LA-UR-25-28277

Approved for public release; distribution is unlimited.

Title: Can Pre-Transformer Models with Linguistic Insights Rival Basic Transformers?

Author(s): Wood, Margaret Kathryn
Parikh, Nidhi K.
Miller, Robyn Leigh

Intended for: Report

Issued: 2025-08-11

Los Alamos
NATIONAL LABORATORY

Can Pre-Transformer Models with Linguistic Insights Rival Basic Transformers?

Margaret Wood[1,2], Nidhi Parikh[3], Robyn Miller[3]

[1]Office of Classification and Controlled Information (ALDDP-OCCI), Los Alamos National Laboratory
[2]Center for Nonlinear Studies, Los Alamos National Laboratory
[3]Information Systems and Modeling (A-1), Los Alamos National Laboratory

Date: 04/17/2025

1. Introduction

The advent of large language models (LLMs) has significantly advanced the field of natural language processing (NLP). However, the research community has expressed concerns about some of their limitations; in particular, the substantial computational power required to run highly parameterized models and their lack of interpretability. At the same time, it remains unclear whether we have fully explored potential methodological enhancements for traditional machine learning (ML) models. In this work, we present a preliminary study with two overarching goals: (1) To test the hypothesis that incorporating grammatical features into traditional ML models can improve their performance, and (2) To compare these grammatically enhanced ML models to transformer-based models, such as BERT. We use these models for the detection of public health misinformation, defined as information that is false or inaccurate, but not intended to deceive or harm. This is distinct from disinformation, which is disseminated with malicious intent. We treat "fake news" as a type of misinformation, consistent with other literature on this subject (Ni et el., 2023; Liu et al., 2024).

One of the major limitations of LLMs, including transformer-based models, is their lack of interpretability. Without knowledge of the linguistic characteristics that inform model output, it is difficult to validate model predictions. This undermines trust, especially in contexts where accuracy and accountability are essential. It may obscure key insights about the target domain, or worse, lead to misguided actions or reinforced bias in high-stakes contexts. This opacity also hampers error diagnosis, and limits progress towards more refined, state-of-the-art research. Additionally, LLMs are computationally intensive, requiring substantial hardware resources, large datasets and significant energy consumption. This hinders the reproducibility of results and limits broader participation in LLM-based research.

Well-established principles of corpus linguistics and grammatical variation have historically been overlooked in the development of automated text classification models. Amongst these are: (a) The importance of preserving naturally occurring language, which is often compromised by the disadvantageous use of text pre-processing, and (b) The role of the functional relationship between situational characteristics, pervasive grammatical features, and the discourse function of those features at the text level (Biber and Egbert, 2023). Decades of grammatical variation research have demonstrated that different text types will display unique combinations of grammatical features based on their situational characteristics. For example, persuasive, expository, and narrative discourse will display unique grammatical patterns due in part to variation in their communicative functions. The pervasiveness of this finding in linguistic research, across all text types, suggests that grammatical patterns could be a key discriminative feature in distinguishing between misinformation and non-misinformation in mainstream U.S. news media.

Recent research on medical and public health misinformation has demonstrated that linguistic insights can play an important role in text classification tasks. For example, in analyzing the effects of text preprocessing on model performance, Siino et al. (2024: 12) found that support vector machine (SVM) models without any form of text pre-processing outperformed their counterparts that applied the full battery of pre-processing techniques (e.g., lower-casing, stop word removal, stemming, etc.). Notably, they reported that stemming led to the poorest performance in all models evaluated. Recent work has also substantiated the importance of considering the grammatical features in misinformation. In a study evaluating the influence of isolated feature types on model performance for text classification tasks, Di Sotto and Viviani (2022: 15) found that of all the isolated features considered in the study (e.g., medical terminology, sentiment), the model relying solely on grammatical features[1] for misinformation detection achieved the highest performance[2].

In this work, we incorporate grammatical features into traditional ML models to evaluate the effect they have on model performance for detecting public health misinformation. We then compare the performance of these grammatically enhanced models to commonly used transformer-based models.

2. Methods

2(a). Corpus Selection

We sought an open-source public health misinformation corpus that met the following criteria:

a) The corpus represents a narrowly defined source domain (e.g., news media articles, Reddit posts, blogs, etc.).
b) Access to the full texts is provided.
c) Annotations represent ground-truth labels of misinformation or non-misinformation, or their equivalents (e.g., fake or real news).
d) Annotations are assigned by subject matter experts using principled and consistent criteria throughout the annotation process.

There were two significant challenges identifying open-source public health misinformation corpora with respect to the criteria above. First, many of the open-source corpora that we identified did not assign class labels based on the content of the text, but on the source domain. For example, the description of a public health misinformation corpus released in 2023 reads: "*(...)* the truthful articles were obtained by crawling articles from Reuters.com *(...)* The fake news articles were collected from unreliable websites that were flagged by Politifact and Wikipedia *(...)*" (Hemina et al., 2023). In addition, many open-source corpora merged multiple source domains to form a single corpus (e.g., social media posts and news articles). For the purposes of this study, combining texts from distinct registers[3] into a single corpus makes it exceedingly difficult to identify distinctive grammatical features of misinformation, as it introduces a number of confounding variables to the model, such as the target audience and a variety of other register-specific conventions.

Despite these challenges, we were able to identify a single corpus that met our criteria. We selected the HealthStory sub-corpus from the greater FakeHealth corpus, which was compiled through a web-based project that ran from 2005 – 2018. The goal of this project was to critically analyze claims about health care interventions to improve the quality of health care information (Dai et al., 2020: 2)[4]. The project was funded by the Informed Medical Decisions Foundation, which is free from industry influence and does not advertise or accept funding from any entities with potential conflicts of interest.

The HealthStory sub-corpus represents public health misinformation broadly, covering a range of health interventions (e.g., procedures, treatments, products) for a variety of health ailments (e.g., cancer, surgery, heart attacks, mental health, pregnancy). The HealthStory sub-corpus was principally compiled in terms of both source domain and public health topics. The source domain is narrow, representing news stories published by a set of mainstream U.S. news media outlets (e.g., *Associated Press, NPR, FoxNews*).

[1] The authors referred to these as "linguistic-stylistic" features, rather than "grammatical" features.
[2] This outcome was specific to the HealthRelease sub-corpus.
[3] In this study, *register* is defined as a culturally recognized text variety with overt external indicators (Biber, 2019: 44).
[4] This paper can be accessed through the archive link: https://arxiv.org/pdf/2002.00837. The full FakeHealth corpus is available for download via Dai's GitHub: https://github.com/EnyanDai/FakeHealth/tree/master/dataset

The “fake” and “real” classes are represented comparably across the news media outlets, as indicated by the average misinformation rating of the texts from each news source (Figure 1). To obtain this average, texts were rated on a scale ranging from 0 – 5 (0 = fake; 5 = real) based on a set of select criteria. As shown in Figure 2, texts in both classes cover a range of public health ailments and interventions.

