Diagnosis Detection from Medical Transcripts Using Machine Learning Algorithms

Khushal Paresh Thaker
1106937
MSc. in Comp. Sci.
Lakehead University, Thunder Bay
kparesh@lakeheadu.ca

***Abstract*—The ultimate purpose of the study of Artificial Intelligence (AI) is to allow the machines to comprehend and interact at human level. Technologies such as Machine Learning (ML), Natural Language Processing (NLP) which are a subset of AI, can be used in many fields. One major industry is the medical field. When a patient visits the doctor to be diagnosed, the information that is discussed between them is noted in a raw file called a medical transcript. This project aims to transform this unstructured raw file into a structured information by using NLP to perform necessary preprocessing steps and build supervised machine learning models that can automate the process of detecting the diagnosis. The dataset is procured from the Medical Transcriptions (MT) Samples. Two supervised ML models, Logistic Regression and Convolution Neural Network (CNN) are implemented, scoring an accuracy of 0.54 and 0.85, F1 score of 0.59 and 0.88 respectively. While implementing the baseline logistic regression model, a minimum accuracy of 0.38 is achieved and so the model is run with under and over sampling techniques and compared as well.**

Keywords–NLP, DL, CNN, MTSamples, Undersampling, Oversampling, Word Vector

I. Introduction

While extending the use of AI in the field of Medical domain can be helpful, it must be well within the boundaries and should be efficient as it deals with human lives. There are many patients who visit the doctor every day to find a diagnosis. During these visits, the information gathered by the staff and the doctors are noted down in a raw file which is referred to as a Medical Transcript. With an expansive rise in such medical records, there is a growing need for automation. To comprehend these vast documents manually could be tiresome and might reduce the efficiency of a physician. Using NLP and DL techniques to transform this document and then provide a means to predict the diagnosis, could be of assistance to the physician by reducing their efforts and improving the efficiency.

The MT Sample dataset [18] is procured from kaggle, is a csv file that contains (i) Description, (ii) Medical Speciality, (iii) Sample Name, (iv) Transcription and (v) keywords. This file contains a lot of information and therefore becomes necessary to segregate the required data from this corpus. The difficulty in transforming the necessary information from this document lies in its unstructured format, grammatical issues and abbreviations that are not always comprehensible. Also, the data type of the context raises issues, with some of them being (i)Date Time object types, (ii) Numeric, (iii) Categorical Data (iv) free form of text that include discharge summary [1]. Another form of challenge is understanding the context and extracting the meaning.

Rule based, ML and Hybrid are the technological ways to identify the diagnosis automatically. Rule based include matching the string and patterns set by the developer and the ML approach is to look at the issue as a classification problem. Hybrid is a combination of both, where rule based decisions help in creating a dictionary from which ML models help predicting.

To predict and evaluate the diagnosis automatically, there has to be an understanding of the previous diagnosis made. Also, necessary data should be extracted from the medical transcripts that can be used in detection. Figure 1 shows how information is retrieved and processed to make a prediction.

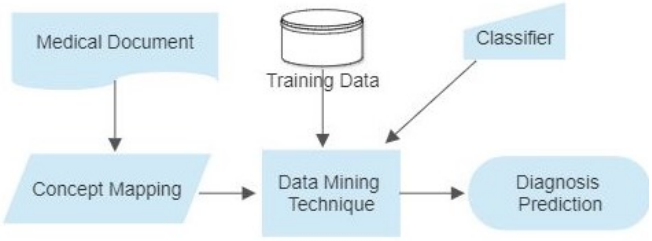

Fig. 1: Methodology for Diagnosis Prediction.

This project uses natural language processing to process the dataset and different Machine Learning (ML) techniques in order to detect a diagnosis using the MT sample dataset. As detection is a classification problem, the models are accordingly chosen. Logistic Regression model is a supervised machine learning algorithm used to classify discrete number of possibilities. This is a generalized linear model and a special case of a neural network with an exception of hidden layers. In order to under sample the data, a technique called Near miss is used, where distance of majority class to minority class is taken into consideration to select the examples to

undersample. Naive Undersampling is a random sampling method where no assumptions about the data is made, and allows for a quick implementation on large dataset. The undersampling method is compared with oversampling method using (Synthetic Minority Oversampling Technique) Smote. This method duplicates the data provided in the minority class before the model is fitted. Although, it balances the dataset, but doesnt provide any new information to the model.

Word embeddings are the means to use CNN for NLP. NLP is a technique to process the data so that relevant textual information can be retrieved. This project will make use of few of the NLP techniques such as lemmatization, normalisation that involves Stop word removal, shifting the letters to lower case, and also unnecessary characters are removed.

