Uncertainty Quantification of Text Classification in a Multi-Label Setting for Risk-Sensitive Systems

Jinha Hwang, Carol Gudumotu, Benyamin Ahmadnia
Department of Computer Engineering and Computer Science
California State University, Long Beach, United States
jinha.hwang01@student.csulb.edu, caroleunice.gudumotu01@student.csulb.edu,
benyamin.ahmadnia@csulb.edu

Abstract

This paper addresses the challenge of uncertainty quantification in text classification for medical purposes and provides a threefold approach to support robust and trustworthy decision-making by medical practitioners. Also, we address the challenge of imbalanced datasets in the medical domain by utilizing the Mondrian Conformal Predictor with a Naïve Bayes classifier. Our findings are expected to complement the risk-aware decision-making process in the medical field.

1 Introduction

This paper focuses on developing a novel method based on a robust conformal framework for a confidence-based classification for better decision-making. Our project aims to develop methods for uncertainty quantification in text classification for risk-sensitive systems. Using medical transcription data from Kaggle, we assign patients to specific labels based on their medical history. With a better understanding of the uncertainty associated with our predictions, we aim to enable more reliable and robust decision-making in the medical domain.

To address the limitations of traditional NLP techniques in the medical domain, our paper proposes a novel framework for uncertainty quantification in text classification for risk-sensitive systems. We highlight the existing problems in text classification and why uncertainty quantification is essential for evaluating the models. We review the previous works on uncertainty quantification in ML and emphasize the need for a reliable decision-making framework. We propose a three-step methodology that involves training and testing data sets, calibration sets, and classification engines. In the first step, we use a medical transcription data set and obtain a confusion matrix using the Naïve Bayes classifier. In the second step, we use conformal prediction with a calibration set and create another confusion matrix to observe a decrease in error rate in most cases. We assign $p - values$ to labels based on the confusion matrix output, which gives us the confidence level and credibility score, decided by the $p - value$. Our main novelty is the integration of existing conformal prediction with text similarity. Our proposed framework gives a classification and provides two evaluation metrics, confidence and credibility, which offer helpful insights instead of just giving binary classification labels. In conclusion, our proposed framework can be used for reliable decision-making in risk-sensitive systems such as the medical domain.

2 Related Work

Text classification has been widely explored in the field of NLP, and it has found applications in various domains such as finance (Ablad et al., 2020), military (Gunasekara et al., 2021), and medical (Lederman et al., 2022; Li et al., 2023), among others. Most of the research in this field has focused on developing algorithms that can improve accuracy while keeping the computational cost low (Li et al., 2022). However, achieving high accuracy alone cannot ensure a reliable system in risk-sensitive domains like medical applications. A framework is required to address the uncertainty associated with the predictions made by ML models to enable trustworthy decision-making (Psaros et al., 2023).

Recently, there has been growing interest in designing novel metrics for the medical applications of Artificial Intelligence (AI) (Hicks et al., 2022; Cheung et al., 2022). 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.

Kuleshov et al. (2018) suggests a technique called "Calibrated Regression" to estimate uncer-

Proceedings of Recent Advances in Natural Language Processing, pages 541–547
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_059

tainty in Deep Learning models accurately. The method involves training a regression model to predict the variance of the model's output given the input data. The regression model is trained on a validation set to ensure it is well-calibrated, meaning that the predicted variance values accurately measure the model's uncertainty. They show that their approach can accurately estimate uncertainty in various Deep Learning models, including those used for Image Classification and NLP.

Another proposed method for estimating predictive uncertainty in deep neural networks is called "Deep Ensembles", where multiple networks with the same architecture but different random initializations are trained to estimate uncertainty. The authors demonstrate that their approach is simple, scalable, and effective in estimating uncertainty in various benchmark datasets, which can be utilized to detect out-of-distribution examples and improve model calibration (Lakshminarayanan et al., 2017).

In this paper, we overcome the above limitations, and with the proposed method, we conclude multifold benefits. We provide complementary metrics to quantify the uncertainty and provide the outcome to the decision maker to make a robust and trustworthy decision.

3 Methodology

The methodology used in this study complements the existing ML classification algorithms for NLP techniques by incorporating Conformal Prediction (CP) as an uncertainty quantifier to reduce the false discovery rate and make the model robust and reliable.

