International Journal of Multimedia and Ubiquitous Engineering
Vol.15, No.2 (2020), pp.49-60
http://dx.doi.org/10.21742/ijmue.2020.15.2.05

Clinical Narrative Summarization based on the MIMIC III Dataset

Jugal Shah[1] and Sabah Mohammed[2*]

Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada
[1]jshah5@lakeheadu.ca, [2]mohammed@lakeheadu.ca*

Abstract

With the increasing technology in the field of healthcare, there is a substantial increase in the amount of clinical data produced for each patient, which makes it difficult for physicians to review all the information. Also, the produced clinical data are found in a multimodal format, making it difficult to interpret and review. It has always been a common practice for healthcare professionals to document patient health data in a non-structured natural human language to communicate specifics accurately without any loss of knowledge. However, reading through all the needless data regularly decreases the optimum usage of doctor-patient time and increases the risk of error. Generally speaking, text summarization is characterized as the creation of a subset of original data containing only relevant information. It is a significant research topic in the field of NLP. Despite this, not much research focuses on summarizing the text data collected in the healthcare sector. This paper represents the application of Bio_Clinical Bert in conjunction with Bert extractive summarizer to shorten clinical data. Additionally, it also implements topic modelling using LDA to assign relevant topics to the summary text. The MIMIC III dataset is used as a source of clinical notes for this project. The paper also presents key concepts revolving around text summarization and topic modelling.

Keywords: *POS, NER, Summarization, TF-IDF, BOW, Topic modeling, BERT*

1. Introduction

Over the last decades, there has been a notable rise in the clinical data available at medical professionals' disposal. Medical professionals play a vital role in the generation, collection, and filtration of this data—the interpretation obtained from the data helps to provide optimal care to the patients [1]. However, with the advent of digitization and wide adoption of Electronic Health Record (EHR), the data collection is growing exponentially on a day-to-day basis. The clinicians are overwhelmed with data generated from different sources and available in a wide variety of formats. With this evergrowing data, today, it is becoming increasingly difficult for the clinicians to separate ground noise from important clues efficiently [1], which in turn affects patient care. Thus, it is extremely essential to extract knowledge from vast structured and unstructured clinical data.

A clinical dataset can consist of a wide variety of data including, structured or coded data, free texts, an entire document and multimedia content [2]. Since structured data such as International Classification of Diseases (ICD) codes, laboratory results and medications may

Article history:
Received (August 27, 2020), Review Result (September 28, 2020), Accepted (November 3, 2020)

Print ISSN: 1975-0080, eISSN: 2652-1954 IJMUE

result in loss of medically relevant information, clinical datasets always consist of free text such as clinical narratives [2][3]. This paper focuses on summarizing such unstructured clinical narratives that are represented in natural human language. Clinical narratives represent the clinical practice performed by the medical professional and the results achieved from it. Educated medical practitioners, after much thought and reasoning, document these narratives. They are known to be the most significant source of effective patient care [4]. Therefore, this paper researches different Natural Language Processing methods (NLP) that can be applied to extract meaningful information and generate knowledge from the broad clinical narratives.

According to Joshua et al., clinical narrative summarization is defined as the process of collecting, distilling and synthesizing patient medical data to support medical experts in patients care [1][5]. In general, text summarization falls under two categories extractive summaries and Abstractive summaries. Extractive summaries are produced by selecting significant phrases from the original text. The frequency parameter is commonly used to identify significant phrases. On the other, abstract summaries synthesize new text that represents essential information from the original text in a concise format. It is accomplished through knowledge of the key concepts in the original text and then expresses those concepts in plain natural language [6][7][8]. While abstractive summaries are more informative and natural compared to ex- tractive summaries, generating abstractive summaries requires techniques like information fusion, sentence compression, and reformulation, thus making it more complicated to develop [8]. This paper focuses on creating extractive summaries for clinical narratives present in the MIMIC III database.

Clinical Narrative Summarization can be performed following three different approaches- statistical learning approach, machine learning approach and deep learning approach. Since interpreting and explaining the statistical method is problematic for clinical data, it is not preferred for summarization [4]. Deep learning is considered a state-of-art approach for text summarization in the field of NLP. A deep learning model requires a massive amount of data for training. However, the medical information mart for intensive care (MIMIC-III) clinical dataset that would be used for research in this paper has less than 60000 clinical summaries making the deep learning method a non-optimal solution for the given clinical dataset [4][6].

