Investigating the Use of Statistical Tests in Natural Language Processing Machine Learning Models

Jule Valendo Halim -1425567
University of Melbourne

Abstract

Objective

To investigate possible roles of statistical tests on natural language processing tasks through models for multi-label and multi-class classifiers.

Materials

Dataset containing online patient reviews of prescribed medication, provided by Zoolnori et al[1 2]. The dataset contains annotated labels for each review sentence, along with the medication being reviewed.

Methods

Various multi-label and multi-class classification tasks were created. Afterwards, each model was evaluated using traditional machine learning metrics and statistical tests.

Results and Discussion

When comparing traditional machine learning metrics, transformer models generally outperform their logistic regression, except for one task-feature pair. Statistical test results were reported and interpretations were provided.

Conclusion

Statistical tests provide valuable insight into model behaviour outside of traditional machine learning metrics. However, additional work is needed in creating a pipeline to investigate causes of possible model behaviour, as well as integration of assumption testing.

Introduction

Statistical tests are a popular and long-standing method in multiple fields of science. However, studies in natural language processing (NLP) and machine learning (ML) do not generally include statistcal testing as part of their model evaluation. A survey on 233 published papers in the field of NLP showed that 132 of these papers did not report statistical significance[3]. However, more studies have begun advocating for the use of these tests to show that experimental results are not coincidental[3] and argue that a combination of NLP and statistical tests can provide a framework for the development of robust, high-throughput health NLP systems[4].

In this report, I aim to investigate the use of statistical tests on ML models that predict multi-label and multi-class classification tasks using the statistical test workflow suggested by Rainio, Tauho, and Klén[5] and compare different feature inputs.

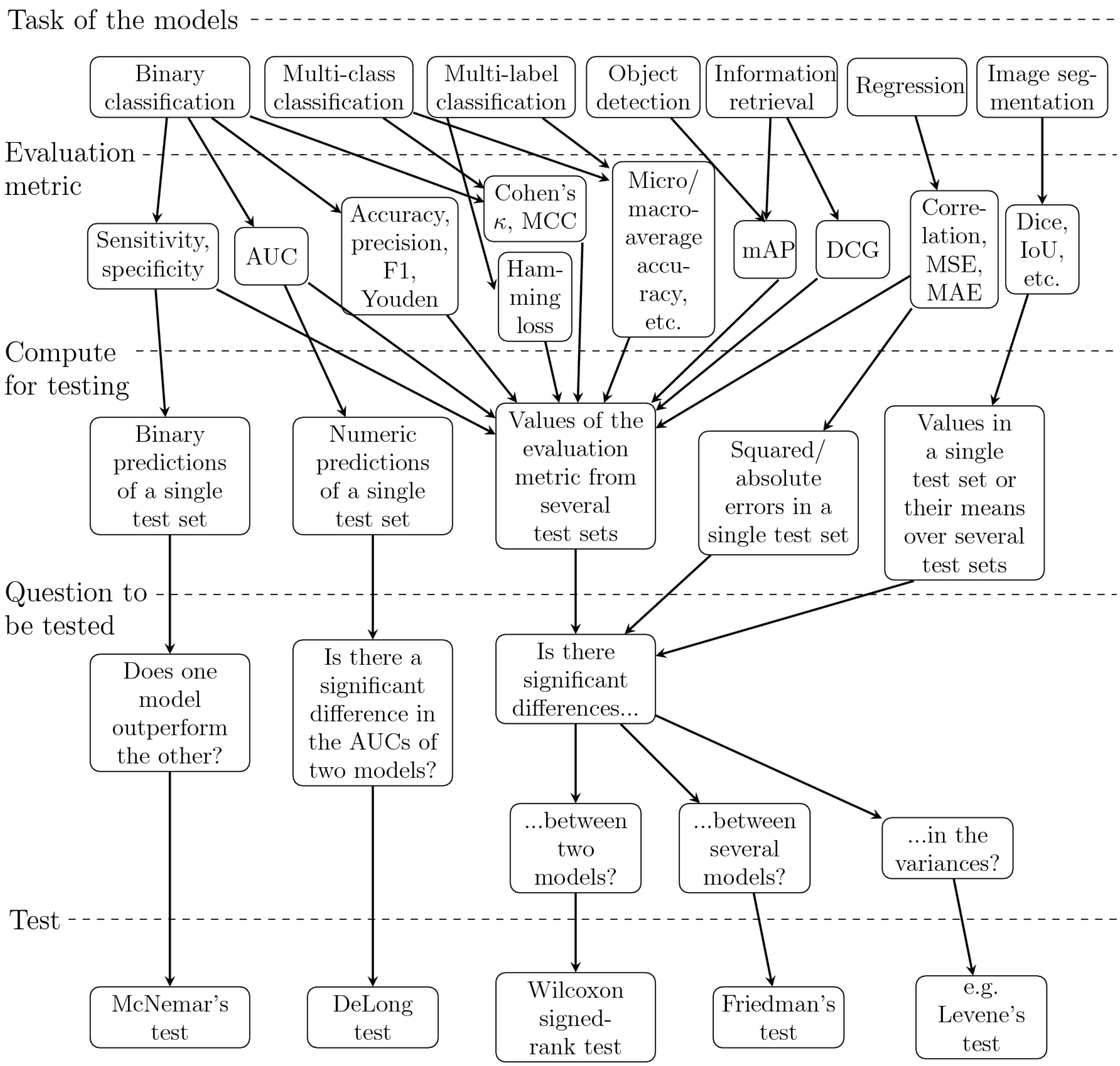

Figure 1. Statistical Workflow Provided by Rainio, Tauho, and Klén[5]

The Sentence_Labeling sheet of the PsyTAR dataset used in this study is provided by Zoolnoori et al[1 2]. This sheet contains three sections of interest. The first is sentences from online patient reviews of certain drugs. The second is annotations of certain labels, described below in table 1. The third is drug labels, which indicates which drug each sentence is reviewing.