Figure 1
Average rating for each news source, 95% confidence interval error bars (Dai et al., 2020).

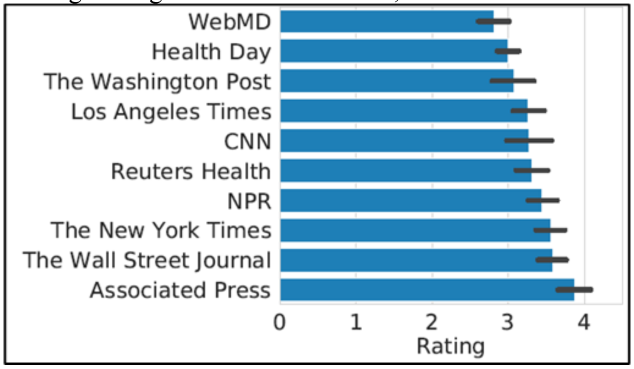
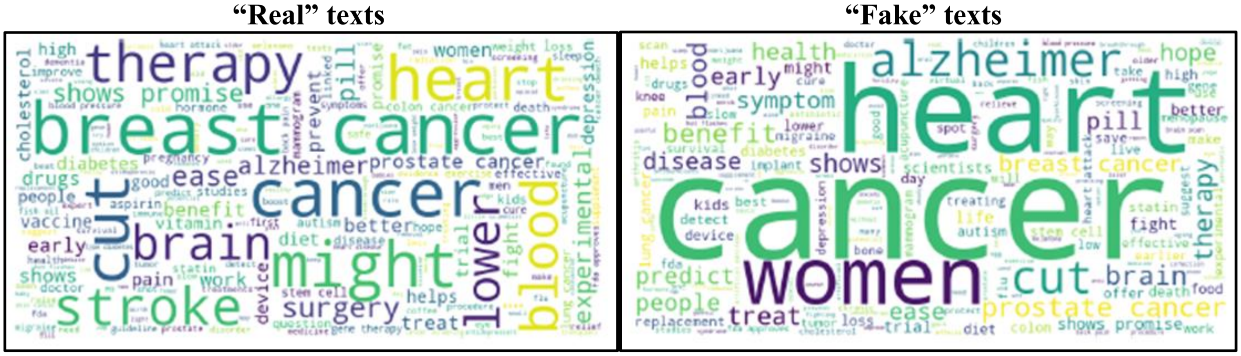


Figure 2
Word clouds of news headlines in the “real” and “fake” classes (Dai et al., 2020)



The texts were annotated by medical and health service experts, all of whom signed industry-independent disclosure agreements. The annotation procedure followed a standard rating system, consisting of 10 criteria used to assess various aspects of the text, such as overclaiming, missing information, and conflicts of interest (Dai et al., 2020: 3). A score between 0 and 5 was assigned to each text based on the proportion of the criteria satisfied, and texts were binned into classes representing ground-truth labels. Texts receiving scores of 0 – 2 were assigned to the “fake” class and texts receiving a score of 3 – 5 were assigned to the “real” class.[5]

A drawback of the HealthStory sub-corpus is that it is relatively small and exhibits class imbalance, comprising 1,218 “real” texts and 472 “fake” texts[6]. We took this into consideration in the fine-tuning and evaluation of both the ML and transformer-based classification models.

2(b). Grammatical Annotation and Feature Selection

We used the Biber Tagger for grammatical annotation (see Biber, 2006). This tagger identifies over 150 features and is currently the most widely used tagger in corpus linguistic research due to the rich

[5] For further details about the design and compilation of the FakeHealth corpus, please see Dai et al. (2020), or visit the official project website: www.HealthNewsReviews.org

[6] Through the process of cleaning and grammatical annotation, a final dataset of 1,102 “real” and 431 “fake” texts were retained for analysis.

syntactic and semantic detail it identifies. To select a subset of grammatical features through quantitative means, we employed a key feature analysis (KFA) (Biber and Egbert, 2018; Egbert and Biber, 2023). A KFA measures the degree to which a grammatical feature is used with a markedly higher or lower frequency in a particular text type relative to another. Relative frequency is measured through a Cohen's *d* effect size; derived from the mean and standard deviation of each feature, in each text type. A feature is considered "key" to a text type if it meets Cohen's (1977) small effect size benchmark of *d* = ±.20. In this study, we considered features with a *d*-value of ≥ .20 to be distinctive of "fake" texts, and features with a *d*-value of ≤ -.20 to be distinctive of "real" texts. Features with a *d*-value between -.20 and .20 were not considered key to either class, and thus dropped from consideration.

Informed by the results of the KFA, we selected nine of the 150 features to incorporate into the traditional ML models: all nouns, modal verbs of possibility (e.g., *may, might*), attributive adjectives, contractions, verbs of communication (e.g., *propose, inform, respond*), word count, and Dimension 1 and 4 scores derived from Biber's 1988 multi-dimensional analysis of spoken and written language (calculated by applying the factor loadings extracted in Biber's 1988 study).

2(c). Model Fine-tuning and Evaluation

We evaluated three traditional ML models that are frequently used for binary text classification: random forest (RF), logistic regression (LR), and support vector machine (SVM). For these ML models, we employed term frequency-inverse document frequency (TF-IDF) for feature extraction. Three transformer-based models were obtained from the Hugging Face Model Hub: *bert-base-uncased, distilbert-base-uncased* and *roberta-base*. Pre-trained BERT models are well-suited for text classification tasks due to the bidirectional, encoder-only architecture. A simple classification head can be added on top of the pre-trained model in the fine-tuning stage for downstream classification tasks (Delvin et al., 2019). All layers of the BERT models accessed through Hugging Face are unfrozen, allowing the model weights to be updated during fine-tuning. Each model includes a classification head implemented as a fully connected layer applied to the [CLS] token representation.

We fine-tuned and evaluated all models under maximally similar conditions. We employed nested cross-validation (CV), where the inner CV is used to select optimal model hyperparameters, and the outer CV is used to evaluate the model, trained with the selected parameters on the held-out data. This is used to test for overfitting. We used 5-fold cross-validation for both the inner and outer CVs, and explored 3 – 4 influential hyperparameters for each model (see Appendix A). The "best" model hyperparameters within the nested-cross validation were selected by maximizing the F1 score for the "fake" class.

