EasyChair Preprint

№ 5361

Classification of Medical Transcriptions with Explanations

Abdul Razak Zakieh and Adil Alpkocak

April 24, 2021

Classification of Medical Transcriptions with Explanations

Abdul razak Zakieh[1][0000-0003-4974-9795] and Adil Alpkocak[2][0000-0001-7695-196X]

Dokuz Eylül University, Izmir, Turkey

Abstract. Researchers and developers are using neural networks and deep neural networks widely in many fields. Even though those artificial intelligent models provide high performance, the way they work is not clear and users cannot understand its logic behind a specific decision. That is why we cannot use AI models in real applications in the medical field for example. In this paper, we focused on the importance of providing explanations, provided a brief review about the field of Explainable AI, XAI, and used three different ways to provide explanations for users by doing experiments on a medical-transcriptions dataset. We used the self-explainable decision trees, different neural network models with separate explainers, and lastly, we used bidirectional LSTM model with attentions as explanations.

Keywords: Explainable Artificial Intelligence, Explainability with Attentions, Self-Explainable Models

1 Introduction

These days, we are using machine learning in most of the fields and the results we are getting are excellent. However, the way most machine learning models work, especially neural networks and deep neural networks, is not clear and researchers do not know the reasons beyond its success or failure. Besides, the end-users cannot trust the decision of AI programs because they do not provide explanations especially in sensitive cases such as the medical or security field. That is why researchers are focusing on explaining the behavior of machine learning models to make them able to provide explanations for their decisions. This field of study is called Explainable AI, or XAI. Researchers have defined explainability in different ways and some of them have divided it into more than one term. Miruna A. Clinciu and Helen F. Hastie distinguished between four terms; transparency, intelligibility, interpretability, and explainability, and tried to establish a set of standard terms [1]. Without digging into the details of those terms, we can say that the ultimate goal of XAI is to make the artificial intelligence model understandable by the user.

1.1 Types of Explainable Models

According to [2] we can divide explainable models by two factors: (i) the scope of the explanation and (ii) the source of it. Local scope means the explanation is for a single prediction while global means the explanation is for the whole prediction scope. The source of explanation can be the model itself, we call it self-explaining model, or from further post-processing where we call it posthoc-explaining model. Therefore, we have four types of explanation models: (I) Local Post-Hoc, (II) local Self-Explaining, (III) Global Post-Hoc, and (IV) Global Self-Explaining. L. A. Hendricks et al. [3] proposed a model that provides explanations of a visual classifier. We can consider their model as a local posthoc explaining model where explanations are for each example and the model generates explanations by a novel model after a visual classifier. Rule-based models such as Decision Trees are self-explaining models and we can say that they are global and local since we can understand the generated model and we can generate an explanation for a specific prediction. N. Liu et al. proposed a scheme to interpret any type of embedding method and the scheme they proposed is global posthoc scheme [4]. H. Liu et al. made a novel model that make classification and provide fine-grained explanations as well [5]. In this paper, we provide explanations for medical transcription classification problem using three methods: decision trees as self-explaining model, attention values as local self-explaining method and lastly we used LIME [6] that explains the prediction of any classifier by learning an interpretable model locally around the prediction. We wanted to apply the novel method by [5] but their model depends on using fine-grained information and we do not have them in our dataset.

1.2 Visualization Techniques

Visualizing the explanation is so important in explainable models. Visualizing the explanations depends on the model we have and the data we treat. White-box models can provide visualization easily. In decision trees, for example, we can plot the tree as an explanation. In other cases, we can plot a heatmap of specific features, like plotting the heatmap of attention values. For NLP models, we can also highlight the important words with high attention instead of plotting the heatmap [7].

In the following sections, we talk about the materials and methods we used, how we preprocessed the dataset, and the technical details about the models we used. Then we talk about the conclusions and results of our experiments and the notices we found.