Two classification tasks were performed using a baseline logistic regression model and a transformer model. The multi label class predicts six binary annotations, while the multi-class predicts four different drugs. For each model, different features will be used as inputs. The resulting predictions will be used for statistical tests in order to determine their statistical signifance.

Methods

Preprocessing

Two main preprocessing steps were done. Firstly, as the drug labels in the data were concatenated with review ID, the review ID was stripped and only the drug label was added to a new column. Secondly, the review text was preprocessed through tokenization, as well as stopword and punctuation removal. The dataset consists of 6009 records. The data was split into train, validation, and test sets using a 45:45:10 ratio. Additional feature processing specific to each task is described in their respective sections.

Multi-Label Classification

Multi-label classification involves predicting six labels, as shown in table 1. Each label is a binary task (1 for present and 0 for absent), which has been manually annotated. This task aims to predict annotations for a given review text.

Predicted Class	**Description**
Adverse Drug Reactions (ADR)	-
Withdrawal Symptom (WD)	-
Effective (EF)	-
Ineffective (INF)	-
Sign/Symptom/Illness (SSI)	Text contains explicit SSI as a result of the drug
Drug Indication (DI)	Text contains SSI that is currently being addressed by the drug

Table 1. The six labels predicted by multi-label classifiers

Selected Features Two features were used for multi-label classification. The first is the preprocessed review texts without any additional features. From here on, references to this task-feature pair will be referred as task 1 feature 1. The second contains the drug name added to the start of the preprocessed review text. This will be referred to as task 1 feature 2.

Feature 1	**Feature 2**
extreme weight gain short-term memory	lexapro extreme weight gain short-term memory

Table 2. Sample Inputs for Each Feature in Multi-Label Classification

Logistic Regression The logistic regression model performed Term Frequency Inverse Document Frequency (TF-IDF) vectorization on the input text. A multi-output classifier was built upon a logistic regression model, which was then tuned using hyperparameter tuning.

Transformer The transformer model uses a pre-trained BERT model, which was trained on a downstream task in order to create a multi-label transformer classifier. Binary cross-entropy loss (BCE) with logits (a sigmoid layer) loss was used for loss calculation. BCE is well-suited for binary classification[6].

Parameter	Value
hidden_size	768
num_hidden_layers	24
num_attention_heads	12
intermediate_size	3072
num_labels	6
optimizer	AdamW
loss_calculation	BCE With Logits Loss
epochs	15

Table 3. BERT Configuration for Multi-Label Classification

Statistical Tests Multi-label classification will be tested using macro averaged precision, recall, and F1-scores. The choice to use macro instead of micro was to give equal weights to each class, as some labels were more prevalent than others. Accuracy will also be reported.

The Hamming Loss (HL) of the predictions are also calculated. HL measures the fraction of labels that are incorrectly predicted, on average, across all samples and is used to evaluate multi-label classification tasks [7]. HL ranges from 0 to 1, where 1 means all predictions are erroneous, while 0 means perfect predictions.Appendix A contains statistical test equations.

Multi-Class Classification

Multiclass classification aims to take in varying inputs and predict the drug that is being reviewed. There are four classes of drugs to predict; lexapro, cymbalta, effexorxr, and zoloft.

Selected Features Three features were selected for multi-class classification. The first is the preprocessed sentence without any additional features, referred to as task 2 feature 1.

The second is the preprocessed text along with its annotations. For the logistic regression model, the binary annotations are simply added onto the end of the sentences as 1s and 0s.

The transformer's input has the review text concatenated with the predicted class names. If the class is labelled 1, a [POS] token was placed in front of it. If the class is labelled 0, a [NEG] token was placed instead. These tokens identify positive(1) and negative(0) labels respectively. This feature will be referred to as task 2 feature 2.

The third feature is to only use the annotation inputs. This will be referred to as task 2 feature 3. The logistic regression model takes in only the binary annotations, while the transformer only takes in the predicted class name along with the described tokens.

Feature 1 Transformer	**Feature 2 Transformer**	**Feature 3 Transformer**
extreme weight gain short-term memory	[POS] adverse drug reaction [NEG] withdrawal symptoms...extreme weight gain short-term memory	[POS] adverse drug reaction [NEG] withdrawal symptoms...[NEG] drug indication
Feature 1 Logistic Regression	**Feature 2 Logistic Regression**	**Feature 3 Logistic Regression**
extreme weight gain short-term memory	extreme weight gain short-term memory 1 0 1 1 0 0	1 0 1 1 0 0

Table 4. Sample Inputs for Each Feature in Multi-Class Classification

Logistic Regression TF-IDF vectorization was used on the input text. However, in contrast to using a multi-label classifier built on top of a logistic regression model, multi-class classification uses logistic regression directly.

Transformer The transformer model is identical to the multi-label transformer, except it was trained on a multi-class downstream task. The number of hidden layers was also decreased due to long training times. The loss calculation was also changed to cross-entropy as it is a better fit for classification tasks[8].

Parameter	**Value**
hidden_size	768
num_hidden_layers	8
num_attention_heads	12
intermediate_size	3072
num_labels	1
optimizer	AdamW
loss_calculation	Cross-Entropy Loss
epochs	20

Table 5. BERT Configuration for Multi-Class Classification

Statistical Tests Multi-class classification will be tested using macro averaged precision, recall, and F1-scores as described previously. The use of macro averaging was due to class imbalance (zoloft had a test count of 565 while cymbalta had a test count of 791). Accuracy will also be reported.

Additionally, following figure 1, Cohen's Kappa (Cohen's K) and Matthews Correlation Coefficient (MCC) will be used. Cohen's K is used to calculate inter-model agreement on predictions. It returns a value between 1 and -1, where 1 means a perfect agreement, 0 means no agreement above chance, and -1 indicating less agreement than random chance.

