Task-Specific or Task-Agnostic? A Statistical Inquiry into BERT for Human Trafficking Risk Prediction

Ana Paula Arguelles Terron* **Jorge Yero Salazar*** **Pablo Rivas**
Department of Computer Science
Baylor University
Waco, TX 76798
{Ana_Arguelles1,Jorge_Yero1,Pablo_Rivas}@Baylor.edu

Ernesto Quevedo Caballero **Alejandro Rodriguez Perez**
Department of Computer Science
Baylor University
Waco, TX 76798
{Ernesto_Quevedo1,Alejandro_Rodriguez4}@Baylor.edu

Abstract

The pervasive issue of human trafficking has increasingly manifested through digital platforms, particularly in the form of textual online advertisements. Leveraging Natural Language Processing (NLP) for risk assessment in this domain has garnered significant attention. This study presents a comprehensive empirical evaluation of machine learning models fine-tuned for emotion and sentiment analysis tasks, specifically utilizing the *BERT-Base Uncased* and *DistilBERT* architectures. These models are rigorously compared against a baseline model, also fine-tuned on the *BERT-Base Uncased* architecture, for the task of human trafficking risk prediction. Employing robust statistical methodologies, namely the Friedman and Nemenyi tests, we scrutinize the performance metrics of these models. Our findings indicate that while task-specific fine-tuned models exhibit promising results, only the fine-tuned model in the emotion analysis task statistically outperforms the baseline model in the human trafficking risk prediction task. This research not only contributes to the growing body of work in NLP applications for social good but also provides valuable insights for future research directions in the field.

1 Introduction

Human trafficking remains a critical issue in contemporary society, posing significant ethical and humanitarian concerns. According to the Global Slavery Index, an estimated 30 million individuals were subjected to involuntary servitude as of 2013, underscoring the magnitude of the problem [1]. A substantial portion of this trafficking involves the exploitation of women in the sex industry, with the United States emerging as a prevalent destination for trafficked individuals [10].

The digital landscape has further complicated the issue, as online platforms have become a prevalent medium for the advertisement of sexual services, often masking underlying trafficking activities [5].

Previous analyses of these online advertisements have identified distinct characteristics that can aid law enforcement agencies in the detection of trafficking-related content [22]. However, the sheer volume of such advertisements presents a formidable challenge for manual processing and intervention.

[*] Equal contribution

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

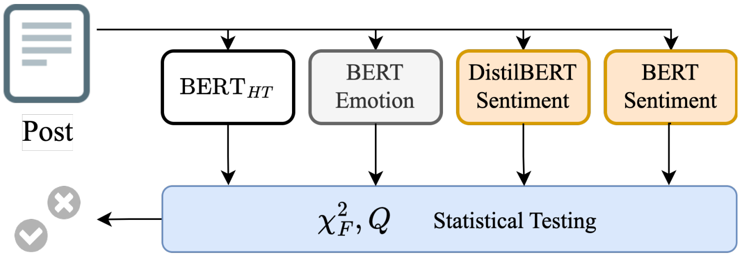


Figure 1: Statistical assessment of the impact of different fine-tuned BERT models on classifying trafficking risk based on text sequences from commercial sex advertisements.

In response to this challenge, Natural Language Processing (NLP) techniques have been increasingly employed to automate the identification of trafficking-related advertisements [21]. Recent advancements in the field have focused on feature extraction from online text to distinguish advertisements with sexual content and further classify their risk of being related to human trafficking [15]. Concurrent research efforts have been directed towards the development of predictive models for risk assessment based on these extracted features [24].

Situated within this context, the present study aims to investigate the impact of task-specific features on the predictive accuracy of human trafficking risk assessment models, as depicted in Figure 1. The *BERT* language model has demonstrated notable proficiency across various Natural Language Processing (NLP) tasks, including sentiment and emotion analysis, as well as question answering, among other tasks. In light of its great performance, we employ fine-tuned versions of the *BERT* and *DistilBERT* models, originally pre-trained for emotion and sentiment classification tasks, to evaluate their efficacy in this domain [13, 14]. To rigorously assess the performance differences among these models, we employ Friedman and Nemenyi statistical tests, thereby providing a robust evaluation of their relative effectiveness [12]. Therefore, our specific contributions are the following:

- We conduct a comprehensive empirical evaluation to assess the performance of task-specific fine-tuned models vis-à-vis a baseline model in the domain of human trafficking risk prediction, leveraging advanced statistical tests such as Friedman and Nemenyi.
- We provide critical insights into the advantages and limitations of sentiment, emotion, and tone analysis for human trafficking risk prediction, thereby directing future research towards potentially more impactful NLP tasks like Named Entity Recognition.

