A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

Improving Patient Referral Classification Through Deep Learning and Natural Language Processing (NLP)

Marouan Kamoun

Work project carried out under the supervision of:

Prof. Patricia Xufre

Prof. Susana Lavado

11/12/2023

Abstract

The health sector is showing interest in machine learning techniques to enhance various aspects of patient care. In this paper, the application of deep learning and Natural Language Processing (NLP) techniques to improve patient referral classification at Hospital Garcia De Orta (HGO) is explored. Results showed that these advanced techniques outperformed traditional machine learning models, highlighting the effectiveness of combining deep learning and NLP.

Keywords: Business analytics, Machine learning, NLP, Deep Learning, Classification, Prediction

Acknowledgments

I would like to express my gratitude and my warmest thanks to my Work Project advisors, Professors Patricia Xufre and Susana Lavado, for their guidance and support provided throughout this master's thesis. I am also grateful to my colleagues for their collaborative efforts during the Project Based Learning and to my PBL mentors, Gustavo Brito and Susana Lavado, for their guidance. Finally, I would like to express my deepest gratitude to my parents for their support during my academic journey.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Table of Contents

Introduction

Hospital Garcia De Orta (HGO), EPE is a hospital located in Almada, Portugal and is a legal entity under public law of business nature, with administrative, financial and patrimonial autonomy. HGO resulted from the replacement of the former Hospital da Misericórdia de Almada/Hospital Distrital de Almada in 1991. It covers 350,000 people in the municipalities of Almada and Seixal with more than 2900 employees and has a capacity of about 590 beds ("Institution Category - Hospital Garcia de Orta" n.d.). Additionally, some services of the Garcia de Orta Hospital respond to the populations of the entire Setubal Peninsula, which indicates the importance of this Hospital.

HGO is facing a problem in the clinical screening process within the Neurology Department. Since 2013, accessibility to the hospital appointments has been managed through a nation-wide integrated system created by the ministry of health for referencing and managing access to the first hospital specialty consultation, called Consultation on Time (CTH), which has largely improved the process. However, there are still constraints that reduce overall efficiency. Specifically, the clinical screening phase is heavily reliant on specialized human resources, primarily the medical staff. This reliance is taking away valuable time that could be allocated to other clinical activities. It is in that context that HGO felt the need to test a new technique that could improve this process. The primary objective was to develop a machine learning model capable of predicting whether a referral should be accepted or rejected, offering doctors doing the screening such prediction. This suggestion would then be validated by the doctors.

Through the implementation of this model, HGO would be able to improve its referral screening process and elevate the overall performance of the neurology department. In fact, by offering doctors suggestions for decision making, the model would allow the hospital to avoid the waste of time caused by the screening of the free text, which can often lack information or contain incomplete and vague information.

It is important to note that a machine learning model to address this issue was already developed during a Project-Based Learning (PBL) project. More details about this model will be displayed in the next sections. The goal of this thesis is to assess whether it is possible to develop a more effective model by applying deep learning techniques to the data.

Deep Learning

Introduction to Deep Learning

Deep Learning is a subset of machine learning techniques capable of self-learning hidden patterns in data in order to make predictions (Madan and Madhavan 2020). Deep learning systems are based on artificial neural networks, which architecture is inspired by the structure of the human brain (see Appendix 1). Just like how the "biological neuron" in a human brain gets information, analyses it and gives an output, a "node" in an artificial neural network, which is the artificial neuron, receives inputs and turns them into an output (Madan and Madhavan 2020).

The history of deep learning started in 1943 when Walter Pitts and Warren McCulloch developed the first mathematical model of a neural network. In 1957, Frank Rosenblatt introduced the perceptron, which allowed computers to learn from the past and was an important step in the development of neural networks (Dukes 2023). After that, the first AI winter took place, in the 1970-1980s, due to computational limitations that halted further developments in the technology. It took until 1986 for neural networks to regain life with the introduction of backpropagation, which improved their training (Beam 2017). But again, because backpropagation couldn't handle large problems, it was time for a second AI winter (Beam 2017). It persisted until 2006 when Hinton published a paper that improved backpropagation and marked the creation of the "deep learning" concept (Lv and Lei 2020). An overview of the timeline is represented in Appendix 2.

Nowadays, thanks to the presence of high-performance computers and the support of big data resources, deep learning is used by many businesses and is present in daily applications in our everyday life (Lv and Lei 2020). It is used in many fields, such as fraud detection, automatic facial recognition systems, driverless cars and language translation, which were before considered tasks exclusive to human intelligence.

Deep Learning Techniques

Having established the definition of deep learning and its importance, some of most prevalent and influential types of neural networks in the field that can be relevant for the health sector will be explored in the following section.

Convolutional Neural Networks (CNN)

A convolutional neural network (CNN) is a feedforward neural network, meaning that the input data travels in a single direction from the input layer to the output layer. It is mainly used for image recognition as it mimics the visual perception of the human brain (Li et al. 2022). However, it is important to note that CNN can also be applied to text analysis tasks. It typically has three types of layers: convolution and pooling layers which are responsible for features extraction, and fully connected layers that have the role of transforming the extracted features into the ultimate result (Yamashita et al. 2018).

A convolution layer is the key element of the CNN network. It performs a convolution operation by taking a small matrix, known as 'filter' or 'kernel', and applying it to the input data in order to detect complex patterns such as shapes or edges and generate feature maps as shown in Appendix 3. These extracted features are then passed through a pooling layer in order to reduce the dimensionality. This type of layers has the ability to reduce the size of feature maps without losing its important information (Qiwei 2023). The most common pooling operation is the Max Pooling which consists of taking the maximum value in a patch of the feature map (Yamashita et al. 2018). An example is shown in Appendix 4. Finally, the fully

connected layers use the features learned by the previous layers to create the final output such as classification or prediction.

Although regular neural networks, that consist only of fully connected layers, could effectively handle small images, they faced significant challenges when dealing with larger ones due to the huge number of parameters they need (Qiwei 2023). CNN was able to handle those challenges thanks to the significant role of the convolution and pooling layers. These architectural made CNNs the currently predominant neural network for image processing and a good choice for text processing.

Recurrent Neural Networks (RNN)

A recurrent neural network (RNN) is a type of artificial neural network mainly used with sequential and time series data ("What Are Recurrent Neural Networks? | IBM" n.d.). Unlike feedforward neural networks, such as the CNN, RNN has connections that can loop back on themselves making it suitable for problems such as natural language processing (NLP) and speech recognition (Prathap 2020). The most popular applications where RNN is used are Apple's Siri and Google's translation.

RNN is composed of three layers: input, recurrent hidden, and output layers. The secret of this model lies in the hidden layers. They perform computations on the input data and the outcome is called 'hidden states' which has a memory containing information retained from previous time steps of a sequence (Kalita 2023). This ability to remember past information is what makes RNN relevant for tasks involving sequential data. For example, to predict the translation of a word, the previous words are necessary so we can know the context. An example of RNN is shown in Appendix 5.

However, RNN suffers from two major problems related to the size of the gradient. The gradient, acting as a vector pointing in the direction of the greatest rate of increase of a function,

symmetrically points in the direction of the greatest rate of decrease of the loss function, which quantifies the dissimilarity between the predicted output and the actual target. The first problem, known as ‘vanishing gradients’ occurs when the gradient is too small resulting in the lack of capturing dependencies from previous long-term time steps. The second problem, ‘exploding gradients’, occurs when the gradient is too large resulting in the instability of the model (“What Are Recurrent Neural Networks? | IBM” n.d.).