Traditionally, classification algorithms for NLP use descriptive text data (x) as input data to predict the output label (y), such as positive or negative sentiment. This prediction is made by feeding x into a function $f(x)$, which returns a label (y) based on the features in x. In this paper, we take a step further by incorporating CP into our approach.

3.1 Conformal Prediction

CP is a method that yields prediction intervals with guaranteed coverage associated with a confidence level, $1 - \alpha$, where α is a predetermined value between 0 and 1 (Chernozhukov et al., 2021). The algorithm aims to compose a function f that can accurately predict the label y for a new feature vector X in a given set of training data consisting of feature vectors x_i and their corresponding labels y_i.

CP generates prediction sets $\Gamma(x)$ for each feature vector X, such that the probability of the true label being in the prediction set is at least $1 - \alpha$ for all x and y. This framework can use different algorithms, including the nonconformist and transductive conformal prediction methods. When the predefined significance level cannot eliminate any of the labels, CP has the potential to generate a prediction set of multiple possible values, which makes the predictions uncertain.

CP is a technique that can produce prediction sets containing multiple possible labels, meaning that the confusion matrix generated differs slightly from the conventional confusion matrix. When using CP for multi-label classification, we must pay attention to the number of correctly predicted examples containing all the correct labels and the number of incorrectly predicted models where the prediction set includes at least one incorrect label. This helps to accurately assess the performance of a conformal predictor in multi-label classification while considering the possible labels of the prediction sets.

The total number of empty prediction sets is another crucial factor in evaluating a conformal predictor for multi-label classification. This occurs when no labels can be rejected at the predefined significance level. In such cases, it is essential to provide a single-point prediction by selecting the labels with the highest p-values. However, this approach can be more complicated to interpret in multi-label classification than in binary classification, as it does not provide information on the relative importance of each label. Hence, it is often better to provide a prediction set or interval that encompasses all the possible labels, along with a measure of the uncertainty associated with each label.

3.2 Proposed Framework

In Figure 1, we provide a broad outline of our solution comprising three key components. As with any ML-based approach, the initial stage involves preprocessing the dataset. For this purpose, we obtained medical transcriptions for diverse medical specialties sourced from Kaggle. Accessing medical data is challenging due to the privacy regulations imposed by HIPAA. However, this dataset presents a viable alternative by providing medical transcription samples, which we utilized in our

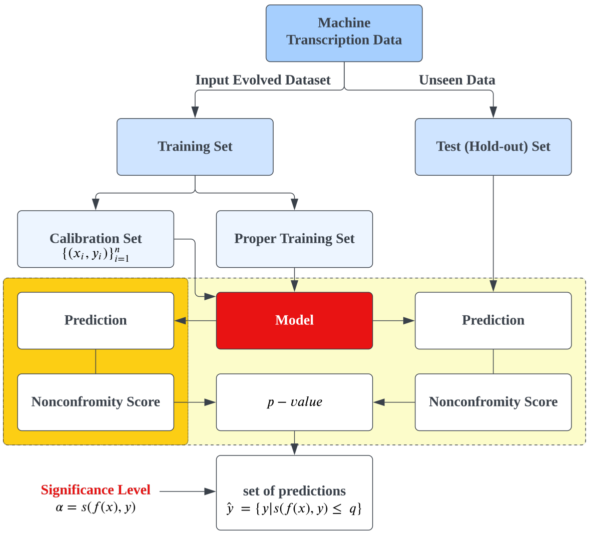
Machine Transcription Data
Input Evolved Dataset
Unseen Data
Training Set
Test (Hold-out) Set
Calibration Set $\{(x_i, y_i)\}_{i=1}^{n}$
Proper Training Set
Prediction
Model
Prediction
Nonconfromity Score
$p - value$
Nonconfromity Score
Significance Level $\alpha = s(f(x), y)$
set of predictions $\hat{y} = \{y|s(f(x), y) \leq q\}$

Figure 1: Proposed framework for uncertainty quantification.

work.

The preprocessed dataset is input into the conformal inference engine, which outputs a set of predictions based on the significance level rather than a single-point prediction. Unlike the traditional approach of splitting a dataset into a train and test set, our method divides the dataset into training, calibration, and test sets. The training set is utilized for training a base learning algorithm on the dataset, resulting in an approach that is *algorithm agnostic*. This implies that any ML classifier, whether statistical or Deep Learning-based, can be used with the conformal inference framework acting as a wrapper over the base algorithm.