Although, CNNs were initially created to be implemented over images, it has been proven by a research conducted by Kim [15] in 2014 that CNNs are efficient with text as well. Sliding window, a concept of convolution, is passed over an image using CNNs to extract features one at a time, for multiple periods. In the case of text, CNN has an input which needs to be split into words or word embeddings, mostly implemented using word2vec or Global Vectors for word representation (GloVE). These words which are broken into features, are sent to the convolution layer, result of which are aggregated to form a representable number. This is later fed to connected layers according to the needs of the problem. Similar to the images, when CNN is used, matrix is the representation format for text where word embedding is represented by each row. The convolution layers then scans the text, breaking it into features.

Comparison is made between Logistic Regression method and CNN. While using Logistic regression, the dataset is transformed using word2vec approach. As the dataset is unbalanced, techniques such as under-sampling and over-sampling were applied to improve the performance. While implementing the CNN model, the keywords are tokenized and the dataset is split to training, validation and testing. GloVE 6B 100D is the word vector used as reference to input for the embedding layer of the model.

The remainder of paper is organized into 4 sections. Section II introduces various research carried out on diagnosis detection, different machine learning algorithms used. Section III involves the methodology of the models for both Logistic Regression and CNN, highlighting the use of pre preprocessing steps, and word vectors. Experimental results and Comparison is is shown in section IV while, Section V concludes the work and provides the possible future enhancements.

II. Literature Review

To solve the problem of diagnosis detection, numerous supervised and unsupervised algorithms can be used by identifying the medical relationship within the document.

Bryan et al., [2] proposed a supervised approach that made use of Support Vector Machine called LibLINEAR as the classifier. Cross Validation was performed to tune the parameters and the F1 score achieved was 48.4. The dataset used was from the i2b2 challenge.

In a predictive model where matrices are sparse and contain large, unbalanced features, specialised learning techniques are required to provide detections. Vidrighin et al., [3] developed a ProICET methodology to tackle this. The model performs heurestic search and manages to perform well with changes made to the existing model. Lee et al., [4] develops a Recurrent Neural Network (RNN) with Medical context attention that makes use of conditional variational autoencoders to derive individual patient information. Sequential patterns of disease were summarized by the help of a GRU. The model is run on National Patient Sample that allows them to produce more sparse weights and increases interpret-ability and performance. The highest recall score of 0.63 was from the 2011 dataset.

Jason et al., [5] proposed a prediction based supervised LDA. The idea was to improve the topic coherence by using vocabulary selection model while not affecting the predictive efficiency. The model managed to learn most coherent topics and Area Under the Curve (AUC) score of 0.748, higher, when compared to other models. The idea to extract word level matrix from each sentence is proposed by Mark et al., [6] where by word2vec is used and after training, the model performs better in accuracy than mean word embedding with Bag-of-Words (BOW).

Qu et al., [7] proposed a prediction at character level by implementing a CNN and bi-RNN model. It negates the problems of unregistered words which arise at the word level implementation and also improves the robustness of the model. Bi-RNN is used to classify the text by getting the contextual information. The model achieves an accuracy of 90.41 with Sougon News dataset.

A method proposed by Yao et al., [8] involved combining the rule-based features and CNN to predict the diagnosis. The method used rules to identify the trigger phrases initially and then predict the classes. These trigger phrases are then given as input to the CNN to train along with the word embeddings and unified medical language. This model secured 96.12 as the Micro F1 score when implemented on i2b2 dataset.

Multiple frameworks have been developed in order to make use of the clinical notes. Fodeh et al., [9] proposed a MedCat framework. Although, the framework was applied to clinical notes for Post Traumatic Stress Disorder (PTSD), the approach was a lot more useful. It included annotations done manually, metamap to expand knowledge base and NLP to generate bag of concepts. This research introduced extracting features from the samples and used specialized concept-category hierarchy to transform it into less granular features. The sensitivity was greater than 90% which was greater than other systems at the time.

Barbantan et al., [10] made use of Electronic Health Record (EHR) samples from the MTSamples dataset and proposed sequence of models to analyse any medical document by making use of NLP and semantic analysis to suggest a diagnosis for the patient and achieved an accuracy rate of 81.25%. Rodica et al., [14] proposed a similar technique whereby

NLP is used to preprocess unstructured medical transcript documents and then concept identifiers were implemented by using MedCIM mapping strategy to firstly, identify the medical related candidates and the other is to map the candidates to medical terms.

MedEx was a system proposed by Xu et al., [11] that relied on coded data. The extracted information from the medical notes were converted into codes and run against discharge summaries proving effective not only in predicting the drug names, but also with signature details such as strength, route and frequency outscoring other models with F1 score of 95.5%, 93.9% and 96% respectively.

Detection of diagnosis from clinical summaries can be extraction and abstraction based. Gehrmann et al., [12] proposed a deep learning model that outperformed extraction based model. The model proposed was CNN with phenotyping tasks using 1610 MIMIC-III dataset. The model outscored the concept based model by 26% when measured in F1 score.