Long Short-Term Memory Networks (LSTM)

Long Short-Term Memory (LSTM) network is a type of RNN and is a remarkable advancement in this field. It was mainly introduced by Hochreiter and Schmidhuber to overcome the limitations of traditional RNN consisting of vanishing gradients and capturing long-term dependencies (Lipton, Berkowitz, and Elkan 2015).

A LSTM network has the same architecture as standard RNN, with the introduction of additional memory cells. While the hidden states are short term memory, these additional memory cells serve as the long-term memory which is the fundamental concept for capturing extended dependencies. They serve as an internal memory bank which can store and retrieve information over time (Saxena 2023).

To control the flow of information over the memory cell, LSTM incorporates a gating mechanism consisting of three gates. The first one is the ‘Forget gate’, which is responsible for making the decision of whether to keep information from the previous time step or to ignore it. The second gate is the ‘input gate’, which chooses what new information is relevant and should be added to the memory cell. The final gate consists of the ‘output gate’, which decides what information should be passed to the next timestamp (Saxena 2023). These significant gates are very important as they allow LSTMs to effectively capture long-term dependencies, address the challenge of vanishing gradients, and model intricate sequential patterns. Appendix 6 presents how LSTM works.

Literature Review of Deep Learning in the Health Sector

In recent years, the healthcare sector became one of the most important fields that got interested in machine learning techniques, and especially deep learning, to achieve better clinical results and improve efficiency of the medical units. As of that, many studies were conducted over the years in order to understand the benefits of using deep learning in the medical sector.

Deep learning can be used for medical images analysis. It helps identifying hidden patterns and anomalies and predicting diseases through analyzing images. While the analysis of images was typically executed by doctors and specialists, the huge amount of data resulted in the overwhelming of medical staff and as a result, the need for assistance of machine learning programs and techniques (Shen, Wu, and Suk 2017). In certain fields, those techniques achieved 91% accuracy in terms of right diagnosis compared to only 79% achieved by humans (Hamed et al. 2020). In 2021, deep learning was expected to reach $300 million of investment in the clinical imaging market, which is more than the total amount spent by the analysis industry in 2016 (Razzak, Naz, and Zaib n.d.).

Clinical decision making using Electronic Health Records (EHR), which consists of different electronic medical records of a patient, is another use case for deep learning. In fact, the complexity of EHR data and its exponential growth is one of the primary reasons that explain the need for deep learning techniques (Poongodi et al. 2021). Using those techniques to predict clinical outcomes surpassed traditional machine learning predictive models (Rajkomar et al. 2018). It registered 0.94 AUROC in predicting in-hospital mortality compared to only 0.86 AUROC using a logistic regression model (Rajkomar et al. 2018). AUROC stands here for area under the receiver operator curve which calculates the model's ability to distinguish between classes.

Moreover, deep learning, and particularly CNNs, have shown a good performance for the health sector text classification. This technique for classifying medical text at the sentence level remarkably registered 15% better results compared to other Natural Language Processing (NLP) techniques (Hughes et al. n.d.). Another example is text classification of adverse nursing events (Lu et al. 2021). The study highlighted the superior performance, in terms of different evaluation metrics but especially the accuracy metric, of the CNN model which achieved an accuracy of 78% while other machine learning algorithms accuracies were below 70% (Lu et al. 2021).

Embeddings in Deep Learning

Embeddings has emerged as one of the most used tools in deep learning enhancing the capability of models to effectively capture intricate patterns within textual data. Embeddings, in the context of NLP, refer to the transformation of words or phrases into continuous vectors representation to detect similarities between words. It maps a word in a vocabulary to a point in a high-dimensional vector space, where the spatial proximity between vectors preserves semantic relationships between corresponding words (Wang, Nulty, and Lillis 2020).

The common method for generating word embeddings involves using pre trained models over vast corpora like Fast Text, GloVE and Word2Vec (Wang, Nulty, and Lillis 2020). In the context of deep learning, embeddings are usually implemented using the embedding layers. These layers adeptly map input words to continuous vectors, offering flexibility tailored to specific tasks.

Recap From PBL

Dataset

All the information displayed in this section was the product of the work developed during the PBL phase and will serve as the fundamental building block for the upcoming deep learning models.

Data Used

The original dataset used in this study was shared by HGO and derived from the CTH - Alert P1 software designed for the management of information related to initial specialty consultation requests. CTH – Alert P1 role is requesting, screening, and forwarding information related to requests for first specialty consultations, from primary health centers to hospitals, allowing the establishment of priorities based on clinical criteria. From Figure 1, it can be visualized the typical working process of this system, showing how the pathway varies depending on whether a referral is accepted or rejected.

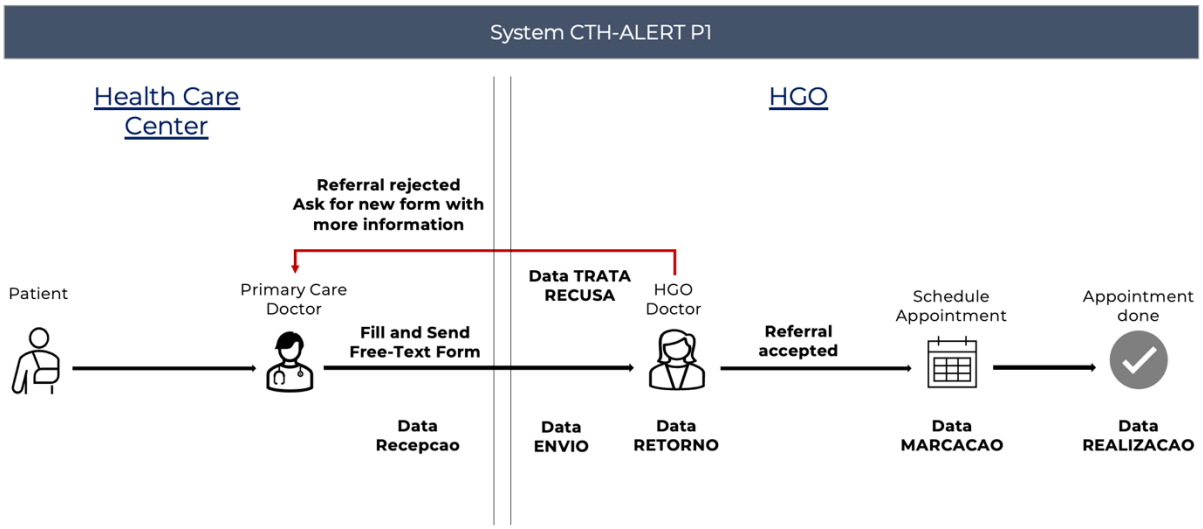


Figure 1: CTH – Alert P1 Process

The data used for the study represents 6 years, from 2012 to 2017 offering a historical view of the referral and consultation procedures recorded within the CTH -Alert P1 system. It consists of 10150 observations containing 31 variables with diverse data types such as text, integer and varchar. The dataset captures essential information such as provenance codes to track the provenance of patients, reference codes to identify and distinguish different referrals, and specialty codes and descriptions to indicate the specific medical specialty to which the patient is referred. Out of these records, we are going to focus only on the ones that contain the free text variable which represents 1771 observations.