In the diagram, the base algorithm is labeled as the “Model”. The conformal inference framework can then be assigned as a wrapper over the base algorithm, denoted as the Model in the diagram. The non-conformity score is calculated for each prediction, and a p-value is assigned based on the significance level. The p-value indicates the probability that the prediction is correct and is used to determine the guaranteed coverage for the prediction. In a high-risk sensitive domain where even a single incorrect decision is intolerable, the most critical aspect of the solution is interpreting the results.

We derive three different inferential use cases based on conformal inference. The motive is to quantify the uncertainty associated with each prediction and reduce the False Discovery Rate (FDR) for medical transcription data. Considering the degree of risk, associated with the prediction, a significance level is defined and applied to the p-values of each label for the data point of -. This results in a set prediction with all the labels, a combination of labels, a single label, or a NULL set, indicating that the model cannot output the prediction. Finally, we calculate the confidence of each prediction and use it to rank the severity of -. The purpose of ranking is to prioritize which one to take action on first.

4 Experimental Framework

This section shows the experimental results of the medical transcriptions dataset from Kaggle. The experimental results with source code and dataset are provided on GitHub [1]

4.1 Dataset

This section details the dataset used for our work, the conducted experiments, and the results. The dataset contains sample medical transcriptions scraped from mtsamples[2]. It includes transcriptions from various medical specialties and can be used for classification tasks to identify the specialty based on the transcription text.

Table 1 shows the column names and descriptions for the medical transcription dataset obtained from Kaggle[3]. The dataset includes sample medical transcriptions for various medical specialties and their titles, relevant keywords, and other relevant information.

We split the dataset into training and test sets, as shown in Table 2. Additionally, we used a calibration set for CP. To divide the data into these three sets, a common practice is randomly splitting the available data into two sets using the train_test_split function from the scikit-learn library. This function divides the data into two sets based on a specified proportion. The first split creates a test set, typically containing around 20% of the available data. The remaining data is then combined into a training and calibration set.

Next, the training and calibration set is divided into two subsets using train_test_split again. This time, the calibration set typically contains around 20% of the available data, while the remaining data is assigned to the training set. By splitting the combined data again, we can obtain a dedicated subset of data for model calibration that is not used for training. Additionally, the random splitting process

[1] https://anonymous.4open.science/r/textconformal

[2] https://mtsamples.com

[3] https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions

should be repeated with different random seeds to assess the robustness of the model's performance estimates.

The dataset split into a training, calibration, and test set for medical specialty features is shown in Table 2.

Column Name	Description
Unnamed (ID)	Unique identifier for each transcription
description	Short description of transcription
medical_specialty	Medical specialty classification of transcription
sample_name	Transcription title
transcription	Sample medical transcriptions
keywords	Relevant keywords from transcription

Table 1: Table description for the Kaggle medical transcription dataset.

	train	cal	test
Cardiovascular/Pulmonary	162	55	64
Consult History and Phy.	137	55	42
Others	1623	554	530
Gastroenterology	118	39	44
General Medicine	88	25	33
Neurology	102	26	40
Obstetrics/ Gynecology	89	22	24
Surgery	39	10	10
Count Total	**2358**	**786**	**787**

Table 2: Dataset split for medical specialty model input.

4.2 Experiments on Medical Transcription Data

For the collected data set, we applied the nonconformist library to perform Inductive Conformal Prediction (ICP) with a Naïve Bayes model on a dataset of patient descriptions. Our goal was to predict the patient's disease based on their description while calculating a prediction interval that measures uncertainty associated with the predicted output.

We selected medical specialty as the target variable (y) for the medical transcription data set and used the remaining columns as features (x). To preprocess and analyze the data, we created five files, one for each feature column, and set the target variable (y) for each file as a medical specialty. Then, we processed and analyzed these files to investigate the features and target variables' relationship. This approach allows us to identify patterns or correlations between the patients' features and medical specialty.