This paper focuses on the BERT pre-trained state-of-art language representation model and how it can be applied to the clinical domain for the MIMIC III dataset. It highlights how we can use Bert extractive summarizer with Bio Clinical Bert to summarize the text in the clinical domain. It also implements unsupervised machine learning approaches like topic modelling on the summarized text to identify topics representing the note. The paper explores a wide range of concepts such as tokenization, Part of Speech (POS) tagging, Named Entity Recognition, etc. The MIMIC III critical care database, created by Johnson et al. [6], is used for this research.

2. Critical review

In their paper on clinical text summarization, Wei-Hung Weng et al. mentioned some crucial additions such as syntax-based negation and semantic concept identification to improve the clinical flow [4]. The complexity of statistical learning approaches being high, the authors focused on using computational linguistics with human extracted biomedical knowledge to achieve meaningful clinical text summarization. The process was initiated by implementing the classic NLP approaches such as sectioning, tokenization, clinical text analysis, and knowledge extraction. Categorization of negation terms such as ‘no, not, rule

out, etc.' was executed in the first stage leading to the identification of negated concepts using sentence pruning, syntactic analysis and parsing using Apache NLP tools. This project's primary objective was to reduce the clinicians' time to read the medical information and dedicate more time and attention to treating the patients instead. Therefore, an integrated method consisting of syntax-based detection and semantic concept identification for text summarization has been proposed. The future work for the same project has plans to improve the parser for clinical narrative texts and design more rules for fragmented sentences. In their research review, Alexandra Pomares-Quimbaya et al. described the current approaches to identify sections within clinical narratives from EHR's (Electronic Health Records) [2]. The authors mentioned that extracting clinical information sections from the digital records is efficient and emphasized the automatic and semi-automatic methods for doing the same. Thirty-nine studies were selected for the project experimentation analysis, out of which 57 percent of surveys proposed formal procedures for identifying sections, and 43 percent adapted a previously studied method. The machine learning methods used for identifying clinical sections are Conditional Random Fields (CRF's) and Support Vector Machine (SVM). The Viterbi algorithm is also viral to detect optimal section labels where the clinical section identification problem is observed as a sequence labelling problem. Apart from ML approaches, the rule-based approaches implemented comprised of exact matching, regular expressions, and probabilistic rules. Some studies already consist of an existing pattern and rely on the knowledge extracted from the text corpora. The common features can be classified such as formatting features, concepts and headings. Other features extracted were lexical, syntactical, semantic and contextual features. Given the group studies on using ML methods in the paper, the hidden Markov model attained 94.6 percent accuracy, and the runner-up was SVM with 94 percent accuracy. The research emphasized the analysis of methods published in the research papers available in the selected databases.

The authors of the paper 'finding clinical knowledge from MEDLIN abstracts by text Summarization Technique' illustrates the work of clinical knowledge extraction from MEDLINE abstracts [8]. MEDLINE is a huge repository consisting of 26 million citations in the areas of medicine and healthcare. Going through millions of pages is a tedious and time-consuming task to extract and acquire knowledge. This project represents a method of summarizing clinical experience from MEDLINE abstracts. The technology used for extracting knowledge from the repositories is based on NLP and filtering techniques. The project is merely a case study where the abstracts related to cervical cancer are experimented on. The predicted results are compared with the actual results obtained from the domain experts in the preliminary stage. The initial stages of project methodology start with collecting and understanding the data for clinical knowledge. Five hundred abstracts aligning with cervical cancer were extracted with PubMed in the intermediate. The secondary stages consist of text processing and text filtering, where the 1-Model SVM (Support Vector Machine) algorithm was chosen as a text filtering model on the training set, i.e. 300 relevant MEDLINE abstracts. The results returned from the experimentation analysis being satisfactory. The average obtained for precision, recall and F-score were 1.00, 0.84 and 0.91. This project's future vision deals with medical term variation, where there are multiple versions of the same word. The automatic analysis of these terms can cause a significant problem, and several other errors in the same system.

