Goldsmiths
UNIVERSITY
OF LONDON

IS71130A- Natural Language Processing (2023-24)

Project Report of

Clinical Notes Summarisation

Sandor Kanda

Table of Contents

I. Introduction

Automatic text summarisation (ATS) is a leading topic in information retrieval research, particularly in the medical and biomedical domains. It offers an efficient solution to access the ever-growing scientific and clinical literature by summarising the source documents while maintaining their most informative contents. [1] The scientific and clinical literature is rapidly growing. The increasing number of patients in the healthcare system requires new data entry, and each hospital visit requires a new clinical note during the assessment. Reducing the note size while retaining the important information enables doctors and clinicians to grasp the critical points, save search time, and focus on the patients. The acquired data does not consist of identifiable patient data, ensuring compliance with privacy regulations and ethical standards for handling sensitive health information. This approach underscores commitment to maintaining the confidentiality and integrity of patient information while leveraging synthetic clinical note summarisation for advancing NLP tasks within healthcare, thereby promoting innovations that respect and protect individual privacy.

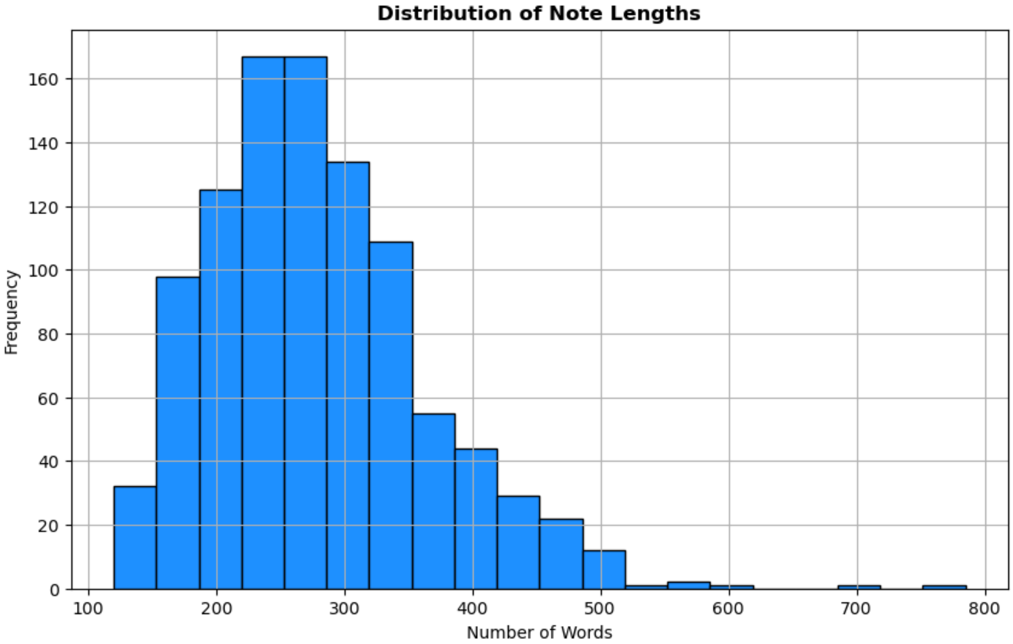
II. Methods

The dataset is sourced from the Hugging Face domain in comma-separated values (CSV) format. The data can also be accessed via a shared Google Drive link. The data file must be in the same directory as the Jupyter program file written in Python. The requirements.txt file is created at the root of the directory for reproducibility. The tqdm progress bar was used to use the progress meter for the iterative processes. The GitHub repository contains the license file, the *NLP_CW1_Text_Summarisation.ipynb* and the *requiremets.txt* file.

2.1 Exploratory Data Analysis

Each record contains the patient's ID, note, question, answer and task. The dataset includes 158 thousand rows of data, filtered to where the Task is Summarisation only (n = 19756). To preserve computational resources, the first one thousand records were used during the execution. No missing values are present. The filtered data notes are of various lengths. The distribution of the notes in terms of word count is represented in Figure 1 below.

Figure 1. Bar Chart of the Distribution of Note Lengths



Most notes are between 200-300 words, while the lengthiest notes (800 words) are minor.