For both CVs, we used stratified cross-validation to ensure balanced class distribution for each fold. To further address class imbalance in the traditional ML models, each sample was assigned a weight that was inversely proportional to its class frequency. In the transformer-based models, class imbalance was addressed by computing class weights and using the WeightedRandomSampler to generate class-balanced batches. These weights were also incorporated into the loss function to reinforce balanced learning.

We evaluated the models using the following performance metrics: F1 macro, AUC, and precision, recall, and F1 scores for both the "fake" and "real" classes. Given the small, imbalanced dataset, we bootstrapped the pooled sample of predictions from the outer folds to calculate standard error and 95% confidence intervals.

3. Results

3(a). Comparison of Unigram-only and Unigram-grammar ML Model Performance

Results indicate that grammatical features have the potential to enhance the performance of traditional ML models. With the addition of only the nine grammatical features selected via the KFA, all traditional ML models outperformed their counterparts, albeit modestly. In the ML models with grammatical features, F1 scores for the "fake" class showed improvement ranging from .02 – .03. While this increment falls within the margin of error, similar improvement was found across all performance metrics reported for the UG models (see Table 1). There are only two deviations from this pattern: the random forest F1 and recall scores for the “real” class, and the logistic regression recall score for the “real” class. The consistency with which grammatical features improved the performance of the traditional ML models suggests that it would be advantageous to consider grammatical features in similar use cases.

Table 1 displays the contrast in performance between unigram-only (U) and unigram-grammar (UG) models. Where the UG model outperformed the U model, the performance metric is marked in bold red font. Where the opposite is true, the performance metric is marked in bold blue font.

3(b). Comparison of Transformer-based and Unigram-grammar ML Model Performance

Given that the grammatically enhanced models largely outperformed their unigram-only counterparts, we chose to compare the performance of the enhanced models to the transformer-based models. In our use case, the two model types achieved comparable performance, with grammatically enhanced ML models yielding slightly higher scores. While all scores fell within the margin of error, the ML models consistently outperformed the transformer-based models across all performance metrics, with the exception of the AUC score (see Table 2).

Table 2 presents the performance metrics of the grammatically enhanced ML models and the transformer-based models. For each of the performance metrics reported, the highest score amongst all models is marked in bold red font. Note that the highest score was achieved by one of the grammatically enhanced ML models on seven of the eight performance metrics. The random forest model recorded the highest F1 score for the “fake” class (.542 ±.017), though only marginally outperforming DistilBERT (.539 ±.018) and well within the margin of error. Figure 3 depicts the performance of all models with respect to F1 macro scores.

Table 1

Unigram-only (U) and unigram-grammar (UG) model comparison: mean, standard error, 95% confidence intervals. Models incorporating grammatical features appear shaded gray. For each ML model, the U and UG performance metrics are compared to one another. Where the UG model outperformed the U model, the performance metric is marked in bold red font. Where the opposite is true, the performance metric is marked in bold blue font.

				“Real” Class			“Fake” Class	
Model	**F1 Macro**	**AUC**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
RF (U)	.660 ± .013	.732 ± .013	.817 ± .009	**.776** ± .013	**.800** ± .009	.493 ± .018	.556 ± .024	.523 ± .019
	[.635, .687]	[.705, .760]	[.801, .835]	[.753, .802]	[.780, .814]	[.460, .531]	[.508, .603]	[.486, .559]
RF (UG)	**.668** ± .013	**.740** ± .014	**.828** ± .009	.765 ± .013	.795 ± .009	**.498** ± .018	**.594** ± .023	**.542** ± .018
	[.644, .694]	[.715, .766]	[.812, .844]	[.740, .790]	[.778, .813]	[.464, .532]	[.550, .636]	[.507, .575]
LR (U)	.604 ± .013	.681 ± .016	.816 ± .010	**.636** ± .015	.714 ± .011	.405 ± .013	.633 ± .024	.494 ± .013
	[.578, .628]	[.649, .711]	[.796, .836]	[.607, .665]	[.692, .737]	[.377, .431]	[.587, .677]	[.463, .525]
LR (UG)	**.618** ± .012	**.685** ± .016	**.833** ± .011	.630 ± .015	**.717** ± .012	**.417** ± .013	**.677** ± .023	**.516** ± .015
	[.592, .641]	[.654, .716]	[.812, .853]	[.601, .659]	[.694, .739]	[.391, .443]	[.631, .724]	[.486, .546]
SVM (U)	.590 ± .012	.653 ± .015	.806 ± .010	.616 ± .014	.700 ± .011	.388 ± .013	.621 ± .024	.477 ± .016
	[.565, .612]	[.626, .683]	[.677, .721]	[.786, .827]	[.589, .645]	[.364, .413]	[.573, .666]	[.448, .507]
SVM (UG)	**.624** ± .012	**.690** ± .014	**.821** ± .009	**.674** ± .014	**.740** ± .010	**.429** ± .013	**.623** ± .022	**.508** ± .015
	[.601, .647]	[.657, .713]	[.803, .839]	[.649, .702]	[.721, .759]	[.403, .456]	[.580, .666]	[.478, .537]

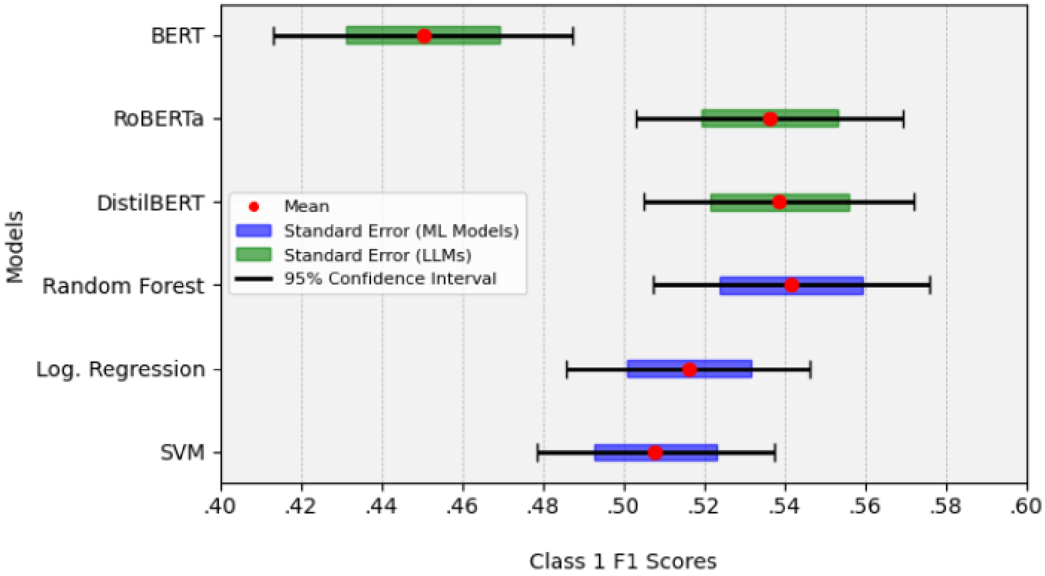
As seen in Table 1, the incorporation of grammatical features improved performance for most traditional ML models, on most metrics.