Yet another research aimed at reducing the time taken by the doctors to parse through the medical records or reports daily. Pooja Vinod et al. in their research focus on reducing time and increasing efficiency of the physical doctor-patient interaction where all the relevant details are discussed [21]. The paper involves a large area occupied by extractive text

summarization in contexts of clinical information. BERT (Bidirectional Encoder Representation from Transformers) is an existing NLP model known or its pre-training tendency. BERTSUM is a fine-tuned version of BERT and BERTSUMEXT being the extractive summarization. The models pre-trained on a big scale with large amount of data. The authors have expanded the scope of BERTSUMTXT knowledge to give it cutting-edge. For the results of this project, the training strategy that modified the parameter values of extractive summarization layers of the model shows improvement on all layers of ROGUE, where ROGUE stands for recall oriented understudy for gusting evaluation. The fine-tuning for BERTSUMEXT model architecture was constructed in three stages such as dataset preparation, creation of two models A and B: BERTSUM be the 'A' model and 'BERT-SUMEXT' be the 'B' model and the last stage before results were 'k-cross fold validation and model selection'. The two models A and B were fed to the k-cross validation module and the result obtained was in favor of model 'A', where it showed improvement on all nine layers of ROGUE model and performed better than the other model in terms of clinical report dataset.

Hardik Gunjal et al. in their research on classifying clinical discharge summaries using deep learning had an intention of segregating all the important details of a clinical summary to have an efficient classification [22]. The important details being the history, previous treatment, symptoms and other important details that are prone to be left out during a clinical discharge summary analysis. The project has the potential concept of automating the system for classification using deep learning algorithms. The deep learning algorithm used for this approach was CNN-2D for classifying and representing complex features from clinical summaries. The methodology construction for this discharge summary project starts with data fetching, followed by data pre-processing. The dataset selected for this experiment consisted of 4999 medical transcription samples of 40 different medical specialties. The extracted keywords are further tokenized and then the training, testing and validation split are made. Glove, standing for global vector for word representation is an unsupervised machine learning technique that uses global word to word statistics from word corpus. The vector representations of words obtained from the word can be further used for training on new dataset and acquiring predictions. CNN algorithm is focused on categorizing medical transcription to their categories. CNN on MT dataset gives out an accuracy of 87 percent for training and a validation accuracy of 89 percent. The other acquires such as precision, recall and F1-score revolve around the same figure and hence, makes CNN a favourable algorithm to perform classification task for clinical discharge summary dataset. The authors in the paper [5] proposed a novel system able to acquire and present legit references to medical disorders from a patient summary. The base technology used for this kind of project is NLP (Natural Language Processing) for extraction of relevant medical entities stored in narrative health records which are necessary for scanning of structured documents. The web-application designed for showcasing the analytics for this project visualizes patient's data and displays the summary the same to the health practitioner operating at the front-end. The main system architecture consists of three phases starting with the extraction subsystem, indexing and storage and the last phase being the summarization subsystem. The first phase is a combination of two modules such as query module and NLP module. After passing through the intermediate phase, the last summarization subsystem comprises of filtering and data aggregation. The NLP module of phase one has two stages such as language identification and tokenization. After the language of the document has been identified, the character sequence is cut down into stream of tokens. The tokens are put through lexical analysis and parsing rules, highlighting medical entities in the text. The use case mentioned in the paper is a real one in which the clinicians make use of the implemented system to examine the

medical problems of a patient starting from a variety of health records. The system can receive any kind of text file written as medical report and goes through the NLP module explained above. The summarization module allows an analysis of processed documents and present concepts of interest to the user through a suitable interface.

3. The key concept

Clinical narrative summarization problems fall under the NLP field of research as the dataset is composed of natural human language. A generalized solution to an NLP problem involves a multi-stage process [9]. It consists of building a language processing pipeline that includes the following components.