The remainder of this paper is structured as follows: Section 2 provides an overview of the relevant literature, Section 3 delineates the methodological approach employed, Section 4 outlines the experimental methodology, Section 5 presents an analytical discussion of the findings, and Section 6 offers concluding remarks.

2 Related Work

The application of Natural Language Processing (NLP) techniques for the detection of suspicious advertisements indicative of human trafficking has garnered considerable scholarly attention. Whitney et al. [28] conducted an empirical investigation into the role of emojis as potential markers for trafficking-related advertisements. Their work provides valuable insights into the nuanced linguistic features that may be leveraged for risk assessment.

In a similar vein, Tong et al. [25] employed deep multimodal models to classify advertisements, integrating both textual and visual elements. Their work culminated in the development of the Human Trafficking Deep Network, a specialized model designed to identify high-risk advertisements. This approach underscores the importance of multimodal data in enhancing predictive accuracy.

Szekely et al. [23] took a different approach by constructing a comprehensive knowledge graph from a large corpus of online advertisements. Their methodology enables sophisticated data visualization and querying capabilities, thereby facilitating more subtle analyses.

Esfahani et al. [8] proposed an automated system for ad classification, focusing on the extraction of contextual features from the text. They employed a variety of machine learning models, including

Latent Dirichlet Allocation (LDA) [3] for topic modeling, FastText [4] for word vector representation, and the pre-trained Bidirectional Encoder Representations from Transformers (*BERT*) [7] language model. Their work aims to evaluate the efficacy of these models in analyzing unstructured data for human trafficking risk assessment.

Li et al. [11] explored the identification of massage businesses potentially involved in human trafficking through NLP techniques. Utilizing data from the *Yelp.com* platform, they developed multiple models based on lexicon-based approaches as well as embeddings (*BERT* and *Doc2Vec*). Their findings suggest that ensemble methods significantly outperform individual models, thereby advocating for a more integrative approach.

The proliferation of online platforms has unfortunately provided human traffickers with new avenues to disseminate their illicit activities. As such, the scholarly contributions in this domain are critical for developing effective computational methods for identifying and mitigating the risks associated with human trafficking.

3 Methodological Approach

The principal objective of this research endeavor is to empirically accept or reject the following hypothesis:

> *The generically pre-trained language model BERT exhibits comparable performance in the task of human trafficking risk detection as its specialized counterparts fine-tuned for emotion and sentiment analysis tasks.*

3.1 Baseline Model

As a point of reference, we employ a fine-tuned version of the *BERT-Base Uncased* model [26] specifically adapted for the human trafficking risk detection task. The architecture of this baseline model, denoted as $BERT_{HT}$, is illustrated in Figure 2 (a). This model comprises a total of 109,483,778 parameters, aligning with the original *BERT* architecture [7].

The *BERT* model employs the Transformer architecture, which is fundamentally based on the attention mechanism. The attention function can be described as mapping a query and a set of key-value pairs to an output. Mathematically, the scaled dot-product attention is given by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the keys [27].

The *BERT* architecture utilizes multi-head attention, which can be expressed as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O,$$

where each $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, and W_i^Q, W_i^K, W_i^V, and W^O are parameter matrices.

The model is trained using the masked language model (MLM) objective, which can be formalized as:

$$\mathcal{L} = -\log P(w_i|\text{context}; \theta).$$

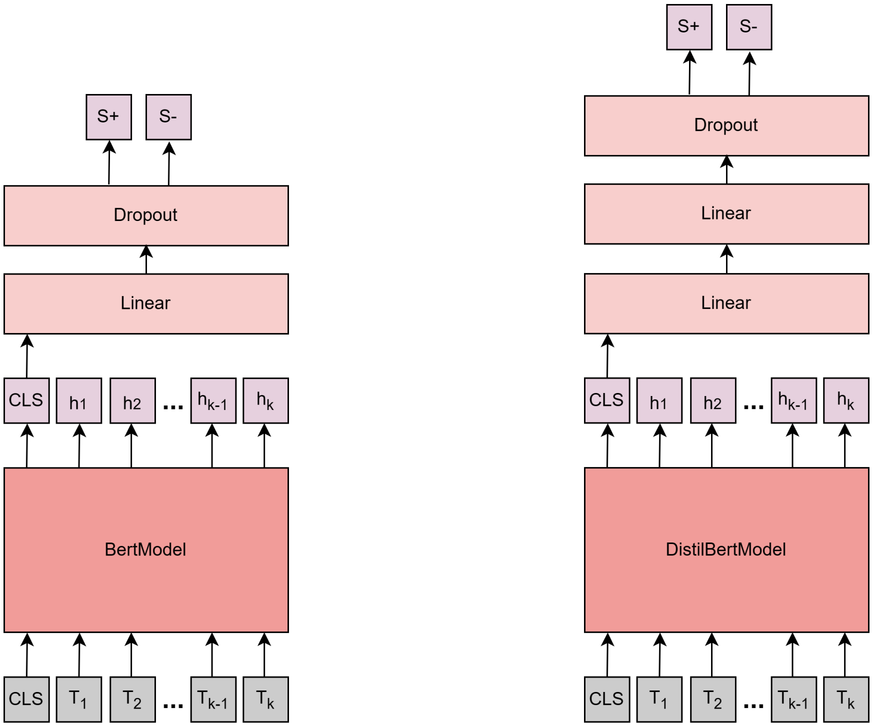
where w_i is the masked word, context is the surrounding words, and θ are the model parameters.

By fine-tuning this architecture on the human trafficking risk detection task, we aim to evaluate its efficacy as a baseline model for comparison with other task-specific models.

3.2 Experimental Models

We subject three distinct models to evaluation for the human trafficking risk classification task. The first model is a fine-tuned version[2] of a sentiment classification model [20], based on the *DistilBERT*

[2] `https://huggingface.co/sbcBI/sentiment_analysis_model`



(a) Architecture of the Baseline Model (*BERT*$_{HT}$), based on classic BERT. 110M parameters.

(b) Architecture of the Fine-Tuned Sentiment Analysis Model Based on *DistilBERT*. 67M parameters.

Figure 2: Comparative Architectures of Baseline and Fine-Tuned Models.

architecture. The architecture of this model is depicted in Figure 2 (b). The subsequent models are fine-tuned sentiment[3] and emotion[4] classification models based on the original *BERT* architecture [7]. The architectures of these latter models are analogous to that of the baseline model (Figure 2 (a)). *BERT* is a more comprehensive model, typically consisting of 12 layers (or 24 for its larger variant), 12 attention heads, and approximately 110 million parameters for its base version. It employs a bidirectional Transformer encoder to understand the context and semantics of the words in a given text. On the other hand, *DistilBERT* is a distilled version of *BERT*, designed to be smaller, faster, and more efficient. It retains 95% of *BERT*'s performance while being 40% smaller, encompassing only six layers and six attention heads. The model achieves this efficiency by forgoing certain architectural elements, such as the token-type embeddings and the pooler, and employing knowledge distillation techniques during training. The distillation process involves training *DistilBERT* to predict the output of *BERT*, effectively transferring knowledge from the larger model to the smaller one. Despite these reductions, *DistilBERT* maintains competitive performance across various natural language understanding tasks, making it a suitable alternative for resource-constrained environments [20].

3.3 Evaluation Metrics

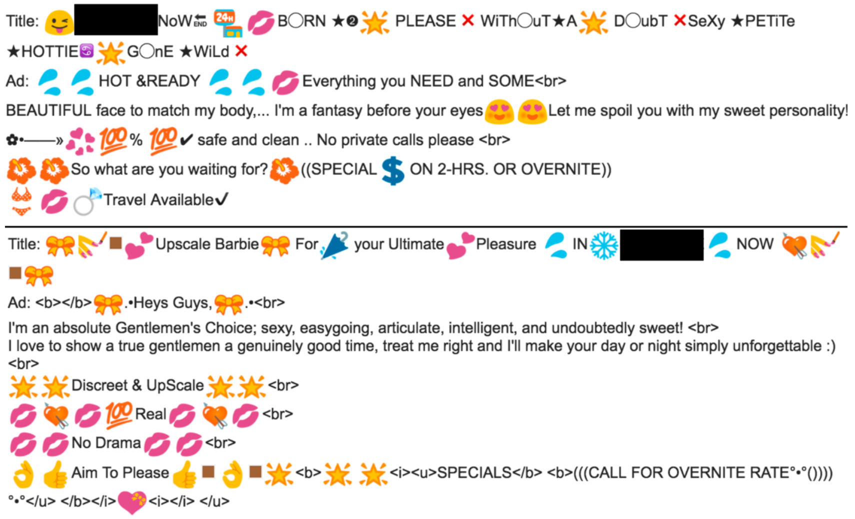
The models were rigorously trained and evaluated using comprehensive evaluation metrics, including the F_1 score, Accuracy, Precision, and Recall [16]. These metrics provide a multi-faceted view of each model's performance, enabling a detailed comparison.