4.2.1 Preprocessing

First, we plotted a pie chart to visualize the frequency distribution of medical specialties in the dataset. Next, we removed rows containing missing values in the keywords column, as these samples would not provide helpful information for our analysis. Then, we used the "LabelEncoder" function to convert the values in the medical specialty column to integers as shown in Table Table 3, allowing us to use this column as a feature in our analysis. The LabelEncoder assigns a unique integer code to each unique label in the input data. So, if a medical record uses the Encoded Label and the value assigned to a particular record is 4, the record is related to the General Medicine specialty. Similarly, a value of 3 would indicate a record related to Gastroenterology, and so on. After that, we replaced values in the medical specialty column that were greater than or equal to 8 with 8, representing "others". After cleaning and reducing the number of categories in the medical specialty column, we plotted a bar chart to visualize the frequency distribution of medical specialties in the cleaned dataset.

We defined a function that performed the following steps to preprocess the keywords column. We first removed punctuation and digits - any non-alphabetic characters from the keywords, such as numbers, symbols, and punctuation marks. Next, we converted all of the keywords to lowercase to ensure consistency and to prevent duplication of keywords that only differ in case. After that, we removed stop-words unlikely to be useful for analysis, such as "the", "and", and "a" to reduce noise in the data. Lastly, we used Stemming. This allows us to group related words and reduce the number of unique words in the dataset. We used the Porter Stemmer algorithm to perform stemming on the keywords. We then applied the text cleaning and preprocessing function to the keywords column and stored the cleaned keywords in a new column called cleaned keywords. Finally, we saved the cleaned dataset with the added cleaned keywords column to a new CSV file for a more straightforward implementation.

Label	Encoded Label
Cardiovascular/Pulmonary	0
Consult History and Phy.	1
Others	2
Gastroenterology	3
General Medicine	4
Neurology	5
Obstetrics/ Gynecology	6
Surgery	7

Table 3: Encoded labels.

5 Results Analysis

5.1 Baseline Model

We can choose any classification algorithm as a baseline model because the framework we will compare in Section 5.2 is model agnostics. Here, we have used Multinomial Naïve Bayes as a classifier for classifying the various medical_specialities mentioned in Table 3 and corresponding confusion matrix as a performance metrics is shown in Table 4.

Multinomial Naive Bayes								
	0	1	2	3	4	5	6	7
0	57	0	1	1	0	0	1	3
1	0	28	2	0	0	0	0	2
2	32	7	404	22	18	17	30	15
3	0	0	0	34	0	0	2	0
4	0	0	2	0	32	0	0	1
5	1	1	5	0	0	29	2	0
6	0	0	0	1	0	0	21	0
7	0	1	3	0	0	0	0	12

Table 4: Confusion matrix of the base model.

5.2 Conformal Inference

Table 5 is the confusion matrix for conformal inference. One observation that can be seen here is that the number of true positives in the confusion matrix for the conformal inference, as shown in Table 5, is lower than the number of true positives in the confusion matrix for the Multinomial Naïve Bayes model as shown in Table 4. Conformal inference is a method for estimating the reliability of predictions made by a model, and it may result in less confident predictions (based on the significance level 1-alpha) compared to the Multinomial Naïve Bayes model. As a result, the model may make fewer optimistic predictions, leading to fewer true positives in the confusion matrix.

Conformal Inference								
	0	1	2	3	4	5	6	7
0	8	2	34	5	4	1	4	5
1	3	2	20	0	1	5	0	1
2	68	34	290	39	27	33	17	37
3	4	4	17	3	1	3	1	3
4	2	5	19	4	1	1	1	2
5	5	0	23	3	0	3	0	4
6	2	2	10	2	0	0	3	3
7	2	0	11	1	1	0	0	1

Table 5: Confusion matrix of conformal inference.

This part shares the results in Table 6. Each row represents seven test instances. The values in columns named p_0, p_1, p_2, p_3, p_4, p_5, p_6, p_7 represent the p-value columns of Cardiovascular/Pulmonary, Consult-History and Phy., Others, Gastroenterology, General Medicine, Neurology, Obstetrics/Gynecology, and Surgery. Algorithm **??** outlines the process for implementing p-values. The p-value is a metric for measuring the confidence of an ML model's predictions. It represents the model's accuracy when making predictions for new data. The p-value 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. Therefore, caution must be exercised when interpreting model predictions with low p-values.

5.3 Performance Metrics

Precision and recall are helpful measures for evaluating the accuracy of a classifier when the classes are well-defined and there is no uncertainty about the labels. However, in CP, there is always some uncertainty about the labels, which needs to be quantified as a prediction interval.

