

**KULLIYYAH OF INFORMATION & COMMUNICATION TECHNOLOGY**

CSCI 4342 - NATURAL LANGUAGE PROCESSING

SEMESTER 1 - 2021/2022

SECTION 1

TEXT SUMMARIZER

TECHNICAL REPORT

|  |  |
| --- | --- |
| **NAME** | **MATRIC NO** |
| Mohamad Faisal Bin Mohd. Hanafi | 1915045 |
| Muhammad Syazmi Bin Suhaidi | 1814573 |
| Md. Rakibul Hassan | 1720465 |
| Md. Najmul Huda | 1627521 |

**LECTURER NAME:** Dr. Suriani Bt. Sulaiman

**SUBMISSION DATE:** 28/01/2022

**1.0 Introduction 3**

**2.0 Problem Statement** [**4**](https://docs.google.com/document/d/1ptjxkBVQBJR2fTCB0rtatDyiZwcbC1pmqq7HW1GaaDM/edit#heading=h.819ptobh6l2q)

[**3.0**](https://docs.google.com/document/d/1ptjxkBVQBJR2fTCB0rtatDyiZwcbC1pmqq7HW1GaaDM/edit#heading=h.fg9rxroqmvyq) **Motivation** [**5**](https://docs.google.com/document/d/1ptjxkBVQBJR2fTCB0rtatDyiZwcbC1pmqq7HW1GaaDM/edit#heading=h.fg9rxroqmvyq)

[**4.0**](https://docs.google.com/document/d/1ptjxkBVQBJR2fTCB0rtatDyiZwcbC1pmqq7HW1GaaDM/edit#heading=h.f7n0jfe797yk) **Related Work** [**6**](https://docs.google.com/document/d/1ptjxkBVQBJR2fTCB0rtatDyiZwcbC1pmqq7HW1GaaDM/edit#heading=h.f7n0jfe797yk)

[**5.0**](https://docs.google.com/document/d/1ptjxkBVQBJR2fTCB0rtatDyiZwcbC1pmqq7HW1GaaDM/edit#heading=h.oqukh79085ae) **Technical Background** [**7**](https://docs.google.com/document/d/1ptjxkBVQBJR2fTCB0rtatDyiZwcbC1pmqq7HW1GaaDM/edit#heading=h.oqukh79085ae)

[**6.0**](https://docs.google.com/document/d/1ptjxkBVQBJR2fTCB0rtatDyiZwcbC1pmqq7HW1GaaDM/edit#heading=h.xk3xvls8cgkm) **Approach Description** [**8**](https://docs.google.com/document/d/1ptjxkBVQBJR2fTCB0rtatDyiZwcbC1pmqq7HW1GaaDM/edit#heading=h.xk3xvls8cgkm)

[**7.0**](https://docs.google.com/document/d/1ptjxkBVQBJR2fTCB0rtatDyiZwcbC1pmqq7HW1GaaDM/edit#heading=h.v5m9grhuro6l) **Experimental Setup 10**

[**8.0**](https://docs.google.com/document/d/1ptjxkBVQBJR2fTCB0rtatDyiZwcbC1pmqq7HW1GaaDM/edit#heading=h.ut2wmrwf7no5) **Experimental Implementation** [**15**](https://docs.google.com/document/d/1ptjxkBVQBJR2fTCB0rtatDyiZwcbC1pmqq7HW1GaaDM/edit#heading=h.ut2wmrwf7no5)

[**9.0**](https://docs.google.com/document/d/1ptjxkBVQBJR2fTCB0rtatDyiZwcbC1pmqq7HW1GaaDM/edit#heading=h.57oh37fi7eqd) **Results 16**

**10.0 Error Analysis 19**

**11.0 Reference 23**

## **Introduction**

Summarization's objective is to extract critical information from enormous amounts of text and present it in a concise, representative, and consistent summary. A well-written summary may greatly lessen the effort required to absorb massive quantities of material on a particular subject. Manually summarizing lengthy text documents is a significant challenge for humans. Text summarizing is the act of spontaneously developing and compressing the format of a given record while preserving its information source into a more compact adaptation of significant length. Text summarization is one of the most important study areas in natural language processing today and is likely to get further interest from NLP researchers in the future.