3.4 Statistical Hypothesis Testing

To ascertain the validity of the research hypothesis, we conducted statistical hypothesis testing to compare the performance of the experimental models against that of the baseline *BERT*$_{HT}$

[3] https://huggingface.co/Ghost1/bert-base-uncased-finetuned_for_sentiment_analysis1-sst2

[4] https://huggingface.co/amitkayal/bert-finetuned-sem_eval-english


Title: 😜 [redacted] NoW🔚 🏪 💋 B◯RN ★❷ 🌟 PLEASE ❌ WiTh◯uT★A 🌟 D◯ubT ❌SeXy ★PETiTe ★HOTTIE♋ 🌟 G◯nE ★WiLd ❌

Ad: 💦 💦 HOT &READY 💦 💦 💋 Everything you NEED and SOME
BEAUTIFUL face to match my body,... I'm a fantasy before your eyes😍 😍Let me spoil you with my sweet personality!
✿•——» 💕 💯% 💯✔ safe and clean .. No private calls please
🌺 🌺 So what are you waiting for?🌺((SPECIAL 💲 ON 2-HRS. OR OVERNITE))
👙 💋 💍Travel Available✓

Title: 🎀 💅 🟫 💕 Upscale Barbie🎀 For ☔ your Ultimate 💕 Pleasure 💦 IN ❄ [redacted] 💦 NOW 💘 💅 🟫 🎀

Ad: <b></b>🎀.•Heys Guys, 🎀.•
I'm an absolute Gentlemen's Choice; sexy, easygoing, articulate, intelligent, and undoubtedly sweet!
I love to show a true gentlemen a genuinely good time, treat me right and I'll make your day or night simply unforgettable :)
🌟 🌟 Discreet & UpScale 🌟 🌟
💋 💘 💋 💯 Real 💋 💘 💋
💋 💋 No Drama 💋 💋
👌 👍 Aim To Please 👍 🟫 👌 🟫 🌟 <b> 🌟 🌟 <i><u>SPECIALS</b> <b>(((CALL FOR OVERNITE RATE°•°()))) °•°</u> </b></i>💝<i></i> </u>

Figure 3: Exemplary posts illustrating the heterogeneity of character elements, thereby posing challenges for conventional NLP methodologies. Personally identifiable information has been redacted.

model. Specifically, we employed statistical tests to determine whether the observed differences in performance metrics were statistically significant, thereby providing empirical evidence to either support or refute the initial hypothesis [6].

4 Experimental Methodology

This section delineates the experimental setup and methodologies employed to evaluate the performance of the proposed models for human trafficking risk prediction. We commence by detailing the dataset characteristics and partitioning strategy, followed by the evaluation metrics and statistical tests utilized to validate the models. Subsequently, we elaborate on the specific configurations and computational resources involved in the training process. Finally, the results are presented and analyzed to draw meaningful conclusions on the efficacy of the models in the task at hand.

4.1 Data

The dataset under investigation comprises 128,182 textual posts, each of which is categorically labeled as either posing a risk of human trafficking or not. These posts are characterized by their free-text nature and may include a diverse range of elements such as emojis and other non-ASCII characters. Due to ethical considerations and privacy constraints, the dataset is not publicly accessible. However, its validity has been corroborated through multiple peer-reviewed studies [19, 2, 9, 17, 18]. Figure 3 provides illustrative examples of posts, highlighting the complexities associated with the textual and character elements commonly encountered in the dataset.

Before model training, the dataset was partitioned to create a testing subset. Specifically, a random subsampling strategy was employed to allocate 10% of each class to the testing set, resulting in a total of 12,819 posts. The residual dataset, consisting of 115,363 posts, was further divided into training and validation sets. The training set encompasses 95% (109,594 posts) of the remaining data, while the validation set constitutes the remaining 5% (5,769 posts). In terms of class distribution, the dataset is imbalanced, with the “high risk” class accounting for 31.88% and the “low risk” class making up the remaining 68.12% (approximately twice the “high risk” sample size).