The significance level determines the frequency at which the ML model produces inaccurate predictions. When the significance level is set to 0.05, we expect the model to make errors 5%

From Table 6, we can infer that as conformal predictors ensure validity, the main factor affect-

sig	mean err	avg c	n correct
0.01	0.013977	6.984752	776
0.05	0.053367	6.129606	746
0.1	0.100381	3.97967	708
0.2	0.194409	1.03939	634
0.3	0.297332	0.867853	557
0.4	0.376112	0.757306	491
0.5	0.506989	0.604828	388
0.6	0.583227	0.505718	328
0.7	0.700127	0.371029	236
0.8	0.80432	0.251588	156
0.9	0.894536	0.115629	83

Table 6: Performance metrics of conformal inference.

ing their performance is efficiency, which refers to the size of the label sets. Smaller sets are considered more informative. The performance of the conformal predictor can be evaluated by measuring $AvgC$ as it is the measure that represents the average number of class labels present in the prediction sets. This directly indicates how well the conformal predictor can reject inappropriate class labels.

5.4 Risk Aware Ranking

The p-value of an ML model indicates the probability of obtaining a similar outcome under the NULL hypothesis, which determines the confidence level in its prediction. A higher level of confidence indicates greater accuracy.

This metric is defined as:

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

Credibility in models refers to the degree to which we can trust the predictions made by a model. A credible model is one that accurately reflects the underlying data-generating process and produces predictions that are reliable and accurate.

This metric is defined as:

$$\text{Credibility}(x) = \max_{i \in \{0,1,\ldots,7\}} p_i$$

Table 7 shows the confidence and credibility score of the predicted labels. For example, in the first test instance, the confidence score is, the credibility score is, and the predicted label is 5. Here 5 represents the $medical\_specialty$ - '$Neurology$'.

In CP, a NULL set refers to a situation where the algorithm cannot confidently assign any label to a new test instance based on the available training data. This can occur when the new instance differs from any instances seen during training or when there is insufficient information to make a reliable prediction.

One way to obtain a NULL set is to set the significance level too high, which can make the algorithm overly conservative and less likely to make a prediction. For example, In the 8-label multi-classification problem, the conformal prediction algorithm is set with a significance level 0.05. Suppose a new test instance differs from any instances seen during training or has insufficient information. In that case, the algorithm may return a NULL set, indicating that it cannot make a confident prediction for that instance.

	Confidence	Credibility	y_pred
1	0.962	0.831	5
2	0.996	0.948	3
3	0.897	0.537	2
4	0.914	0.672	4
5	0.894	0.496	3
6	0.863	0.358	2
7	0.999	0.997	0

Table 7: Adoption of confidence for risk-aware ranking.

	train	cal	test
Cardiovascular/Pulmonary	162	55	64
Consult History and Phy.	137	55	42
Others	1623	554	530
Gastroenterology	118	39	44
General Medicine	88	25	33
Neurology	102	26	40
Obstetrics/ Gynecology	89	22	24
Surgery	39	10	10
Count Total	**2358**	**786**	**787**

Table 8: Dataset split for model input.

6 Conclusions

This paper introduced an algorithm-agnostic framework that quantifies uncertainty associated with new, unseen data points in the medical domain. The proposed approach is evaluated on the medical transcription dataset. We also showed how the risk-aware ranking of the Labels could help prioritize the treatment in a large-scale setting.

Acknowledgments

The authors thank the CSULB College of Engineering and the CSULB Department of Computer

Engineering and Computer Science for their support.

References

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

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. 2021. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118.

Ronald Cheung, Jacob Chun, Tom Sheidow, Michael Motolko, and Monali S Malvankar-Mehta. 2022. Diagnostic accuracy of current machine learning classifiers for age-related macular degeneration: a systematic review and meta-analysis. *Eye*, 36(5):994–1004.

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

Steven A Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A Riegler, Pål Halvorsen, and Sravanthi Parasa. 2022. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1):5979.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Asher Lederman, Reeva Lederman, and Karin Verspoor. 2022. 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.

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

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

Apostolos F Psaros, Xuhui Meng, Zongren Zou, Ling Guo, and George Em Karniadakis. 2023. Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *Journal of Computational Physics*, page 111902.