

MLRG-JBTTM at MEDIQA-Sum 2023: Dialogue2Topic Classification

Harshida Sujatha Palaniraj, Keerthan Vinod, Mohith Adluru, Bhuvana Jayaraman and Mirnalinee TT

Sri Sivasubramaniya College of Engineering, Chennai, Tamil nadu, India

Abstract

In medical settings, effective organization and retrieval of information from doctor-patient conversations are essential for providing quality healthcare. This paper focuses on dialogue-to-topic classification, specifically the task of identifying the topic (associated section header) in a conversation snippet between a doctor and a patient. Accurate classification of these topics can significantly improve the accessibility and efficiency of medical records, facilitating medical research, clinical decision-making, and patient care. We used the SVM, Random Forest, and KNN models for evaluation. Among these models, the SVM achieved a ranking of 18, followed by the Random Forest model at 21, and the KNN model at 22, based on the obtained accuracies.

Keywords

Clinical dialogues, Doctor-patient conversation, Medical records, Section headers

1. Introduction

Doctor-patient conversations serve as a primary source of valuable information in the field of healthcare. Also, there is significant research to show that capacity to remember medical records is different among young and adults [1]. The ability to efficiently organize and retrieve information from these conversations is crucial for effective medical record management, knowledge extraction, and medical decision-making. However, extracting meaningful insights from large volumes of conversational data remains a challenging task. By some estimates, doctors may spend as much as two additional hours of administrative work to every one hour spent with patients [2]. This paper focuses on the problem of dialogue-to-topic classification in the context of doctor-patient conversations.

The dialogue-to-topic classification task involves automatically identifying the topics associated with a given conversation snippet [3]. The topic headers serve as labels or categories representing the main topics discussed in the conversation, such as medical history, symptoms, diagnosis, treatment, and follow-up [4]. Accurate classification of section headers enables

CLEF 2023 – Conference and Labs of the Evaluation Forum, September 18-21, 2023, Thessaloniki, Greece

✉ harshida2110349@ssn.edu.in (H. S. Palaniraj); keerthan2110685@ssn.edu.in (K. Vinod);

mohith2110799@ssn.edu.in (M. Adluru); bhuvanaj@ssn.edu.in (B. Jayaraman); mirnalineett@ssn.edu.in (M. TT)


🌐 <https://www.ssn.edu.in/staff-members/dr-j-bhuvana/> (B. Jayaraman);

<https://www.ssn.edu.in/staff-members/dr-t-t-mirnalinee/> (M. TT)

🆔 0000-0002-9328-6989 (B. Jayaraman); 0000-0001-6403-3520 (M. TT)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

efficient indexing, retrieval, and analysis of medical records, supporting various healthcare applications.

The primary objective of this paper is to develop and evaluate advanced techniques for dialogue-to-topic classification in doctor-patient conversations [5]. By accurately assigning topics to conversation snippets, medical practitioners, researchers, and healthcare systems can efficiently navigate and access specific information within medical records, ultimately improving patient care and medical outcomes.

The motivation behind this research stems from the increasing digitization of medical records and the growing volume of doctor-patient conversations recorded in electronic formats. Traditional manual methods of organizing and categorizing these conversations are time-consuming and error-prone, hindering the efficient retrieval and analysis of crucial medical information. By automating the dialogue-to-topic classification process, we aim to enhance the organization and accessibility of medical records, promoting more effective healthcare management.

Section 1 provides an overview of the need for medical dialogue summarization. Section 2 discusses existing research work in this area. Section 3 presents the methodology of the proposed models. Section 4 showcases the results and includes a discussion. Finally, Section 5 concludes the paper, summarizing the key findings and highlighting potential future research directions.

2. Survey of existing systems

The research paper focuses on clinical note summarization, which falls under Subtask A. The medical dialogues are captioned and summarised using machine learning and deep learning models. The dialogue summarization pipeline [6] combines AI and computational linguistics algorithms to automatically generate medical reports. The processes involve speech transcription, triple extraction, triple matching, and report generation .

The authors [7] developed an original automatic synthesis method for medical conversations between a patient and a healthcare professional. Some of the machine learning approaches used are naïve bayes, SVM and genetic algorithms .

The increasing size of language models raises great research interests in parameter-efficient fine-tuning such as LoRA that freezes the pre-trained model, and injects small-scale trainable parameters for multiple downstream tasks (e.g., summarization, question answering and translation). The paper uses a framework that integrates LoRA and structured layer pruning. By tuning 0.6% parameters of the original model and pruning over 30% Transformer-layers, our framework can reduce 50% of GPU memory usage and speed up 100% of the training phase, while preserving over 92% generation qualities on free-text sequence-to-sequence tasks [8].