A method using graph convolution and RNN was implemented Li et al., [16] that used the i2b2 dataset containing clinical information and concept relation to achieve an F1 score of 0.827. The classifiers developed were using knowledge based annotations using NLP pipelines that involved MetaMap and cTakes.

Tomas et al., [17] made use of multiple machine learning algorithms such as Naive Bayes, Random Forest, Decision Tree, Support Vector Mechanism and Logistic Regression classifiers to classify text reviews. The research compared the techniques on amazon customerś product review data and achieved a maximum accuracy of 58.50% using Logistic Regression classifier, outscoring others by a very good margin.

III. Methodology

Description of the models implemented will be provided in this section. Two models are used and compared in the project. A Logistic Regression Classifier which is built using word2vec approach. To improve the accuracy of the model, varios sampling techniques are implemented and compared with. The CNN model implemented makes use of the GloVE word vector file.

A. Dataset

The MTSamples dataset used is procured from Kaggle and consists of 5000 records of patients varying over 40 different medical specialities. Medical Speciality is the transcriptions's classification, also the model's target variable. Sample name is the title of the transcription, and the Transcription are the indepth content of the information. Keywords are the necessary content relating to the detection task. The dataset is free to use and needs no licensing. After segregating the sentences, Figure 2 shows the number of sentences with the its word length as the size.

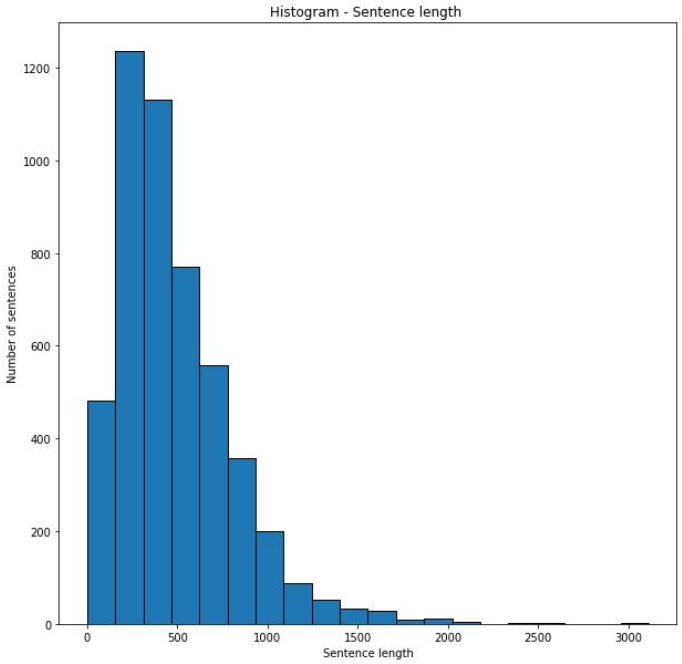

Fig. 2: Histogram - Sentence Length.

B. Logistic Regression

1) Preprocessing: As the dataset is unbalanced, necessary NLP tasks are implemented to preprocess. The dataset is tokenized using 'RegexTokenizer' which helps in splitting it using regular expressions. The target variables are encoded using 'LabelEncoder'. The dataset is normalised by removing the unnecessary symbols and stop words. Stop words make the amount of words redundant and can reduce the accuracy of the model. Normalisation also involves the basic step of shifting the letters to lower case so that there are no different vectors created because of different casing. Lemmatization, a process of analysing the words morphologically and reducing it to their root form, is performed using Spacy.

2) Model Working: The dataset contains many categories, out of which 'transcription' and 'medical speciality' is required. Word2Vec is applied to the dataset to create embeddings. Google's pretrained word vector model is used. It contains vectors for about 3 million words, which is trained over a 100 billion words from its own news dataset.

The baseline model is created with a 70:30 train:test split. Once this is done, the model seems to have a very low F1 score which is because of huge unbalance. A stratisfied k fold corss validator (k = 3) is applied while running the model so that splits are made and folds created, dont alter the sample percentage in each class. The logistic regression classifier model is created with multi_class parameter set to 'multinomial' so that cross entropy loss is attained. Model solver is chosen as 'saga' because it optimal for big datasets and more over it supports elasticnet penalty and handles L1.

The dataset is adjusted by dropping the objects with less than 100 observations. The model seems to still have very

low F1 score. Imbalance still prevails and hence, techniques such as Undersampling and Oversampling are used. In order to perform Under sampling, few observations are cut off from the major class so as to even it out with the minor class. Near miss and Naive approach are used to perform Under sampling. With less data, it shows that it isn't effective. Over sampling is the technique by which the observations in the minor classes are increased to even it out with the major class. SMOTE is used for oversampling. This seemed to provide the best result of all the techniques.

C. CNN Model

The CNN model is implemented to predict the diagnosis from the MTSample dataset. The approach makes use of GloVE "glove.6b.100d" pretrained weights to get the word vectors of the key words which are provided as the inputs. Flow of the proposed CNN model is shown in figure 3.