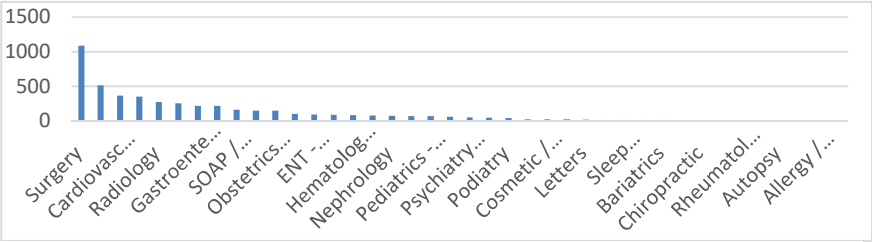


Fig. 1. Number of transcriptions for each category

2 Material and Methods

The dataset we have consists of textual medical transcriptions and the category each transcription belongs to. In our tests, we did not use accuracy as a metric to compare different models because the data is imbalanced. We used F1-Score that represents the harmonic mean of precision and recall. Equation (1) shows the formula of F1-Score.

$$F_1 = 2 \times (Precision \times Recall) / (Precision + Recall) = 2TP / (2TP + FP + FN) \quad (1)$$

2.1 Dataset

Medical Transcriptions[1] is a dataset that offers medical transcription samples with their categories. The dataset has originally 40 different categories with 2348 transcripts.

2.2 Preprocessing

First, we processed the data and removed any invalid entries that have either of the transcription and/or the category empty. Then we transformed all the texts to lower case, deleted punctuations, and removed stop words. For representing the textual data, we chose Word2Vec [8] with an embedding dimension equals to 100. The dataset was highly imbalanced as we can see in Fig.1 so we deleted the categories that have less than 100 transcriptions and then oversampled the data using SMOTE [8] over-sampler

2.3 Classification

To classify the transcriptions we used four different models; decision trees, artificial neural network (ANN), convolutional neural network (CNN), and Bidirectional-Long-Short-Term-Memory (Bi-LSTM). The ANN, CNN, and Bi-LSTM models are the same ones proposed in [10] except that we used Bi-LSTM cells instead of LSTM ones. We used Bi-LSTM because it gives better performance with textual data. Figures 2, 3, and 4 show the ANN, CNN, and Bi-LSTM models respectively.

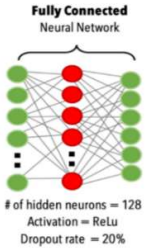


Fig. 2. ANN Model

[1] https://www.kaggle.com/tboyle10/medicaltranscriptions

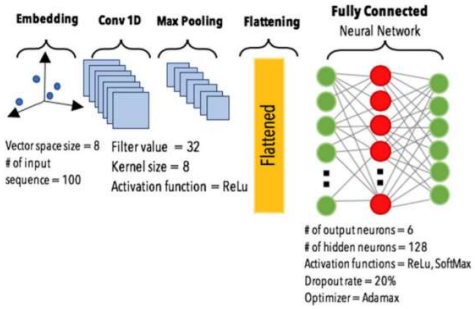


Fig. 3. CNN Model

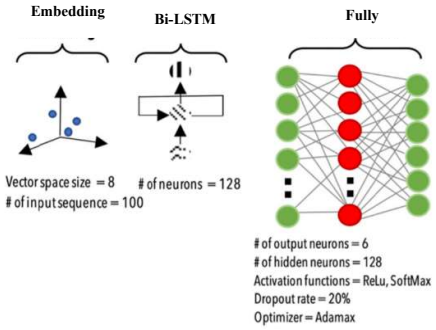


Fig. 4. Bi-LSTM Model

2.4 Explanations and Visualizations

We aimed in our experiments not only to do classification but to make explanations as well. Decision trees are open-box models, self-explained, so we did not have to do extra processing. For visualizing the explanations, we plotted the decision tree. Fig 5 shows the plotted tree. For the Bi-LSTM model, we used LIME [6], which produces a local explanation for a specific prediction by training a separate surrogate model. Fig. 6 shows an explanation for correctly classified transcriptions and fig. 7 shows an explanation for wrongly classified transcriptions where we used text highlighting to visualize the explanations. Finally, we used attentions with the Bi-LSTM model to enable explanations. For each word in the input, we have a value that represents how it affects the output. We used those attentions values as local explanations and plotted them as heatmap. Fig. 8 shows the attention model we used and fig. 9 shows parts of two heat maps for correctly and incorrectly classified examples.