The authors [9] utilized deep convolutional neural networks to learn complex features and demonstrate through evaluation that their method outperforms existing natural language processing approaches by approximately 15% in terms of accuracy. The study focuses on categorizing medical text fragments and compares their CNN-based approach with three other methods: Sentence Embeddings, Mean Word Embeddings, and Word Embeddings with BOW. The CNN-based approach achieves the highest accuracy due to its ability to capture more complex features.

Manual labeling is a time consuming and errorprone task. One possible solution to this issue is to exploit the large number of unlabeled samples that are easily accessible via the internet. The proposed method [10] selects a batch of informative samples using the posterior probabilities provided by a set of multi-class SVM classifiers, and these samples are then manually labeled by an expert. Experimental results indicate that the proposed active learning method significantly reduces the labeling effort, while simultaneously enhancing the classification accuracy.

3. Methodology

The paper proposes three classifiers for text categorization namely Support Vector Machines (SVM), Random Forest, and K-Nearest Neighbors (KNN) classifiers. Each model offers unique strengths and employs different techniques, as shown in Figure 1, to accurately categorize text sections, providing researchers and practitioners with diverse options for their specific text categorization tasks.

3.1. Classification Methods

3.1.1. SVM

SVM is a suitable choice for dialogue-to-topic categorization from doctor-patient discussions due to its ability to handle high-dimensional data and flexibility in kernel selection. With doctor-patient discussions covering a wide range of subjects, SVM's capability to handle high-dimensional data is valuable for capturing complex correlations between characteristics and topics. Additionally, SVM's support for various kernel functions enables the translation of input data into a higher-dimensional feature space, aiding in capturing nonlinear patterns in dialogue features.

3.1.2. Random Forest

Because of the casual character of the communication, doctor-patient talks may involve noise or outliers. The ensemble technique of Random Forest helps to lessen the influence of noisy or outlier data points on overall classification performance. Furthermore, Random Forest gives a measure of feature importance, which can be helpful in determining the most significant conversation characteristics that contribute to topic categorization.

3.1.3. KNN

In addition to being reasonably simple to build without the need for a lengthy training procedure, KNN makes localised decisions by classifying new instances based on the majority class of its k closest neighbours. This method to localised decision-making might be useful for capturing the local structure and linkages inside doctor-patient dialogues.