Absolute Cohen's Kappa Range	**Interpretation**
$\|\kappa\| \leq 0$	No agreement
$0.01 \leq \|\kappa\| \leq 0.20$	None to slight agreement
$0.21 \leq \|\kappa\| \leq 0.40$	Fair agreement
$0.41 \leq \|\kappa\| \leq 0.60$	Moderate agreement
$0.61 \leq \|\kappa\| \leq 0.80$	Substantial agreement
$0.81 \leq \|\kappa\| \leq 1.00$	Almost perfect agreement

Table 6. Cohen's Kappa Cutoff Points Based on McHugh's Research[9]

MCC has been adapted for multi-class classification by taking into account all the true and false positive as well as true and false negatives for each class[10]. This adaptation of the MCC can indicate whether model performance across all classes.

MCC returns a value betwen 1 and -1, where 1 means perfect predictions, 0 means a prediction that is no better than random, and -1 means total disagreement between predictions and ground truth. MCC follows cutoff point selection of graphs such as an area under a recciever operating characteristic (AUROC) curve[11 12]. However, this report will follow arbitrary cutoff points.

Absolute MCC Values	**Interpretation**
$0 \leq \|MCC\| \leq 0.1$	Very Poor Predictions
$0.1 < \|MCC\| < 0.3$	Poor Predictions
$0.3 \leq \|MCC\| < 0.5$	Moderate Predictions
$\|MCC\| \geq 0.7$	Good Predictions

Table 7. MCC Cutoff Points

Friedman's Test of Significance

In order to investigate whether the predictions are significantly different from each other, Friedman's test will be attempted to be performed on both tasks. Friedman's test is a hypothesis testing method, where the null hypothesis is that there is no significant difference between two samples of predictions[13]. Meanwhile, the alternative hypothesis is that a significant difference does exist. A cutoff point (α value) of 0.05 will be used, meaning that if the p-value is <0.05, the null hypothesis will be rejected.

This test was found to not be suitable for the multi-class classification task due to the output. Additional discussion on the impact of this will be discussed in the discussion and conclusion section. For multi-label classification, each label had the Friedman's test performed.

- **Null Hypothesis (H_0)**: There is no significant difference between the models' predictions for the label.
- **Alternative Hypothesis (H_1)**: There is a significant difference between the models' predictions for the label.
- $\alpha = 0.05$

Ethics Statement

One ethical consideration for this study is that interpretations of statistical test results should only be seen as possible recommendations and indicators of possible model behaviour. As such, additional testing and observations need to be done to confirm these interpretations.

Secondly, this study proposes some potential issues regarding the statistical workflow by Rainio, Tauho, and Klén[5]. However, this should not be seen as a direct criticism of their study. Rather, it serves to highlight possible difficulties in the interpretations of how to apply statistical tests to natural language processing models.

Finally, this study proposes changes to the way machine learning models are evaluated by the wider scientific community. However, these are proposed changes require significantly more robust investigation into their advantages before being taken as the norm for evaluating machine learning tasks. Instead, this study serves as an investigation into current methods and pipelines suggested by other studies in the field of combining statistical tests with traditional machine learning metrics.

Results and Analysis

Multi-Label Statistical Tests

Accuracy, Precision, Recall, and F1-Scores Table 8 shows the accuracy along with the macro precision, recall, and F1 scores. Transformer models also show training and validation loss.

Transformers show much higher performances compared to logistic regression for both features. Change in features does not appear to significantly affect model performance for all models, with the highest change being accuracy in logistic regression models (1.66% difference).

Hamming Loss The results of calculating HL for multi-label classification are shown in table 9. Each model was compared to the ground truth.

The transformer models have significantly lower HL than the logistic regression models. This is as expected, as a higher accuracy indicates a higher chance of correct predictions. However, HL can show a more detailed view on model performance. For example, feature 1 using logistic regression has an accuracy of 50.83%. This accuracy metric is strict as it only considers an instance as correct if all labels were correctly predicted. However, the HL for this model suggests that only about 11.16% of the predictions were incorrect, indicating that the model performs better when considering individual label predictions.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Training Loss	Validation Loss
Task 1 Feature 1 Transformer	0.98	0.95	0.91	0.93	0.0134	0.0789
Task 1 Feature 1 Logistic Regression	0.5083	0.64	0.38	0.47	N/A	N/A
Task 1 Feature 2 Transformer	0.98	0.92	0.93	0.92	0.0115	0.0772
Task 1 Feature 2 Logistic Regression	0.5249	0.65	0.38	0.47	N/A	N/A

Table 8. Accuracy and Macro-Averaged Precision, Recall, F1-Scores, Training Loss, and Validation Loss for Task 1

Model	Hamming Loss Against Ground Truth
Task 1 Feature 1 Transformer	0.016057586
Task 1 Feature 1 Logistic Regression	0.11517165
Task 1 Feature 2 Transformer	0.019379845
Task 1 Feature 2 Logistic Regression	0.109634551

Table 9. Hamming Loss for Multi-Label Task (Task 1)

Friedman's Test of Significance Friedman's test was done on all the models simultaneously. It returns two values, a test statistic and a p-value. Results are shown in table 10.

Label	Statistic	p-value
ADR	16.373	0.003
WD	10.667	0.031
EF	64.681	3.00×10^{-13}
INF	12.653	0.013
SSI	34.712	5.32×10^{-7}
DI	27.855	1.33×10^{-5}

Table 10. Friedman's Test of Significance for Multi-Label Task (Task 1)

The results of the Friedman's test statistic shows the difference in predictions for all models. These differences are statistically significant, as each of the p-values are less than the determined cutoff point (0.05). As such, we reject the null hypothesis. These results provide evidence in favor of the alternative hypothesis that each model's predictions are significantly different.