We demonstrate how to use the spacy library and extractive summarization to the issue of multi-document summarizing in this project. Spacy library and extractive summarization, abbreviated for frequency-inverse document frequency, is a numerical metric used to quantify the value of a word in a document based on its frequency of occurrence in that document and a particular collection of documents. The logic behind this metric is that if a term occurs often in a text, it must be significant, and so deserves a high score. However, if a term occurs in an excessive number of other texts, it is unlikely to be a unique identifier, and hence deserves a lower score. Our tests will demonstrate that optimizing the performance of a state-of-the-art summarizing framework strongly implies the spacy library and extractive summarization are advantageous for this project.

## **Problem Statement**

Nowadays, everyone is aware of text summarization and understands how critical it is to their everyday life. When someone attempts to create it manually, it is not a simple process. Utilizing text summarization, you may quickly do a literature review, an analysis, or a paper review. It is a well-established fact that present abstractive text summarization methods often create incorrect data. This may occur at the entity level (new entities are created) or at the entity relation level (the context in which entities occur is incorrectly generated). This article assesses factual consistency exclusively at the entity level, leaving relationship-level consistency to future research. They suggest a metric for quantifying the model's hallucinations, as well as a few measurements and training techniques for improving the model's performance and producing factually true entity-level summaries. Customer testimonials are often lengthy and comprehensive. As you may guess, carefully analyzing these evaluations is somewhat time intensive. This is where natural language processing's genius may be employed to provide a summary for lengthy evaluations.

## **Motivation**

## Text summarizing (TS) is the process of extracting the most critical information from a document or collection of related documents and expressing it in a fraction of the space (usually by a ratio of five to ten) of the original text. Utilizing the background and related work, it was determined that a critical component of current research and projects was a lecture summarizing service that could be used by students with variable lecture sizes while using the most recent deep learning technology. This finding inspired the creation of the lecture summary service, a cloud-based service that performed inference from the spacy library and extractive summarization model in order to provide dynamically scaled lecture summaries. Several causes for text summarizing include the following:

1. Individuals become informed about international events through listening to the news.
2. Individuals make investing selections based on market updates.
3. Even still, many go to movies primarily based on the reviews they've read.
4. Summaries enable consumers to make more informed judgments in less time.

The goal here is to develop a program that is computationally efficient and automatically generates summaries.

## **Related Work**

To contextualize the suggested solution of automated lecture summarizing, it is useful to review prior research, noting the advantages and disadvantages of each technique. Many multimedia apps provided manual summaries for each lecture in the early days of lecture searching. One such example comes from M.I.T.'s lecture processing project, which uploaded a significant number of lectures, complete with transcripts for keyword searching and a synopsis of the course's content (Glass, Hazen, Cyphers, Malioutov, Huynh, & Barzilay, 2007). This strategy may serve for small amounts of material, but as the data becomes larger, the human summary process may become wasteful.

In the mid-2000s, one reason for manual summation was the low quality of extractive summary tools. In 2005, researchers developed a technique for automatically extracting business meeting summaries using basic probabilistic models but rapidly discovered that the output was much inferior to summaries provided by humans (Murray, Renals, & Carletta, 2005). Due to the methodology's poor performance, multiple research articles were written to enhance the procedure.

## **Technical Background**

There is a massive amount of textual information available, and it continues to develop every day. In the world of the Internet, where web pages, news articles, status updates, blogs, and so much more. Because the data is unstructured, the best way is to go through and search all the data just to find the results. Therefore, it is a great method to reduce much of this text data to short, focused summaries that capture the key features so that users can explore it more easily and ensure that the larger documents include the information needed. Stated by (), the 6 reasons why the community needs automatic text summarization tools

1. Summarizing reduces reading time.
2. When researching documents, summaries make the selection process easier.
3. Automatic summarization improves the effectiveness of indexing.
4. Automatic summarization algorithms are less biased than human summarizers.
5. Personalized summaries are useful in question-answering systems as they provide personalized information.
6. Using automatic or semi-automatic summarization systems enables commercial abstract services to increase the number of texts they are able to process.

While the world keeps on evolving, text summarization has also been widely used. Deep learning algorithms for text summarization have recently shown promising results. Text summarization has been framed as a sequence-to-sequence (Seq2Seq) algorithm, which has been motivated by the application of deep learning approaches for automatic machine translation.