2.2 Text Pre-Processing

The pre-processing text included removing the placeholders and normalising case and whitespace. The next step was to tokenise the words and sentences and produce the Part of Speech Tags (PoS). The feature extraction process extracted the medical entities, and the TF-IDF vectors fitted the notes' vocabulary in a numerical representation suitable for the summarisation model. The snippets of the medical entities and the TF-IDF vectors are listed below:

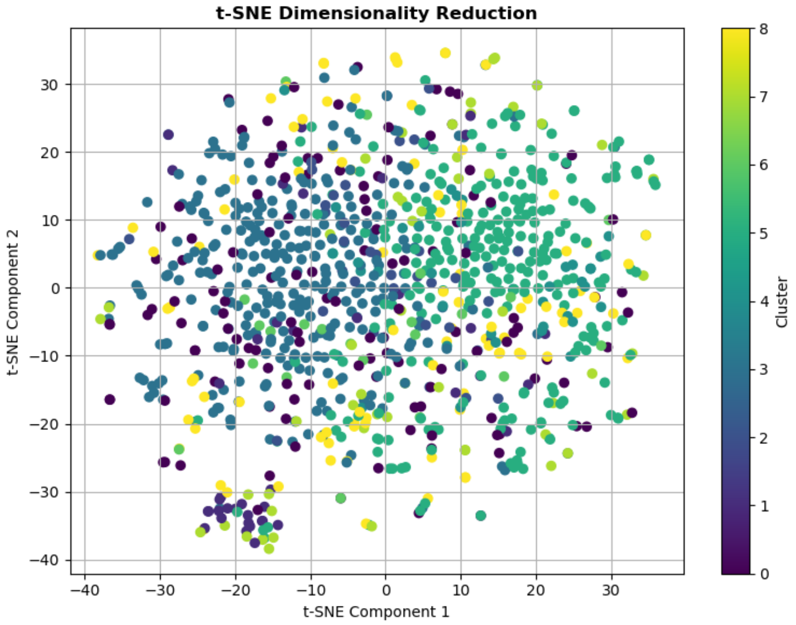
```
Medical Entities for the first note:
[('hospital course', 'ENTITY'), ('admission', 'ENTITY'), ('discharge', 'ENTITY'), ('patient', 'ENTITY'), ('sex', 'ENTITY'),
('male', 'ENTITY'), ('age', 'ENTITY'), ('years', 'ENTITY'), ('admission', 'ENTITY'),
```

```
TF-IDF Vector for the first note (in sparse matrix format):
  (0, 13983)	0.027979924576191733
  (0, 13895)	0.01334408502622908
  (0, 13852)	0.020694906707484452
  (0, 13850)	0.05312252999366043
  (0, 13836)	0.018049562427279186
  (0, 13822)	0.04184042942602779
  (0, 13784)	0.09294317182700323
  (0, 13778)	0.06565131102470032
  (0, 13772)	0.1299459508951531
```

2.3 Clustering

The medical entities and the TF-IDF vectors were combined to sidestep dimensionality issues. The fuzzy c-means clustering function set to three clusters produced a tuple containing the cluster centres, membership values and the number of iterations. The extracted feature visualisation is represented in Figure 2 below.

Figure 2. PCA t-SNE Dimensionality Reduction Heatmap



The visualisation in Figure 2 is a t-SNE (t-distributed Stochastic Neighbour Embedding) plot, a dimensionality reduction technique used to represent high-dimensional data in a lower-dimensional space. Due to this plot's overlapping and non-distinct clusters, the clustering algorithm could not locate well-defined clusters in the high-dimensional space. In addition to showing a lack of distinct borders, the figure also indicates outliers, changes in density, and smooth colour gradients across clusters. The Silhouette Analysis determined the optimal number of clusters for fuzzy clustering. The silhouette score measures how well the data points fit into the assigned cluster, where a higher score indicates a greater quality of clustering. In this case, the optimal number of clusters is 2. The visualisation of the Silhouette Analysis is shown in Appendix 1.

2.4 Summarisation Model

This project used the graph-based PageRank algorithm with the TF-IDF vectors as a summarisation model. The summarize_note function takes the note text and the summary ratio or the number of sentences. The code included reducing the summary to five sentences or reducing the text size to a ratio of 25%, which can be adjusted.