3 Experiments and results

We did all the experiments using Keras framework. We split the data randomly as 80% for training and 20% for testing. After preprocessing the data, we generated the

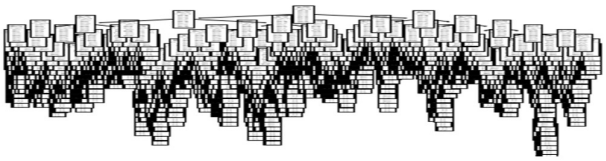


Fig. 5. The decision tree used to classify transcriptions

chief complaint abdominal pain history present illness patient year old female patient dr x patient presented emergency room last evening approximately day history abdominal pain persistent seen days ago abc er underwent evaluation discharged ct scan time told normal given oral antibiotics cipro flagyl nausea vomiting persistent associated anorexia passing flatus obstipation symptoms last bowel movement two days ago denies bright red blood per rectum history recent melena last colonoscopy approximately years ago dr definite fevers chills history jaundice patient denies significant recent weight loss past medical history significant history atrial fibrillation good control normal sinus rhythm metoprolol also premarin hormone replacement past surgical history significant cholecystectomy appendectomy hysterectomy long history known grade bladder prolapse seen past dr chip winkel believe consulted allergies allergic sensitive macrodantin social history drink smoke review systems otherwise negative recent febrile illnesses chest pains shortness breath physical examination general patient elderly thin white female pleasant acute distress vital signs temperature vital signs stable within normal limits heent head grossly atraumatic normocephalic sclerae anicteric conjunctivae non injected neck supple chest clear heart regular rate rhythm abdomen generally nondistended soft focally tender left lower quadrant deep palpation palpable fullness mass focally tender rebound tenderness cva flank tenderness although minimal left flank tenderness pelvic currently deferred history grade urinary bladder prolapse extremities grossly neurovascularly intact laboratory values white blood cell count hemoglobin platelet count normal alkaline phosphatase elevated liver function tests otherwise normal electrolytes normal glucose bun creatinine diagnostic studies ekg shows normal sinus rhythm impression plan year old female greater one week history abdominal pain localized left lower quadrant currently nonacute abdomen working diagnosis would sigmoid diverticulitis history distant past sigmoid diverticulitis would recommend repeat stat ct scan abdomen pelvis keep patient nothing mouth patient seen years ago dr colorectal surgery consult also evaluation patient need repeat colonoscopy near

Fig. 6. Explanation for correctly classified transcription by Bi-LSTM model (Correctly predicted as "Consult - History and Phy")

embedding vectors using Word2Vec class from Gensim library with min_count = 1 and embedding size equals 100. Then we converted the input texts to their relevant word indices that map each word to its embedding vector. We vectorized the labels and binarized them to make them ready to use with SMOTE oversampler. The next step was oversampling the data where we got 4902 examples. For the neural network models, we used categorical cross-entropy as a loss function and Adam optimizer with learning rate equals to 0.001. Table 1. shows the accuracy, recall, precision, and f1-score for each of the classifiers we have in addition to the Bi-LSTM with Attention model. We see that the Bi-LSTM model has the best f1-score. However, in our evaluation, we need to consider the explainability. Decision Trees are straightforward and clear but their performance in classification is very low. ANN, CNN, and Bi-LSTM models are like black boxes and are not understandable at all but they provide very good performance compared to decision trees. However, using decision trees to make a prediction and provide an explanation as well, the user will have more confidence in the model if he sees a logical explanation. On the other hand, even if the Bi-LSTM model has very good performance it still has an error rate and the user will not know whether the model is giving correct or incorrect prediction because it has no explanation. The attention model, which we can apply to ANN, CNN, or Bi-LSTM, can provide high performance with good and logical explanations. When providing a prediction with an explanation, the end-user can decide whether to trust the prediction or not based on the provided explanation. Finally, we can use a posthoc model and get it trained to provide explanations for any model we have. However, we believe that explanations provided by a classifier itself are more accurate than explanations provided by other posthoc functions.

Table 1. Performance metrics for decision tree, ANN, CNN, Bi-LSTM, and Bi-LSTM with Attention models

Model	Precision	Recall	F1-Score	Accuracy
Decision Tree	0.271464	0.286241	0.277782	0.238974
ANN	0.998974	0.228718	0.372216	0.228718
CNN	0.762953	0.518974	0.612205	0.518974
Bi-LSTM	0.712139	0.571282	0.617014	0.571282
Bi-LSTM with Attention	0.841395	0.559633	0.654027	0.559633

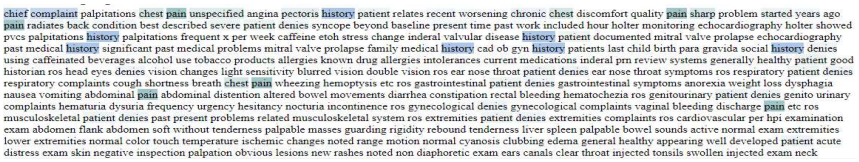


Fig. 7. Explanation for incorrectly classified transcription by Bi-LSTM model (Predicted as "Consult - History and Phy" instead of "Cardiovascular / Pulmonary")