Multi-Class Statistical Tests

Accuracy, Precision, Recall, and F1-Scores Table 11 shows the accuracy along with the macro precision, recall, and F1 scores. Transformer models also show training and validation loss.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Training Loss	Validation Loss
Task 2 Feature 1 Transformer	0.84	0.84	0.84	0.84	0.044	0.286
Task 2 Feature 1 Logistic Regression	0.3688	0.78	0.35	0.29	N/A	N/A
Task 2 Feature 2 Transformer	0.834	0.84	0.84	0.83	0.031	0.289
Task 2 Feature 2 Logistic Regression	0.3688	0.78	0.35	0.29	N/A	N/A
Task 2 Feature 3 Transformer	0.322	0.33	0.31	0.26	0.696	0.696
Task 2 Feature 3 Logistic Regression	0.3422	0.28	0.31	0.24	N/A	N/A

Table 11. Accuracy and Macro-Averaged Precision, Recall, F1-Scores, Training Loss, and Validation Loss for Task 2

The resulting predictions of the model for multi-class classification generally show a high accuracy for transformers, compared to their logistic regression baselines. However, feature 3 showed a different trend, where the logistic regression baseline performed better. There was also a large change in accuracy compared to other transformer models (with feature 3 having roughly 50% lower accuracy than other transformer models). This could be due to how transformers require rich language information, which feature 3 does not provide, as the inputs are only a series of token-word pairs.

Cohen's Kappa A heatmap of Cohen's K for each model tested against each other is shown in figure 2.

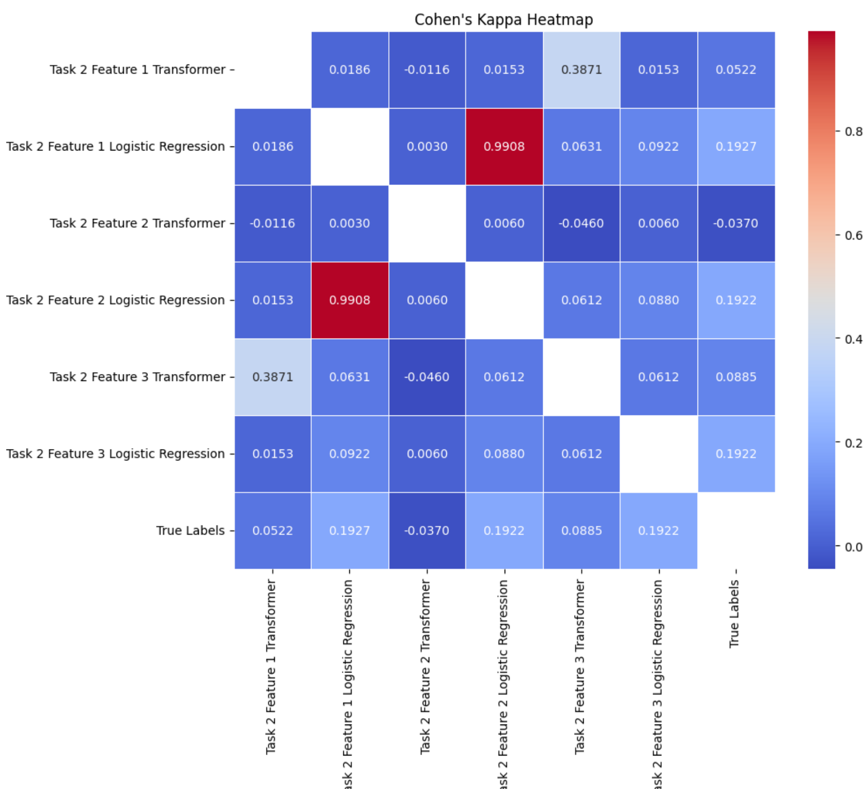

Figure 2. Heatmap of Cohen's Kappa Results on Multi-Class Classification Task

The results of Cohen's K shows that for most models, the agreement between predictions are low. This is true even for the models with high accuracy. For example, the transformer models for features 1 and 2 have similar accuracies (84% and 83.4% respectively). However, they have a Cohen's K value of -0.0116. This indicates that on the wrong predictions, these models do not predict the same incorrect class even if they have high accuracy on correct predictions.

Two Cohen's K values are of note. The first is between the logistic regression models of feature 1 and 2, which has a Cohen's K of 0.99, indicating an almost perfect positive agreement. This could be due to how the inputs for feature 1 and 2 are similar, and logistic regression models predict similar patterns for each of these. As such, the models predict similar classes, even for incorrect predictions.

The second is between the transformer models of feature 1 and feature 3 with a Cohen's K of 0.39, indicating substantial positive agreement. Upon inspection, predictions of the transformer models for feature 3 only predict mainly one class. This could cause a higher than expected agreement with the transformer model for feature 1, depending on how the model for feature 1 predicts incorrect predictions.

Matthews Correlation Coefficient Each model was compared against the ground truth to calculate their MCC values. The resulting values are shown in table 12.

	Task 2 Feature 1 Transformer	**Task 2 Feature 1 Logistic Regression**	**Task 2 Feature 2 Transformer**	**Task 2 Feature 2 Logistic Regression**	**Task 2 Feature 3 Transformer**	**Task 2 Feature 3 Logistic Regression**
MCC Score	0.066509038	0.194445905	-0.043092423	0.19418937	0.103772985	0.124725893

Table 12. MCC Scores for Each Model, Compared Against Ground Truth

The results of the MCC indicate that most models are not able to correctly predict across all classes evenly. For example, the transformer models for features 1 and 2 have MCC scores that indicate very poor predictions when considering performance across all classes. This could suggest that the model performs significantly worse when considering a class-to-class basis.

However, for some models that do not have high accuracy, they appear to have higher MCC scores. For example, the transformer for feature 3 has an MCC score of 0.10. This particular transformer model predicted only the majority class, allowing it to minimize false positives and false negatives. As such, although they do not have a good accuracy, they are able to minimize misclassifications, leading to a decent MCC score.