Data Preprocessing

During the data preprocessing phase, several crucial steps were taken to enhance the quality and uniformity of the dataset. Firstly, the six dates variables displayed in Figure 1 were converted into datetime format. Additionally, missing values were identified on 'Provenance' and 'Priority' variables and were replaced by 'unknown' labels for easy interpretation. Furthermore, some entities were identified to share the same name in the dataset but with different codes, so these codes were updated, ensuring each entity to be represented by a unique code. Finally, a new column representing the length of the text variable, which will be useful for further analysis, was added.

Moreover, the target variable, which reflects the outcome of the referral triage process, did not exist in the original dataset. It was constructed based on a variable indicating the reason why a referral did not happen. In collaboration with HGO, we constructed a binary variable, our target variable, by going through each reason and assigning it a value of 1 if accepted or 0 if refused by the hospital. Cases, such as instances where the patient had an appointment registered with another reference number or where the patient died before the referral decision, were considered as 'unknown'. They totaled 162 referrals and were removed from the dataset. A comprehensive correspondence table detailing these classifications can be found in Appendix 7. In conclusion, after preprocessing data, we had 1609 observations containing free text and representing 1320 patients to be studied.

Main Conclusions from EDA

During the exploratory data analysis (EDA) phase, we observed that the distribution of appointment requests displayed interesting patterns. Approximately 68,3% of the cases were accepted while 31,7% were rejected, indicating a slight imbalance. We also noticed that more than 87% of the cases requested a general neurology appointment; in the remaining 13% of cases where a specific neurology subspeciality was requested, only 3 out of 198 referrals were

rejected. Moreover, over 18% of patients had multiple appointment requests, with the majority of these patients requesting two appointments. We observed that the majority (98%) of the appointment requests originated from Health Centers (Centro de Saúde) or Family Health Units (Unidades de Saúde Familiar) entities, making it beneficial to conduct a more granular analysis by dividing them into subcategories. Finally, we discovered that more than 59% of the cases have CTH as the originating system.

Concerning the text variable, we found that the length of the text provided by primary care physicians plays a role in the acceptance rate of referrals. The analysis revealed that referrals with longer text tend to have a higher acceptance rate. A plot of the acceptance rate by text length is shown in Appendix 8. Additionally, the analysis revealed that the length of the text and acceptance rate may vary depending on the health institution that the referral is coming from.

Another analysis was conducted on the frequency of words used in accepted and rejected cases (excluding stop words). It showed that certain words or phrases appear to be more prevalent in accepted referrals as compared to rejected referrals. These observations suggest that the presence of these specific words or phrases may be considered as important indicators for acceptance in the referrals process, perhaps because these words convey important clinical information. A word cloud with frequent words across accepted and rejected cases is shown in Appendix 9.

Variables Used for Modelling

During this process, we undertook a dual approach using both the free text variable and the more standardized data available. On one side, we used NLP techniques to extract valuable information from the text variable. Simultaneously, we created features based on the non-textual variables present in the dataset. These features represent the characteristics of a referral, and we will call them baseline features.

After modifications like the creation of new variables or the subdivision of existing ones, baseline features were divided into 4 groups, each representing a specific characteristic and containing a list of variables. All the categorical variables in this list were transformed into dummies. A table in Appendix 10 shows the baseline features used.

For NLP features, diverse techniques were employed including TF-IDF (Karabiber n.d.), LDA (Tharwat et al. 2017), Word2Vec (Meyer 2016), and BERT (Devlin et al. 2018).

Additionally, a Word Search approach was developed using a tailor-made dictionary of relevant medical terms and concepts, built with the help of the hospital's doctors. The concepts consisted of 3 groups containing specific words and their synonyms as shown in Appendix 11. For the terms, they were divided into 4 groups: Medication, Symptoms, Exams and Comorbidities. Additionally, within "Medications" and "Symptoms" categories, each word is associated with a relevance level in the Neurology-appointment's environment. For example, symptoms are divided into symptom level 1 and symptom level 2.

However, there are more different ways to write referrals than there are doctors. To deal with the different ways to write the words and account for possible typos and mistakes, referrals are processed using Levenshtein distance which calculates the similarity between referral's words and the dictionary's words. If the dictionary word distances from the referral word less than a pre-defined threshold value, it is considered a match.

After classifying the words, we create functions that generate new features for each referral based on each group of words and relevance. Those features consist of "word count" indicating the total count of words of a given group in a referral and "word concentration" which informs about the ratio of medical terms groups within all words in a referral.

Furthermore, we explored the Chi-Squared approach to identify relevant words which lead to acceptance. Chi-squared (χ^2) is a statistical test used to determine the association between two categorical variables. The test compared the association between the presence of a word in

a given referral and its outcome. After the analysis was performed, the most relevant words were selected (71 words) and a new feature was created: the ratio of the number of matches and the total number of words in a referral. Appendix 12 shows the list of these relevant words.

Methodology and Results

In this chapter, results and work from PBL will be firstly displayed. Then, deep learning models and their results will be explored.

Train/Test Split

Since the dataset contained only 1609 observations, which is a low amount of data, it was divided into train and test sets only, skipping a validation set. The first step involved sorting the data by chronological order based on the "DATA_RECEPCAO" column which indicates the date when HGO received a referral. The goal of using this chronological order is to mimic a real use case for the model, as it is predicting the outcome of new instances based on historic information, and therefore, providing a good foundation for the model evaluation. Subsequently, the dataset was divided into two subsets: train set and test set. The train set, representing the 80% oldest observations, was used to train the models. The remaining 20% of the data constituted the test set, which served as an independent dataset to assess the model's performance on new and unseen data.

Performance Measures Considered and Threshold Adaptations

During the model evaluation process, precision, accuracy, and recall metrics were examined as they offer valuable insights into the model's performance, highlighting areas where optimization and improvement are necessary. Figure 2 below shows the formulas for these three metrics. However, as agreed with HGO, recall was established as the evaluation metric for the performance of the model due to the assumption that in healthcare, it is often more important to correctly identify all the positive cases, even if it means having a higher false positive rate.

This is because missing a potentially critical referral can have serious consequences for patient outcomes.