## **Approach Description**

For this project, we decided to use Text Summarization through the use of the spaCy library. SpaCy is a free and open-source advanced natural language Python and Cython. The main reason we used spaCy is that spaCy offers a syntactic analysis that is fast and accurate, as well as named entity recognition and access to word vectors. The spaCy library comes with:

* tokenization,
* sentence boundary detection,
* POS tagging,
* syntactic parsing, integrated word vectors,
* alignment into the original string with accuracy.

Usually, projects of text summarization can largely be divided into two categories; **Extractive Summarization** and **Abstractive Summarization.**  We have decided to approach our project with Extractive Summarization. These methods rely on extracting numerous components from a piece of text, such as phrases and sentences, and stacking them together to form a summary. As a result, in an extractive technique, identifying the appropriate sentences for summary is essential.

## **Experimental Setup**

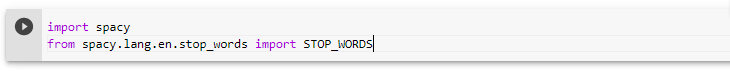
1. System Development

The tools that we decided to use for this project are Google Colab or Jupyter Notebook. These two are original web applications for creating and sharing computational documents. Both websites allow anybody to write and execute arbitrary python code through the browser and are especially well suited to data analysis.

To kick off the project the first thing is to install the spaCy Library.



After completing the installation process of the spaCy make sure to import the library and also import stop words



1. Article Selection

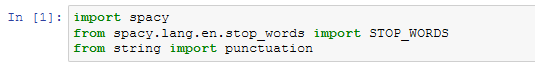
We believed that in the pursuit of producing a good model the utilization of high-quality data is essential in the training and testing phase. For this process, we had obtained six different literatures in terms of the type, genre, complexity, size, and the source which are all obtained from. Here are the differences of the article:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Title | Type | Complexity | Genre | Size | Source |
| Automatic text summarization techniques & methods | Web Article | Medium | Education | 1545 words | <https://www.sciencedirect.com/science/article/pii/S1319157820303712> |
| Who are the best up-and-Coming Coaches in European Football? | Blog | Low | Sport | 3209 words | <https://www.sportsblog.com/uksoccer/who-are-the-best-up-and-coming-coaches-in-european-football/> |
| History of light novel | Encyclopedia | Medium | Fiction | 7797 words | <https://en.wikipedia.org/wiki/Light_novel> |
| ‘Janet Jackson’ tells the singer's story, but it's clear who's in control | News Article | Low | Entertainment | 3217 words | <https://edition.cnn.com/2022/01/28/entertainment/janet-jackson-documentary-review/index.html> |
| Classification of Ice in Lutzow-Holm Bay | Journal Article | High | Education | 13214words | <https://www.mdpi.com/2072-4292/12/19/3179> |
| State and local public policies | Research Paper | High | Education | 15326 words | <https://ivypanda.com/essays/state-and-local-public-policies/> |

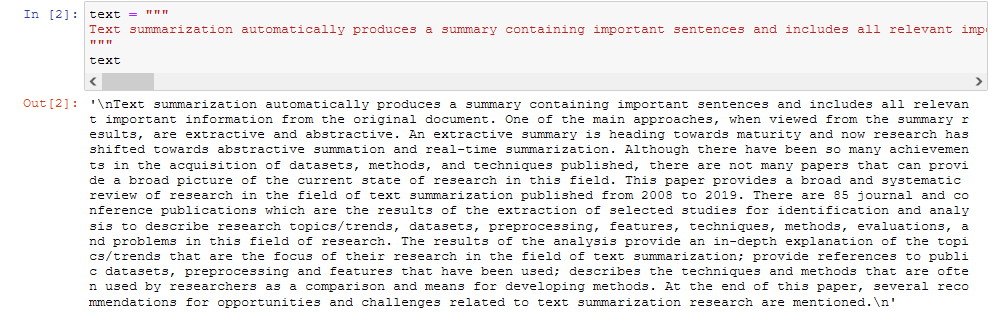
Table 1: Articles Difference

1. **Experimental Implementation**
2. *Data Cleaning*

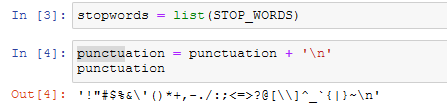
The first step is to import the library that will be utilized, in this case, we will be importing the spaCy library.