1) Preprocessing: Keywords from the dataset are the input sequence shown in the figure 3. Preprocessing takes place, converting the alphabets to lower case and removing unwanted stop words and punctuations. The stop word corpus is procured from the NLTK library. Characters are transformed to lower case so that the integer value isn't different for the same word if in capitals. The dataset is tokenised and split into 70:30 train ratio. The pretrained GloVE file is loaded to the embedding layer. Output of which is fed to the 1D convolution layer.

2) GloVE Word Vector: Global Vector (GloVE) for word representation is an unsupervised machine learning algorithm that makes use of global co-occurrence statistics from a word corpus. This was developed by Jeffrey Pennington, Richard Socher and Christopher D. Manning in 2014 [13]. Available in different dimensions: 50D, 100D, 200D, and 300D these files can be used in any model to make predictions.

In order to make the probability of word's co-occurrence similar to the order of dots, the necessary GloVE file is used. This ratio is then used to provide the data encoded as difference of vectors and that is why final word vectors are suitable to extract word analogies.

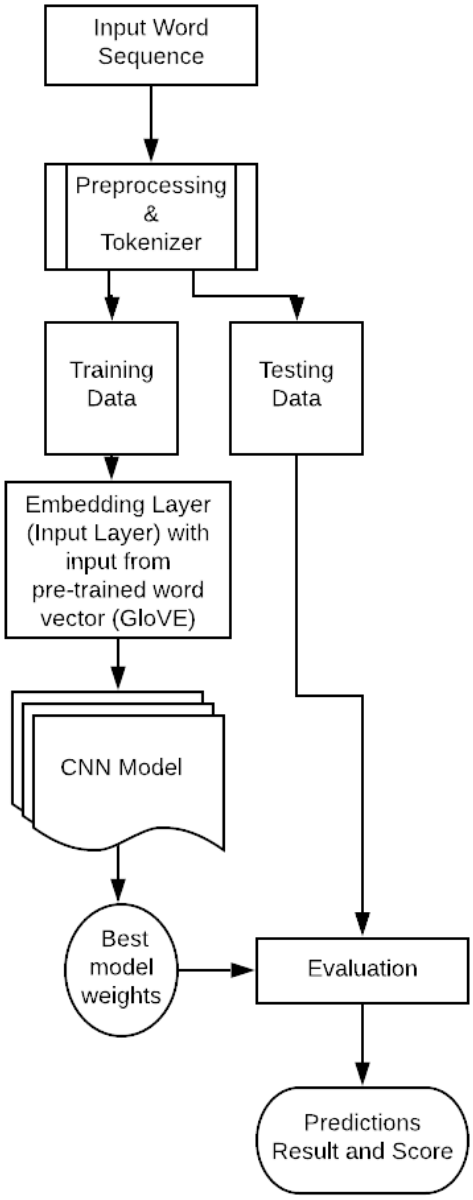

Fig. 3: CNN Model Flowchart.

3) Model Working: The text is converted into sequence of indexes (integer ID for the word). The embedding vectors are presented in the embedding matrix. The embeddings are mapped to the index mapping words once the data dump of pre trained embeddings are parsed. The embedding matrix is the input to the Keras model with embedding layer as its input layer. The CNN model as shown in figure 4, contains a hidden layer that runs along a 1D sequence. As the input sequence is long, an additional layer of 1D convolutional layer is added. Then, a max pooling layer is placed in order to filter the output of the previous convolution layer. to reduce the extent of overfitting because of large dataset, a dropout layer is added as with every training iteration, it assigns the outward edges to zero randomly. A flatten layer is added at the end to change the output shape so that it is suitable for the final dense layer. Rectified Linear Unit (ReLU) is used as the activation function in the initial and intermediate layers of the model as it activates only the necessary neurons at a time because of which weights and biases are not updated during backpropogation. Finally, the softmax activation function is used at the last dense layer as it provides a vector of categorical probabilities making the output a probability distribution because the vector range is in (0,1). The model makes use of categorical cross entropy as the loss which is also called as the softmax loss. This is used as the problem is multi class classification problem. Also, RMSprop is used as the optimization technique. This technique