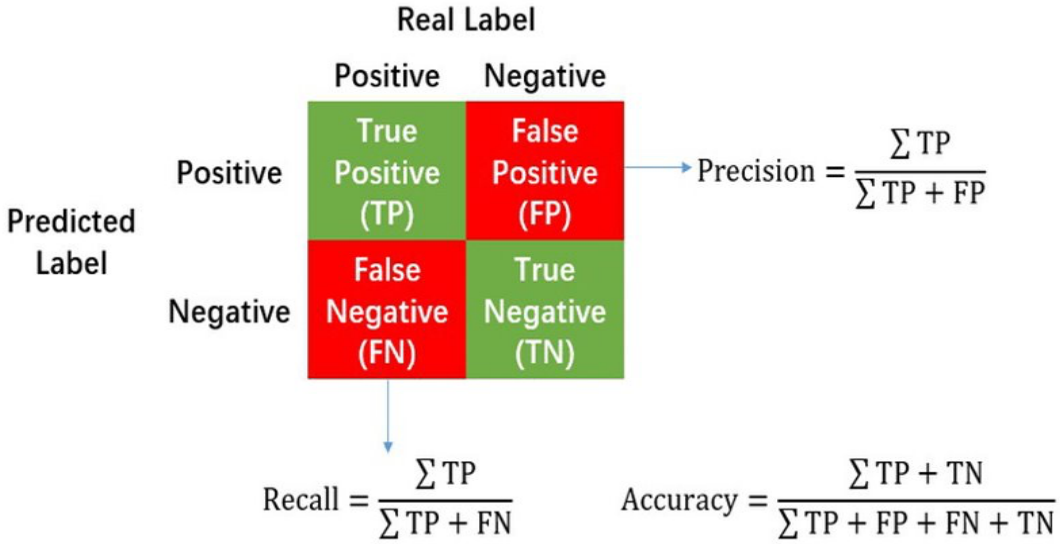


Figure 2: Metrics Formulas

Additionally, different threshold adaptations were employed in order to find the best modelling results. A threshold refers to the decision boundaries set by a model. The initial adaptation, which served as the baseline for comparison, was the default one with 0.5 threshold, where predicted probabilities above classify as one class, below as the other. Subsequently, precision-recall Trade-off, which involves finding the right balance between precision and recall was explored. The third strategy involved Balanced Accuracy which maximizes the arithmetic mean of sensitivity (recall) and specificity (true negative rate). Finally, we utilized Cohen's Kappa to measure the agreement between predicted and actual referral outcomes, taking into account the agreement that would be expected by chance.

Models Fine Tuning

Machine learning algorithms involve many hyperparameters that should be set before running the models. The choice of these hyperparameters can influence the results of the algorithms. To identify the most effective combination and get the best results, a Grid Search

method was employed. This approach involves exhaustive testing of different configurations, ultimately selecting the combination that yields the best metric on the training set(Belete and Huchaiah 2022). By leveraging Grid Search, our work ensures a meticulous and data-driven optimization of model's performance.

Baseline Model

In order to fairly evaluate the performance of the models, it is essential to have a baseline model that serves as a benchmark. For this purpose, a logistic regression model was established, to compare the performance of other models. The choice of this model was driven by its high interpretability, as it provides a straightforward understanding of variable importance in the decision-making process, making it an ideal benchmark for our study. The variables used as features for the logistic regression were the baseline features mentioned previously.

Baseline Model Results

The logistic regression model achieved a high recall of 97% indicating that it correctly identified the majority of actual referrals that require an appointment. However, it showed a limitation on classifying the cases that should be rejected as it achieved only 69% of precision. Additionally, the model showed that some features were more powerful than others. For a detailed list of coefficients of the features please refer to Appendix 12.

Even if the overall results were encouraging, it would be preferrable to achieve a more balanced model by increasing precision scores.

Best Modelling Results

Different models were explored during the modelling phase including Random Forest and XGBoost with the goal of obtaining a better model compared to the baseline one. After extensive experimentation and evaluation, including the examination of 39 different combinations of feature groups, the best-performing model was found to be a XGBoost model. This model was trained using baseline features, which includes structured data, in combination

with NLP features derived from Word Search approach and Chi-Squared Scores techniques mentioned in the previous section.

In terms of performance, after fine-tuning the model, the developed algorithm achieved a good performance in predicting whether a referral should be accepted or not using Cohen's Kappa as the threshold adaptation method. A table with the optimal hyperparameters for this model is presented in Appendix 13. As shown in Figure 3, the confusion matrix demonstrates how the model effectively identified most referrals that should be accepted. It registered a recall rate of 95% on the test set. It also achieved a precision score of 72% indicating that the model performed better in predicting the referrals that should be refused compared to the logistic regression model.

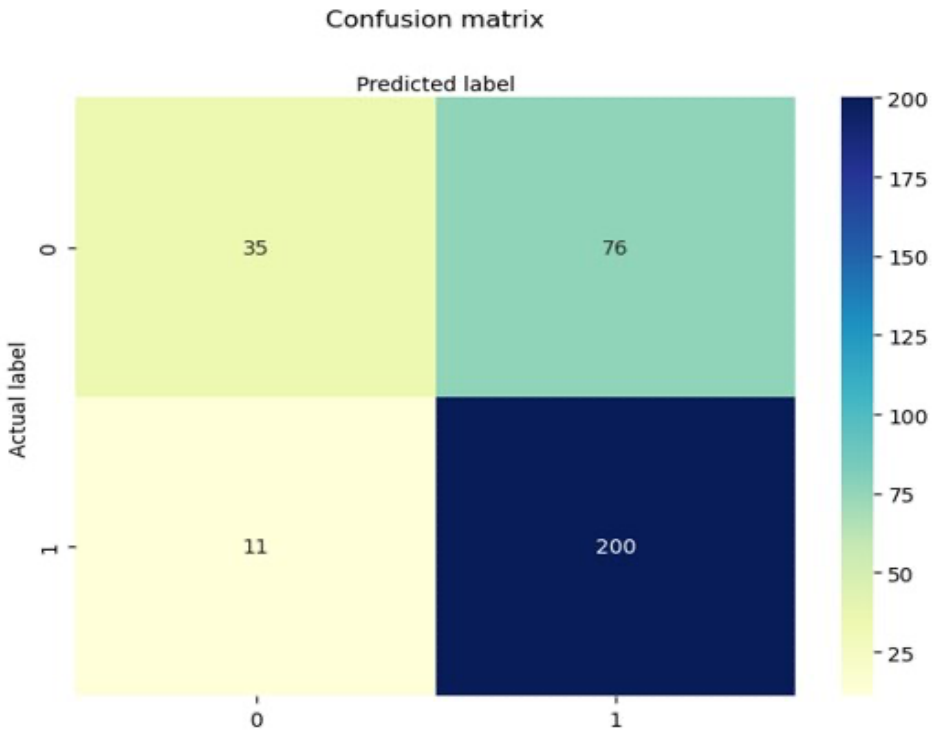


Figure 3: XGBoost Confusion Matrix

Despite the recall lower than the baseline model, we considered the XGBoost model to have better results, as it achieved a high recall and was also able to get a higher precision. This difference of results between the XGBoost and the baseline reflects the importance of including

features derived from the NLP techniques.

Current Thesis Work

Deep Learning Models

Features Used for Modelling

Deep learning models were trained using the same set of features used in the XGBoost model. Additionally, and recognizing the power of deep learning models to detect hidden patterns within textual data, word embeddings of the text variable were included to capture the semantic richness within this type of data. The text variable used for generating embeddings consists of the original text which was preprocessed by removing stop words and adding the wordsearch approach words. This update on the free text variable empowered the models to extract valuable information and achieve better results.

Models

All the parameters mentioned in this section and their significance can be found in table 1.

Three deep learning models were meticulously explored including CNN, RNN and LSTM. To exploit the patterns within the textual data, the CNN architecture included an embedding layer followed by a one-dimensional convolution layer with 128 filters and a kernel size of 3, employing the rectified linear unit (ReLU) activation function (Krishnamurthy 2022). Max pooling with a pool size of 5 was then employed to capture essential features and flattening was done to facilitate the integration of the other features. For RNN and LSTM models, an embedding layer was first introduced then an RNN and LSTM layers respectively with 64 filters and ReLU activation function were employed.

