SSNdhanyadivyakavitha at MEDIQA-Sum 2023: Medical Dialogue Summarization using Linear Support Vector Classification Technique

Dhanya Krishnan , Divya Srinivasan and Kavitha Srinivasan

Department of Computer Science Engineering, Sri Sivasubramaniya Nadar College of Engineering, Rajiv Gandhi Salai, Kalavakkam- 603110, India

Abstract

This research paper proposes a working model for the "ImageCLEFmed MEDIQA-Sum task" of clinical note summarization in healthcare using machine learning algorithms. The model is developed using a "Linear Support Vector Classification (SVC) algorithm" with TF-IDF features to perform text classification on the doctor-patient conversation snippets given in the dataset. A training accuracy of 0.99 and a validation accuracy of 0.69 were obtained. Linear SVC results in improved accuracy when the classes are distinguishable and separated into the respective section classes. In addition, TF-IDF is used to efficiently convert and extract the information given in the dataset. The model also employs several preprocessing techniques to improve the accuracy and random oversampling is performed to combat the heavy imbalance between classes. Other models using CART and CNN are also analyzed. Moreover, this paper discusses the importance and need for text summarization in the medical field with the scope for improving diagnosis and treatment.

Keywords

Doctor-patient conversation, MEDIQA-Sum task, Linear Support Vector Classification, Section headers

1. Introduction

ImageCLEF, an evaluation platform originally proposed by Mark Sanderson and Paul Clough from the Department of Information Studies, University of Sheffield seeks to provide a cross-language annotation and retrieval [1]. This forum creates the necessary infrastructure for the evaluation of visual information retrieval systems operating in monolingual, cross-language, and language-independent contexts. ImageCLEF's main objective is to support the development in the field of visual media analysis, indexing, classification, and retrieval [2]. This objective was motivated by the requirement to serve multilingual users from a global community who wanted to access the constantly expanding visual data and text data to build efficient models for real-time data analysis.

Text summarization is the practice of reducing lengthy texts into manageable paragraphs or sentences. Text summary seeks to extract key details while keeping the overall meaning of the paragraph

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
EMAIL: dhanya2010402@ssn.edu.in(A. 1); divya2010335@ssn.edu.in(A. 2); kavithas@ssn.edu.in (A. 3)
ORCID:0000-0003-3439-2383(A. 3)

intact. Different methods exist in the literature for keyword and semantic-based text summarization, where the medical data in the healthcare domain plays a vital role in the analysis [3].

The clinical summary can be defined as a process of collecting, reviewing, and analyzing patient data to assist in the conclusion of different clinical needs. Existing automated text summarization techniques are particularly relevant and useful for medical teams and researchers when there is a need for research evidence of the COVID pandemic [4,5]**.** When analyzing patient data, clinicians are frequently given abundant data from several sources that must be examined independently. Patients have the chance to elaborate on the story by employing natural language processing techniques to paraphrase the text, which also improves and increases accuracy. Text summarisation helps in summarizing medical records and improving diagnosis. Scaling up to large collections and diversifying applications, text summarization is a useful tool to be incorporated [6]. The task chosen from the CLEF forum is ImageCLEFmed MEDIQA-Sum 2023. This task aims to generate a summarization of clinical notes between the clinician and the patient via three subtasks, out of which we have participated in Subtask A and the implementation with results are submitted for the same. The volume of published medical research continues to grow rapidly and staying updated with the best available evidence is a challenge for clinicians [7]. Many text summarization systems are difficult to customize or are deployed in low-resource settings. This paper aims at an efficient yet accurate summarisation approach. Given the number and complexity of medical text records, it has been realized that there is no added value for the larger quantities of data. Easier access to required information through models that facilitate information retrieval increases the scope of research and the medical domain [8]. A model taking user interests into account, and presenting findings about a user model based on an existing patient record is highly recommendable [9]. Researchers examined a great variety of techniques and applied the same in different domains to yield practical research[10].