- Tokenization
- POS tagging and Parser
- NER

(1) Tokenization

According to [9], tokenization is defined as breaking down of long texts into a list of sentences or words. Each item in the list is referred to as tokens. This breaking down the text into words helps in understanding the context of the text. Tokenization can be achieved using a wide range of methods. Primitive methods involve splitting the text using space or dot as a delimiter or using a regular expression. Additionally, due to NLP and machine learning advancements, different packages like NLTK, Spacy, Keras, and Genism are available now to directly perform tokenization on a given text.

(2) POS tagging and Parser

According to [10], Part-of-Speech (POS) Tagging is the method of assigning different labels known as POS tags to terms in a phrase that informs us about the term part-of-speech. POS tagging can be of two types universal POS tagging and detailed POS tagging. The table below shows the universal POS tags. Detailed POS tagging involves further dividing the universal tags into subtabs; for example, Noun can be divided into NNS (common plural Noun) or NN (common singular Noun). POS tagging helps to tag individual words; however, to analyze and understand the given text's grammar, it is essential to identify the relationship between a single word and its neighbours that modifies the word. This relationship between the words is identified using dependency parser. Independency parsing, a word and its dependent neighbouring word are linked with a dependency tag for a given sentence. The python Spacy package is well known for identifying the pos tag and dependency tag for the words in a given text.

(3) NER

Referencing from [11], NER is defined as the process of extracting entities from the text and assigned them to a set of pre-defined categories. Some of the pre-defined categories are person, organization, data, location, quantity, work-of-art, etc. You can obtain key information (entities) with named entity recognition to recognize what a subject is regarding. NER is widely used for the recommendation system. The spacy package is one of the well-known packages for performing NER on unstructured data.

Table 1. Universal pos tags [10]

Tag	Description
ADJ	Adjective
ADV	Adposition
ADP	Adverb
AUX	Auxiliary
CCONJ	Coordinating Conjunction
DET	Determiner
NOUN	Noun
PART	Particle
INTJ	Interjection
NUM	Numerical
PRON	Pronoun
PROPN	Proper Noun
PUNCT	Punctuation
SCONJ	Subordinating Conjunction
SYM	Symbol
VERB	Verb
X	Other

4. Clinical information extraction tools

The previously mentioned key concepts can be implemented by collection of text document [13][14]. Here each topic represents a set of words that co-occur together. In general, in topic modelling, classification is performed, and each document is represented using a particular topic. Topic modelling is of great use in the clustering of documents and information retrieval. One of the well-known topic modelling techniques is Latent Dirichlet Allocation (LDA), provided by the genism library. According to N K Nagwani [16], LDA models each document as a probability distribution over topics, and each topic is a vector of terms with a probability between 0 and 1. Before passing the corpus to the LDA model, build using the genism library, the documents' collection must be converted into the document vector matrix and then passed as input. Converting text data into numeric matric representation is called as data vectorization.

Data vectorization can be done using a variety of NLP techniques. Some of the well-known techniques are Bow, TF- IDF, and Word2Vec. In Bow, the frequently occurring words are given more importance. It forms a matrix of documents against tokens, each cell storing the count of the token in the document. This count is defined as Term Frequency (TF). In TF-IDF, the relevant words are given more importance than the frequent words. Equation (1) denotes the TF-IDF formula. In the formula t denotes the terms; d denotes each document; D denotes the collection of documents.

using pre-defined models from packages like Spacy or using existing Information Extraction (IE) tools such as cTAKES, which are specific to the clinical domain. Referencing the literation review by Yanshan et al. [12], cTAKES, MetaMap, and MedLEE are the most known IE tools in the clinical domain.

$$tf - idf(t, d, D) = tf(t, d) \times idf(t, D) \quad (1)$$

Equation (2) shows the expansion of *idf(t,D)*

$$idf(t, D) = log \frac{|D|}{1+|\{d\ D{:}t\ d\}|} \quad (2)$$

cTAKES: It is based upon open-source apache projects and is utilized for numerous case studies such as smoking status extraction, genome-wide association studies, risk stratification, and risk factor identification.

MetaMap: Developed by the National Library of Medicine, it is used to perform a mapping between biomedical text and the Unified Medical Language System (UMLS). It can also help in fragment recognition in clinical documents and the extraction of patient-related attributes.

MedLEE: It is the oldest NLP tool and is used for pharmacovigilance and pharmacoepidemiology.