Simultaneously, all the other features created during the PBL project were processed through densely connected layers with 128 filters and 64 filters and the ReLU activation function. For all these models, the outputs of the textual data were then concatenated with the output of the non-textual features and processed using connected layers comprising 128 and 64 filters,

followed by an output layer with 1 neuron and the sigmoid activation function (Topper 2023), enabling binary prediction for referrals outcome. The final models were optimized using the binary_crossentropy loss function and the ADAM optimizer (Agarwal 2023) with a learning rate of 0.001. The models were finally trained using 10 epochs and 32 batch size.

Parameter	Significance
filters	Number of convolution filters in the CNN layer
Kernel size	Scope of information captured by each filter during convolution
Pool size	Summarizes the features present in a region
Epochs	Determines how many time the model is exposed to the entire train set
Batch size	Number of samples per gradient update

Table 1: Parameters and Significance

Models Results

Before delving into the models' results, it's crucial to note that each model had different tokenization and embedding values to achieve its best results. In fact, after trying different combinations, the optimal values that registered the best metric results were determined as shown in table 2.

Model	**Tokenization value**	**Embedding value**
CNN	12	100
RNN	13	90
LSTM	32	100

Table 2: Tokenization and Embedding Values

In terms of performance, and as shown in Table 3 below, LSTM model achieved the best results in the test set with a recall rate of 95% and a precision of 74% using Cohen's Kappa as threshold adaptation. For CNN and RNN, the models achieved robust results using the default threshold of 0.5 with recall scores of 97% and 95%, and a precision score of 71% and 73%, respectively.

Model	Recall	Precision	Threshold Adaptation	Threshold Value
CNN	0.97	0.71	Default	0.5
RNN	0.95	0.73	Default	0.5
LSTM	0.95	0.74	Cohen's Kappa	0.59
XGBoost (PBL)	0.95	0.72	Cohen's Kappa	0.6
Baseline (PBL)	0.97	0.69	Default	0.5

Table 3: Models Results

Discussion and Conclusions

This study investigated the role of deep learning and NLP techniques consisting of embeddings in improving patient referral classification. By looking at the results obtained, it is evident that LSTM model outperformed traditional machine learning. Despite the constraint of having a small dataset comprising only 1609 observations, the model was able to improve the precision score by 2% on the test set. Even though the increase was not very large, it confirms the power of this model when dealing with textual data. Since deep learning models require significant data volumes, additional data could potentially result in a larger increment in performance, compared with the models that did not use deep learning.

Achieving a recall rate of 95% and a precision of 74% makes the model a promising tool for doctors. Its use in conjunction with their clinical judgment offers an opportunity to enhance the decision-making process in patient referrals. Additionally, from a resource management perspective and in a field where medical staff and facilities are at a premium, this increase in precision ensures the reduction of unnecessary diagnostic procedures on cases likely to not necessitate intervention. It also aligns with cost-effectiveness goals, offering potential cost savings by avoiding false necessary diagnosis.

Limitations

This study faced different limitations. In fact, machine learning models require a remarkable amount of data to achieve robust model training and perform well. However, in our case, we had only 1609 observations making it hard for the models to capture relevant patterns in this referral classification task.

Additionally, due to computational limitations and to the memory required by deep learning models, it was not possible to better explore the hyperparameters. The nature of these models' architecture would necessitate more computational resources and time for an exhaustive search through different hyperparameters combinations.

Future Work

For future work, it would be beneficial to run a pilot study incorporating a control and experimental group to compare the impact of the model on referral outcomes and decision-making. The control group would serve as a reference, experiencing the current referral screening process without the intervention of the model. On the other hand, the experimental group would experience the intervention of the model during the same process. This approach will allow for a more rigorous evaluation and assessment of the model's effectiveness in improving this referral screening process.

Additionally, an interface for deploying the model that provides suggestions to doctors

should be implemented to facilitate efficient decision making. A prototype of the interface, developed during PBL is shown in Figure 4.

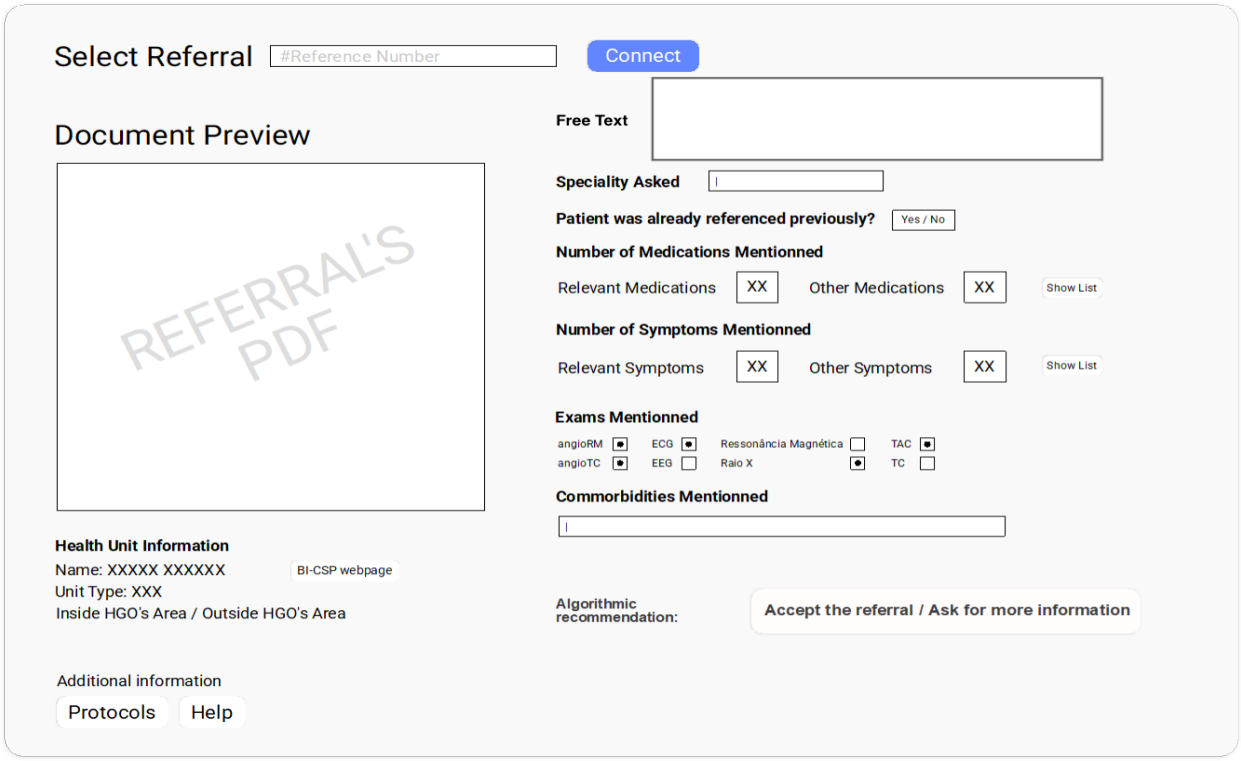


Figure 4: Interface Design

References

Agarwal, Rahul. 2023. “Complete Guide to the Adam Optimization Algorithm | Built In.” September 13, 2023. https://builtin.com/machine-learning/adam-optimization.

