People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
M'hamed Bougara University - Boumerdès

Classification Of Medical Records By NLP

By
ZORGANI Alaaeddine
BENALLAL Amira
TOULMOUTINE Merouane

A thesis submitted to the School of M'hamed Bougara University - Boumerdès in partial fulfillment of the requirements for the degree of Bachelor of ...

Boumerdes City, Algeria
May 2025

Contents

Chapter 1

INTRODUCTION

1.1 Introduction to Artificial Intelligence Concepts

1.1.1 Artificial Intelligence (AI):

Artificial Intelligence is a wide and multi-faceted subject area of computer science that attempts to make systems that can carry out tasks normally demanding human intellect. These tasks have a very broad spectrum, which includes difficult problem-solving, learning algorithms, visual and audio perception, knowledge of natural language, and complicated decision-making. AI research explores various approaches from logical-based systems, probabilistic paradigms, and biological computing, all driving towards a common objective of realizing machines with the ability to reason, learn, and act with a human-like autonomy level [**?**].

1.1.2 Machine Learning (ML):

Within the general realm of artificial intelligence, Machine Learning (ML) has emerged as an extremely powerful and innovative area of research. Rather than depending on explicit coding for each eventuality, ML algorithms provide systems with the ability to learn to identify patterns, make predictions, and enhance performance through purely data inputs. By analyzing large datasets, the algorithms determine inherent structures, patterns, and interdependencies and refine their internal models progressively without ongoing human input. Some typical paradigms are supervised learning (learning from labeled examples), unsupervised learning (discovery of patterns in unlabelled data), and reinforcement learning (error and trial learning) [**?**].

1.1.3 Artificial Neural Networks (ANNs):

One prominent class of machine learning models, drawn from the complex structure and operation of the biological brain, is Artificial Neural Networks (ANNs). ANNs are composed of networked processing units, commonly referred to as neurons, which are usually stacked in layers with input, hidden, and output layers. Each connection among neurons transmits a signal and has a corresponding weight that depicts the strength or significance of the connection. Through the learning process, usually with the assistance of algorithms such as backpropagation, the weights are iteratively tuned based on the performance of the network on the training data. Information flows through the network, being transformed through activation functions within each neuron, thereby

enabling the artificial neural network to learn and depict intricate, non-linear relationships among inputs and outputs [**?**].

1.1.4 The Perceptron:

The Perceptron is a starting point for a wide variety of complex neural networks and was initially proposed by Frank Rosenblatt in the late 1950s. The Perceptron model essentially captures a single artificial neuron. It receives a quantity of binary inputs, each multiplied by its corresponding weight, sums these weighted inputs, and finally applies a basic threshold or step function to give a single binary output. Although a single Perceptron can only solve linearly separable problems, its creation was a breakthrough, illustrating the power of learning algorithms and establishing the fundamental foundation on which more advanced, intricate multi-layered network structures developed. [**?**].

1.1.5 Deep Learning (DL):

Following the foundational concepts of artificial neural networks (ANNs), deep learning (DL) has achieved remarkable progress in various domains of artificial intelligence (AI) in recent years. Deep learning has a structure defined by the utilization of neural networks with more than one hidden layer between the input and output layers. This depth enables the model to discover hierarchical representations of data, automatic discovery, and extraction of intricate patterns and features at multiple levels of abstraction – from low-level simple features to high-level complex concepts. Driven by the availability of large datasets and high-performance computing hardware (e.g., GPUs), DL has achieved state-of-the-art performance in challenging fields like computer vision (image classification), natural language processing (machine translation, sentiment analysis), and speech recognition [**?**, **?**].

1.1.6 Transformers:

The Transformer model, introduced by Vaswani et al.[?] [**?**] in 2017, replaces recurrence and convolution with a purely attention-based mechanism. In this architecture, each input element computes a weighted sum over all other elements via a scaled dot-product role of attention, which enables the model to particularly discover long-range dependencies in a single step. To strengthen this capacity, a number of attention "heads" project the inputs into multiple subspaces—allowing the model to simultaneously learn a vast range of relationships. Since attention alone does not encode order, learned or fixed positional encodings are embedded in the input embeddings, providing hints regarding the relative or absolute the token configuration. Such a combination of self-attention and positional information offers a flexible, highly parallelizable framework which has emerged as the foundation for a large number of sequence modeling applications

1.1.7 Transfer Learning and Fine-Tuning

Fine-tuning is a specialized type of transfer learning in which a pre-trained model, initially developed on a large general-purpose corpus, is subsequently trained on a small task-specific corpus to retune its learned representations for a new task. Rather than initializing parameters randomly, fine-tuning starts with the pre-trained weights and

continues training with a smaller learning rate, enabling the model to tune its internal features to the nuances of the target task while retaining the general knowledge that it has acquired thus far. This method generally decreases the amount of data needed and the computational expense relative to training a model from scratch, usually resulting in improved performance when there is little data available that is relevant to the task.