Table 2

ML and transformer-based model comparison: mean, standard error, 95% confidence intervals. Grammatically enhanced ML models appear shaded gray. For each performance metric, the highest value is marked in bold red font.

			“Real” Class			“Fake” Class		
Model	**F1 Macro**	**AUC**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
RF (UG)	**.668** ± .013	.740 ± .014	.828 ± .009	**.765** ± .013	**.795** ± .009	**.498** ± .018	.594 ± .023	**.542** ± .018
	[.644, .694]	[.715, .766]	[.812, .844]	[.740, .790]	[.778, .813]	[.464, .532]	[.550, .636]	[.507, .575]
LR (UG)	.618 ± .012	.685 ± .016	**.833** ± .011	.630 ± .015	.717 ± .012	.417 ± .013	**.677** ± .023	.516 ± .015
	[.592, .641]	[.654, .716]	[.812, .853]	[.601, .659]	[.694, .739]	[.391, .443]	[.631, .724]	[.486, .546]
SVM (UG)	.624 ± .012	.690 ± .014	.821 ± .009	.674 ± .014	.740 ± .010	.429 ± .013	.623 ± .022	.508 ± .015
	[.601, .647]	[.657, .713]	[.803, .839]	[.649, .702]	[.721, .759]	[.403, .456]	[.580, .666]	[.478, .537]
BERT	.612 ± .013	.655 ± .016	.785 ± .008	.761 ± .013	.773 ± .009	.434 ± .018	.468 ± .024	.450 ± .019
	[.586, .637]	[.626, .684]	[.770, .801]	[.735, .788]	[.756, .791]	[.398, .473]	[.422, .515]	[.413, .487]
RoBERTa	.657 ± .013	**.741** ± .014	.830 ± .090	.732 ± .014	.778 ± .100	.474 ± .016	.617 ± .024	.536 ± .017
	[.631, .681]	[.712, .768]	[.813, .847]	[.704, .758]	[.759, .795]	[.443, .504]	[.568, .664]	[.502, .568]
DistilBERT	.664 ± .012	.731 ± .015	.829 ± .090	.752 ± .013	.788 ± .090	.487 ± .024	.603 ± .016	.539 ± .017
	[.638, .687]	[.702, .759]	[.811, .846]	[.727, .776]	[.770, .806]	[.455, .518]	[.555, .650]	[.505, .572]

Figure 3
Comparison of F1 scores for the "fake" class across grammatically enhanced ML models and transformer-based models



4. Conclusion

In this work, we explored the extent to which incorporating grammatical features into traditional ML models (e.g., logistic regression, random forest, support vector machine) can improve model performance for misinformation detection. To identify grammatical features that may help distinguish texts containing public health misinformation ("fake" texts) from those that do not ("real" texts), we used a key feature analysis, which serves to identify grammatical features that are frequent and pervasive in each group of texts relative to the other. We then compared the performance of the grammatically enhanced ML models to the performance of LLMs on the same task. Preliminary results of our study show that incorporating grammatical features into traditional ML models led to modest improvements in performance, with the top-performing ML model achieving results comparable to those of the LLMs.

Time limitations hindered our ability to optimize both the traditional ML and transformer-based models. All models were fine-tuned and evaluated with a maximum allowable runtime of 48 hours. While the traditional ML models took anywhere from five minutes to seven hours to run, the transformer-based models frequently approached or exceeded the 48-hour runtime limit. As a result, we attempted to train the transformer-based models multiple times with slightly altered hyperparameters to avoid early termination. This required that we weigh the importance of various hyperparameters for tuning, such as batch_size, learning_rate, epochs, and max_token_length. While BERT models have a max_token_length of 512, we were unable to run the *bert-based-uncased* model at the maximum token length without cutting numerous other hyperparameters from the grid. Future iterations of this study should employ a standardized method of hyperparameter specification in order to properly optimize the models.

The limitations of the transformer-based models also affected our ability to optimize the traditional ML models. To prevent providing the ML models with a performance advantage due solely to the lack of constraints relative to the transformer-based models, we standardized the training conditions of all model types. While we could have carried out a far more extensive exploration of optimal hyperparameters for the traditional ML models, we chose to avoid introducing this potential bias.

While we saw a modest increase in ML model performance when grammatical features were incorporated into the models, and comparable performance between grammatically enhanced ML models and transformer-based models, it is important to note that the overall performance of all models in this study was relatively poor. Future studies should consider exploring a broader range of LLMs and traditional ML models, as well as additional features that have shown promise in detecting misinformation (e.g., sentiment and polarity). It would also be valuable to conduct similar studies using misinformation corpora from diverse source domains to assess the generalizability of these findings.

Overall, our preliminary results indicate that traditional models still hold untapped potential. With further methodological enhancements, we may be able to narrow the performance gap between traditional ML models and more advanced AI models. Given their advantages in computational efficiency and interpretability, traditional ML models remain a compelling alternative to LLMs, especially in contexts where resources are limited or model transparency is critical. Continued exploration and refinement of traditional ML models, alongside ongoing comparisons with LLMs, will be important as AI technologies continue to advance.

Acknowledgments

We gratefully acknowledge the U.S. Department of Energy through the Los Alamos National Laboratory/Laboratory Directed Research and Development Program and the Center for Nonlinear Studies for supporting this work.

Sources

Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (2006). *University language*. John Benjamins Publishing Company.

Biber, D. (2019). Text-linguistic approaches to register variation. *Register Studies*, *1*(1), 42-75.

Biber, D., & Egbert, J. (2018). *Register Variation on the Web*. Cambridge University Press.

Biber, D., & Egbert, J. (2023). What is a register?: Accounting for linguistic and situational variation within–and outside of–textual varieties. *Register Studies*, *5*(1), 1-22.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* Academic Press.

Dai, E., Sun, Y., & Wang, S. (2020, May). Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. *Proceedings of the International AAAI Conference on Web and Social Media, 14*(1), 853-862. https://doi.org/10.1609/icwsm.v14i1.7350

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers).* Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Di Sotto, S., & Viviani, M. (2022). Health misinformation detection in the social web: an overview and a data science approach. *International Journal of Environmental Research and Public Health*, *19*(4), 2173. https://doi.org/10.3390/ijerph19042173

Egbert, J., & Biber, D. (2023). Key feature analysis: a simple, yet powerful method for comparing text varieties. *Corpora*, *18*(1), 121-133. https://doi.org/10.3366/cor.2023.0

Hemina, K., Boumahdi, F., Madani, A., & Remmide, M. A. (2023, April). A cross-validated fine-tuned gpt-3 as a novel approach to fake news detection. *Proceedings of the International Conference on Applied Cybersecurity (ACS).* Springer Nature Switzerland.