Aggarwal, Tavish. 2020. “Neural Networks and Deep Learning - Blog - ACS Solutions.” 2020. https://acsicorp.com/blogs/fundamentals-artificial-neural-networks-are-to-deep-learning-what-atoms-are-to-matter/.

Beam, Andrew. 2017. “Deep Learning 101 - Part 1: History and Background.” 2017. http://beamlab.org/deeplearning/2017/02/23/deep_learning_101_part1.html.

Belete, Daniel Mesafint, and Manjaiah D. Huchaiah. 2022. “Grid Search in Hyperparameter Optimization of Machine Learning Models for Prediction of HIV/AIDS Test Results.” *International Journal of Computers and Applications* 44 (9): 875–86. https://doi.org/10.1080/1206212X.2021.1974663.

Dettmers, Tim. 2015. “Understanding Convolution in Deep Learning.” 2015. https://timdettmers.com/2015/03/26/convolution-deep-learning/.

Devlin, Jacob, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (October): 4171–86. https://arxiv.org/abs/1810.04805v2.

Dukes, Kais. 2023. “The History of Neural Networks Part 2: Rosenblatt’s Perceptron.” 2023. https://www.linkedin.com/pulse/history-neural-networks-part-2-rosenblatts-perceptron-dr-kais-dukes/.

Hamed, Ghada, Mohammed Abd El Rahman Marey, Safaa Amin, Mohamed Tolba, Mohammed Abd El-Rahman Marey, Safaa El-Sayed Amin, and Mohamed Fahmy Tolba. 2020. “Deep Learning in Breast Cancer Detection and Classification,” March. https://doi.org/10.1007/978-3-030-44289-7_30.

Hughes, Mark, Irene Li, Spyros Kotoulas, and Toyotaro Suzumura. n.d. “Medical Text Classification Using Convolutional Neural Networks.” Accessed October 13, 2023. http://www.merckmanuals.com/.

“Institution Category - Hospital Garcia de Orta.” n.d. Accessed October 26, 2023. https://www.hgo.min-saude.pt/category/institucional/instituicao/.

Kalita, Debasish. 2023. “A Brief Overview of Recurrent Neural Networks (RNN) - Analytics Vidhya.” 2023. https://www.analyticsvidhya.com/blog/2022/03/a-brief-overview-of-recurrent-neural-networks-rnn/.

Karabiber, Fatih. n.d. “TF-IDF — Term Frequency-Inverse Document Frequency – LearnDataSci.” Accessed November 27, 2023. https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/.

Krishnamurthy, Bharath. 2022. “ReLU Activation Function Explained | Built In.” October 28, 2022. https://builtin.com/machine-learning/relu-activation-function.

Li, Zewen, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2022. “A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects.” *IEEE Transactions on Neural Networks and Learning Systems*. Vol. 33. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/TNNLS.2021.3084827.

Lipton, Zachary C, John Berkowitz, and Charles Elkan. 2015. “A Critical Review of Recurrent Neural Networks for Sequence Learning.”

Loye, Gabriel. 2019. “Beginner’s Guide on Recurrent Neural Networks with PyTorch.” 2019. https://blog.floydhub.com/a-beginners-guide-on-recurrent-neural-networks-with-pytorch/.

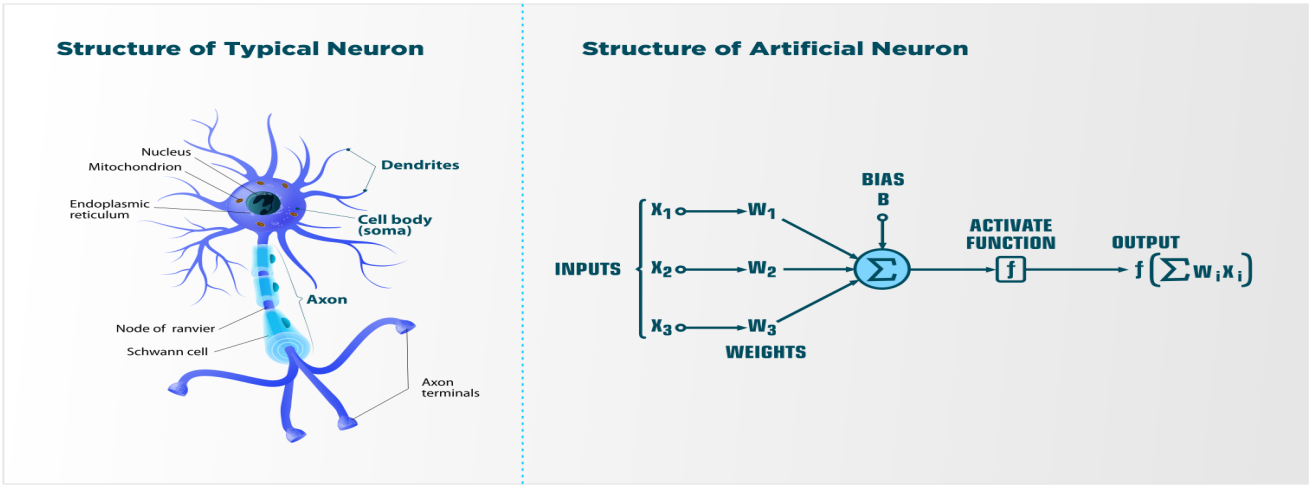
Lu, Wenjing, Wei Jiang, Na Zhang, and Feng Xue. 2021. “A Deep Learning-Based Text Classification of Adverse Nursing Events.” https://doi.org/10.1155/2021/9800114.

Lv, Wen Hao, and Ju Yang Lei. 2020. “Deep Learning Development Review.” *Proceedings - 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software*