Additionally, tools like mallet and weka help in performing machine learning techniques like clustering and topic modelling, which will be explored in the coming section.

5. Topic modeling and Latent Dirichlet Allocation (LDA)

Topic modelling is an NLP based unsupervised machine learning algorithm that identifies abstract topics from the given. The numerator in the above equation represents the number of documents in the corpus and denominator represents the number of documents in which the term t occurs. Word2Vec is distinct from the previous two vectorization technique. Bow and TF-IDF perform frequency-based vectorization, whereas Word2Vec is two layers neural network that can store the context of the given text with the help of weights. The output of the Word2vec model can have several dimensions.

6. The methodology

The below methodology is inspired by the work in [18]. Additional work done by [13][14][19] is used as a reference to gain a better understanding. The overall methodology can be divided into three phases

Preprocessing the MIMIC III dataset.

Summarize the clinical notes in the dataset using BERT

Apply topic modeling on the summary data.

(1) Dataset

The MIMIC III is a single large relational database that contains information of patients admitted to the critical care unit at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. The database consists of 26 tables. These tables include various patient-related data such as vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, and survival data more. All the tables in the database are by identifiers having the suffix 'ID'. The database is also de-identified following the Health Insurance Portability and Accountability Act (HIPAA) standards. Additionally, all dates attribute in the database were shifted by a random offset to a future date while preserving the intervals

between dates. This paper will be working with only two tables the Admission table and the Note Events table.

The Admission table contains entries of every unique hospitalization, identified by HAMID, for each patient in the database, identified by subject.

The Note Events table contains de-identified notes, including nursing and physician notes, ECG reports, radiology reports, and discharge summaries.

To access the MIMIC III database, a request can be raised at the link (https://mimic.physionet.org/gettingstarted/access/). Due to the sensitive nature of medical data and privacy permission, I cannot include the raw data openly in this paper.

(2) Data Pre-processing

In the first step of data processing we would be filtering the Note Events table based on the category of the note. There are total of 15 categories. [Figure 1] shows the total number of notes in each category. For this paper to reduce the computation time we would be focusing only on the Discharge summary. Total of 59652 belongs to the Discharge summary category.

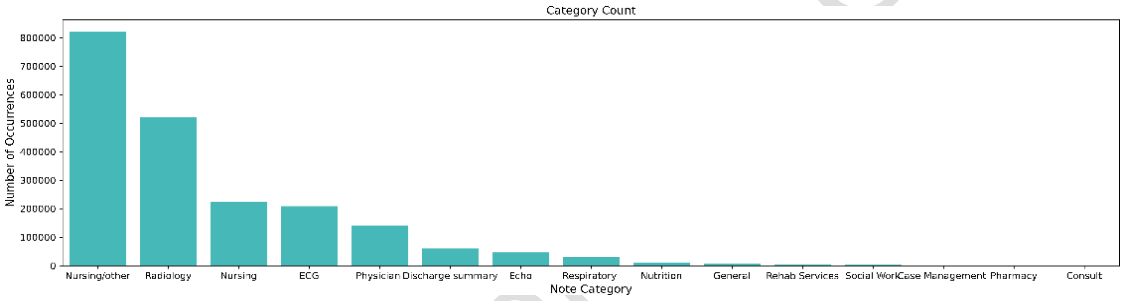

Figure 1. Note category count

The second step involves cleaning the text. Referring from [18], based on the analyses of the clinical notes, data cleaning focuses on things such as:

- Removing characters like '?', '*', '=', ';' and other punctuations
- Combining hyphenated characters like 'Follow-up'.
- In the clinical notes patients, names are de-identified and replaced with a random pseudo name like FirstName8 LastName4. These names are removed.
- Remove Dr. abbreviation as we are removing all the names.
- Removing consecutive punctuations.
- Removing duplicate words.

All of the cleanings would be performed with the help of regular expression marks.

(3) Text Summarization

This paper focuses on generating extractive summaries for the given Notes in the MIMIC III dataset. There are numerous ways to achieve the above objective; for example, the graph-based method - Text Rank algorithm provided by the genism package. The author in [18] performs text summarization by identifying dependency in the unstructured text using POS tagging. The Summarized text is generated by performing dependency tracking preprocessed text. Dependency tracking focus on the extraction of subjects and objects from the text.