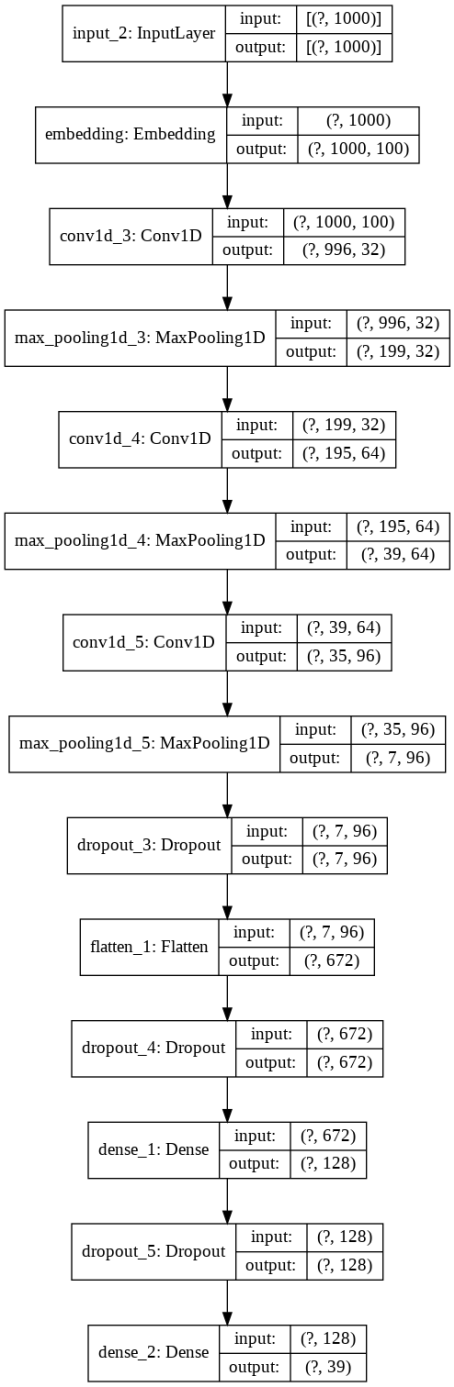

Fig. 4: CNN Model.

is gradient based normalisation which increases the steps for small gradients which in turn reduces the chances of issues like vanishing and also exploding.

The purpose of drop outs is that with each training pass, while switching off few random input neurons, there are less chances of overfitting. By learning improved representations of the patterns, it would generalize better on unseen data. The dropout rate is chosen to be 0.2, which allows 80% of embeddings of each training sample to go to the next layer. This is why it is an hyperparameter.

D. Performance Metrics

This project makes use of F1 score, Precision, Recall as the standard for evaluating the performance of the models. As this is a multi class problem, this metric is quite suitable. F1 Score is the (weighted) average of Recall/Sensitivity and Precision, where Recall is the proportion of effectively anticipated positive observations to all observations in real class and Precision is the proportion of effectively anticipated positive observations to all the anticipated positive observations.

$$Precision = TP/(TP + FP) \tag{1}$$

$$Recall = TP/(TP + FN) \tag{2}$$

$$F1Score = 2 * (Precision * Recall)/(Precision + Recall) \tag{3}$$

where, TP is True positive, TN is True Negatives, FP is False Positives and FN is False Negative. These values help build a confusion matrix that indicates the performance of the model on test data provided the true values are known.

IV. Experimental Results and Discussions

The models are implemented on MTSamples dataset. The Logistic Regression model is implemented on the dataset after performing the necessary preprocessing steps and produces an F1 Score of 0.31 with accuracy as 0.38. It is also interesting as the major classes of the dataset get predicted quite well, but the minor classes were confused a lot. This can be because of imbalanced data. In order to reduce the imbalance, the classes with more than 100 objects were only retained and the model was run again. Accuracy in this run increased to 0.48 and F1 score to 0.52. Although, there was an increase, it wasnt much. The confusion matrix for which is shown in the figure 5. Precision is the highest, at 0.61.

To reduce the imbalance problem, the dataset is under sampled using Naive and near miss method. Accuracy and F1 Score of 0.49 and 0.53 respectively were achieved for naive and near under sampling with a negligible difference.

Smote oversampling is performed and measured. The model is saved as 'smote_log.pkl' and loaded to test against the test data achieving a higher accuracy and F1 score with Smote over sampling. The model performing better than previous times with accuracy at 0.54 and F1 score at 0.59. The related confusion matrix is shown in the figure 6.

CNN model with the same dataset performed exceedingly well achieving a training accuracy of 0.85 and an F1 Score of 0.88 while the validation accuracy was considerably higher, at 0.90, the validation F1 score achieved 0.92. The training and validation results gradually increased over the 7 epochs as shown in the figure 7. When validated with the test data, the accuracy remained similar as validation, with 0.89, while, F1 score increased to 0.92 as shown in table 1.

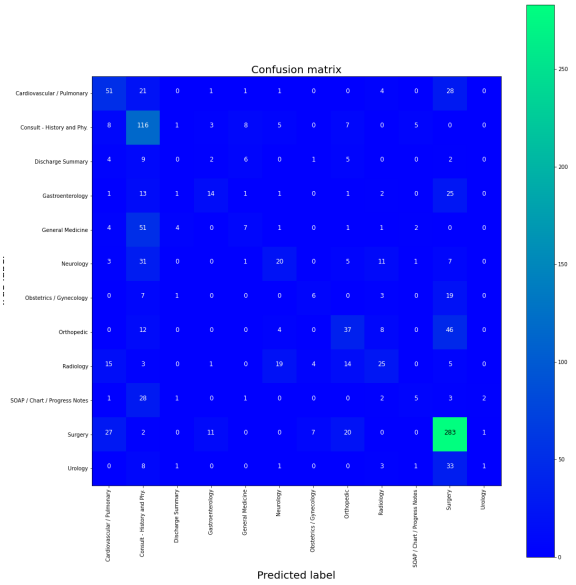

Fig. 5: Confusion Matrix Before Sampling.

