Enhancing Risk Aware Decision in Healthcare through Probabilistic Modeling of Uncertainty

Rahul Vishwakarma*
rahuldeo.vishwakarma-sa@csulb.edu
California State University
Long Beach, California, USA

Jinha Hwang*
jinha.hwang01@student.csulb.edu
California State University
Long Beach, California, USA

Benyamin Ahmadnia
benyamin.ahmadnia@csulb.edu
California State University
Long Beach, California, USA

ABSTRACT

In domains with high stakes, like healthcare and medicine, trustworthy and robust decision-making is crucial due to the potential risks associated with misclassification. However, many traditional machine learning classifiers lack calibrated predictions, and reliable uncertainty estimates for new unseen data. This paper addresses the challenge of uncertainty quantification in text classification in healthcare and proposes a three-fold approach to *support* robust and trustworthy decision-making by medical practitioners. To evaluate our solution, we implement it on a multi-label medical transcription dataset from Kaggle. Our study demonstrates three significant results: the ability of our model to reject uncertain predictions by providing a null set, the provision of set predictions with guaranteed coverage for further investigation, and the prioritization of decision-making based on confidence levels of predictions with the same label. Additionally, we tackle the issue of imbalanced datasets in the medical domain by employing the Mondrian Conformal Predictor with a Naïve Bayes classifier. Our findings are expected to enhance the risk-aware decision-making process in the medical field.

CCS CONCEPTS

• **Computing methodologies** → **Probabilistic reasoning**.

KEYWORDS

uncertainty, conformal inference, risk-aware decision

1 INTRODUCTION

Machine learning has emerged as a powerful tool in the medical domain, providing healthcare professionals with a means to make more informed decisions and improve patient outcomes [1]. Medical transcript analysis, in particular, holds the potential to assist in diagnostics [2], treatment planning [3], and patient monitoring [4]. However, traditional natural language processing techniques have limitations that can result in inaccuracies in classification [5, 6].

Natural Language Processing (NLP) techniques, specifically text classification, can be leveraged in various industry applications to overcome these limitations. Several methods have been proposed to achieve this goal. [3, 7] However, they are often under-utilized due to the criticality of the decisions involved and a lack of confidence in individual decisions. Currently, with text classification, we don't have a mechanism to 'tell' the model to be more strict or lenient while making a decision. Hence, the accuracy of the model solely depends on the classifier.

1.1 Related Work

Text classification has been widely explored in the field of NLP, and it has found applications in various domains such as finance [8], military [9], and medical [10, 11], among others. Most of the research in this field has focused on developing algorithms that can improve accuracy while keeping the computational cost low [12]. A few works [13] have used aleatory and epistemic uncertainty. Still, they do not quantify the prediction of each new prediction, while [14] focuses on the profound learning aspect of text classification. However, we still see a gap in the practical realization and the applicability of the metrics for confident decision-making for a text classification system.

2 NOVEL CONTRIBUTIONS

The novelty of this work is that the proposed mechanism not only gives classification out, but two measures, i.e., confidence and credibility, to tune those decisions based on the criticality and control the algorithmic decision-making. The main contributions of this paper are as follows:

- Uncertainty quantification for each prediction for improved and robust decision-making for the medical domain with imbalanced data.
- An algorithm-agnostic framework with an option to reject predictions made by the model and the reduction of false positive rate when the model says for a particular prediction for which it is unsure.
- Disease severity rating mechanism helps the decision makers prioritize individualized treatment.

3 PROPOSED SOLUTION

We use the dataset from Kaggle Medical transcription data scraped from mtsamples.com, which has used only 8 Medical specialties for the classification task. The experimental results, source code, and the dataset is hosted on GitHub [1]. We used Mondrian conformal predictor as an uncertainty quantifier as a wrapper on top of the Naïve Bayes classifier; however, any other classification algorithm can also be used. Designing an efficient non-conformity score is of prime importance and shall be taken care of by the designer. The solution provides the three contributions as discussed in the results section.

*Both authors contributed equally to this research.

[1] https://github.com/rahvis/CECS590

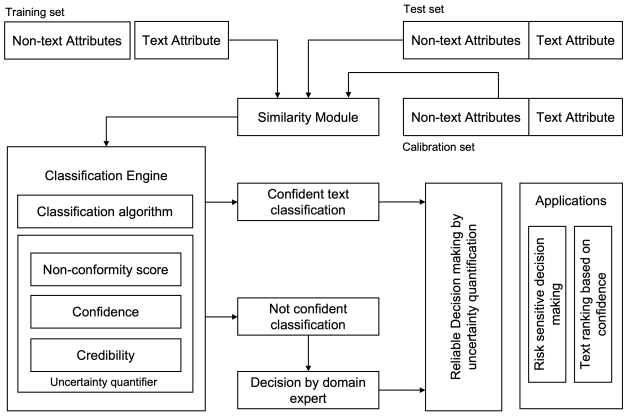

Figure 1: Conformalized text classification for implementing a reliable decision making

4 RESULTS

4.1 Conformal Inference

Conformal prediction is a framework for constructing prediction intervals for machine learning models. The main goal of conformal prediction is to provide a measure of confidence for the predictions made by a model. Traditional performance metrics like precision and recall are insufficient in conformal prediction because they do not capture the uncertainty in the predictions.