Next, we will be storing the text that will be summarized later into a variable, for this we had chosen to name the variable as “text” variable.



Then, we will import the punctuation marks from the string and also add the additional next line tag into it.



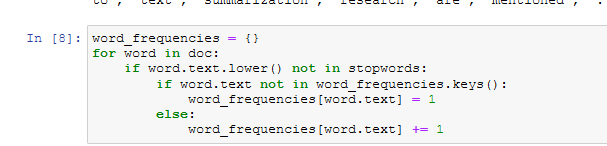
1. *Initialize Tokenization*

We will then start the tokenization process by tokenizing the words from the sentences in the “text” variable.

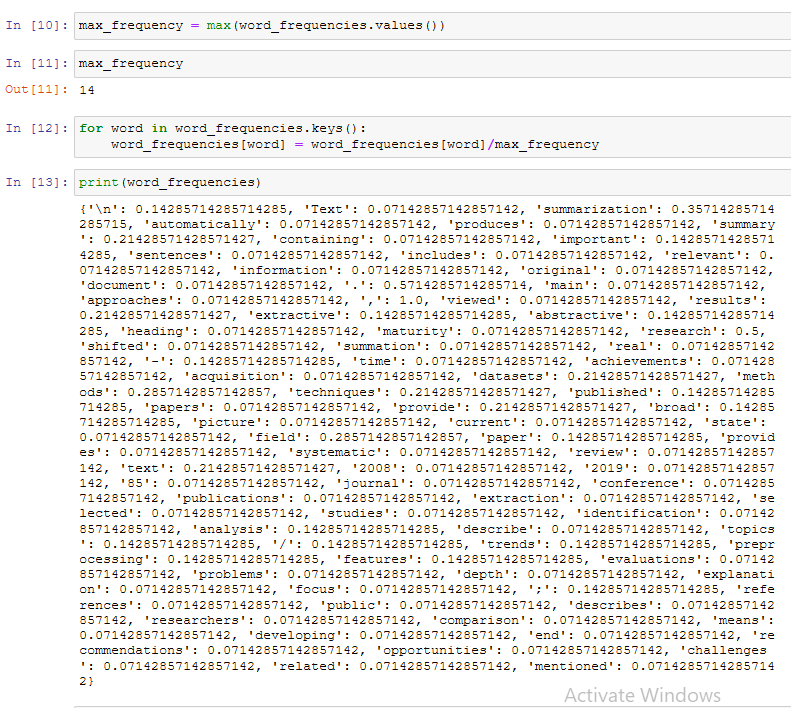


1. *Word Frequency Table*

Next, we will calculate the word frequencies from the text after removing stop words and punctuations.

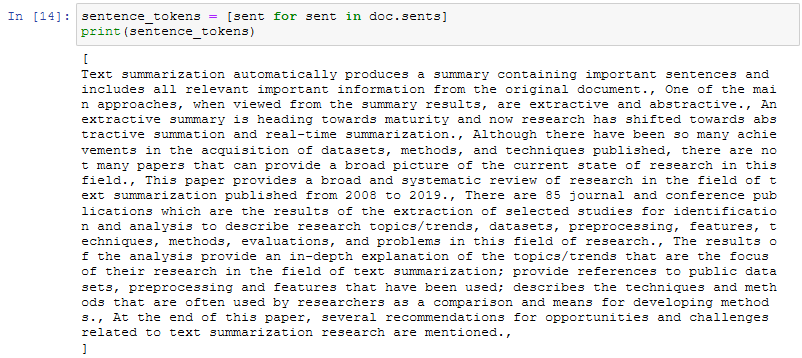


We will print the word frequencies to know the important words in the text.

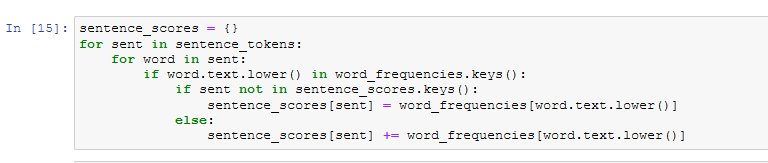


1. *Sentence Tokenization*

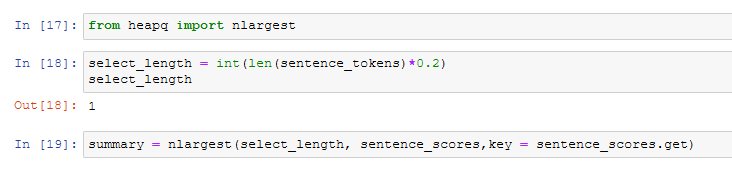
After knowing the important words in the text, we will get the sentence tokens in the text.