The pre-processed sentences are transformed into a TF-IDF matrix using the TfidfVectorizer from sci-kit-learn. TF-IDF assigns weights to each word in the sentences based on frequency and importance. Based on the TF-IDF matrix, the cosine similarity matrix, which measures the similarity between each pair of sentences, is computed. The PageRank algorithm is applied to the constructed graph using the nx.pagerank() function. The algorithm assigns scores to each sentence based on its importance and connection to the graph's other important sentences. The top-ranked sentences are selected based on the desired summary length. The model's performance was evaluated using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation metric.

III. Results

The model's performance was measured using a text ratio of 25%. The ROUGE score metrics are Precision, Recall, and F1 Score. Precision is the ratio of true positive predictions to the total number of positive predictions (true positives + false positives). It measures the model's ability to avoid false positives. Recall, also known as sensitivity or true positive rate, is the ratio of true positive predictions to the total number of actual positive instances (true positives + false negatives). It measures the ability of the model to find all the positive instances. The F1 Score is the harmonic mean of Precision and Recall. It provides a balanced measure of the model's performance, considering Precision and Recall.

The results of the summarisation by 25% ratio, using 1000 clinical notes, are represented in Table 1.

Table 1. Mean of Rouge Scores for 1000 records

Metric	Value
Precision	0.896540
Recall	0.290230
F1 Score	0.430317

The original note is compared with the summarised note below.

Hospital Course Summary:

Admission Date: [Insert date]
Discharge Date: [Insert date]

Patient: [Patient's Name]
Sex: Male
Age: 57 years

Admission Diagnosis: Oxygen Desaturation

Hospital Course:

The patient was admitted to the ICU one week after a positive COVID-19 result due to oxygen desaturation. Physical therapy was initiated promptly after admission, which helped improve the patient's breathing frequency and oxygen saturation. The patient was guided to achieve a prone position resulting in a significant increase in oxygen saturation from 88% to 96%. The patient continued to receive intensive physical therapy, positioning, and oxygen therapy for the next few days. Although there were challenges in achieving the prone position due to the patient's profoundly reduced respiratory capacity and high risk of symptom exacerbation, the medical team succeeded in implementing a safe and individualized approach.

After three days with this regime, the patient was transferred to the normal ward, where physical therapists continued his rehabilitation, including walking and strength training. However, the patient's severe instability remained a challenge. Nevertheless, after nine days from ICU admission, the patient was successfully discharged from the hospital as a pedestrian.

Discharge Condition:

At the time of discharge, the patient's medical condition had significantly improved, and he was considered stable enough to be discharged from the hospital. The patient's oxygen saturation had returned to normal limits, and his breathing frequency had decreased significantly.

Summary:

This course summary demonstrates that the patient responded positively to a physical therapy treatment regimen, including positioning, deep-breathing exercises, and walking. Although the patient's medical condition was quite severe during the initial ICU admission, his rehabilitation resulted in marked improvement, leading to a successful discharge from the hospital

Summarized Note:
Physical therapy was initiated promptly after admission, which helped improve the patient's breathing frequency and oxygen saturation. Although the patient's medical condition was quite severe during the initial icu admission, his rehabilitation resulted in marked improvement, leading to a successful discharge from the hospital. The patient continued to receive intensive physical therapy, positioning, and oxygen therapy for the next few days.

Notably, the key information of COVID-19 or the oxygen saturation levels, known as phenotypes, have yet to be captured in the summaries.