However, this paper utilizes a fine-tuned versions of BERT's (Bidirectional Encoder Representations from Transformers) to summarize clinical notes.

BERT is a state-of-art language representational model that has been successfully applied to various NLP tasks. It improves the accuracy of general NLP tasks like Question Answering by applying bi-directional training to the Transformers. Transformers are a popular deep learning model for transferring one sequence to another in the domain of NLP. Since BERT has also been trained on a significantly large corpus, it is considered a primary choice for NLP tasks [21][28].

To summarize the clinical notes using BERT, we pip in- stalled the package proposed in [26]. The author of the paper used the pre-trained BERT model to generate text embedding, which is a N matrix where N is the number of sentences, and E is the embedding dimensions. Then using the embeddings, K-mean clustering is performed, and the sentence closest to the centroid is selected for summarization. However, to fine-tune the above summarization technique for generating clinical domain embeddings, we have passed the Bio Clinical Bert Model [27] as a custom model to the summarization function. Bio Clinical Bert is a variant of the BERT model that is pre-trained on the MIMIC III dataset. [Table 2] below showcases sample of Clinical note summarization.

(4) Topic Modeling

In this step, the aim is to assign a topic to each clinical summary. To achieve the below steps are performed.

- Tokenize the summary using the simple process from the genism package
- Remove stop words from the list of tokens generated in the previous step.
- Lemmatize the tokens using the WordNetLemmatizer from NLTK
- Create a dictionary mapping each token to an ID.
- Vectorized the summarized text using the genism bag of word modules.
- Passed the vectorized document to the genism LDA model. For experimentation, the number of topics extracted from the corpus is set to 15
- Generate topics for the vectorized document.

Each topic is defined by the set of words with associated weights. A single clinical note can have multiple topics assigned to it, along with the note's probability to fall under a particular topic. In this paper, we have assigned the topic with maximum probability to the notes. For instance clinical note at index 1 is assigned topic 8 and topic 12. Topic 12 has a higher probability for the note and the representation for topic 12 is shown in [Table 2] below. Using the pyLDAvis package the relationship between the topics is shown in [Figure 2] above. Each circle in the figure represent a topic and the distance between circle represents which all topics are related to each other. The main aim here is to tune model parameters to generate distinct topic such that the topic don't overlap. From the figure it can be seen that the topic 2,4,5,6 represent a similar field, whereas other topics are distinct. For instance topic 13 is focuses on general things like location, medication, telephone, fax, records, etc.

Table 2. Clinical summary

Text	Summary

The patient is a 34 year old male. statuspostmotor vehicle accident on 2132-7-9 > status post Ulextomy：onthe right foot and bilateral here for a ri^ht tibio/r.alc.axxeal fusion a right iliac bone graft and internal hardware placement. The patient tolerated the procedure well. On postoperative day number one. he had a maximum tanperature of 5 that spontaneously defervesced. He was maintained on intravenous antibiotics throughout die course of his stay. His incision was clean, dry and intact. The patient was seen by physical therapy and. after plain films were reviewed, it was deemed appropriate to allow die patient to have weightbearing as tolerated on the left lower extremity with walker boot in place and non weightbearing on the right ankle. He was discharged to rehabilitation on an AFO boot and is to follow up with. In two weeks, he will be following as an outpatient regarding his sciatic nerve issues. Dictated By	The patient is a 34 year old male, status post motor vehicle accident on 2132-7-9, status post talectomy on the right foot and bilateral here for a right tibiocalcaneal fusion and a right iliac bone graft and internal hardware placement. The patient was seen by physical therapy and. after plain films were reviewed, it was deaned appropriate to allow die patient to have weightbearing as tolerated the left lower extremity with a walker boot in place and non weightbearing on the right ankle.