Subtask A: Dialogue2Topic Classification. This task aims at identifying the topic given a snippet of the conversation between the doctor and the patient. The topics have to be identified from the given list of 20 topics or section headers (e.g. Assessment, Diagnosis, Exam, Medications, Past Medical History).
Subtask B: Dialogue2Note Summarization. In this task, the conversation snippet between a doctor and patient with a section header is given.
Subtask C: Full-Encounter Dialogue2Note Summarization. Given a full conversation between a doctor and patient, participants are tasked to find the complete clinical note summarizing the conversation.

The remaining part of the paper spans across following subsections. Section 2 of this paper describes the dataset provided by the organizers. Further in Section 3, the proposed methodologies including Linear Support Vector Machine, Classification and Regression Trees and Convolutional Neural Networks are described in detail. The results section of the paper analyzes the differences in accuracy between the different models. Section 5 provides a brief description of the System Specifications required to implement the developed models. A summary of the inferences is presented in Section 6. The conclusion and future work are summarized at the end.

Researchers and healthcare professionals often need to stay up-to-date with the latest medical literature. Text summarization algorithms can summarize complex research articles, enabling researchers to quickly note the key findings, methodologies, and implications without reading the entire paper. Overall, text summarization in healthcare has the potential to save time, improve information retrieval, support evidence-based decision-making, and enhance patient care by distilling vast amounts of textual data into concise and meaningful summaries.

2. Dataset Description

Clinical conversation between a doctor and a patient has been recorded by an expert in the field is the foundation of the dataset for subtasks A and B. The subtask C dataset consists of doctor-patient conversations and corresponding notes written by medical scribes.

This dataset consists of 1200 conversations between doctors and patients as given in Table 1 and Table 2. Each sample is a conversation snippet between a doctor and a patient wherein the patient details his symptoms and the doctor presents a diagnosis. The tasks are divided into 20 different section headers or classes. These headers help to arrange and classify the conversations according to the subject of interest.

The dataset headers are Allergy, Assessment, CC, Diagnosis, Disposition, Exam, Fam/Sochx, Genhx, Medications, PastmedicalHx, PastSurgical, Plan, Ros, Other History, ED Course, Immunisations, Labs, Imaging, Procedure and GynHx. An example conversation snippet is given in Fig 2.1

Table 1.
Dataset Description

Dataset	Training	Validation	Testing
Clinical conversation exchanged between a doctor and a patient	1201	100	200

Table 2.
Brief inference from the dataset

Criteria	Section Header name	Frequency
Maximum occurring Header	FAM/SOCHX	351
Minimum occurring Header	GYNHX	2

Doctor: Is everything fine?
Guest_family: My mom is not well.
Doctor: When did this start?
Guest_family: I don't know but she is not in her correct state of mind.
Doctor: Okay let's see what we can do, how old is she?
Guest_family: She is around seventy four years old.
Doctor: Okay. Don't worry, we will see what we can do.
Guest_family: Thanks!
Doctor: Of Course.

Figure 2.1 Conversation snippet example

The dataset is analyzed and a text summarization algorithm is used to create a system that can automatically produce brief overviews of doctor-patient conversations, to gain valuable insights and save the doctor's time.

3. Proposed Methodologies

The ImageCLEFmed MEDIQA-Sum 2023 aims to simplify the tedious task of clinical note-writing and summarizing clinician-patient conversations. In subtask A, categorizing the given snippets of doctor-patient conversation into appropriate headers falls under the problem of Text Classification. Several algorithms can be used for this purpose, including but not limited to Naive-Bayes Algorithm, Support Vector Machines, ID3 Algorithm, Latent Derelict Allocation, etc.

A flow diagram for the developed model is given in Fig. 2. Since there is only a limited training dataset (of 1201 samples), this study has chosen to apply the following methodologies to achieve the best results:

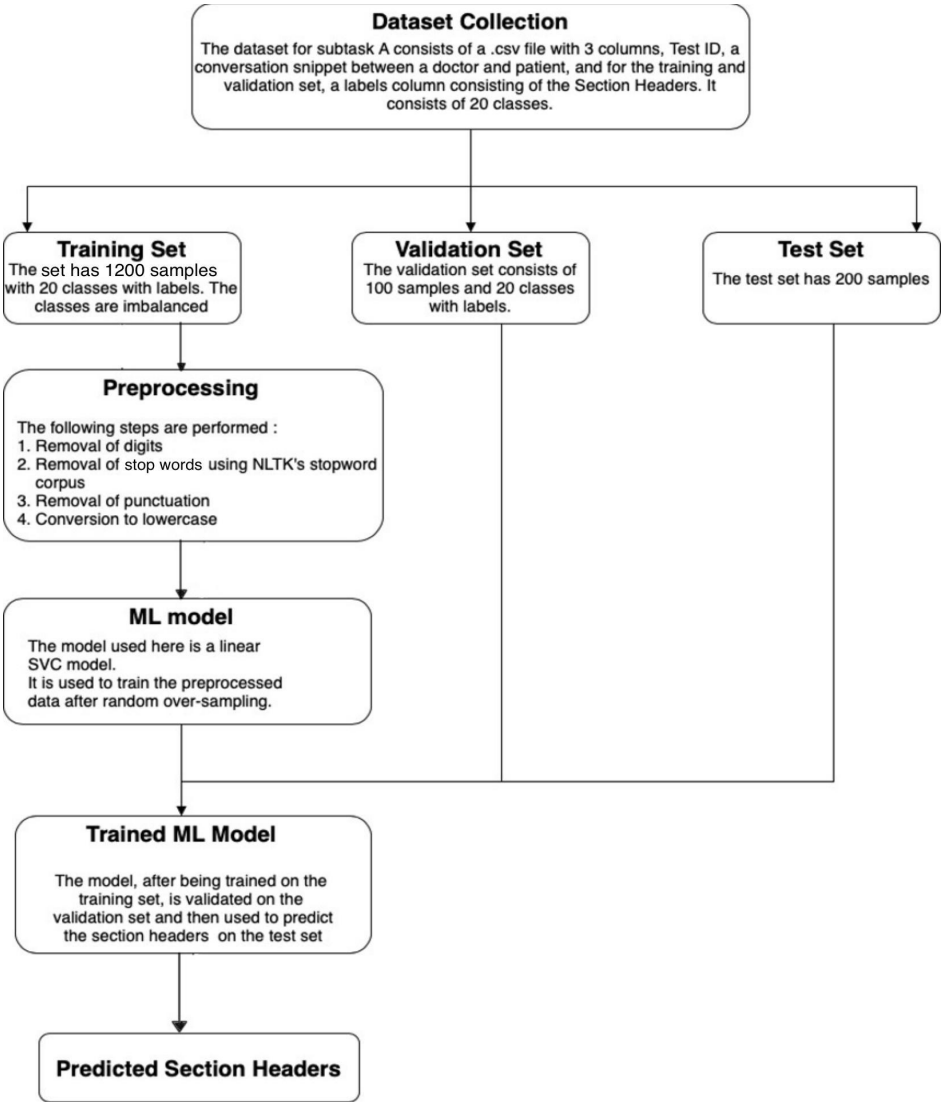

Fig. 2: Proposed system design

3.1. Model 1: Linear SVC with Random Oversampling

Linear SVC is a classification algorithm, used for text classification, image classification, natural language processing, etc. It aims to find a linear decision boundary which effectively separates the different classes present in the data set.

This model uses a Linear Support Vector algorithm along with TF-IDF features (Term Frequency Inverse Document Frequency). It employs several pre-processing techniques, including removing digits, punctuations, and stopwords from NLTK's stopword corpus, as well as the conversion of the text to lowercase. The TF-IDF vectorizer class from Scikit-learn converts text data into a numerical representation.

Due to severe imbalances between the different classes, the accuracy of the model might be affected. To combat this, this technique proposes to use the RandomOverSampler class from the imbalanced-learn library, which helps even out the imbalances between classes and improves the performance of the model. The Linear SVC is then trained on the training dataset, and validated using the validation set. Finally, we use the model to predict the “Section Headers” for the conversation snippets in the test data.