Figure 3. Interactive Graph Snippet of the ROUGE-L metrics

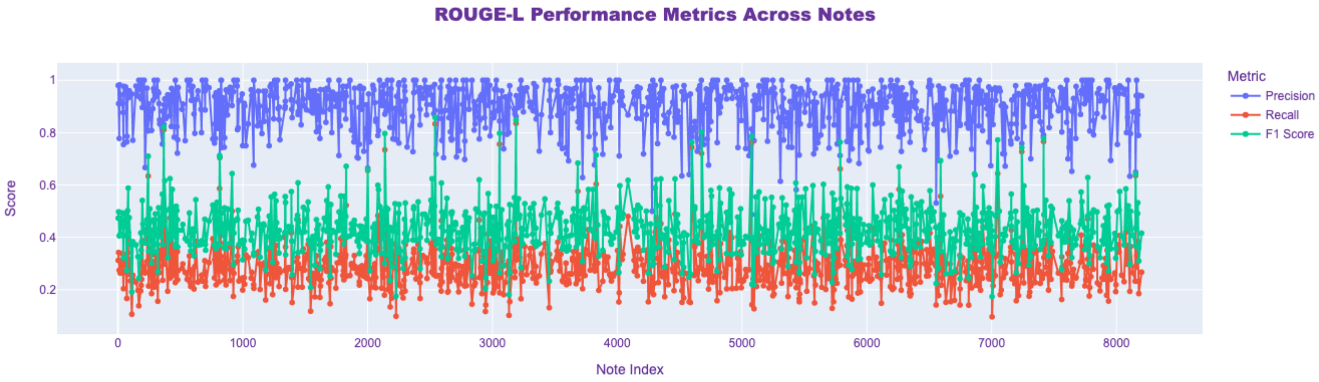


Figure 3 above provides a visual representation of the ROUGE-L model performance. The numerical representation is shown in Table 1.

IV. Conclusion

This work investigated how automated text summarisation (ATS) methods may be used for artificial clinical notes. The main goal was to create a summarisation model that could efficiently condense clinical notes while preserving crucial details, allowing medical staff to concentrate on patient care and rapidly understand significant aspects. The dataset consisted of 19,756 synthetic clinical notes. It was pre-processed using several techniques, including tokenising words and phrases, eliminating placeholders, normalising case and whitespace, and creating Part of Speech (PoS) tags. Identifying medical entities and creating TF-IDF vectors to represent the notes numerically were required for feature extraction. To find possible categories within the data, fuzzy c-means clustering was done to the combined medical entities and TF-IDF vectors. The clustering method could not have identified separate clusters in the high-dimensional space. Despite the generally low quality of clustering, the Silhouette Analysis found that two clusters were the ideal amount.

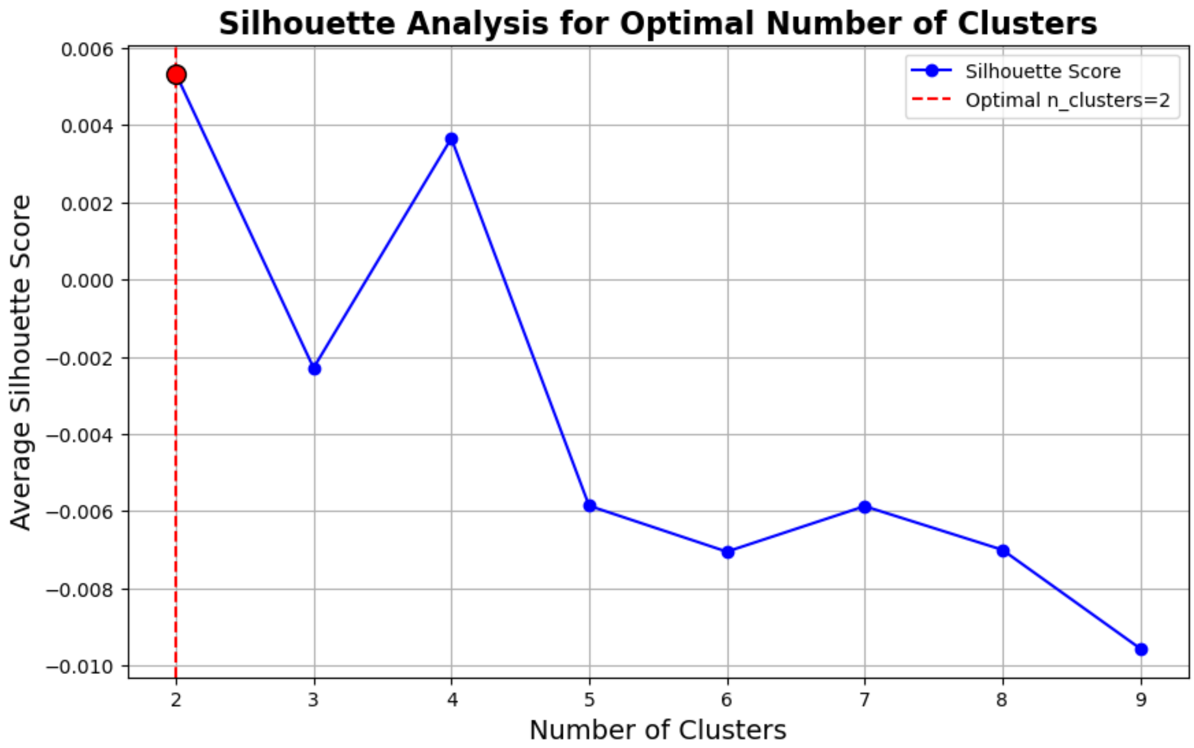
The summarisation model made use of TF-IDF vectors and a graph-based PageRank algorithm. Based on a 25% text ratio applied to 1,000 clinical notes, the results demonstrated a high Precision (0.896540), indicating the model's ability to avoid false positives. However, the Recall (0.290230) and F1 Score (0.430317) were relatively low, suggesting improvement in capturing all relevant information and balancing Precision and Recall. While the model reduces note size and retains important information, the performance metrics suggest room for improvement. Future research should focus on refining pre-processing, exploring alternative clustering approaches, and optimising the algorithm. Collaboration with healthcare professionals is crucial to validate the summaries' clinical relevance and ensure the practical applicability.