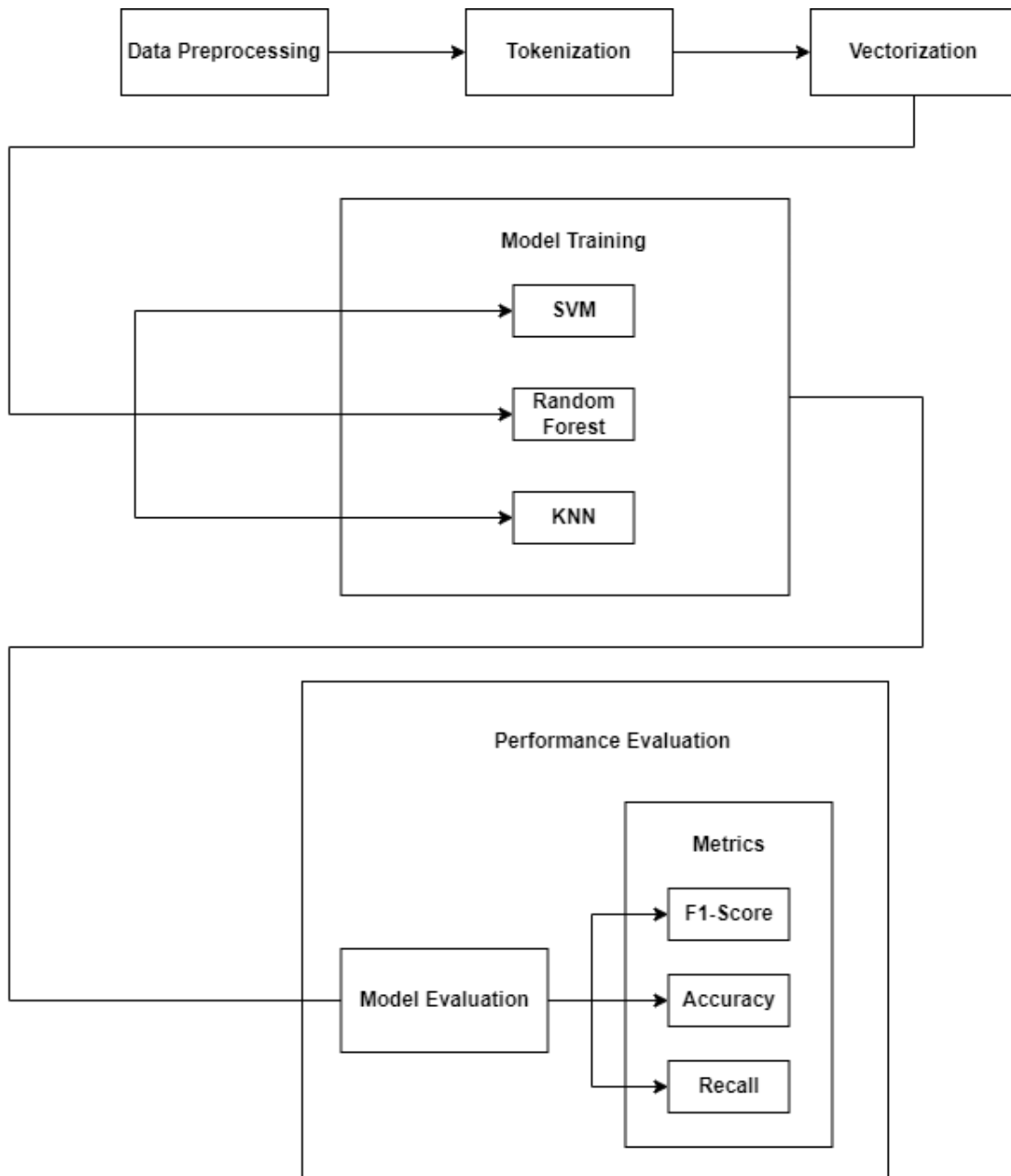


Figure 1: Architecture overview

3.2. Feature Extraction and Classification Pipeline

To convert the textual data into numerical representations, the TF-IDF vectorization technique is applied using the 'TfidfVectorizer' class from scikit-learn. The vectorizer considers unigrams and bigrams and limits the maximum number of features to 999,000. A pipeline is created using

the 'Pipeline' class from scikit-learn.

Each proposed model demonstrates how to preprocess textual data, create a pipeline, and train the classifier on a dataset consisting of dialogues and their corresponding section headers with the use of SVM, Random Forest, and KNN Classifiers respectively.

It consists of two steps: TF-IDF vectorization and the classification. Each classifier is used in the pipeline, with default parameter settings. It aims to find an optimal hyperplane that separates different classes.

3.3. Model Training and Evaluation

The models are trained on the training set, where each classifier learns the patterns and relationships between the text data and their corresponding section headers. The training process involves adjusting the parameters and optimizing the model's performance using techniques such as grid search and cross-validation [11]. To further improve the model's performance, hyperparameter tuning is conducted using a grid search approach. Hyperparameters are parameters that are not learned from the data but are set prior to the training process. Grid search involves specifying a range of values for each hyperparameter and exhaustively searching through all possible combinations to identify the best set of hyperparameters. During the grid search process, different combinations of hyperparameters are evaluated using cross-validation. Cross-validation involves splitting the training set into multiple subsets called folds. The model is trained on a subset of the folds and validated on the remaining fold. This process is repeated for each fold, and the average performance is computed. By performing grid search and cross-validation, the models are trained with the optimal set of hyperparameters, maximizing their performance on the training set. Once trained, the models are evaluated using the test set to assess their accuracy and generalization capabilities.

3.4. Dataset Description

The Medical Conversation Dataset [12] comprises a diverse range of medical scenarios, covering chief complaints, medical history, medications, allergies, family and social history, and general health discussions. Each dialogue snippet within the dataset encapsulates a specific patient's case, providing examples of doctor-patient interactions.

The dataset, named "TrainingSet.csv," consists of a collection of medical conversations and their associated section headers and contents. It is structured in a tabular format with three columns: "id," "section header," and "dialogue."

The dataset comprises 1,201 pairs of conversations in the training set, allowing for extensive training and development of natural language processing (NLP) models [4]. Each conversation pair includes the section headers, which provide a categorical context for the corresponding dialogue content.

Additionally, the dataset includes a validation set with 100 pairs of conversations and their summaries. This subset serves as a valuable resource for evaluating the performance and generalization capabilities of NLP models in summarizing medical conversations.