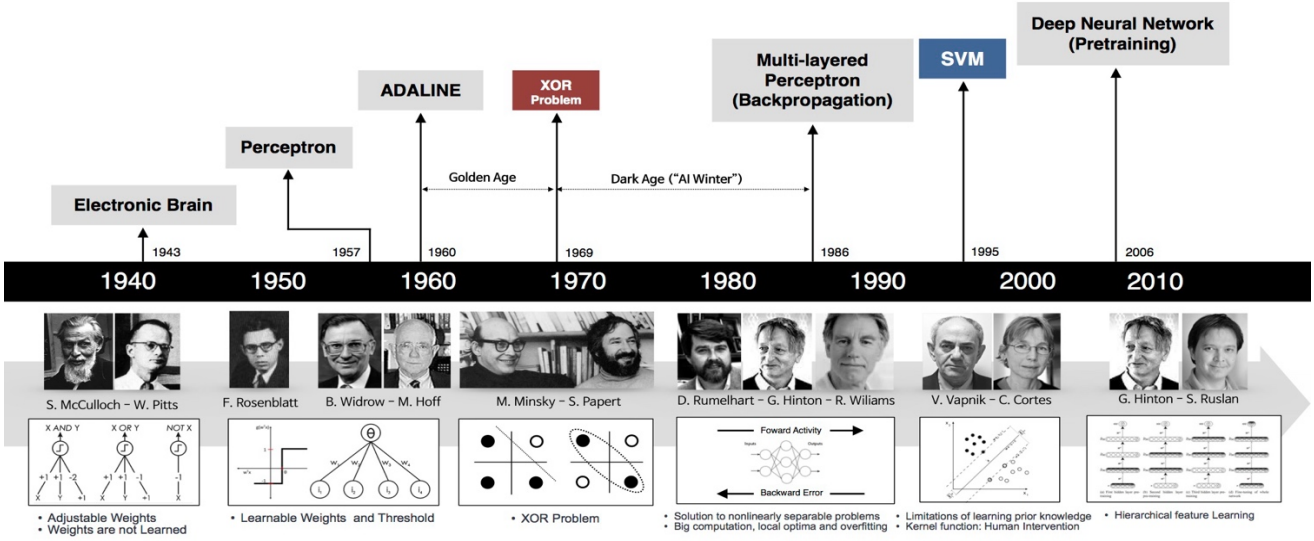
Engineering, AEMCSE 2020, April, 171–74. https://doi.org/10.1109/AEMCSE50948.2020.00043.
Madan, Piyush, and Samaya Madhavan. 2020. “An Introduction to Deep Learning - IBM Developer.” March 3, 2020. https://developer.ibm.com/articles/an-introduction-to-deep-learning/.
Meyer, David. 2016. “How Exactly Does Word2vec Work?”
Poongodi, T., D. Sumathi, P. Suresh, and Balamurugan Balusamy. 2021. “Deep Learning Techniques for Electronic Health Record (EHR) Analysis.” *Studies in Computational Intelligence* 903: 73–103. https://doi.org/10.1007/978-981-15-5495-7_5/COVER.
Prathap, Prajeesh. 2020. “Feed-Forward and Recurrent Neural Networks: The Future of Machine Learning | by Prajeesh Prathap | Medium.” 2020. https://medium.com/@prajeeshprathap/feed-forward-and-recurrent-neural-networks-the-future-of-machine-learning-8b2c1975f0c5.
Qiwei, Han. 2023. “2487-S2 Machine Learning(Data Science for Business 201) Week 11: Convolutional Neural Networks.”
Rajkomar, Alvin, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, et al. 2018. “Scalable and Accurate Deep Learning with Electronic Health Records.” *Npj Digital Medicine 2018 1:1* 1 (1): 1–10. https://doi.org/10.1038/s41746-018-0029-1.
Razzak, Muhammad Imran, Saeeda Naz, and Ahmad Zaib. n.d. “Deep Learning for Medical Image Processing: Overview, Challenges and Future.” Accessed October 13, 2023.
Saxena, Shipra. 2023. “What Is LSTM? Introduction to Long Short-Term Memory.” 2023. https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/#.
Shen, Dinggang, Guorong Wu, and Heung-Il Suk. 2017. “Deep Learning in Medical Image Analysis.” https://doi.org/10.1146/annurev-bioeng-071516.
Tharwat, Alaa, Tarek Gaber, Abdelhameed Ibrahim, and Aboul Ella Hassanien. 2017. “Linear Discriminant Analysis: A Detailed Tutorial.” http://www.egyptscience.net.
Topper, Noah. 2023. “Sigmoid Activation Function: An Introduction | Built In.” July 10, 2023. https://builtin.com/machine-learning/sigmoid-activation-function.
Wang, Congcong, Paul Nulty, and David Lillis. 2020. “A Comparative Study on Word Embeddings in Deep Learning for Text Classification.” https://doi.org/10.1145/3443279.3443304.
“What Are Recurrent Neural Networks? | IBM.” n.d. Accessed October 26, 2023. https://www.ibm.com/topics/recurrent-neural-networks.
Yamashita, Rikiya, Mizuho Nishio, Richard Kinh, Gian Do, and Kaori Togashi. 2018. “Convolutional Neural Networks: An Overview and Application in Radiology.” https://doi.org/10.1007/s13244-018-0639-9.

Appendices

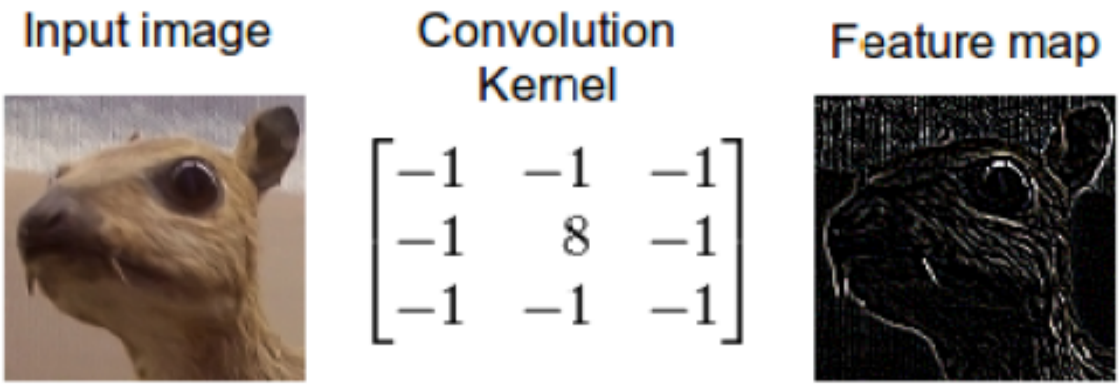
Appendix 1: Biological Neuron VS Artificial Neuron (Aggarwal 2020)



Appendix 2: History of deep learning (Beam 2017)

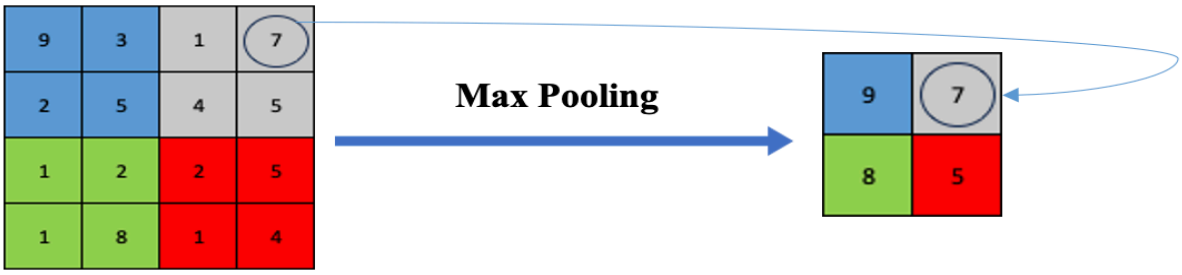


Appendix 3: edge detection with a convolution kernel (Dettmers 2015)