Next, we can calculate the most important sentences by adding the word frequencies in each sentence in the “text” variable.

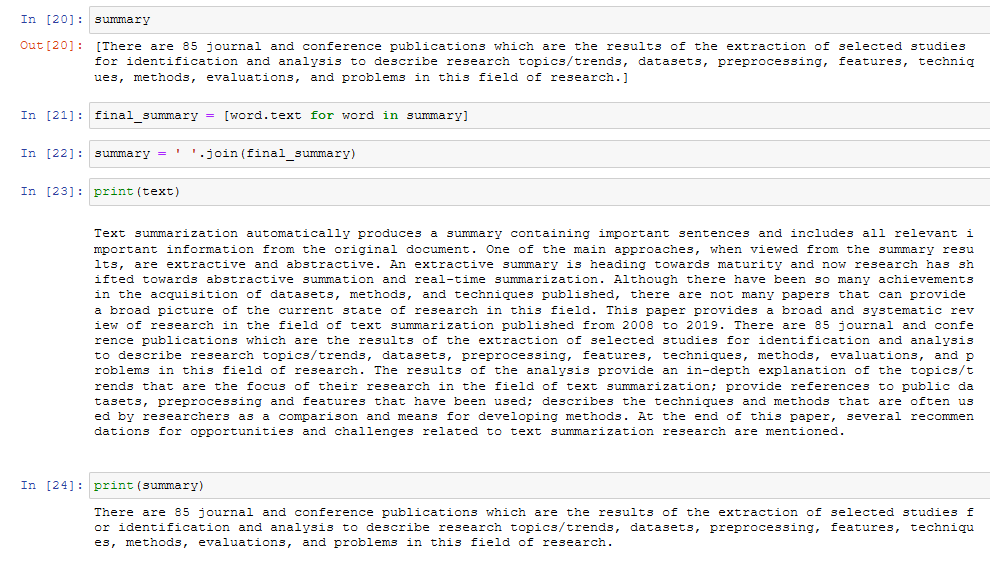


Then, by utilizing nlargest that had been obtained from heapq library, we can calculate the nlargest and calculate 20% of text with maximum score. For your information, after some testing it is recorded that 20% is the lowest that we can summarize this specific text, we believe if the text is longer, sub-20% may be achievable.

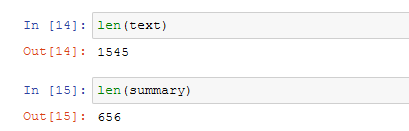


1. *Summarization*

Finally, we can get the summary of the text.



Finally, we can also calculate the length of the summary from the original text.



## **Result**

In this section, in order to evaluate the model’s capabilities, we tested it with three different articles, each of which came from three diversely different genres from each other, and all of the articles had different ranges of words with three different length of summary. Here are the results:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Title | | | | | |
| Length of Summary | Automatic text summarization techniques & methods | ‘Janet Jackson’ tells the singer's story, but it's clear who's in control | Who are the best up-and-Coming Coaches in European Football? | History of light novel | Classification of Ice in Lutzow-Holm Bay | State and local public policies |
| Original | 1545 | 3217 | 3209 | 7797 | 13214 | 15326 |
| 10% | NA | 141 | 412 | 1167 | 2071 | 2491 |
| 20% | 289 | 646 | 838 | 2499 | 3848 | 4499 |
| 30% | 656 | 1039 | 1370 | 3573 | 5902 | 6293 |
| 40% | 879 | 1461 | 1709 | 4681 | 7382 | 8065 |

Table 2: Model’s results

## 

## **Error Analysis**

For this section, we will evaluate the result obtained from the model by comparing the predicted value of the summary with the actual output, hence with it, we can calculate the accuracy of each summary’s length, with the result, we can acknowledge the length of summary with lowest accuracies as the most accurate summary that can produce the most accurate result. Here are the results:

1. 10% length of summary

|  |  |  |  |
| --- | --- | --- | --- |
|  | Length of Summary | | |
| Title | 10%  (Predicted) | 10%  (Actual) | Difference  (Predicted/Actual) |
| Automatic text summarization techniques & methods | 155 | NA | 0% |
| ‘Janet Jackson’ tells the singer's story, but it's clear who's in control | 322 | 251 | 128.3% |
| Who are the best up-and-Coming Coaches in European Football? | 322 | 412 | 78.2% |
| History of light novel | 778 | 1167 | 66.7% |
| Classification of Ice in Lutzow-Holm Bay | 1321 | 2071 | 63.8% |
| State and local public policies | 1533 | 2491 | 61.5% |
| Average Accuracy | 66.4% | | |

Table 3: 10% length of summary accuracies

## 20% length of summary

|  |  |  |  |
| --- | --- | --- | --- |
|  | Length of Summary | | |
| Title | 10%  (Predicted) | 10%  (Actual) | Difference  (Predicted/Actual) |
| Automatic text summarization techniques & methods | 309 | 289 | 106.9% |
| ‘Janet Jackson’ tells the singer's story, but it's clear who's in control | 643 | 646 | 99.5% |
| Who are the best up-and-Coming Coaches in European Football? | 641 | 838 | 76.5% |
| History of light novel | 1559 | 2499 | 62.4% |
| Classification of Ice in Lutzow-Holm Bay | 2642 | 3848 | 68.7% |
| State and local public policies | 3065 | 4499 | 68.1% |
| Average Accuracy | 80.4% | | |

Table 4: 20% length of summary accuracies

1. 30% length of summary

|  |  |  |  |
| --- | --- | --- | --- |
|  | Length of Summary | | |
| Title | 10%  (Predicted) | 10%  (Actual) | Difference  (Predicted/Actual) |
| Automatic text summarization techniques & methods | 464 | 656 | 70.7% |
| ‘Janet Jackson’ tells the singer's story, but it's clear who's in control | 965 | 1039 | 92.9% |
| Who are the best up-and-Coming Coaches in European Football? | 963 | 1370 | 70.3% |
| History of light novel | 2339 | 3573 | 65.5% |
| Classification of Ice in Lutzow-Holm Bay | 3964 | 5902 | 67.2% |
| State and local public policies | 4598 | 6293 | 73.1% |
| Average Accuracy | 73.3% | | |

Table 5: 30% length of summary accuracies

1. 40% length of summary

|  |  |  |  |
| --- | --- | --- | --- |
|  | Length of Summary | | |
| Title | 10%  (Predicted) | 10%  (Actual) | Difference  (Predicted/Actual) |
| Automatic text summarization techniques & methods | 618 | 879 | 70.3% |
| ‘Janet Jackson’ tells the singer's story, but it's clear who's in control | 1287 | 1461 | 88.1% |
| Who are the best up-and-Coming Coaches in European Football? | 1284 | 1709 | 75.1% |
| History of light novel | 3119 | 4681 | 66.6% |
| Classification of Ice in Lutzow-Holm Bay | 5285 | 7382 | 71.6% |
| State and local public policies | 6130 | 8065 | 76.0% |
| Average Accuracy | 74.6% | | |

Table 6: 40% length of summary accuracies

We also prepare the summary of the error analysis in the graph, here is:

Chart, line chart

Description automatically generated

Graph 1: Result accuracies

To conclude, based on the graph above, the best way to produce an accurate summary is to utilize only 20% of the text percentage.

1. **Manual Evaluation**

For this last section, we will be manually evaluating the summaries by scoring them on a range of one to five (1-5) in term of its summarization’s accuracy with 1 being the most inaccurate and 5 being the most accurate summarization. However, for the sake of ease of understanding, we will evaluate it by utilizing a table per article. Here are the tables:

1. Automatic text summarization techniques & methods summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Evaluator | | | |
| Length of Summary | Faisal | Syazmi | Rakibul | Najmul |
| 10% | 2 | 3 | 2 | 2 |
| 20% | 4 | 3 | 3 | 4 |
| 30% | 3 | 3 | 4 | 3 |
| 40% | 3 | 4 | 4 | 4 |