Liu, Z., Zhang, T., Yang, K., Thompson, P., Yu, Z., & Ananiadou, S. (2024). Emotion detection for misinformation: A review. *Information Fusion*, *107*. https://doi.org/10.1016/j.inffus.2024.102300

Ni, Z., Bousquet, C., Vaillant, P., & Jaulent, M. C. (2023). Rapid review on publicly available datasets for health misinformation detection. *Healthcare Transformation with Informatics and Artificial Intelligence*, 305, 123-126. https://doi.org10.3233/SHTI230439

Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, *121*, 102342.

Appendix A

Hyperparameter Grids for Nested Cross-Validation

Table A1

Hyperparameter grids for each text classification model

Model	Hyperparameters
BERT	max_token_length, param_grid, batch_size, learning_rate
RoBERTa	max_token_length, batch_size, learning_rate, epochs
DistilBERT	max_token_length, batch_size, learning_rate, epochs
Random Forest	n_estimators, min_samples_leaf, max_depth
Logistic Regression	C, max_iter, solver, penalty, fit_intercept

Support Vector Machine	C, kernel (linear);
	C, gamma, kernel (rbf);
	C, gamma, kernel (poly), degree

Appendix B

Boxplots Depicting the Mean, Standard error, and 95% Confidence Intervals for Performance Metrics in ML and Transformer-based Models