Discussion

Multi-Label Classification Task

Interpretation of the statistical tests on the multi-label task can be considered in three stages. First, the traditional machine learning metrics show that transformer models are able to outperform their logistic regression baselines consistently. Secondly, the HL values provide further support that transformers are able to outperform logistic regression models, even when considering individual label predictions. However, the HL shows that logistic regression models perform better on a label-by-label basis than suggested by their accuracy.Finally, Friedman's test shows that the difference in model predictions are statistically significant.

Multi-Class Classification Task

Investigation on Cohen's K supports how logistic regression models learn similar patterns given similar inputs, while transformers might learn different weights with similar inputs. Furthermore, Cohen's K can also provide insights into what incorrect predictions are done. While this report does not investigate this behaviour in depth, the results of Cohen's K provides an indicator of possible model behaviour that could be investigated further.

Finally, the results of the MCC scores show that transformer models do not tend to each class equally, which could support the behaviour of the transformer for feature 1 to default to a majority class. While investigating the exact causes of these MCC scores are out of the scope of this report, MCC scores can provide indications on what the model struggles with, such as evenly predicting every class.

Friedman's Test on Multi-Class Classification

The application of Friedman's test on multi-class classification poses a challenge due to the output, which comes in the form of text. Friedman's test calculates differences between a series of predictions. However, text data such as drug names need to be preprocessed into numerical variables. While some studies have used Friedman's test on textual data by converting it into a ranking task [14], there is no consensus on how to handle this change in features. Future work could be done to tackle this issue.

Conclusion and Future Directions

In conclusion, this report has shown how statistical tests, combined with traditional machine learning metrics can help to better understand model behaviour. While statistical tests indicate model behaviour, traditional machine learning models are still important when considering practicality. For example, for multi-class classification, while the transformer models for feature 1 and 2 do not perform evenly on all classes, their accuracy would still make a transformer model preferred in performing predictions. In addition, while some studies have suggested how MCC has advantages over F1-scores and accuracy [11], these metrics should still be understood as providing ideas on model behaviour, instead of replacing traditional machine learning metrics.

Future work can investigate more sophisticated workflows to understand model behaviour based on statistical tests. Additional statistical tests could also be integrated, such as testing assumptions of independence, which could guide what statistical tests fit a specific task and to identify possible problems with a provided dataset [15].

References

1. Zolnoori M, Fung K, Patrick T, Fontelo P, Kharrazi H, Faiola A, et al. The PsyTAR Dataset: From Patients Generated Narratives to a Corpus of Adverse Drug Events and Effectiveness of Psychiatric Medications. Data in Brief. 2019 03;24:103838.

2. Zolnoori M, Fung KW, Patrick TB, Fontelo P, Kharrazi H, Faiola A, et al. A systematic approach for developing a corpus of patient reported adverse drug events: A case study for SSRI and SNRI medications. Journal of Biomedical Informatics. 2019;90:103091. Available from: https://www.sciencedirect.com/science/article/pii/S1532046419300012.

3. Dror R, Baumer G, Shlomov S, Reichart R. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In: Gurevych I, Miyao Y, editors. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 1383-92. Available from: https://aclanthology.org/P18-1128.

4. Wang X, Hripcsak G, Markatou M, Friedman C. Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. Journal of the American Medical Informatics Association. 2009 05;16(3):328-37. Available from: https://doi.org/10.1197/jamia.M3028.

5. Rainio O, Tauho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep. 2024;14(6086).

6. Xu Z, Liu R, Yang S, Chai Z, Yuan C. Learning Imbalanced Data with Vision Transformers; 2023.

7. Díez J, Luaces O, del Coz JJ, Bahamonde A. Optimizing different loss functions in multilabel classifications. Regular Paper. 2015;3:107-18. Available from: https://link.springer.com/article/10.1007/s13748-014-0060-7.

8. Hui L, Belkin M. Evaluation of Neural Architectures Trained with Square Loss vs Cross-Entropy in Classification Tasks; 2021.

9. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/#:~:text=Cohen%20suggested%20the%20Kappa%20result,1.00%20as%20almost%20perfect%20agreement.

10. Jurman G, Riccadonna S, Furlanello C. A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. PLOS ONE. 2012 08;7(8):1-8. Available from: https://doi.org/10.1371/journal.pone.0041882.

11. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining. 2023 02;16.

12. Yang S, Berdine G. The receiver operating characteristic (ROC) curve. The Southwest Respiratory and Critical Care Chronicles. 2017 May;5(19):34-6. Available from: https://pulmonarychronicles.com/index.php/pulmonarychronicles/article/view/391.

13. Hoffman JIE. Chapter 26 - Analysis of Variance II. More Complex Forms. In: Hoffman JIE, editor. Biostatistics for Medical and Biomedical Practitioners. Academic Press; 2015. p. 421-47. Available from: https://www.sciencedirect.com/science/article/pii/B9780128023877000263.

14. Wang J, Kan H, Meng F, Mu Q, Shi G, Xiao X. Fake Review Detection Based on Multiple Feature Fusion and Rolling Collaborative Training. IEEE Access. 2020 01;8:182625-39.

15. Flatt C, Jacobs R. Principle Assumptions of Regression Analysis: Testing, Techniques, and Statistical Reporting of Imperfect Data Sets. Advances in Developing Human Resources. 2019 11;21:484-502.