Table 7: Manual Evaluation Article 1

1. ‘Janet Jackson’ tells the singer's story, but it's clear who's in control

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Evaluator | | | |
| Summary | Faisal | Syazmi | Rakibul | Najmul |
| 10% | 2 | 3 | 2 | 3 |
| 20% | 3 | 4 | 4 | 4 |
| 30% | 3 | 3 | 3 | 3 |
| 40% | 3 | 4 | 3 | 3 |

Table 8: Manual Evaluation Article 2

1. Who are the best up-and-Coming Coaches in European Football?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Evaluator | | | |
| Summary | Faisal | Syazmi | Rakibul | Najmul |
| 10% | 2 | 3 | 3 | 2 |
| 20% | 4 | 4 | 3 | 4 |
| 30% | 4 | 3 | 4 | 3 |
| 40% | 3 | 3 | 4 | 2 |

Table 9: Manual Evaluation Article 3

1. History of light novel

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Evaluator | | | |
| Summary | Faisal | Syazmi | Rakibul | Najmul |
| 10% | 2 | 3 | 3 | 4 |
| 20% | 4 | 4 | 4 | 4 |
| 30% | 3 | 4 | 4 | 2 |
| 40% | 4 | 4 | 3 | 2 |

Table 10: Manual Evaluation Article 4

1. Classification of Ice in Lutzow-Holm Bay

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Evaluator | | | |
| Summary | Faisal | Syazmi | Rakibul | Najmul |
| 10% | 1 | 2 | 2 | 3 |
| 20% | 4 | 5 | 3 | 3 |
| 30% | 3 | 4 | 4 | 3 |
| 40% | 3 | 2 | 2 | 3 |

Table 11: Manual Evaluation Article 5

1. State and local public policies summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Evaluator | | | |
| Summary | Faisal | Syazmi | Rakibul | Najmul |
| 10% | 2 | 2 | 2 | 1 |
| 20% | 4 | 4 | 3 | 4 |
| 30% | 4 | 4 | 3 | 3 |
| 40% | 3 | 4 | 4 | 5 |

Table 12: Manual Evaluation Article 6

## Final Evaluation

After the result of the manual evaluation is produced, we can now calculate the best length for a summary to be understood the most by calculating the average of the summaries by length, similarly like what we had done in the error analysis section. Here is the result:

## 

Chart, line chart

Description automatically generated

Graph 2: Average Manual Evaluation

In conclusion, there are two main observations found, the first is that based on the manual evaluation tables above, it is found that the manual evaluators do not prefer the 10% length of summary in articles that contained a huge number in the size section. Finally, based on the result on the graph above, we can make the same conclusion as in the previous section as our second finding, that is the fact that 20% is still found to be the best length of a summary by both human and computer evaluation.

## **Reference**

[1] H. P. Luhn, "The Automatic Creation of Literature Abstracts," in IBM Journal of Research and Development, vol. 2, no. 2, pp. 159-165, Apr. 1958. DOI: 10.1147/rd.22.0159.

[2] Kupiec J, Pedersen JO, Chen F. A trainable document summarizer. Research and Development in Information Retrieval 1995: 68–73.

[3] S. P. Singh, A. Kumar, A. Mangal, and S. Singhal, "Bilingual automatic text summarization using unsupervised deep learning," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, 2016, pp. 1195-1200.doi: 10.1109/ICEEOT.2016.7754874.

[4] K. Kaikhah, "Automatic text summarization with neural networks," Intelligent Systems, 2004. Proceedings. 2004 2nd International IEEE Conference, 2004, pp. 40-44 Vol.1. doi:10.1109/IS.2004.1344634.

[5] Barzilay R, Elhadad M. Using lexical chains for text summarization. In: Proceedings of the ACL/EACL ‘97 workshop on intelligent scalable text summarization 1997: 10–17.

[6] Ercan G., Cicekli I. (2008) Lexical Cohesion Based Topic Modeling for Summarization. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2008. Lecture Notes in Computer Science, vol 4919. Springer, Berlin, Heidelberg.

[7] Kadhar Batcha, Nowshath & Aziz, N.A.. (2014). An Algebraic Approach for Sentence Based Feature Extraction Applied for Automatic Text Summarization. Advanced Science Letters. 20. 139-143. 10.1166/asl.2014.5258.

[8] T. K Landauer, P. W. Foltz, and D. Laham "An Introduction to Latent Semantic Analysis" Discourse Processes, col. 25, pp. 259-284, 1998