Figure B1

Performance metrics for unigram-only logistic regression model

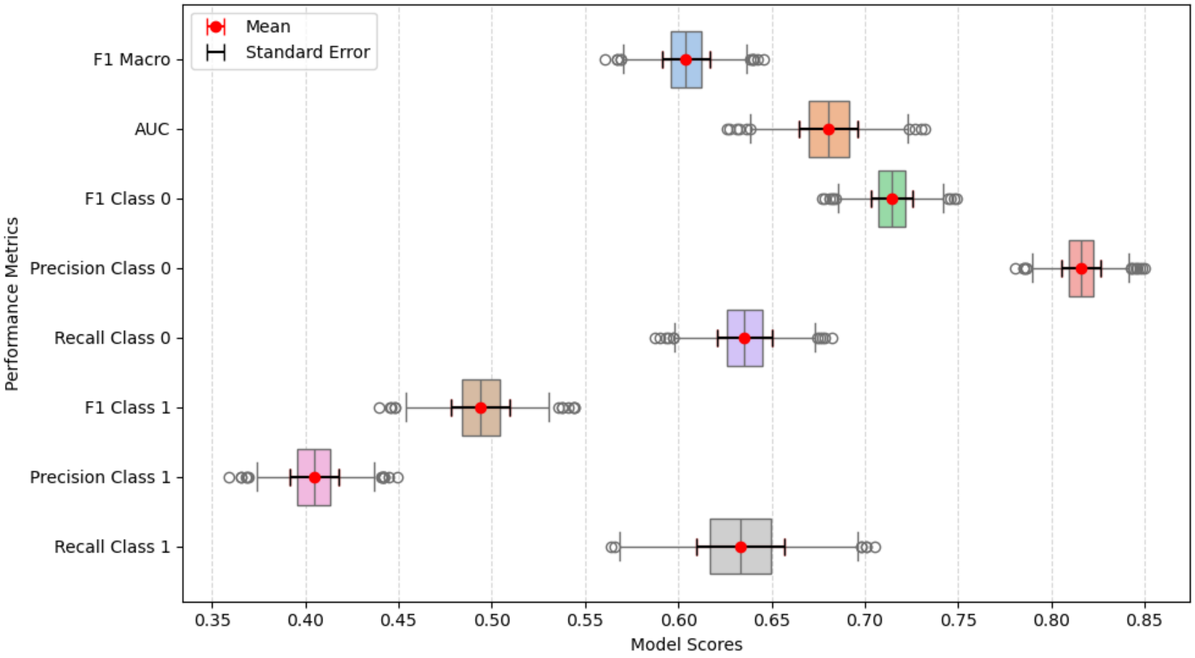


Figure B2

Performance metrics for unigram-grammar logistic regression model

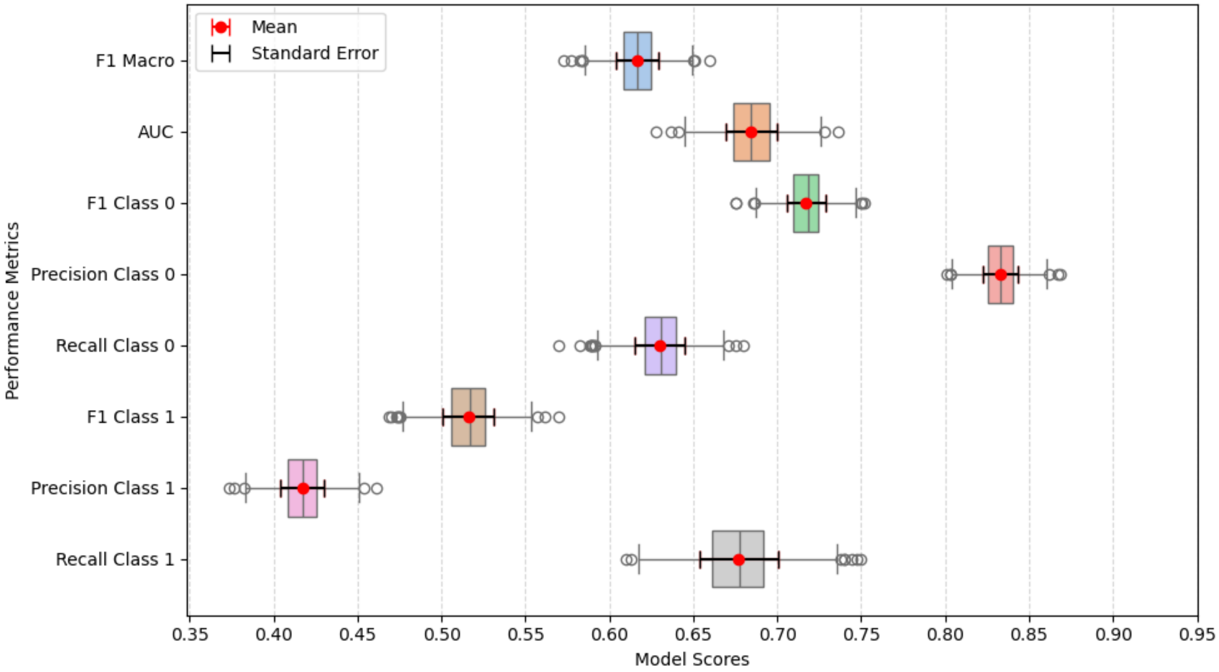


Figure B3

Performance metrics for unigram-only random forest model

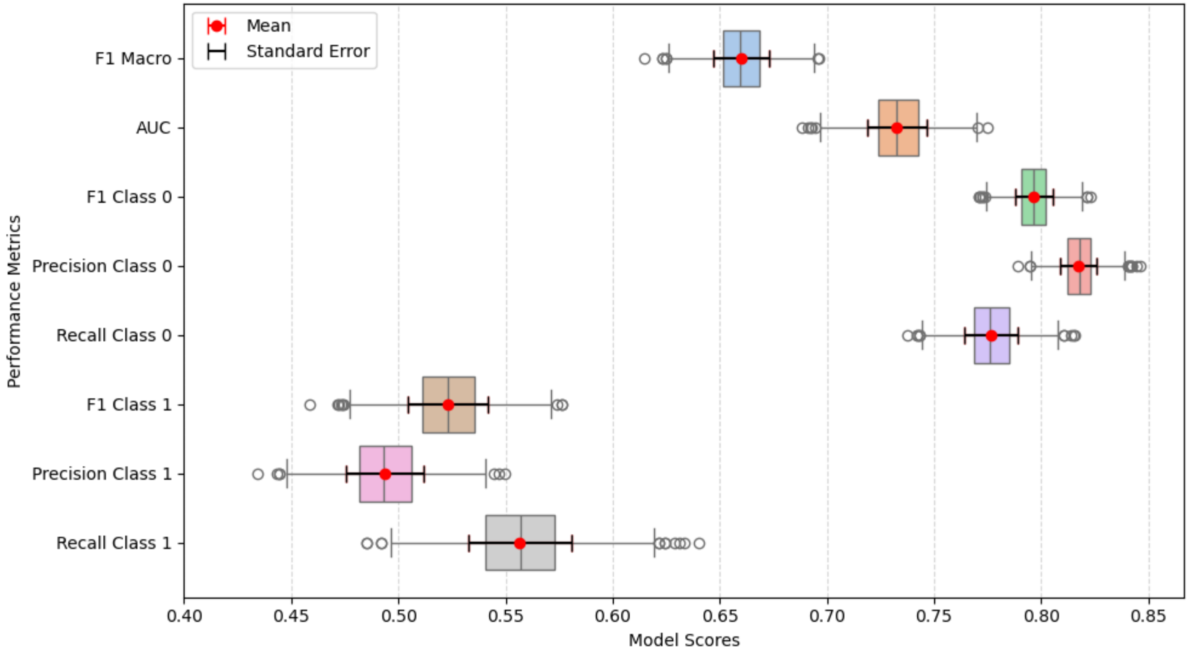


Figure B4

Performance metrics for unigram-grammar random forest model