Appendix

Appendix A - Equations for Statistical Tests

Cohen's K:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where:

$$\begin{aligned} p_o &= \text{relative observed agreement among raters} \\ p_e &= \text{hypothetical probability of chance agreement} \end{aligned}$$

(p_o) is calculated as:

$$p_o = \frac{a + d}{a + b + c + d} \tag{2}$$

(p_e) is calculated as:

$$\begin{aligned} p_e = &\left(\frac{(a+b)}{a+b+c+d} \cdot \frac{(a+c)}{a+b+c+d} \right) \\ &+ \left(\frac{(c+d)}{a+b+c+d} \cdot \frac{(b+d)}{a+b+c+d} \right) \end{aligned} \tag{3}$$

where:

- a = Number of times both annotators agreed the sample was positive
- d = Number of times both annotators agreed the sample was negative
- b = Number of times Annotator A said positive but Annotator B said negative
- c = Number of times Annotator A said negative but Annotator B said positive

Matthews Correlation Coefficient:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

$$\begin{aligned} &\text{where:} \\ &TP : \text{True Positive} \\ &FP : \text{False Positive} \\ &TN : \text{True Negative} \\ &FN : \text{False Negative} \end{aligned}$$

Hamming Loss:

$$\text{HL} = \frac{1}{N \times L} \sum_{l=1}^{L} \sum_{i=1}^{N} (Y_{i,l} \neq X_{i,l})$$

$$\begin{aligned}
&\text{where:}\\
&\quad N : \text{Number of samples}\\
&\quad L : \text{Number of labels}\\
&\quad Y_{i,l} : \text{True label for } i\text{th sample and}\\
&\qquad l\text{th label}\\
&\quad haty_{ij} : \text{Predicted label for } i\text{th sample and}\\
&\qquad l\text{th label}\\
&\quad (Y_{i,l} \neq haty_{ij}) : 1 \text{ if } y_{ij} \neq \hat{y}_{ij} \text{ and 0 otherwise}
\end{aligned}$$

Accuracy, Precision, Recall, and F1-Scores:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{8}$$

$$\begin{aligned}
&\text{where:}\\
&\quad TP : \text{True Positive}\\
&\quad FP : \text{False Positive}\\
&\quad TN : \text{True Negative}\\
&\quad FN : \text{False Negative}
\end{aligned}$$

Friedman's Test:

$$T = \frac{12 \sum S_{ij}^2}{jk(j+1)} - 3k(j+1) \tag{9}$$

where:

$$\begin{aligned}
T &= \text{test statistic}\\
\sum S_{ij}^2 &= \text{sum of the squared sums of ranks for each prediction set}\\
j &= \text{number of prediction sets}\\
k &= \text{number of instances}
\end{aligned}$$

Appendix B - Transformer Model Performances

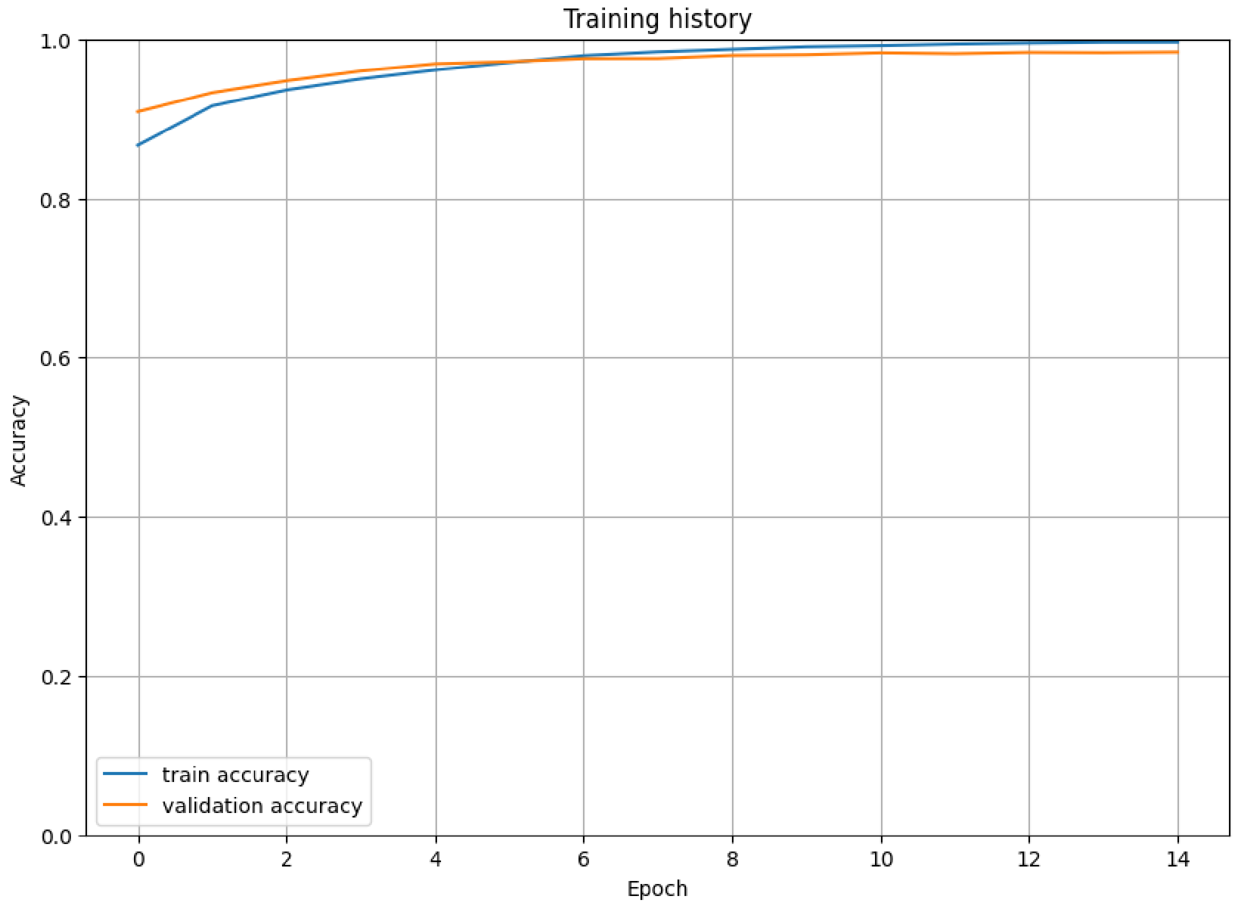

Figure 3. Task 1 Feature 1 Model Training (Total Training Time: 2056.51 Seconds)