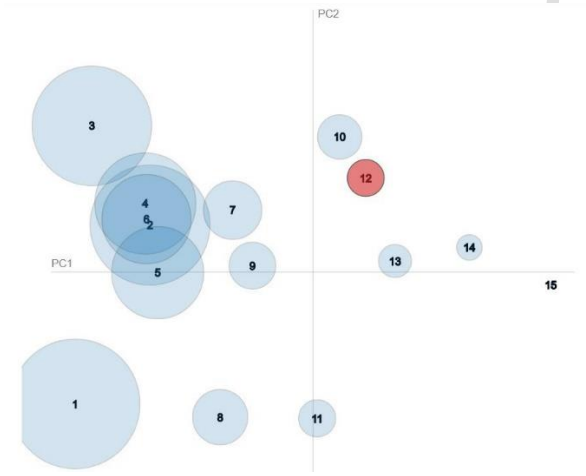

Figure 2. Topic clusters.

Table 3. Topic 12 representation

Word	Meiahtase
patient	0.039600402
history	0.013263325
day	0.009984136
blood	0.009315542
nght	0.008790495
leave	0.008559095
discharge	0.008024054
status	0.007813565
time	0.0077999807
year	0.007628321

7. Conclusions

This paper presents a variety of research on text and clinical data summarization. It explores different approaches to perform clinical narrative summarization and gives an overview of general key concepts that can be applied to achieve a summary of clinical data. The paper also showcases the implementation of a fine-tuned version of the state-of-art language model BERT to summarize clinical notes in the MIMIC III database and topic modelling using LDA. It majorly utilizes the Spacy and genism NLP library to perform most of the tasks. The project's future goals include performing abstractive summarization in the clinical domain using BERT and exploring BerTopic, an external library for performing topic modelling using BERT and c-TF-IDF.

Acknowledgment

This paper is part of the MSc Project of the first author.

References

[1] Feblowitz, Joshua C., Adam Wright, Hardeep Singh, Lipika Samal, and Dean F. Sittig., "Summarization of clinical information: A conceptual model," Journal of biomedical informatics, vol.44, no.4, pp.688-699, **(2011)**

[2] Pomares-Quimbaya, Alexandra, Markus Kreuzthaler, and Stefan Schulz, "Current approaches to identify sections within clinical narratives from electronic health records: a systematic review," BMC medical research methodology, vol.19, no.1, pp.155, **(2019)**

[3] Gehrmann, Sebastian, Franck Dernoncourt, Yeran Li, Eric T. Carlson, Joy T. Wu, Jonathan Welt, and John Foote Jr et al., "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives." PloS one, vol.13, no.2, article no. e0192360, **(2018)**

[4] Weng, Wei-Hung, Yu-An Chung, and Schrasing Tong, "Clinical text summarization with syntax-based negation and semantic concept identification," arXiv preprint arXiv:2003.00353, **(2020)**

[5] Diomaiuta, Crescenzo, Maria Mercorella, Mario Ciampi, and Giuseppe De Pietro, "A novel system for the automatic extraction of a patient problem summary," In 2017 IEEE Symposium on Computers and Communications (ISCC), pp.182-186, IEEE, **(2017)**

[6] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., Mark, and R. G., "MIMIC-III, a freely accessible critical care database," Scientific Data, vol.3, 160035, **(2016)** DOI: 10.1038/sdata.2016.35

[7] Pivovarov and Rimma, "Electronic health record summarization over het- erogenous and irregularly sampled clinical data," Ph. D. dissertation, Columbia University, **(2015)**

[8] Sibunruang, Chumsak, and Jantima Polpinij, "Finding clinical knowledge from MEDLINE abstracts by text summarization technique," In 2018 International Conference on Information Technology (InCIT), pp.1-6, IEEE, **(2018)**

[9] Shubham Singh, "How to get started with NLP - 6 unique methods to perform tokenization," Analytics Vidhya, Available: https://www.analyticsvidhya.com/blog/2019/07/how-get-started-nlp-6- unique ways perform-tokenization/, Sept., **(2020)**

[10] Abhishek Sharma, "How Part-of-Speech Tag, dependency and constituency parsing aid in understanding text data?" Analytics Vidhya. Available: https://www.analyticsvidhya.com/blog/2020/07/part- of-speechpos-tagging-dependency-parsing-and-constituency-parsing-in- nlp/ Sept., **(2020)**