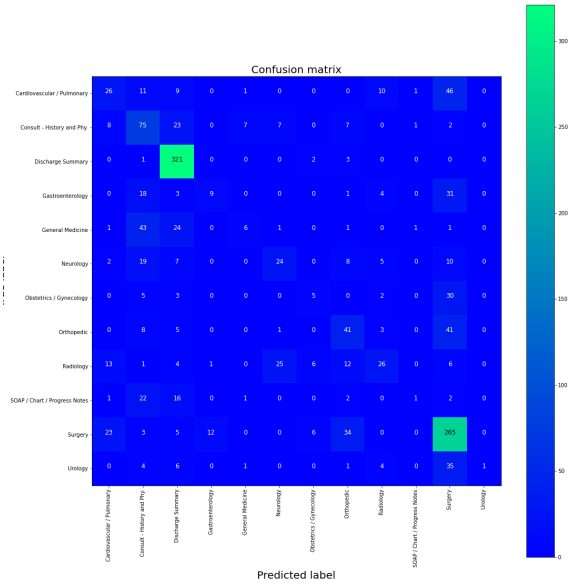

Fig. 6: Confusion Matrix After Smote OverSampling.

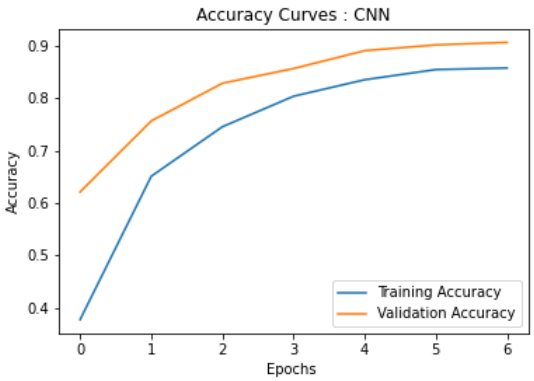

Fig. 7: CNN Accuracy.

	Accuracy	F1 Score	Precision	Recall
BaseLine LogReg	0.38	0.31	0.34	0.39
LogReg with Naive Under Sampling	0.49	0.53	0.61	0.49
Log Reg with Near Under Sampling	0.49	0.53	0.61	0.49
Log Reg with SMOTE Over Sampling	0.54	0.59	0.67	0.54
CNN + GloVE	0.89	0.92	0.98	0.86

TABLE I: CNN vs Logistic Regression

allows the understandability of a normal model run. Accuracy achieved was 0.89, and F1 score was at 0.92.

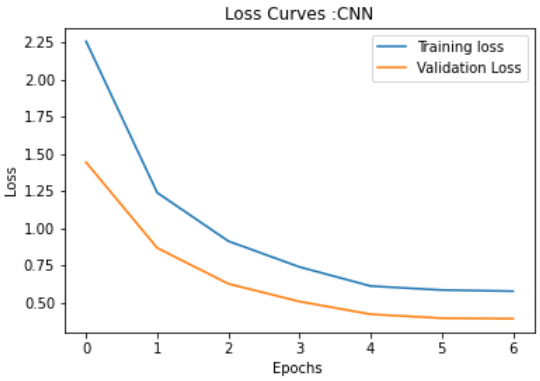

Fig. 8: CNN Loss.

Figure 8 represents the Loss incurred to the model. The training loss reduces gradually from 2.25 in the first epoch to 0.57 in the last epoch. The validation loss reduces gradually from 1.44 to 0.39 in the 7th epoch as shown in table 2. Checkpoints are created after every increase in validation accuracy for each epoch and model weights are saved as 'model_cnn.hdf5'. These are then loaded to evaluate the test data and the loss achieved was lesser, 0.47. There wasnt a major difference between the testing and validation phase with regards to Accuracy, F1 score and loss for every epoch which

While comparing the various sampling techniques in order to reduce the imbalance issue in the dataset, it is understood that SMOTE oversampling produces a higher accuracy and F1 score as seen in Table 1. Although, there is a major improvement, it could be attributed to a random doubling of data in minority class without adding further information to the dataset as a whole. The undersampling techniques of near miss and Naive (random undersampling) produce almost similar results except for negligible changes.