1.2 Introduction to Natural Language Processing Concepts

1.2.1 Natural Language Processing (NLP)

NLP is a field of artificial intelligence that deals with the interaction of computational systems with humans using natural everyday language. The aim ive is to teach computers to analyze and process large amounts of natural language data. NLP can alows machines and create intelligent systems that can understand, Interpret, generate text, and develop human language in a way that is meaningful. and meaningful.

1.2.2 NLP Tasks

it help process human data to make it understandable by machines ,It's included in:

- Coreference resolution: identifying if and when two words refer to the same entity. (such as "she" = "Mary"), or identifying metaphors.
- Named entity recognition: identifies words or phrases as useful entities. NER identifies "NewYork" as a location or "Mary" as a person's name.
- Part-of-speech tagging: based on its use texts and context. Like "make" as a verb in "I can make a paper plane," and as a noun in "What make of car do you own?"
- Word sense disambiguation: Determining the correct meaning of a word with multiple meanings based on context .

1.2.3 How NLP Works

NLP make combines computational techniques to analyze, generate and understand human language in a way that machines can process. Here is an overview:

Text Preprocessing

prepares raw text for analysis by transforming it into a format that machines can more easily understand. First tokenization, by simplifying the text by dividing to sentences and words. Next, lowercasing is applied to standardize the text by converting all characters to lowercase. Then removing common words , symbols, punctuation, while Stemming or lemmatization reduces words to their root .

Feature Extraction

converting raw text into numerical representations that machines can analyze and interpret by using Bag of Words and TF-IDF, the most advanced methods include word embeddings like Word2Vec or GloVe.

Text Analysis

it involves interpreting and extracting meaningful information from text data . Such as Part-Of-Speech (POS) tagging (identifies grammatical roles of words), NER (detects specific entities like names, locations and dates...).

Model Training

Processed data is then used to train machine learning models, which learn patterns and relationships within the data.

1.2.4 Word embedding in NLP

In NLP, word embedding is a representation of a word in vectors . to facilitate the process of classification and collecting more information . The development of embedding to represent text has played a crucial role in advancing NLP and machine learning (ML) applications. Word embeddings have become integral to tasks such as text classification, sentiment analysis, machine translation and more.

1.2.5 Text Representation Techniques

Text representation is a fundamental task in NLP that transforms raw text into vector representations with numerical values that can be readily used by machine learning models.

Traditional Approach

like BoW, which represents a document as an unordered set of words with their frequencies, ignoring word order.

Neural Approach

Such as Word2Vec is to capture semantic and syntactic relationships between words based on their co-occurrence patterns in a large corpus of text.

Transformer-Based Contextual Embeddings

Transformer-based models produce contextual embeddings by encoding an entire input sequence by stacked self-attention and feed-forward layers, which outputs a sequence of hidden-state vectors—assigned to each token—that capture both proximal context and distant dependencies. dependencies. Practically, one usually takes advantage of the final hidden state of a specific class- classification token (e.g., [CLS]) or uses a pooling procedure (mean or max pooling) Aggregate the token vectors in order to obtain a uniform-length representation for the whole sequence. Or, intermediate layers can be merged (through weighted sums or concatena- dimension) to emphasize features

learned at different depths. Because these embeddings dynamically adjust according to context words, they provide more informative, task-oriented representations compared to static vectors. As input to downstream classifiers, transformer-based embeddings were found to improve performance on di- verse NLP tasks—particularly in technical fields where fine-tuning on domain data additionally refines the representations.

Chapter 2

RELATED WORK

Chapter 3

METHODOLOGY

3.1 Introduction

In this chapter, we will explain how to use word embedding and machine learning on an existing dataset to predict medical orientation (medical class or medical service) from a natural language medical description. We will describe in detail the dataset used, the type of word embedding used, the machine learning algorithms applied, and the results obtained with our learning model.