4.1.1 Class imbalance mitigation

To mitigate and verify the model wasn’t biased toward the majority class, we divided the larger "low risk" group into two parts to create two balanced training sets. We then duplicated the "high risk" entries in both sets to maintain balance. As a result, each training set now contains 76,068

Table 1: Performance metrics for the trained models, where bold font indicated the top performing model

Model	F_1	Accuracy	Precision	Recall	AUC
Fine-tuned Emotion Model	**0.75**	**0.84**	**0.76**	**0.74**	**0.90**
Fine-tuned Sentiment Analysis (*DistilBERT*)	0.72	0.83	0.77	0.67	0.89
Fine-tuned Sentiment Analysis (*BERT*)	0.73	0.83	0.77	0.69	0.89
Baseline Model ($BERT_{HT}$)	0.73	0.83	0.74	0.72	0.89

advertisements. The proportion of the number of examples per class is similar, with "low risk" at 52% and "high risk" at 48%. For the actual training phase, we utilized 95% of the data (72,265 ads) for training purposes, while the remaining 5% (3,803 ads) were set aside for validation. To test the models, we employed the original split testing set, which comprises 12,819 posts.

4.2 Evaluation Methodology

The baseline model, denoted as $BERT_{HT}$, served as a comparative standard and achieved an accuracy of 83%, an F_1 score of 73%, a precision of 74%, and a recall of 72%.

To statistically validate the efficacy of the proposed models, hypothesis testing was conducted. Specifically, non-parametric statistical tests, namely the Friedman test and the Nemenyi post-hoc test, were utilized [6]. These tests provide a robust framework for comparing multiple algorithms and determining the statistical significance of observed performance differences.

4.3 Experimental Configuration

To facilitate the training and evaluation of the models, we employed the `transformers` Python library. Specifically, we utilized the `DistilBERTForSequenceClassification` and `BERTForSequenceClassification` classes for the fine-tuned models targeting sentiment analysis and emotion recognition. The `DistilBERTForSequenceClassification` model comprises 66,955,010 parameters, while the `BERTForSequenceClassification` model contains 109,483,778 parameters.

The models were trained throughout three epochs. The fine-tuned emotion and sentiment analysis models based on *BERT* required approximately three hours for training, whereas the *DistilBERT*-based sentiment analysis model was trained in approximately two hours for the original training dataset. For the balanced datasets, the *BERT*-based models required roughly one hour and 35 minutes, and the *DistilBERT* models took about 40 minutes for each set.

The computational resources allocated for this task included machines equipped with 20 CPU cores, 1 GeForce RTX 4060 8GB GPU, and 64GB of RAM.

4.4 Results

In this section, we present the results of the evaluation of the different models. In addition, we provide the results obtained for the statistical tests.

4.4.1 Model evaluation

The performance of the trained models was evaluated using the metrics specified in Section 3.3 and the Area Under the Receiver Operating Characteristic Curve (AUC). Table 1 presents the quantitative results for each model.

The fine-tuned Emotion model based on *BERT* demonstrated superior performance across all metrics, outperforming the baseline model. To further elucidate these findings, Figure 4 illustrates the Receiver Operating Characteristic (ROC) curves for the evaluated models. Notably, the AUC for the *Emotion* model based on *BERT* was the highest, registering at 0.90. The curve corresponding to the *Emotion* model based on *BERT* most closely approximates the ideal curve, further substantiating its efficacy.

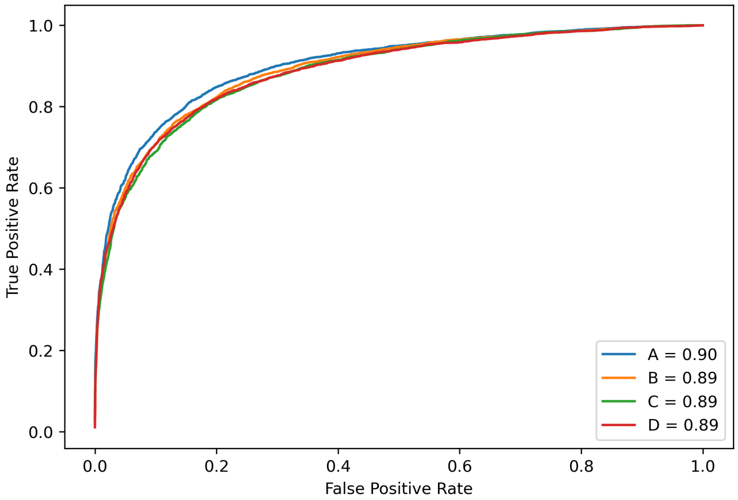


Figure 4: ROC curve for the trained models and the baseline. A is the emotion model based on BERT. B is the sentiment analysis model based on *BERT*, while C is based on *DistillBERT*. D is the baseline model based on *BERT*.