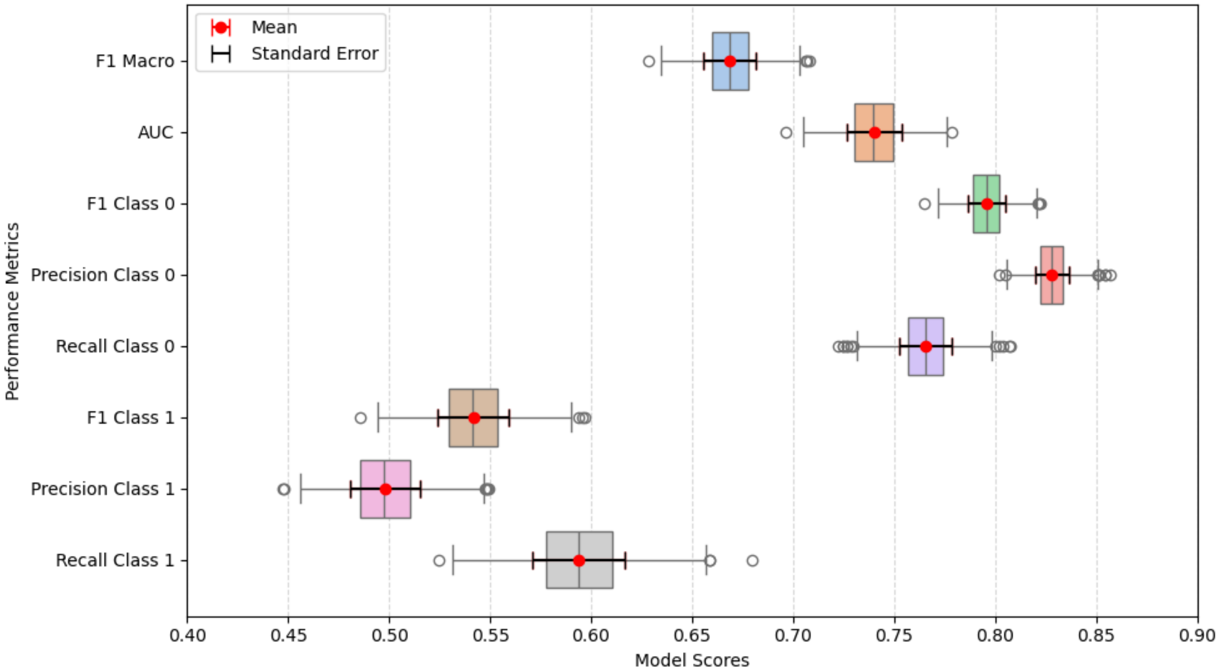


Figure B5

Performance metrics for unigram-only support vector machine model

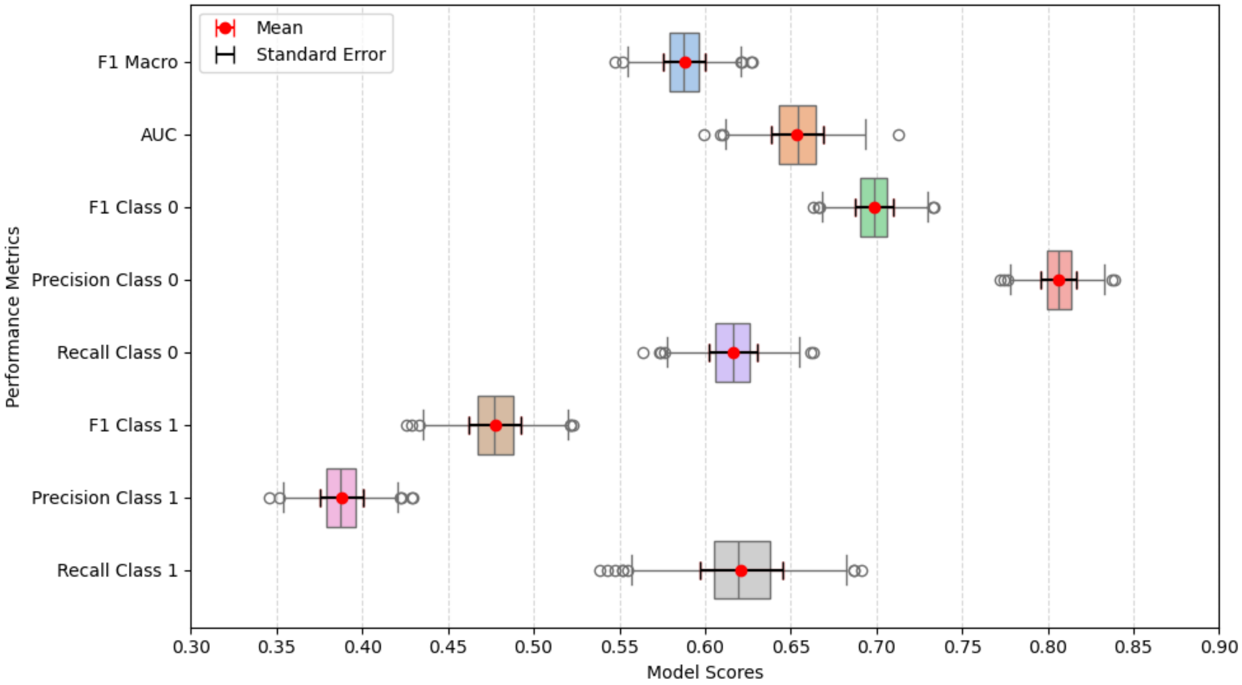


Figure B6

Performance metrics for unigram-grammar support vector machine model

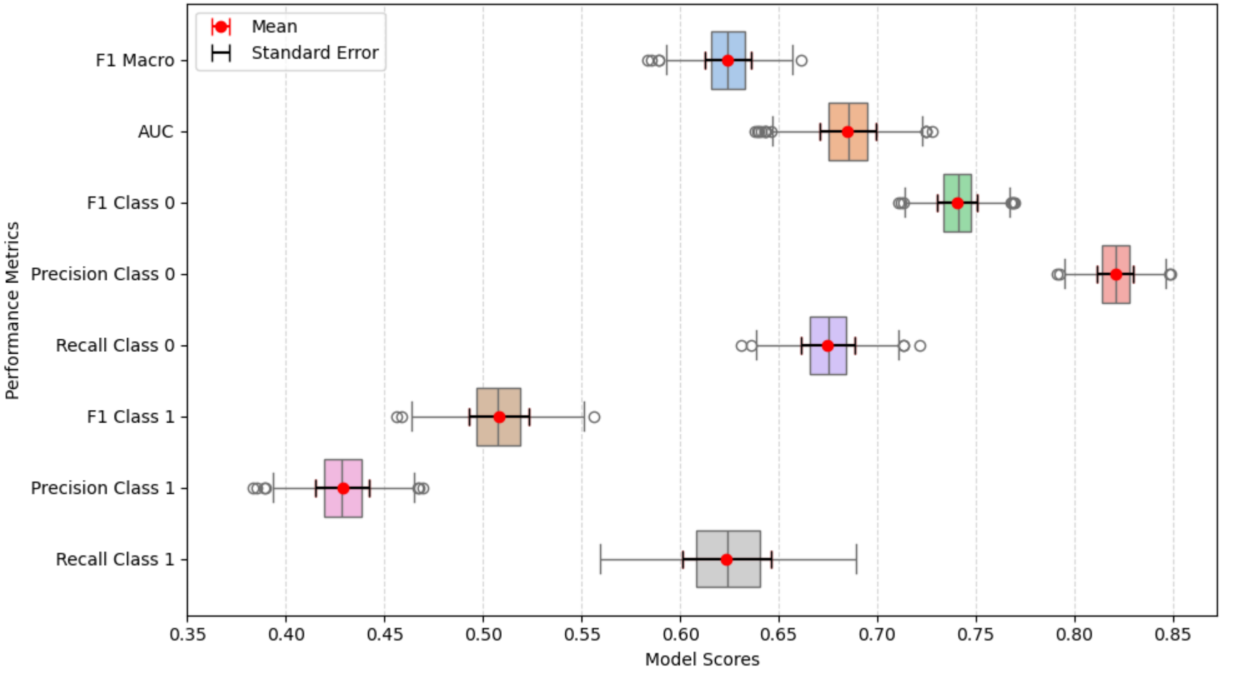


Figure B7
Performance metrics for the *bert-base-uncased* model