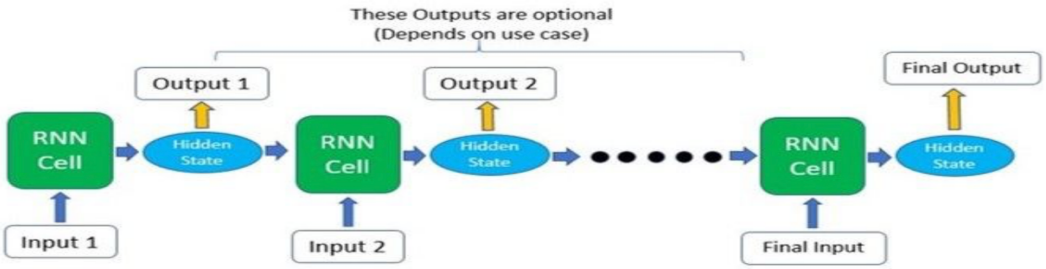


$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

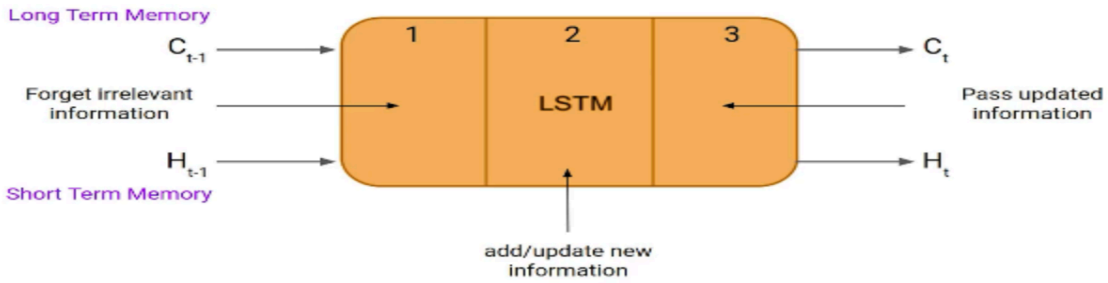
Appendix 4: Max Pooling



Appendix 5: Recurrent Neural Network (Loye 2019)



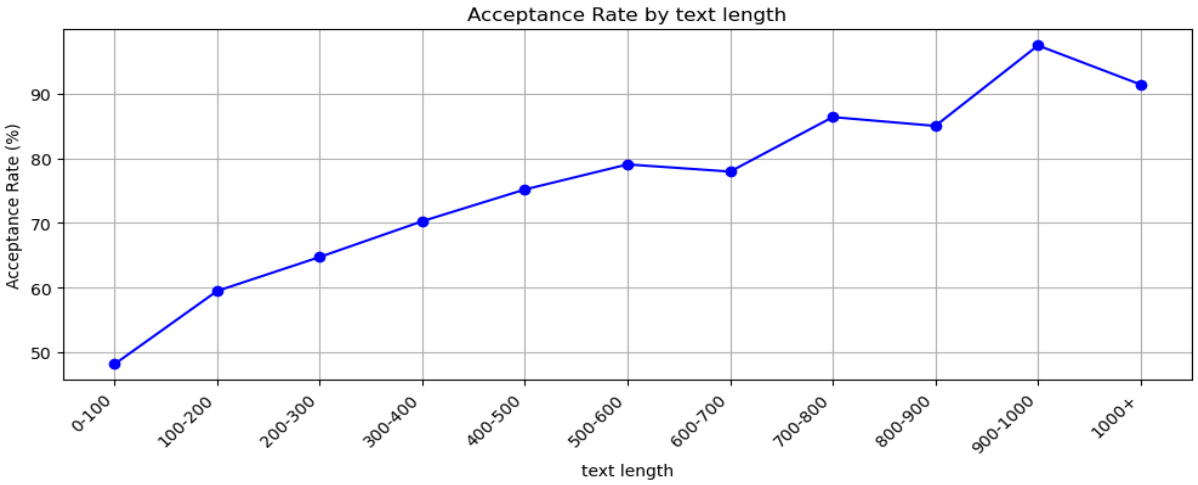
Appendix 6: LSTM (Saxena 2023)



Appendix 7: Target variable binary correspondence

COD_MOTIVO_RECUSA	Code Description	Triage Outcome	Explanation
1	Sem relatorio clinico	Negative	Referral was rejected, lacked a clinical report
2	Utente fora da area	Negative	Referral was rejected, the patient was outside hospital area. [This restriction is no longer applicable for current referrals]
3	Especialidade nao existente	Negative	Referral was rejected, the requested speciality doesn't exist
6	Sem Exames Complementares	Negative	Referral was rejected, lacked complementary exams
7	Sem marcacao	Negative	Referral was rejected
10	Enviado ao Medico Assistente	Negative	Referral was rejected and sent back to the referral doctor
51	Transferido Para Outra Instituic?o	Negative	Referral was rejected and the patient transfered to another institution
54	Não cumpre os critérios clínicos para a consulta indicada	Negative	Referral was rejected, lacked the clinical requirements for the requested speciality
0	Null	Positive	Referral was accepted
14	Consulta Marcada. O Doente Faltou	Positive	Referral was accepted but the patient missed the consultation
15	Consulta Marcada. O Doente Desistiu	Positive	Referral was accepted but the patient gave up on the consultation
53	Falta Injustificada do Utente	Positive	Referral was accepted but the patient missed the consultation without any notice
12	Marcada em Subsequentes	Positive	Referral was accepted. The Patient already had a neurologist and previous appointments in the HGO/Neurology
13	Marcada sem Referencia	Positive	Referral was accepted but the consultation was scheduled without a reference code
8	Enviado para outra Especialidade	Positive	Referral was accepted but the patient was sent to another speciality
16	O Doente Desistiu	Unknown	Patient gave up before triage
18	O Doente ja teve Consulta com outra referencia	Unknown	Patient already had appointment registered with another reference number
19	Referencia Duplicada	Unknown	Process reference is duplicated
50	Em processamento no SIGLIC	Unknown	Patient referral is processed in different platform. Doesn't apply to Neurology consultations
52	Cancelado pelo SIGLIC - Utente	Unknown	Patient referral is processed in different platform. Doesn't apply to Neurology consultations
20	O Doente Faleceu	Unknown	Referral was accepted but the patient died

Appendix 8: Acceptance rate by text length



Appendix 9: Word cloud with frequency of words

Appendix 10: Baseline features

Group	Variable	Dummies for :					
Referral Origin	Originating system	CTH	SAM	SON	UNKNOWN		
	Type of entity	USF A	USF B	UCSP	OUTRO	HGO	HOSP
	Entity is inside vs. outside HGO's referring catch-area	INSIDE	OUTSIDE				
Referral History	Nth time that patient is referred	1	2	3+			
	Patient was accepted before or not	Accepted before	Not accepted before				
Referral Specificities	Specialty being asked	General Neurology	Other Specialities				
Text	Text length						

Appendix 11: Key concepts in referral's text

Alteração Alterações Mudança Mudanças	Agravamento Agravamentos Piora Pioras	Estável Estáveis Inalterado Inalterados	**Change**
Modificação Modificações Variação Variações	Deterioração Deteriorações Declínio Declínios	Inalterável Inalteráveis Controlado Controlado	**Aggravation**
Transformação Transformações alt.	Degeneração Degenerações Decadência Decadências	Controlados Regulado Regulados	**Stability**

Appendix 12: Coefficients from baseline model

Feature	coefficients
Other specialities	3.291589
USF B	0.75999
outro	0.628784
outside area	0.201002
USF A	0.19818
UCSP	0.097788
text_length	0.001506
2	-0.029162
3+	-0.058123
SON	-0.132115
SAM	-0.141081
HOSP	-0.321639
not accepted before	-0.590184
unknown	-1.070032

Appendix 13: Best model hyperparameters

Hyperparmeter	Value
Max_depth	2
eta	0.4
objective	binary:logistic'
seed	16
num_round	6
threshold	0.6