[11] Pooja Mahajan, "NER tagging in python using spicy," Medium.com. Available: https://medium.com/analytics-vidhya/ner-tagging-in-python- using-spacy-c66cf01d3c7f Sept., **(2020)**

[12] Wang, Yanshan, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu et al., "Clinical information extraction applications: A literature review," Journal of biomedical informatics, vol.77, pp.34-49, **(2018)**

[13] Jonathan Keller, "Building a Topic Modeling Pipeline with spacy and Gensim". TowardsDataScience.com. Available: https://towardsdatascience.com/building a topic modeling pipeline with spacy and genism c5dc03ffc619 Sept., **(2020)**

[14] Susan Li, "Topic modeling and Latent Dirichlet Allocation (LDA) in Python," TowardsDataScience.com., Available: https://towardsdatascience.com/topic-modeling-and-latent-dirichletallocation-in-python-9bf156893c24 Sept., **(2020)**

[15] Shivam Bansal, "Beginners guide to topic modeling in Python," Analytics Vidhya, Available:https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/ Sept., **(2020)**

[16] Nagwani, Naresh Kumar. "Summarizing large text collection using topic modeling and clustering based on MapReduce framework," Journal of Big Data, vol.2, no.1, pp.1-18, **(2015)**

[17] Kenei, Jonah Kipcirchir, Elisha TO Opiyo, Juliet Chebet Moso, and Robert Oboko, "Clinical documents summarization using text visualization technique," International Journal of Computer and Information Technology, vol.7, no.4, pp.139-156, **(2018)**

[18] Gaurika Tyagi, "NLP-Preprocessing clinical data to find sections," TowardsDataScience.com. Available: https://towardsdatascience.com/nlp- preprocessing-clinical-data-to-find-sections-461fdadbec77 Sept., **(2020)**

[19] NSchrading, "Subject object extraction," Github.com, Available: https://github.com/NSchrading/intro-spacy-nlp/blob/master/subject object extraction.py Sept., **(2020)**

[20] Andrew Long, "Machine learning with datetime feature engineering: predicting healthcare appointment no-shows," Medium.com., Available: https://towardsdatascience.com/machine-learning-with- datetime-feature-engineering-predicting-healthcare-appointment-no- shows-5e4ca3a85f96 Sept., **(2020)**

[21] Vinod, Pooja, Seema Safar, Divins Mathew, Parvathy Venugopal, Linta Merin Joly, and Joish George, "Fine-tuning the BERTSUMEXT model for clinical report summarization," In 2020 International Conference for Emerging Technology (INCET), pp.1-7, **(2020)**

[22] Gunjal, Hardik, Preetkumar Patel, KhushalParesh Thaker, Abhishek Nagrecha, Sabah Mohammed, and Alizar Marchawala, "Text summarization and classification of clinical discharge summaries using deep learning," **(2020)**

[23] Alsentzer, Emily, and Anne Kim, "Extractive summarization of EHR discharge notes," arXiv preprint arXiv:1810.12085, **(2018)**

[24] Johnson, A., Pollard, T., Mark, and R., "MIMIC-III clinical database demo (version 1.4)," Physio Net, **(2019)** DOI: 10.13026/C2HM2Q

[25] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... Stanley, H. E., "Physio Bank, Physio Toolkit, and Physio Net: Components of a new research resource for complex physiologic signals," Circulation [Online], vol.101, no.23, pp.e215-e220, **(2000)**

[26] Miller and Derek, "Leveraging BERT for extractive text summarization on lectures," arXiv preprint arXiv:1906.04165, **(2019)**

[27] Alsentzer, Emily, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott, "Publicly available clinical BERT embeddings," arXiv preprint arXiv:1904.03323, **(2019)**

[28] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova., "Bert: Pretraining of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, **(2018)**

[29] Maarten Grootendorst, "Topic modeling with BERT," Towards Data- Science.com. Available: https://towardsdatascience.com/topic-modeling- with-bert-779f7db187e6 Sept., **(2020)**

[30] Nicha Ruchirawat, "6 tips for interpretable topic Models," TowardsDataScience.com. Available: https://towardsdatascience.com/6-tips-to-optimize-an-nlp-topic-model-for-interpretability-20742f3047e2 Sept., **(2020)**