3.2. Model 2: CART with Logistic Regression

Classification and Regression Tree (CART) is a decision tree-based algorithm. It recursively partitions the dataset based on features to construct a tree structure wherein, each internal node represents a feature test, each branch describes a possible outcome of the test, and each leaf node depicts a class label. A CART model with logistic regression was developed for the subtask. The data is first preprocessed using lemmatization which helps in standardising words to their root form. This is followed by label encoding which assigns a numerical value to each class. Countvectorizer then represents the text as a matrix of word counts. The model is then trained and validated on the provided datasets and used to predict section headers of the test dataset.

3.3. Model 3: Convolutional Neural Networks

The data is preprocessed using the Tokenizer class and pad_sequences function from the Keras library to help with standardising it. The embedding layer learns dense vector representations for each word index in the input text, capturing semantic meaning and relationships. It uses 100-dimensional vectors to represent the words. The MaxPooling1D layer downsamples the feature maps by taking the maximum value from each local region, reducing dimensionality while preserving important information. A dropout layer is added to prevent overfitting by randomly setting a fraction of input units to 0 during training, promoting regularization. The Flatten layer converts the feature maps into a 1D vector, followed by a Dense layer with softmax activation for final classification. Softmax assigns probabilities to each class, indicating the model's confidence for each label. The developed model is then trained and validated.

4. Result and Performance Analysis

The Linear SVC model is trained on the training set provided by the organizer. The model displayed a commendable training accuracy of 0.99. This suggests that the model learned the features in the training set effectively. However, there is a drop in validation accuracy, with 0.69. This drop in performance when

evaluated on unseen data may imply that the model is overfitted, ie, the model does not generalize well to new data. Methods to address this issue are discussed in Section 5.
Test accuracies are considered to be representational of how the model performs in real-world situations. Analysis of the model's performance on classifying the conversation snippets in the test dataset shows that it obtained a test accuracy of 0.72. It is also noteworthy that the model trains quickly, in 5 seconds. The current standing of this model after the evaluation of the MEDIQA-Sum subtask A is at rank 10. The results published by the organizers are displayed in Table 4.

Table 3.
Proposed Model and Accuracy

Model	Model used	Training Accuracy	Validation Accuracy	Test Accuracy
Model 1	Linear SVC	0.99	0.69	0.72
Model 2	CART with LR	1.0	0.66	0.69
Model 3(not submitted)	CNN with 17 epochs	0.99	0.63	0.66

5. System Specifications

The hardware and software requirements for the MEDIQA-Sum subtask A of medical dialogue summarization are as follows : (i) Dual-core Intel Core i5, clocked between 2.3GHz and 3.5GHz, 8GB of 2133MHz LPDDR3 onboard memory and 256GB PCIe-based onboard SSD. (ii) MacOS Ventura 13.4 operating system, Python 3.7 package with required libraries like tensorflow, torch, sklearn, nltk, pickle, etc.

6. Inference

The scope for improvement in the task of medical dialogue summarization using Machine Learning models is abundant. The main strengths of our model are that it has high training accuracy and it uses techniques to even out imbalance between classes. Furthermore, it is very time efficient. However, there is reason to believe that this particular model might be overfitted since there is a drop in its validation accuracy.

To address the issue of overfitting, additional training data with more samples, if available, can be used to train the model. Regularization is also an option to be considered, as it adds a penalty to the loss function of the model. It also reduces the impact of less relevant features on the prediction. Early stopping can also be used to identify the optimal point in training. Addressing overfitting could potentially lead to better performance.

It has also been established that there is an imbalance between classes in the training data with some headings having merely 2-8 samples. To mitigate this, random oversampling has been used. However, other techniques such as data augmentation and class weighting can be explored. In data augmentation, new samples for the minority classes are obtained by transforming the existing samples. This may help even out

the imbalance. Class weighting may also prove to be of use, as minority classes can be given more importance during the optimization process.

With further research and experimentation, it seems plausible that accuracy can be improved in the future. The task of manually summarizing clinical notes is a mammoth task. But by implementing the proposed technique, this can be completed in a matter of seconds, which highlights the indispensability of Machine Learning in the medical field.