References

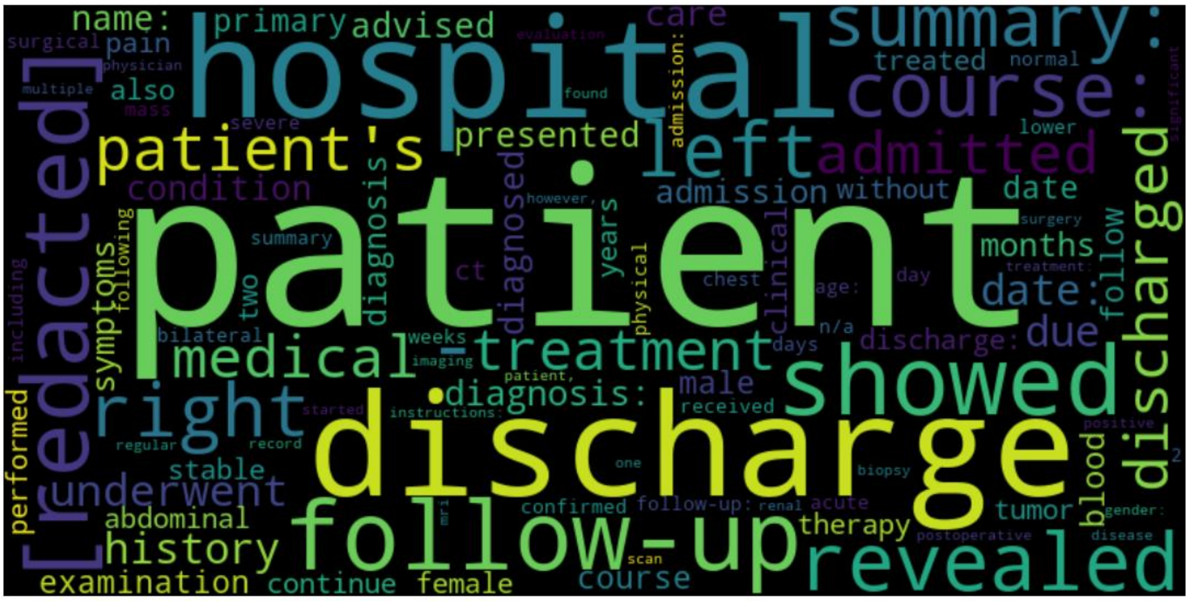
[1] ‘Afzal et al. - 2020 - Clinical Context–Aware Biomedical Text Summarizati.pdf’.

Appendix

Appendix 1. Silhouette Analysis Line Chart



Appendix 2. Word Cloud visualisation of the 100 most common words in the clinical notes


name:
primary
advised
hospital
care
summary:
surgical
pain
evaluation
treated
normal
physician
course:
multiple
also
found
admission:
mass
significant
severe
patient's
presented
left
admitted
lower
[redacted]
condition
admission
without
date
discharged
however,
surgery
patient
summary
diagnosis
diagnosed
months
following
symptoms
two
ct
years
physical
clinical
chest
age:
day
treatment:
including
date:
follow
bilateral
weeks
n/a
discharge:
due
medical
treatment
imaging
days
right
diagnosis:
male
showed
patient,
performed
started
instructions:
received
positive
regular
record
discharge
stable
one
biopsy
underwent
blood
confirmed
follow-up:
renal
acute
tumor
gender:
abdominal
mri
therapy
history
follow-up
postoperative
disease
revealed
scan
examination
continue
female
course