Table 2: Class imbalance mitigation results. S refers to the balanced datasets used during training: 0 or 1. DBERT stands for *DistilBERT*

Model	S	F_1	Accuracy	Precision	Recall	AUC
Fine-tuned Emotion Model	0	**0.71**	**0.80**	**0.65**	**0.80**	**0.88**
	1	**0.72**	**0.80**	**0.65**	**0.81**	**0.88**
Fine-tuned Sentiment Analysis (*DBERT*)	0	0.69	0.78	0.62	0.77	0.85
	1	0.69	0.78	0.62	0.77	0.85
Fine-tuned Sentiment Analysis (*BERT*)	0	0.71	0.79	0.63	0.82	0.87
	1	0.70	0.78	0.61	0.83	0.87
Baseline Model ($BERT_{HT}$)	0	0.70	0.79	0.63	0.79	0.86
	1	0.70	0.79	0.64	0.79	0.86

4.4.2 Class imbalance mitigation

As described in Section 4.4.1, we assessed our trained models using the evaluation metrics outlined in Section 3.3 and the AUC. The results for models trained on balanced datasets, as mentioned in Section 4.1.1, are summarized in Table 2.

The models fine-tuned on the original dataset performed better than those trained on the balanced dataset. We observed an increase in recall but a more significant decrease in precision. This might be because the balanced training sets had 33% less data than the original dataset, which was necessary to achieve balance. Nevertheless, the fine-tuned emotion model, even when trained on the balanced dataset, demonstrated superior performance in all metrics compared to the other models.

4.5 Statistical tests

Statistical hypothesis tests were conducted to validate the observed differences in model performance. The Friedman Test was applied to four subsamples of the test dataset, aiming to ascertain whether the models' performances were statistically indistinguishable. The Friedman estimator, denoted as χ_F^2, is calculated as follows:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right]$$

where N is the number of blocks (subsamples), k is the number of treatments (models), and R_j is the rank sum for the j^{th} treatment across all blocks. The Friedman estimator yielded a value of 10.799, exceeding the critical value of 6.992 at a 99% confidence level, thereby rejecting the null hypothesis H_0. This suggests that the models do not perform equally.

Subsequently, the Nemenyi Test was employed to perform pairwise comparisons between the models. The performance of two classifiers is significantly different if the corresponding average ranks differ by at least the critical difference denoted as CD that is calculated as:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

where critical values q_α are based on the Studentized range statistic divided by $\sqrt{2}$ [6]. The test indicated that the *Emotion* model had ranking differences exceeding the $CD = 2.345$ for the baseline and the *Sentiment Analysis* model based on *DistilBERT*, thereby rejecting the null hypothesis H_0 at a 95% confidence level. This confirms that the *Emotion* model significantly outperforms both the baseline *BERT*$_{HT}$ and *DistilBERT* within the context of human trafficking risk prediction. See Appendix A for more details.

5 Discussion

The empirical evaluation elucidates that the fine-tuned *Emotion* model predicated on the *BERT* architecture outperforms the baseline and the *DistilBERT*-based *Sentiment Analysis* model. Utilizing rigorous statistical hypothesis testing, we were able to substantiate this observation with a 95% confidence interval. However, it is important to note that the statistical tests were inconclusive in establishing the superiority of the *Sentiment Analysis* model over the baseline model.

A salient aspect of the experimentation lies in the class imbalance inherent in the dataset. As delineated in Section 4.1, the dataset exhibits a disproportionate representation of the "low-risk" class relative to the "high-risk" class. Although the results from the class imbalance mitigation didn't show any indication of this imbalance being an issue, it will be very beneficial for future research to count on a bigger and more balanced dataset. Future models could be susceptible to this disproportion and cause model bias. The implications of such bias are far-reaching, especially considering the critical nature of the human trafficking risk prediction task.

More varied data will allow us to further verify our hypothesis. That is, we could deepen our experiments to evaluate the impact of NLP task-specific features in the human trafficking risk prediction task.

6 Conclusion

The primary objective of this research was to rigorously evaluate the efficacy of task-specific fine-tuned models in comparison to a baseline model fine-tuned on the overarching task of human trafficking risk prediction. Our empirical findings indicate that the fine-tuned *Emotion* model, based on the *BERT* architecture, exhibited superior performance metrics. The observed differences in model performance were statistically significant, as corroborated by the Friedman and Nemenyi statistical tests. However, the performance of the *Sentiment Analysis* model was not statistically better than the baseline.