Table 4.
MEDIQA-Sum 2023 Results

Team	Run	Accuracy	Rank
Cadence	run1	0.82	1
HuskyScribe	run1	0.815	2
Tredence	run2	0.8	3
Tredence	run1	0.8	3
StellEllaStars	run1	0.765	5
Tredence	run3	0.755	6
SSNSheerinKavitha	run3	0.74	7
SSNSheerinKavitha	run2	0.735	8
SuryaKiran	run1	0.735	8
SSNdhanyadivyakavitha	run1	0.72	10
ds4dh	run1	0.71	11
Uetcorn	run3	0.71	11
SKKU-DSAIL	run1	0.7	13
StellEllaStars	run2	0.695	14
SSNdhanyadivyakavitha	run2	0.68	15
StellEllaStars	run3	0.675	16
Uetcorn	run1	0.67	17
MLRG-JBTTM	run1	0.665	18
SSNdhanyadivyakavitha	run3	0.66	19

7. Conclusion

The ImageCLEFmed MEDIQA-Sum 2023, subtask A is implemented and the results are analyzed. The following observations were made in the dataset: There are 1200 samples in the training set. There is an imbalance between the classes, where the maximum number of samples in one class is 351, whereas the

minimum is 2. Several models were trained and tested on the given data sets and further analyzed to improve accuracy.

The Linear SVC model developed showed an improvement of 3% in test data when compared to the CART model with Logistic Regression. A model using CNN was also experimented with but needs further refining as it only has an accuracy of 0.66. Thus, this study concludes that the Linear SVC model is the most suitable for the given task out of the models experimented with the given dataset.

8. Acknowledgements

We would like to thank the ImageCLEF 2023 organisers for providing us with the dataset, which was imperative for this study. We would also like to express our gratitude to our professors for extending their support throughout.

9. References

[1] W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, Overview of the MEDIQA-Sum Task at ImageCLEF 2023: Summarization and Classification of Doctor-Patient Conversations in CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[2] Nicolson, A., Dowling, J., & Koopman, B. (2022, August 25). ImageCLEF 2021 Best ofLabs: The Curious Case of Caption Generation for Medical Images. ImageCLEF 2021 Best of Labs: The Curious Case of Caption Generation for Medical Images|SpringerLink. https://doi.org/10.1007/978-3-031-13643-6_15

[3] Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, Martin Potthast, Overview of the MEDIQA-Sum Task at ImageCLEF 2023: Summarization and Classification of Doctor-Patient Conversations CEUR-WS.org/Vol-3180 - (2022, August 9). . http://ceur-ws.org/Vol-3180/

[4] Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, CLEF 2023 Working Notes Experimental IR Meets Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science (LNCS, volume 13390). https://link.springer.com/book/10.1007/978-3-031-13643-6

[5] Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, Nicola Ferro, Experimental IR Meets Multilinguality, Multimodality, and Interaction 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022. https://link.springer.com/book/10.1007/978-3-031-13643-6

[6] Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, De Rosal Ignatius Moses Setiadi, A Review of automatic text summarization techniques & methods. (2020, May 20). Volume 34, issue 4 (Pages 1011-1624). Review of automatic text summarization techniques & methods - ScienceDirect

[7] Sarker, A., Yang, Y. C., Al-Garadi, M. A., & Abbas, A. (2020, November 13). A Light-Weight Text Summarization System for Fast Access to Medical Evidence. BRIEF RESEARCH REPORT article Front. Digit. Health, 04 December 2020, Sec Health Informatics, Volume 2 - 2020. https://doi.org/10.3389/fdgth.2020.585559

[8] Stergos Afantenos Vangelis Karkaletsis Panagiotis Stamatopoulos, Summarization from medical documents: a survey (2004, December). Summarization From Medical Documents Artificial Intelligence in Medicine, Volume 33, Issue 2, February 2005, Pages 157-177. https://doi.org/10.1016/j.artmed.2004.07.017

[9] Noemie Elhadad, User-sensitive text summarization: Application to the medical domain. Columbia University, ProQuest Dissertations Publishing, 2006. 3203749. https://www.proquest.com/openview/c0c29a6f5b3229c21418f355b1c57c28/1?pq-origsite=gscho

[10] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. Garcıa Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brün- gel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A. Andrei, A. Radzhabov, I. Coman, V. Ko- valev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, social media and recommender systems applications, in Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science LNCS, Thessaloniki, Greece, 2023.