The pipeline undergoes training on the provided training data to learn patterns and establish a classification hyperplane. Subsequently, it is evaluated by comparing its predicted section

Table 1
Classification Reports for SVM, Random Forest, and KNN Classifiers

Section Headers	precision	recall	f1-score	support	SVM Accuracy	RF Accuracy	KNN Accuracy
ALLERGY	1.00	1.00	1.00	9	1.00	0.74	0.82
ASSESSMENT	0.00	0.00	0.00	6	0.00	0.00	0.25
CC	0.50	0.33	0.40	15	0.40	0.57	0.37
DIAGNOSIS	0.00	0.00	0.00	4	0.00	0.00	0.00
DISPOSITION	0.67	0.67	0.67	3	0.67	0.00	0.67
EDCOURSE	0.00	0.00	0.00	3	0.00	0.00	0.00
EXAM	1.00	0.75	0.86	4	0.86	0.33	0.86
FAM/SOCHX	0.92	0.96	0.94	71	0.94	0.79	0.90
GENHX	0.62	0.95	0.75	55	0.75	0.87	0.71
GYNHX	0.00	0.00	0.00	4	0.00	0.00	0.00
IMAGING	0.00	0.00	0.00	2	0.00	0.00	0.00
LABS	0.00	0.00	0.00	2	0.00	0.00	0.00
MEDICATIONS	0.92	1.00	0.96	11	0.96	0.82	0.74
PASTMEDICALHX	0.69	0.75	0.72	24	0.72	0.39	0.70
PASTSURGICAL	0.90	0.82	0.86	11	0.86	0.64	0.80
PLAN	1.00	0.25	0.40	4	0.40	0.00	0.40
ROS	0.88	0.54	0.67	13	0.67	0.53	0.80
accuracy			0.77	241	0.77	0.73	0.73
macro avg	0.53	0.47	0.48	241	0.48	0.35	0.47
weighted avg	0.72	0.77	0.73	241	0.73	0.68	0.70

headers against the actual section headers to calculate its accuracy.

The dataset, stored in a CSV file, is split into training and test sets with a ratio of 80% for training data and 20% for testing data.

4. Results and discussion

Google Colab notebook was used to train the model with RAM size of 8GB on a 2.3GHz Intel Xenon CPU.

The three models are trained and accuracy metrics is used to study the performance of the model during training. The validation accuracy of SVM, Random Forest Classifier and KNN Classifier models are 77%, 73% and 73% respectively.

Among the three models, the SVM classifier showed the highest performance, with notable precision, recall, and F1-score values for various section headers. The SVM model's generalization capabilities and handling of high-dimensional feature spaces contributed to its superior performance. However, the Random Forest and KNN classifiers also delivered competitive results, indicating their suitability for dialogue-to-topic classification tasks.

The variation in performance across different section headers suggests that the models excel in classifying certain sections but struggle with others. The inherent complexity and variety of doctor-patient dialogues is an essential element. The language used, the variety of subjects

covered, and the dynamics of the discussion can all have a major influence on the categorization assignment. Some subjects may be simpler to categorise because of different patterns or explicit references, but others may be more difficult because of ambiguity or overlapping language clues. Predictions were biased due to the class imbalance, with the model favouring the dominant class. As a result, the majority class performed poorly whereas the minority classes performed well in terms of precision and recall.

4.1. Submitted runs and results

The proposed models, SVM, Random Forest and KNN, have obtained a testing accuracy of 66.5%, 57% and 56.5% respectively as reported in Table 4 [4]. The variations observed in the analysis of different topics can be attributed to the sensitivity of each model. The models employed in this study incorporate distinct algorithms and make different assumptions, thereby leading to variations in their performance across various topics. Based on these accuracies, the models achieved rankings of 18, 21, and 22 respectively. The observed variations in performance across different section headers are as follows:

4.1.1. Support Vector Machines (SVM) Classifier

The model achieved perfect precision, recall, and f1-score (all 1.00) for the "ALLERGY" section. Poor performance was observed for sections namely, "ASSESSMENT," "DIAGNOSIS," "EDCOURSE," "IMAGING," "LABS," and "GYNHX" with all metrics being 0.00, as shown in Table 1.

4.1.2. Random Forest Classifier

As depicted in Table 1, good performance for sections namely "ALLERGY," "FAM/SOCHX," "GENHX," "MEDICATIONS," and "PASTSURGICAL" with reasonably high precision, recall, and f1-scores and poor performance for sections like "ASSESSMENT," "DIAGNOSIS," "DISPOSITION," "EDCOURSE," "IMAGING," "GYNHX," and "IMMUNIZATIONS" with all metrics being 0.00 contributed to the obtained accuracy.