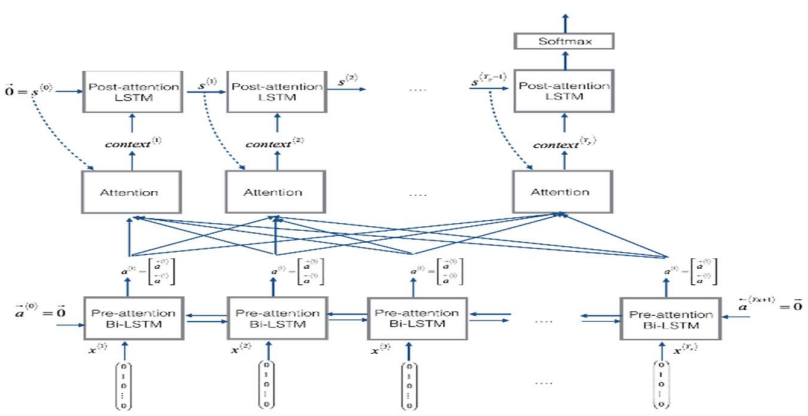
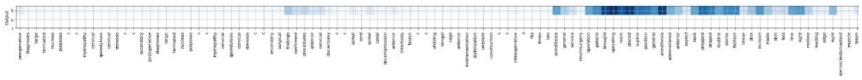


Fig. 8. Bi-LSTM with Attention Model



Incorrectly classified prescription. (Classified as Surgery instead of Orthopedic)

Correctly classified prescription. (Classified as History and Phy.)

Fig. 9. Attention as plotted as heatmap

4 Conclusion

In this paper, we explored different ways to provide and visualize explanations for AI models. We saw that black-box models, neural networks, and deep neural networks, give better performance than white-box models. However, white-box models are more clear and understandable by users that make them more explainable. To overcome this problem, we modify the deep neural networks to make them able to provide explanations for their decisions or we can train another different model to do so. We noticed that most of the previous works focus on making neural models explainable from the end-user point of view and the provided explanations help slightly in making the researcher understand the model he built and how to improve it. For example, we noticed from the explanations provided by Bi-LSTM with the Attention model that

our model provides wrong decisions for short transcriptions and we have also to enhance the classifier part, fully connected layers, of our model rather than the Bi-LSTM and attention part. However, we cannot understand the real functionality of each layer, when to add an extra layer or delete one, whether we have to add more neurons or delete some of them and so on. In order to make the AI field more practical and trustworthy, we encourage working on XAI models and not using AI without explanations at all.

Acknowledgment

The researcher *Abdul razak Zakieh* is pursuing his master's degree as a scholarship student that is provided by "Presidency for Turks Abroad and Related Communities (YTB)". For the experiments, we used the public Python notebook "Oversampling SMOTE and ADASYN" as a starter, and for the attention model, we got great benefit from the "Neural Machine Translation with Attention" lab in "Sequence Models by DeepLearning.AI" course on Coursera where we modified the model and the code to fit our experiment.

References

1. M. A. Clinciu and H. F. Hastie, "A Survey of Explainable AI Terminology," *Edinburgh Centre for Robotics.*
2. M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas and P. Sen, "A Survey of the State of Explainable AI for Natural Language Processing," *IBM Research – Almadem.*
3. L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele and T. Darrell, "Generating Visual Explanations," 2016.
4. N. Liu, X. Huang, J. Li and X. Hu, "On Interpretation of Network Embedding via Taxonomy Induction," 2018.
5. H. Liu, Q. Yin and W. Y. Wang, "Towards Explainable NLP: A Generative Explanation Framework for Text Classification," *Association for Computational Linguistics,* no. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, p. 5570–5581, 2019.
6. M. T. Ribeiro, S. Singh and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," 2016 .
7. J. Mullenbach, SarahWiegreffe, J. Duke and Jimeng, "Explainable prediction," *Proceedings of the 2018 Conference of the North American ,* vol. 1, no. Chapter of the Association for Computational Linguistics: Human Language Technologies, p. 1101–1111, 2018.
8. T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013 .
9. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," 2002.
10. M. A. TOCOGLU, O. OZTURKMENOGLU and A. ALPKOCAK, "Emotion Analysis From Turkish Tweets Using Deep Neural Networks," 2019.