The inconclusive nature of these results raises questions about the inherent utility of sentiment in enhancing the predictive capabilities for human trafficking risk assessment. Specifically, our findings suggest that some task-specific features may not necessarily contribute to a significant improvement in the identification of human trafficking instances from noisy textual data. This has implications for the broader field of human trafficking research, indicating that the focus may need to shift towards other Natural Language Processing tasks, such as Named Entity Recognition, which could potentially offer substantial contributions to the risk prediction task.

We recommend future research to employ a balanced and bigger dataset. This would enable a more robust evaluation of model performance and could potentially yield more conclusive insights into the utility of task-specific fine-tuned models for human trafficking risk prediction.

References

[1] M. Abas, N. Ostrovschi, M. Prince, V. Gorceag, C. Trigub, and S. Oram. Risk factors for mental disorders in women survivors of human trafficking: a historical cohort study. *BMC Psychiatry*, 13, 2013.

[2] Rodriguez Perez Alejandro, Korn Sooksatra, Pablo Rivas, Ernesto Quevedo Caballero, Javier S. Turek, Gisela Bichler, Tomas Cerny, Laurie Giddens, and Stacie Petter. Aligning word embeddings from bert to vocabulary-free representations. In *The 25th International Conference on Artificial Intelligence (ICAI 2023)*, pages 1–8, 2023.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[5] M. Decker, H. McCauley, D. Phuengsamran, S. Janyam, and J. Silverman. Sex trafficking, sexual risk, sexually transmitted infection and reproductive health among female sex workers in thailand. *Journal of Epidemiology & Community Health*, 65:334–339, 2010.

[6] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Saeideh Shahrokh Esfahani, Michael J Cafarella, Maziyar Baran Pouyan, Gregory DeAngelo, Elena Eneva, and Andy E Fano. Context-specific language modeling for human trafficking detection from online advertisements. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1180–1184, 2019.

[9] Laurie Giddens, Stacie Petter, Gisela Bichler, Pablo Rivas, Michael H Fullilove, and Tomas Cerny. Navigating an interdisciplinary approach to cybercrime research. In *Proceedings of the 56th Hawaii International Conference on System Sciences*, 2023.

[10] L. Giommoni and R. Ikwu. Identifying human trafficking indicators in the uk online sex market. *Trends in Organized Crime*, 2021.

[11] Ruoting Li, Margaret Tobey, Maria E Mayorga, Sherrie Caltagirone, and Osman Y Özaltın. Detecting human trafficking: Automated classification of online customer reviews of massage businesses. *Manufacturing & Service Operations Management*, 2023.

[12] X. L'Hoiry, A. Moretti, and G. Antonopoulos. Identifying sex trafficking in adult services websites: an exploratory study with a british police force. *Trends in Organized Crime*, 2021.

[13] N. Mai. The fractal queerness of non-heteronormative migrants working in the uk sex industry. *Sexualities*, 15:570–585, 2012.

[14] J. Mendel and K. Sharapov. Human trafficking and online networks: policy, analysis, and ignorance. *Antipode*, 48:665–684, 2016.

[15] S. Mostajabian, D. Maria, C. Wiemann, E. Newlin, and C. Bocchini. Identifying sexual and labor exploitation among sheltered youth experiencing homelessness: a comparison of screening methods. *International Journal of Environmental Research and Public Health*, 16:363, 2019.

[16] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

[17] Pablo Rivas, Gisela Bichler, Tomas Cerny, Laurie Giddens, and Stacie Petter. Bottleneck-based encoder-decoder architecture (bear) for learning unbiased consumer-to-consumer image representations. In *LXAI Workshop @ International Conference on Machine Learning (ICML 2022)*, pages 1–7, 2022.

[18] Pablo Rivas and Liang Zhao. Clip-acqua: Clip autoencoder-based classic-quantum latent space reduction. In *The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'22)*, 2022.

[19] Alejandro Rodriguez Perez, Korn Sooksatra, Pablo Rivas, Ernesto Quevedo Caballero, Javier S. Turek, Gisela Bichler, Tomas Cerny, Laurie Giddens, and Stacie Petter. An empirical analysis towards replacing vocabulary-rigid embeddings by a vocabulary-free mechanism. In *LXAI Workshop @ International Conference on Machine Learning (ICML 2023)*, pages 1–7, 2023.

[20] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[21] C. Schwarz, D. Alvord, D. Daley, M. Ramaswamy, E. Rauscher, and H. Britton. The trafficking continuum: service providers' perspectives on vulnerability, exploitation, and trafficking. *Affilia*, 34:116–132, 2018.