p0	p1	p2	p3	p4	p5	p6	p7	y
0.03	0.03	0.04	0.03	0.03	0.83	0.02	0.03	5
0	0	0	0.95	0	0	0	0	3
0.1	0.1	0.54	0.1	0.1	0.1	0.1	0.1	2
0.08	0.08	0.09	0.08	0.67	0.08	0.08	0.08	4
0.1	0.11	0.1	0.5	0.11	0.1	0.11	0.1	3
0.12	0.14	0.36	0.12	0.12	0.12	0.12	0.12	2
1	0	0	0	0	0	0	0	0

Figure 2: Conformal inference and associated p-values for each label for 7 test samples

The p-value is a metric for measuring the confidence of a machine learning model's predictions. It is calculated by comparing the model's prediction for a new piece of data with its predictions for the data on which it was trained through hypothesis testing. Suppose the new data differs significantly from the data seen during training. In that case, the p-value will be low, indicating that the model's prediction for the new data may not be as reliable.

4.2 Reject - I don't know

When a model does not get any of the p-values greater than the alpha value, it outputs a NULL set as shown in the Figure 3.

4.3 Disease severity rating

We use the confidence metrics to rank the predicted labels, which are defined as:

$$\text{Confidence}(x) = \sup\{1 - \epsilon : |\Gamma_\epsilon(x)| \leq 1\}.$$

Significance	Confidence	Prediction
0.5	0.5	{NULL}
0.14	0.86	{N}
0.13	0.87	{A, N}

Figure 3: Model says I don't know based on the NULL set.

Here, data point 7 is of high priority compared to data point 6 because of the confidence metrics.

	Confidence	Credibility	y_pred
1	0.962	0.831	5
6	0.863	0.358	2
7	0.999	0.997	0

Table 1: Adoption of confidence for risk-aware ranking

5 CONCLUSION

The medical field poses a challenge when it comes to trusting individual predictions because of its complexity and uncertainty. This research paper introduces a framework not tied to any specific algorithm and aims to quantify the uncertainty associated with new and unseen data points in the medical domain. The proposed approach is tested on a dataset of medical transcriptions and demonstrates promising yet modest results, offering a valuable contribution from a methodological standpoint. Furthermore, we demonstrate how ranking labels based on their associated risks can assist in prioritizing treatment in large-scale scenarios. In future work, we plan to evaluate various customized conformal prediction techniques such as Mondrian conformal predictors, risk-aware prediction sets (RAPS), and top-k methods.

REFERENCES

[1] K. S. Lakshmi and G. Santhosh Kumar. Association rule extraction from medical transcripts of diabetic patients. In *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, pages 201–206, 2014.

[2] Loukas Ilias and Dimitris Askounis. Explainable identification of dementia from transcripts using transformer networks. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4153–4164, 2022.

[3] Marjory Day, Rupam Kumar Dey, Matthew Baucum, Eun Jin Paek, Hyejin Park, and Anahita Khojandi. Predicting severity in people with aphasia: A natural language processing and machine learning approach. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 2299–2302, 2021.

[4] Vitali Loseu, Hassan Ghasemzadeh, and Roozbeh Jafari. A mining technique using *n* n-grams and motion transcripts for body sensor network data repository. *Proceedings of the IEEE*, 100(1):107–121, 2012.

[5] Shang Gao, Mohammed Alawad, M. Todd Young, John Gounley, Noah Schaefferkoetter, Hong Jun Yoon, Xiao-Cheng Wu, Eric B. Durbin, Jennifer Doherty, Antoinette Stroup, Linda Coyle, and Georgia Tourassi. Limitations of transformers on clinical text classification. *IEEE Journal of Biomedical and Health Informatics*, 25(9):3596–3607, 2021.

[6] Oscar NE Kjell, Sverker Sikström, Katarina Kjell, and H Andrew Schwartz. Natural language analyzed with ai-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific reports*, 12(1):3918, 2022.

[7] Anjana S. Nambiar, Kanigolla Likhita, K. V. S. Sri Pujya, Deepa Gupta, Susmitha Vekkot, and S. Lalitha. Comparative study of deep classifiers for early dementia detection using speech transcripts. In *2022 IEEE 19th India Council International Conference (INDICON)*, pages 1–6, 2022.

[8] Mouad Ablad, Bouchra Frikh, and Brahim Ouhbi. Uncertainty quantification in deep learning context: Application to insurance. In *2020 6th IEEE Congress on Information Science and Technology (CiSt)*, pages 110–115, 2020.

[9] Charith Gunasekara, Tobias Carryer, and Matt Triff. On natural language processing applications for military dialect classification. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 211–218, 2021.

[10] Jie Li, Qilin Huang, Siyu Ren, Li Jiang, Bo Deng, and Yi Qin. A novel medical text classification model with kalman filter for clinical decision making. *Biomedical Signal Processing and Control*, 82:104503, 2023.

[11] Asher Lederman, Reeva Lederman, and Karin Verspoor. Tasks as needs: reframing the paradigm of clinical natural language processing research for real-world decision support. *Journal of the American Medical Informatics Association*, 29(10):1810–1817, 2022.

[12] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–41, 2022.

[13] Wenshi Chen, Bowen Zhang, and Mingyu Lu. Uncertainty quantification for multilabel text classification. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6):e1384, 2020.

[14] Jialin Yu, Alexandra I Cristea, Anoushka Harit, Zhongtian Sun, Olanrewaju Tahir Aduragba, Lei Shi, and Noura Al Moubayed. Efficient uncertainty quantification for multilabel text classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.