3.2 Data Preparation and Preprocessing

Data preparation [1] is a critical step in any data analysis or machine learning project. It involves a variety of tasks that aim to transform raw data into a clean and usable format. Properly prepared data ensure more accurate and reliable analysis results, leading to better decision-making and more effective predictive models. To do so, we remove the HTML tags and accented characters and remove the spaces and punctuation slashes. The text data need to be converted to lowercase and the stop words need to also be removed. Once the above steps are completed successfully, they should be stemmed and lemmatized. This is the process that can change the derived words to their root word which helps in maintaining the information. It may include handling missing values, encoding categorical variables, and identifying outliers [2]. Investing time in data preparation can significantly reduce errors and improve the overall performance of analytical models.

3.2.1 About the dataset

Diagnoise me is an LFQA dataset of dialogues between patients and doctors based on factual conversations from blueicliniq.com and blueaskadoctor24x7.com that aims to collect more than 257k different questions and prescriptions from patients.

This dataset is in JSON format and contains multiple columns, but we first need to add a column blue"Specialty" to identify medical specialties for future purposes.

3.2.2 Adding the 'Specialty' column

This column contains string-type values that refer to different specialties used in this dataset, but how can we identify the right one?

	id	Description	Doctor	Patient
0	0	Q. What does abutment of the nerve root mean?	Hi. I have gone through your query with dilige...	Hi doctor,I am just wondering what is abutting...
1	1	Q. Every time I eat spicy food, I poop blood. ...	Hello. I have gone through your information an...	Hi doctor, I am a 26 year old male. I am 5 fee...
2	2	Q. Will Nano-Leo give permanent solution for e...	Hi. For further doubts consult a sexologist on...	Hello doctor, I am 48 years old. I am experien...
3	3	Q. Will Kalarchikai cure multiple ovarian cyst...	Hello. I just read your query. See Kalarachi K...	Hello doctor, I have multiple small cysts in b...
4	4	Q. I masturbate only by rubbing the tip of the...	Hi. For further doubts consult a sexologist on...	Hi doctor, During masturbation I just rub the ...
...	...	...	...	...

Figure 3.1: the 'Diagnoise me' Dataset.

We can do this by extracting them from the text data of the **Doctor**'s column, each row of this column contains a unique URL that can have the medical specialty written in it (mostly at the end of the URL), for example: red"Hi. For further doubts, consult a sexologist online –¿https://www.icliniq.com/ask-a-doctor-online/sexologist"

With a simple function code, we can extract the medical specialty like 'sexologist', normalize it, and then add it in the **Specialty** column. We get the following result:

	id	Description	Doctor	Patient	Specialty
0	0	Q. What does abutment of the nerve root mean?	Hi. I have gone through your query with dilige...	Hi doctor,I am just wondering what is abutting...	neurology
2	2	Q. Will Nano-Leo give permanent solution for e...	Hi. For further doubts consult a sexologist on...	Hello doctor, I am 48 years old. I am experien...	sexology
4	4	Q. I masturbate only by rubbing the tip of the...	Hi. For further doubts consult a sexologist on...	Hi doctor, During masturbation I just rub the ...	sexology
15	15	Q. What does abutment of the nerve root mean?	Hi. I have gone through your query with dilige...	Hi doctor,I am just wondering what is abutting...	neurology
17	17	Q. Will Nano-Leo give permanent solution for e...	Hi. For further doubts consult a sexologist on...	Hello doctor, I am 48 years old. I am experien...	sexology
...	...	...	...	...	...
31059	31059	Q. Brother-in-law got a stent for stomach canc...	Hi. I understand your concern. For further dou...	Hi doctor, My brother-in-law has stomach cance...	surgical-oncology
31060	31060	Q. Having osteoporosis and spine fracture. How...	Hi. I have gone through your message and under...	Hi doctor, I have osteoporosis and L3 spine fr...	orthopaedician-and-traumatology

Figure 3.2: Updated dataset.

After reviewing our dataset, we notice that there are multiple medical specialties with very few representations in the dataset(low-frequency classes), which can make the model overfit or make poor predictions.

By grouping similar ones together, we provide more training data per category, which often improves performance, generalization, and makes it easier for the model to learn patterns. For example: we know that a "neuro-surgeon" and a "spine-health-specialist" likely deal with related conditions, even if they have different titles.

In the end, we end up with a dataset of 18818 rows x 6 columns with no duplicates or missing values. A cleaned dataset ready for the next step of text classification which is **Text Embaddings**.