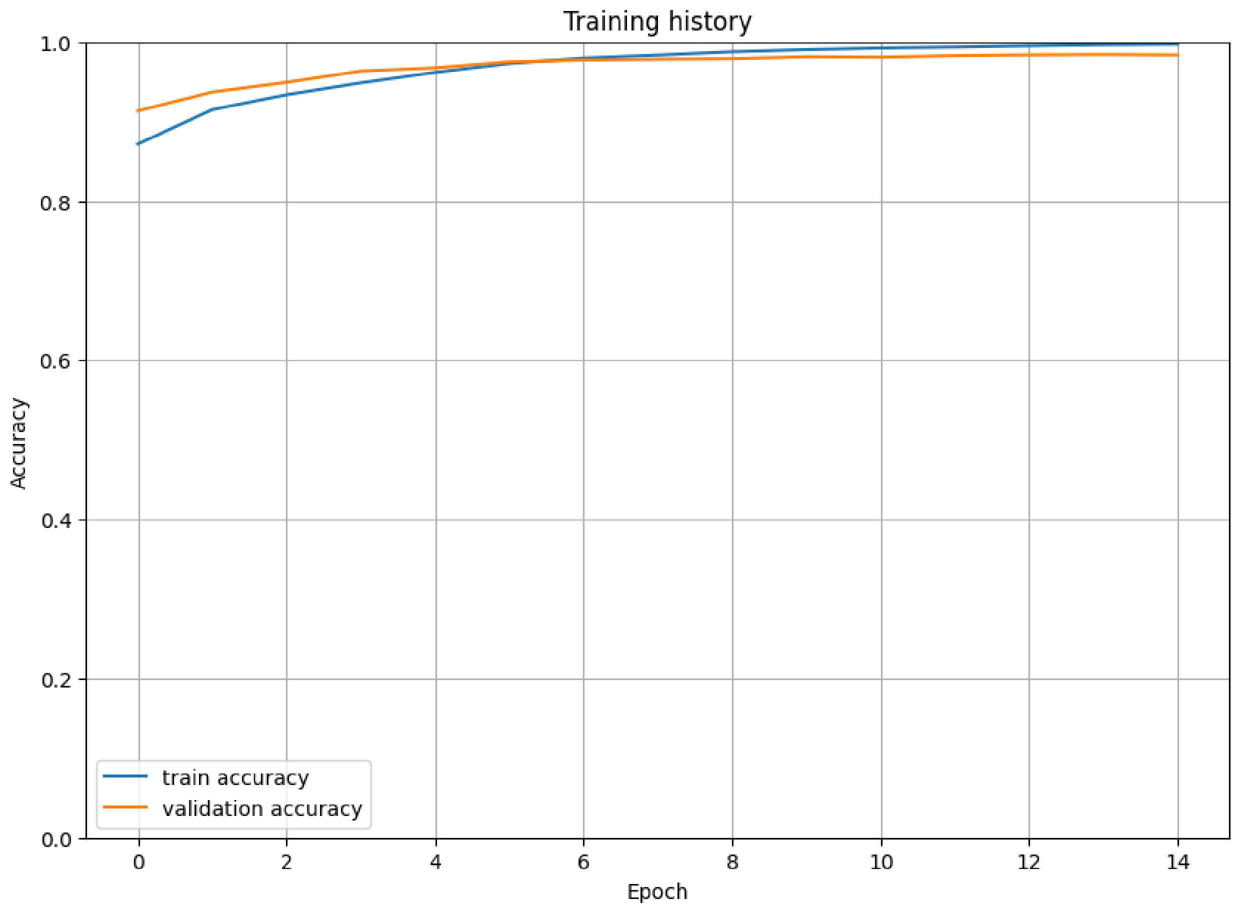

Figure 4. Task 1 Feature 2 Model Training (Total Training Time: 2037.76 Seconds)