[22] D. Shepherd, V. Parida, T. Williams, and J. Wincent. Organizing the exploitation of vulnerable people: a qualitative assessment of human trafficking. *Journal of Management*, 48:2421–2457, 2021.

[23] Pedro Szekely, Craig A Knoblock, Jason Slepicka, Andrew Philpot, Amandeep Singh, Chengye Yin, Dipsy Kapoor, Prem Natarajan, Daniel Marcu, Kevin Knight, et al. Building and using a knowledge graph to combat human trafficking. In *The Semantic Web-ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II 14*, pages 205–221. Springer, 2015.

[24] J. Todres and A. Diaz. Covid-19 and human trafficking—the amplified impact on vulnerable populations. *Jama Pediatrics*, 175:123, 2021.

[25] Edmund Tong, Amir Zadeh, Cara Jones, and Louis-Philippe Morency. Combating human trafficking with multimodal deep models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1547–1556, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[26] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models, 2019.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[28] Jessica Whitney, Murray Jennex, Aaron Elkins, and Eric Frost. Don't want to get caught? don't say it: The use of emojis in online human sex trafficking ads. 2018.

Ethics Statement

This study relies solely on internally scraped and curated datasets and *BERT* models, and does not involve humans as subjects; however, the original data contains personally identifiable information that prevents the investigators from releasing the dataset to the general public. The methodology is supported by internal review board (IRB) approval at Baylor University.

Further, while we have vetted our model regarding ethical considerations, we acknowledge that it may inherit biases in the original *BERT* embeddings. We emphasize the need for further research to mitigate these biases and are committed to methodological transparency.

Acknowledgements

This work is a culmination of efforts originating from a graduate NLP course under the tutelage of Dr. Pablo Rivas at Baylor University, subsequently fostered by a collaborative grant. We thank Dr. Gisela Bichler of California State University San Bernardino, Dr. Laurie Giddens at the University of North Texas, and Dr. Stacie Petter at Wake Forest University for their expertise. Special thanks to Dr. Tomas Cerny of the University of Arizona and Dr. Javier Turek at Intel Labs for their expert contributions in machine learning. The rigorous dedication of Ph.D. students Korn Sooksatra and Maisha Rashid Binte has been exemplary in advancing our research dialogues.

This work is supported by the National Science Foundation under Grant Nos. 2039678, 2136961, and 2210091. The views expressed herein are solely those of the author and do not necessarily reflect those of the National Science Foundation.

A Appendix: Statistical Hypothesis Tests

A.1 Friedman Test

For the Friedman Test, we use a total of 4 subsamples of the testing dataset. The goal is to test whether the four models perform equally or not. Table 3 shows the Accuracy score obtained from the different models on each subsample. For each of the subsamples, we determined the resulting mean ranking. This resulted in the values shown in parentheses.

Table 3: Models Performance Accuracy for the four different subsamples.

Subsamples	Baseline	Emotion	SA (BERT)	SA (DistilBERT)
0	0.8397	0.8525	0.8416	0.8358
1	0.8375	0.8451	0.8426	0.8383
2	0.8403	0.8524	0.8426	0.8381
3	0.8338	0.8481	0.8410	0.8342
Avg. (Rank)	0.8378 (3.5)	0.8495 (1)	0.8419 (2)	0.8366 (3.5)

The Friedman estimator for this ranking is equal to 10.799. The critical value at $\alpha = 0.01$ is 6.992. $10.799 > 6.992$, thus, with 99% confidence, H_0 is rejected and we can say that models don't perform equally.

A.2 Nemenyi Test

The Nemenyi estimator for the four subsamples and four models, with a confidence level of 95%, is equal to 2.345. The critical value for $\alpha = 0.05$ is equal to 2.569. Table 4 shows the ranking differences for every pair of models.

Table 4: Differences of models' ranking. SA: Sentiment Analysis.

	Baseline	Emotion	SA (*BERT*)	SA (*DistilBERT*)
Baseline	0			
Emotion	2.5	0		
SA (*BERT*)	1.5	1	0	
SA (*DistilBERT*)	0	2.5	1.5	0

The ranking differences between the *Emotion* model predicated on the *BERT* architecture and the *Sentiment Analysis* model utilizing *DistilBERT*, as well as between the *Emotion* model based on *BERT* and the baseline model also employing *BERT*, exceed the critical threshold set by the Nemenyi estimator ($2.5 > 2.345$). Consequently, at a 95% confidence level, we reject the null hypothesis H_0, thereby substantiating that the performance of the *Emotion* model is statistically distinguishable from both the baseline model and the *Sentiment Analysis* model based on *DistilBERT*.