	Training	Val	Test
Accuracy	0.85	0.90	0.92
Loss	0.57	0.39	0.47
F1 Score	0.88	0.92	0.92
Precision	0.95	0.98	0.98
Recall	0.84	0.88	0.86

TABLE II: CNN Model Performance Metrics

DCNN + GloVe	DCNN – Wiki News	CNN - Wiki News	DNN - Wiki News	Bidirectional RNN-Wiki News
0.85	0.76	0.78	0.71	0.68

TABLE III: Performance Comparison - proposed CNN model with other similar Models

We can clearly note that CNN outperforms the Logistic regression model. The proposed model is similar to that of traditional CNN classifier with the only difference being the text intput.

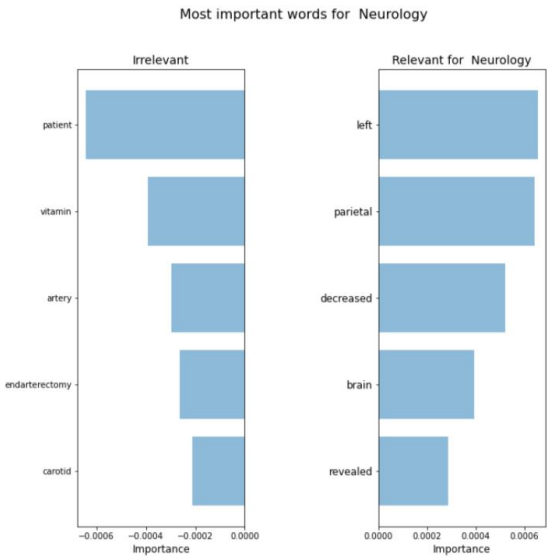

Fig. 9: Important words for Diagnosis - 'Neurology'.

Table 3 represents the model comparison between the CNN with GloVE and other models that have used wiki-news with a dimension of 300 and 1 million vectors although implementing it on another dataset.

Images 8 and 9 represent the important word plot of certain test labels - 'Neurology' and 'Cardiovascular / Pulmonary' respectively.

The CNN model is significantly improved by the GloVE word vector embedding as seen from figure 11. Medical

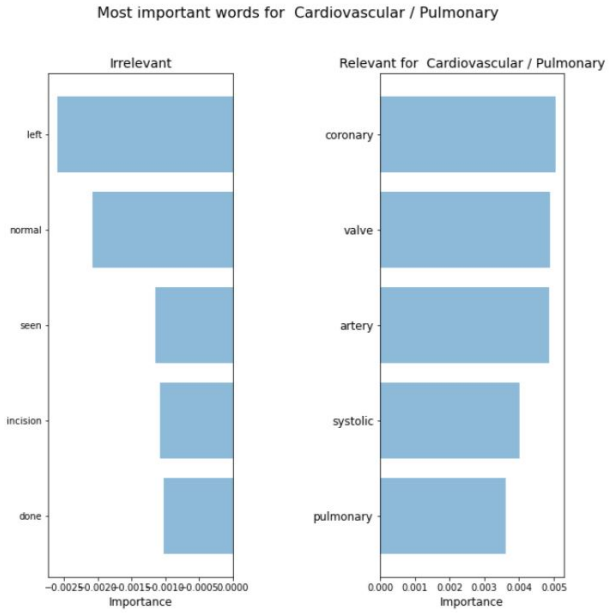

Fig. 10: Important words for Diagnosis - 'Cardiovascular / Pulmonary'.

Transcripts aren't easy to classify and even more difficult is to detect the diagnostics from it as it has specific terminologies.

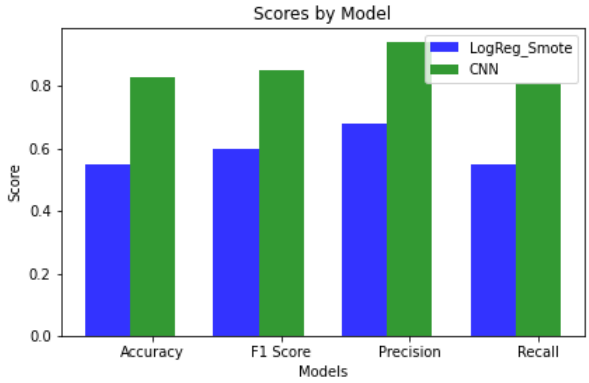

Fig. 11: Comparison of models - CNN vs Logistic Regression.

As shown in figure 12, precision is highest metric measured when compared between the values of the regression models with different sampling. For the near and naive undersampling, it is considered as one as the scores are negligibly different.

