29 google.

Ray, P.P. (2023) *ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope*, Elsevier BV.

Rose, S., Engel, D., Cramer, N., Cowley, W., Berry, M.W. & Kogan, J. *Automatic keyword extraction from individual documents*.

Sukmandhani, A.A., Ramadhan, A., Abdurachman, E. & Trisetyarso, A. (2022) *Single and Multi-Documents Text Summarization Technologies for Natural Language Processing: a Systematic Review on Method and Dataset*, IEEE.

Widyassari, A.P., Rustad, S., Shidik, G.F., Noersasongko, E., Syukur, A., Affandy, A. & Setiadi, D.R.I.M. (2020) *Review of automatic text summarization techniques & methods*, Elsevier BV.

Zhang, H., Liu, X. & Zhang, J. *Extractive Summarization via ChatGPT for Faithful Summary Generation*.

Zohery, M. (2023) *ChatGPT in Academic Writing and Publishing_ A Comprehensive Guide 2023*