Variable name	**Variable description**	**Label**
Description	Short description of the patient's thesis	Feature
Specialty	Medical specialty	Target

Table 3.1: Description of the variables

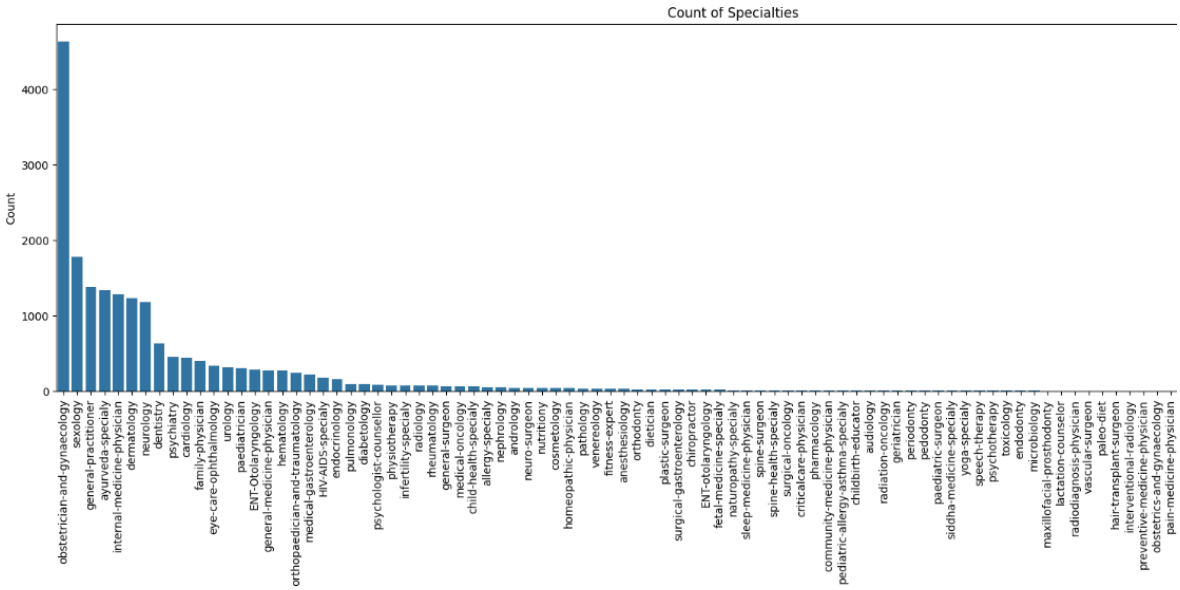

Figure 3.3: Before the grouping.

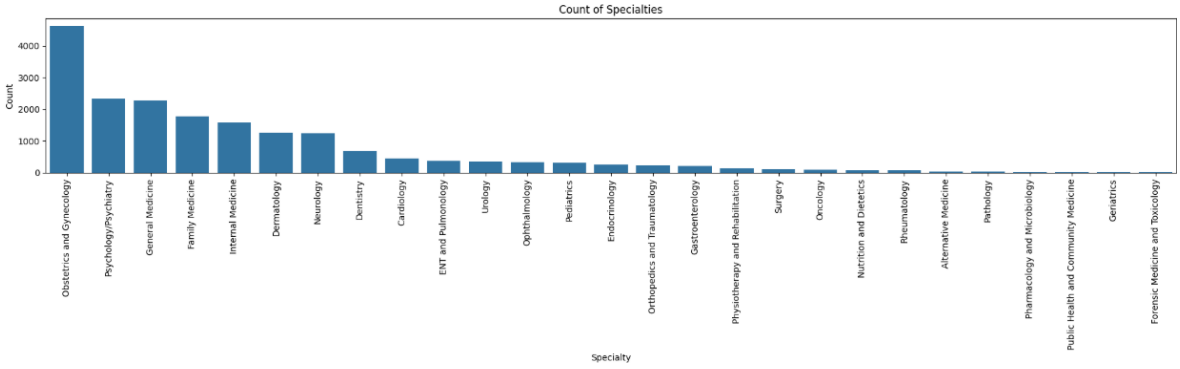

Figure 3.4: After the grouping.

Chapter 4

PROTOTYPING

Chapter 5

IMPLEMENTATION

Chapter 6

RESULT

Chapter 7

DISCUSSION

Chapter 8

CONCLUSION

Bibliography

[1] H. Yang, "Data preprocessing," *Pennsylvania State University: Citeseer*, 2018.

[2] A. Unknown, "The imperative of data cleansing," 2020. Accessed: 2025-04-18.

Appendix A

LISTINGS