V. Conclusion

A Logistic Regression and a CNN model were implemented to detect daignostics from the medical transcripts (MTSamples) dataset. Logistic Regression classifier with SMOTE performs better than other sampling techniques achieving an

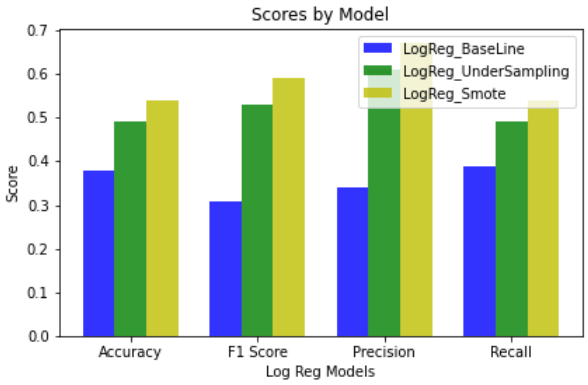

Fig. 12: Comparison of Logistic Regression model - with and without sampling.

accuracy of 0.55 and a F1 score of 0.60. Oversampling with SMOTE, although produces a better accuracy, doesnt provide added information to the dataset. The CNN model with GloVE word vector outscores similar models that use different word vectors to perform detection. The model achieves an accuracy of 0.85 and a F1 score of 0.88. As the validation and testing accuracy is higher than the training values for the CNN model, there might be under fitting issue. The future work includes continuous optimization of the performance and using a bigger dataset to have the metrics evaluated. Although, this work cannot be implemented and used in real life scenario, it could be a small step in the field. Future work can also include a detailed text analysis, a better tuning can be opted. Also, with additional data on the minor classes, those could be involved during the logistic regression model to increase the metric accuracy.

REFERENCES

[1] Murphy, S. F., L. Lenihan, F. Orefuwa, G. Colohan, I. Hynes, and C. G. Collins. "Electronic discharge summary and prescription: improving communication between hospital and primary care." Irish Journal of Medical Science (1971-) 186, no. 2 (2017): 455-459.

[2] Rink, B., Harabagiu, S., Robert, K.: Automatic extraction of relations between medical concepts in clinical texts. J Am Med Inform Assoc. 18(5): 594–600 (2011)

[3] Vidrighin, C., Savin C., Potolea, R.: A Hybrid Algorithm for Medical Diagnosis, Proceedings of EUROCON, Warsaw, 668-673 (2007)

[4] Lee, Wonsung, Sungrae Park, Weonyoung Joo, and Il-Chul Moon. "Diagnosis prediction via medical context attention networks using deep generative modeling." In 2018 IEEE International Conference on Data Mining (ICDM), pp. 1104-1109. IEEE, 2018.

[5] Ren, Jason, Russell Kunes, and Finale Doshi-Velez. "Prediction Focused Topic Models for Electronic Health Records." arXiv preprint arXiv:1911.08551 (2019).

[6] Hughes, Mark, Irene Li, Spyros Kotoulas and ToyotaroSuzumura. "Medical Text Classification using Convolutional Neural Networks." Studies in health technology and informatics 235 (2017): 246-250 .

[7] Hua, Qu, Shi Qundong, Jiang Dingchao, Guo Lei, Zhang Yanpeng and Liu Pengkang. "A Character-Level Method for Text Classification." 2018 2nd IEEE Advanced Information Management,Communicates,Electronic and Automation Control Conference (IMCEC) (2018): 402-406.

[8] L. Yao, C. Mao and Y. Luo, "Clinical Text Classification with Rule-based Features and Knowledge-guided Convolutional Neural Networks," 2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W), New York, NY, 2018, pp. 70-71.

[9] S. J. Fodeh, M. Zirkle, D. Finch, R. Reeves, J. Erdos and C. Brandt, "MedCat: A Framework for High Level Conceptualization of Medical Notes," 2013 IEEE 13th International Conference on Data Mining Workshops, Dallas, TX, 2013, pp. 274-280.

[10] Barbantan, Ioana, and Rodica Potolea. "Learning Diagnosis from Electronic Health Records." In KDIR, pp. 344-351. 2016.

[11] Xu, Hua, Shane P. Stenner, Son Doan, Kevin B. Johnson, Lemuel R. Waitman, and Joshua C. Denny. "MedEx: a medication information extraction system for clinical narratives." Journal of the American Medical Informatics Association 17, no. 1 (2010): 19-24.

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

[13] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

[14] Barbantan, Ioana, and Rodica Potolea. "Knowledge Extraction and Prediction from Unstructured Medical Documents." (2015).

[15] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).

[16] Li, Yifu, Ran Jin, and Yuan Luo. "Classifying relations in clinical narratives using segment graph convolutional and recurrent neural networks (Seg-GCRNs)." Journal of the American Medical Informatics Association 26, no. 3 (2019): 262-268.

[17] Pranckevičius, Tomas, and Virginijus Marcinkevičius. "Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification." Baltic Journal of Modern Computing 5, no. 2 (2017): 221.

[18] Mtsamples.com. 2020. Transcribed Medical Transcription Sample Reports And Examples - Mtsamples. [online] Available at: ¡https://www.mtsamples.com/¿ [Accessed 22 March 2020].