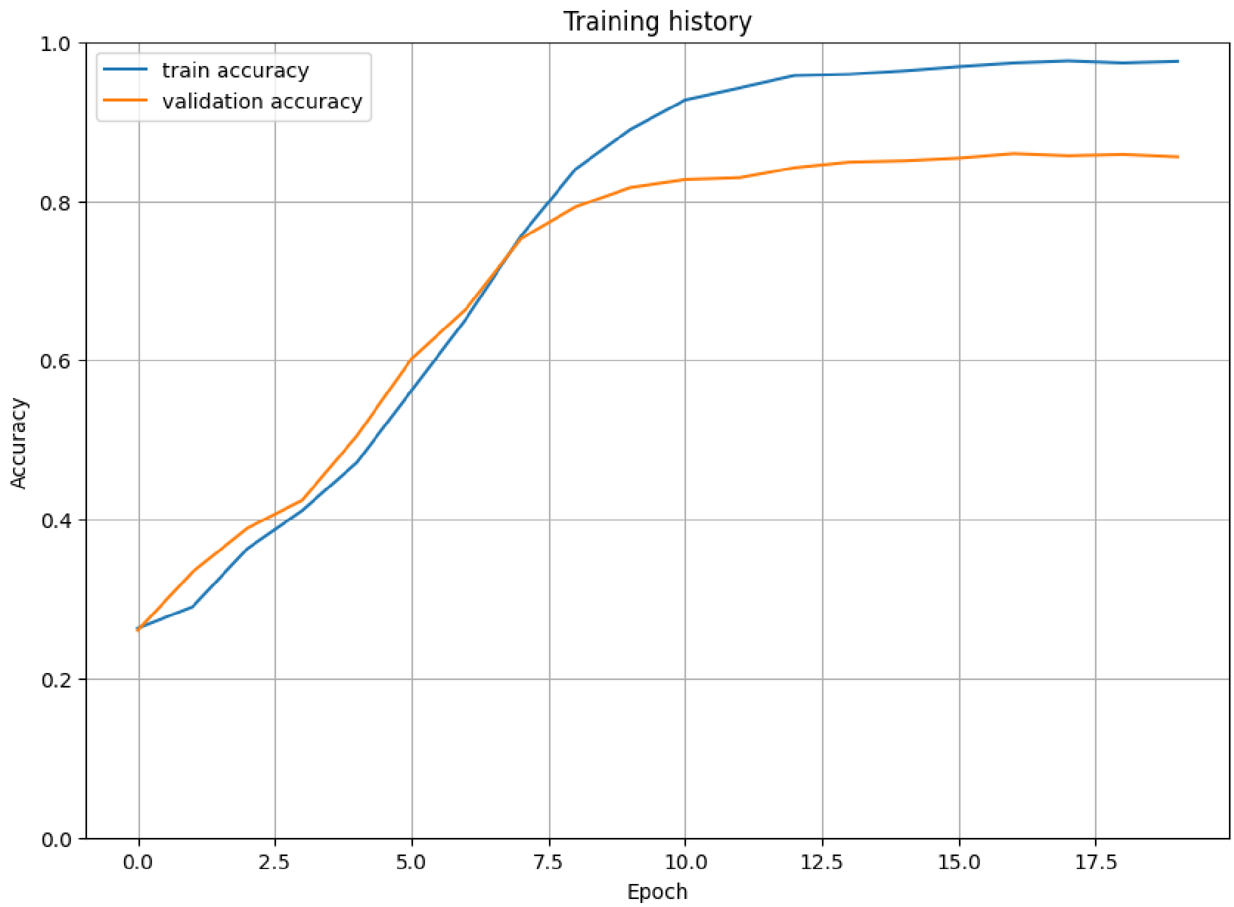

Figure 5. Task 2 Feature 1 Model Training (Total Training Time: 7471.89 Seconds)

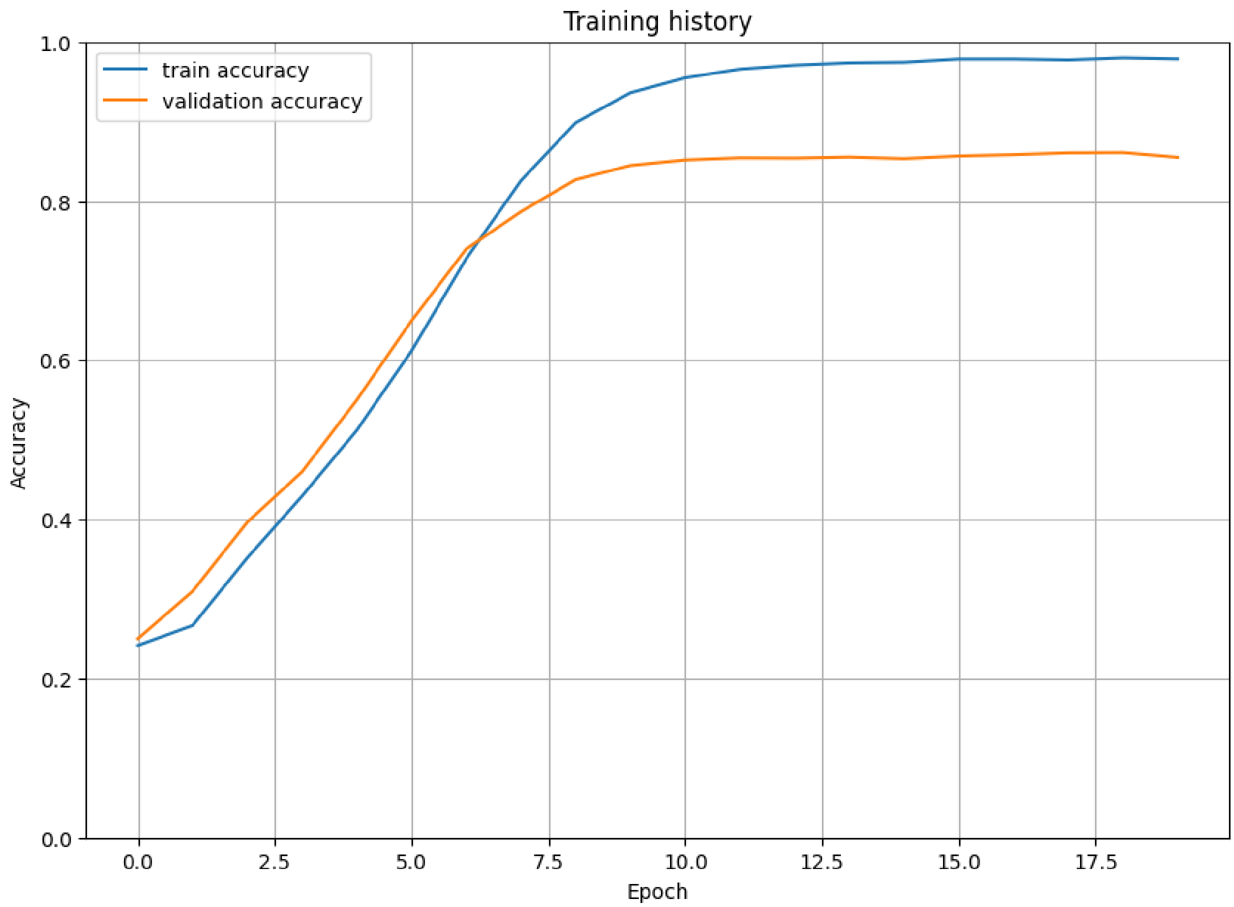

Figure 6. Task 2 Feature 2 Model Training (Total Training Time: 6587.32 Seconds)

Figure 7. Task 2 Feature 3 Model Training (Total Training Time: 6590.17 Seconds)

Appendix C - Source Code