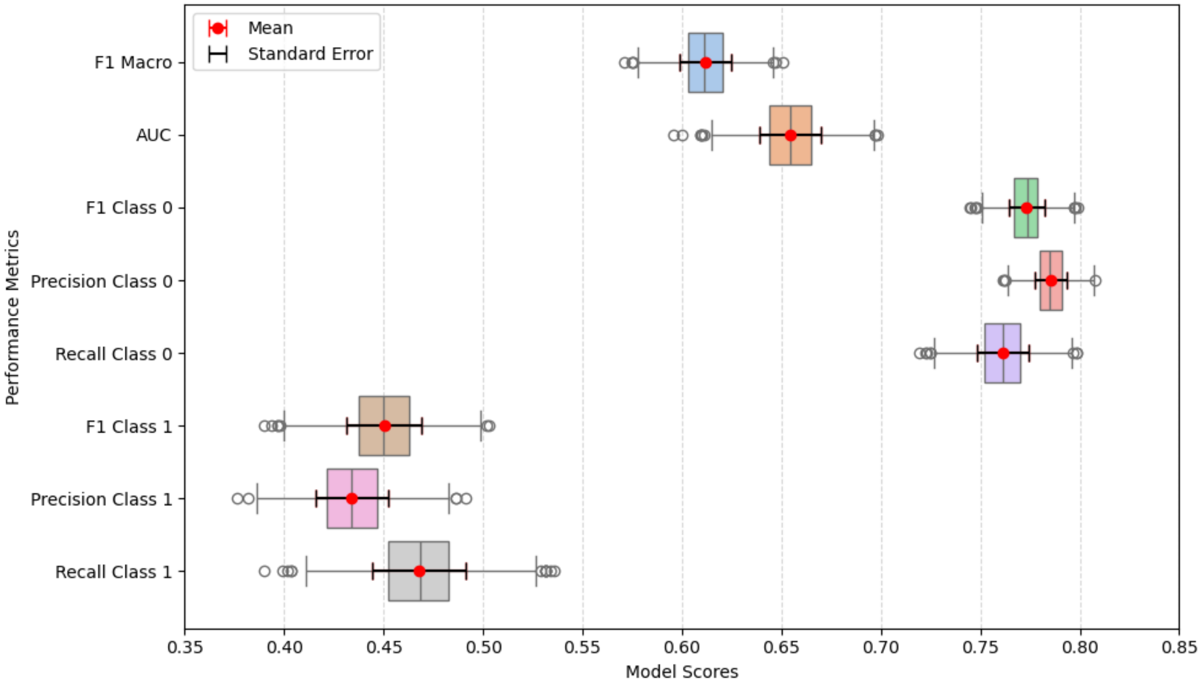


Figure B8
Performance metrics for the *RoBERTa-base* model

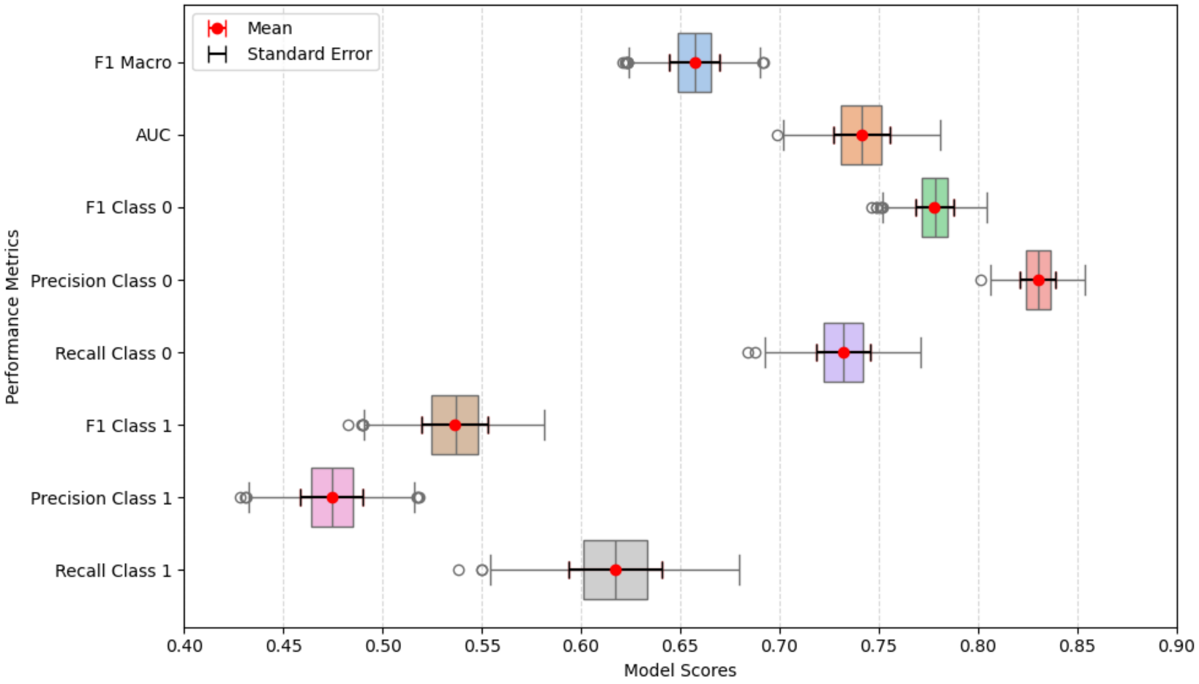
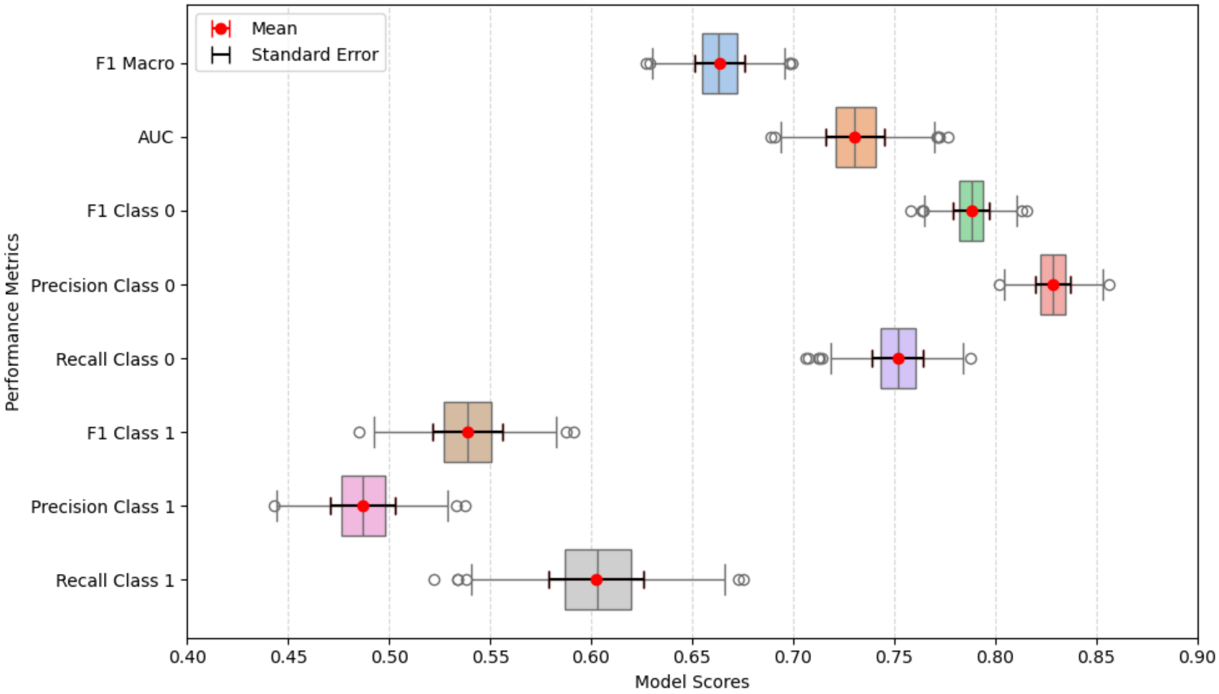


Figure B9

Performance metrics for the *DistilBERT-base-uncased* model


Mean
Standard Error
F1 Macro
AUC
F1 Class 0
Precision Class 0
Recall Class 0
F1 Class 1
Precision Class 1
Recall Class 1
Performance Metrics
Model Scores