4.1.3. K-Nearest Neighbours (KNN) Classifier

Perfect precision (1.00) for sections like "ALLERGY," "EXAM," "FAM/SOCHX," "MEDICATIONS," and "PLAN." and poor performance for sections like "ASSESSMENT," "DIAGNOSIS," "EDCOURSE," "IMAGING," "LABS," "GYNHX," and "IMMUNIZATIONS" with all metrics being 0.00 were observed in Table 1.

4.2. Limitations

SVMs are susceptible to noisy or incorrectly labelled data. SVM performance may suffer in the presence of noisy conversation data or wrongly labelled samples. On large datasets of doctor-patient conversations, SVM model training and tweaking may consume a substantial amount of computing time.

Table 2

Evaluated results of ImageCLEF MediQA

Rank	Participant	Run	Accuracy
18	MLRG-JBTTM	run1	0.665
21	MLRG-JBTTM	run2	0.57
22	MLRG-JBTTM	run3	0.565

Although beneficial in terms of performance, Random Forest models can be challenging to comprehend because to their ensemble nature. It might be difficult to draw conclusions and comprehend the unique feature contributions to subject categorization from the ensemble of decision trees.

KNN takes into account all features equally while determining distances, which might be problematic if the dataset contains irrelevant or noisy characteristics. These characteristics may increase noise and have a detrimental effect on classification accuracy.

5. Conclusion

The focus of MLRG-JBTTM's submission is to accurately assign section headers to conversation snippets by using dialogue-to-topic classification techniques. By addressing this task, we aimed to improve medical record management, knowledge extraction, and medical decision-making processes.

To achieve this objective, three classification models, namely Support Vector Machines (SVM), Random Forest, and k-Nearest Neighbors (kNN) classifiers, were proposed and evaluated. These models were chosen due to their proven effectiveness in various text classification tasks.

The study's results have practical applications in healthcare settings because they enable automated analysis of doctor-patient conversations to get significant information. Future study may concentrate on improving dialogue to topic classification systems. This may entail investigating deep learning-based systems such as recurrent neural networks or transformers, which have showed promise in natural language processing. Model understanding and classification accuracy can be improved by including domain-specific knowledge or contextual information, such as patient demographics or medical data. Integration of domain-specific characteristics such as named entity identification or sentiment analysis may improve the models' capacity to collect complex information in discussions.

References

- [1] L. Mcguire, Remembering what the doctor said: Organization and adults' memory for medical information, *Experimental aging research* 22 (1996) 403–28. doi:10.1080/03610739608254020.
- [2] C. Sinsky, L. Colligan, L. Li, M. Prgomet, S. Reynolds, L. Goeders, J. Westbrook, M. Tutty, G. Blike, Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties, *Annals of Internal Medicine* 165 (2016). doi:10.7326/M16-0961.

- [3] B. Cao, K. Ma, Y. Liu, Y. Xu, L. Zhu, Intention classification in multiturn dialogue systems with key sentences mining, *Computational Intelligence* 37 (2020). doi:10.1111/coin.12345.
- [4] 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*.
- [5] J. Schuurmans, F. Frasincar, Intent classification for dialogue utterances, *IEEE Intelligent Systems PP* (2019) 1–1. doi:10.1109/MIS.2019.2954966.
- [6] S. Molenaar, L. Maas, V. Burriel, F. Dalpiaz, S. Brinkkemper, Medical Dialogue Summarization for Automated Reporting in Healthcare, 2020, pp. 76–88. doi:10.1007/978-3-030-49165-9_7.
- [7] J. López Espejel, Automatic summarization of medical conversations, a review, 2019.
- [8] Y. Zhu, X. Yang, Y. Wu, W. Zhang, Parameter-efficient fine-tuning with layer pruning on free-text sequence-to-sequence modeling, 2023. arXiv:2305.08285.
- [9] M. Hughes, I. Li, S. Kotoulas, T. Suzumura, Medical text classification using convolutional neural networks, *Studies in Health Technology and Informatics* 235 (2017). doi:10.3233/978-1-61499-753-5-246.
- [10] M. Goudjil, M. Koudil, M. Bedda, N. Ghoggali, A novel active learning method using svm for text classification, *International Journal of Automation and Computing* 15 (2016). doi:10.1007/s11633-015-0912-z.
- [11] R. Ghawi, J. Pfeffer, Efficient hyperparameter tuning with grid search for text categorization using knn approach with bm25 similarity, *Open Computer Science* 9 (2019). doi:10.1515/comp-2019-0011.
- [12] B. Ionescu, H. Müller, A. Drăgulescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. Garcia Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, 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. Kovalev, 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, socialmedia 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*.