International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-14 Issue-7, June 2025

RiskBERT: A Pre-Trained Insurance-Based Language Model for Text Classification

Rida Ghafoor Hussain

Abstract: ***The rapid growth of insurance-related documents has increased the need for efficient and accurate text classification techniques. Advances in natural language processing (NLP) and deep learning have enabled the extraction of valuable insights from textual data, particularly in specialised domains such as insurance, legal, and scientific documents. While Bidirectional Encoder Representations from Transformers (BERT) models have demonstrated state-of-the-art performance across various NLP tasks, their application to domain-specific corpora often results in suboptimal accuracy due to linguistic and contextual differences. In this study, I propose RiskBERT, a domain-specific language representation model pre-trained on insurance corpora. I further pre-trained LegalBERT on insurance-specific datasets to enhance its understanding of insurance-related texts. The resulting model, RiskBERT, was then evaluated on downstream clause and provision classification tasks using two benchmark datasets – LEDGAR [1] and Unfair ToS [2]. I conducted a comparative analysis against BERT-Base and LegalBERT to assess the impact of domain-specific pre-training on classification performance. The findings demonstrate that pre-training on insurance-specific corpora significantly improves the model's ability to analyse complex insurance texts. The experimental results show that RiskBERT significantly outperforms LegalBERT and BERT-Base, achieving superior accuracy in classifying complex insurance texts, with 96.8% and 92.1% accuracy, respectively, on the LEDGAR and Unfair ToS datasets. These findings highlight the effectiveness of domain-adaptive pre-training and underscore the importance of specialised language models for enhancing insurance document processing, making RiskBERT a valuable tool for this purpose.***

Keywords: ***Clause, Domain-Specific, Insurance, Legal, Pre-Training,***

Abbreviations:
BERT: Bidirectional Encoder Representations from Transformers
PLL: Pseudo-Log-Likelihood
HER: Electronic Health Records
ASR: Automatic Speech Recognition
LMs: Language Models
WER: Word Error Rates
XAI: Explainable AI
KPA: Knowledge Process Automation
LDMM: Loss Dirichlet Multinomial Mixture
MLM: Masked Language Modelling
RMSE: Root Mean Square Error
ECCO: Eighteenth Century Collections Online

I. INTRODUCTION

The volume of insurance-related literature is rapidly Expanding, incorporating new insights and increasing the complexity of insurance corpora. Consequently, there is a growing need for advanced research and analysis to enhance the accuracy of insurance text classification and data mining. The ability to efficiently extract and categorize information from vast insurance documents is crucial for policy analysis, risk assessment, and compliance monitoring. Recent advancements in natural language processing (NLP) and deep learning have significantly improved text classification across various specialized domains, such as biomedical, scientific and legal corpora extracting valuable insights from textual data.

Transformer-based pre-trained language models [3] have gained considerable attention for their ability to produce state-of-the-art results in various NLP tasks. Models such as Bidirectional Encoder Representations from Transformers (BERT) [4], along with its extensions (Liu et al. [5]; Yang et al. [6]; Lan et al. [7]), have become the foundation for modern text classification research. While these models are publicly available, many studies bypass the computationally intensive pre-training phase and focus on fine-tuning for specific downstream tasks. However, directly applying general-purpose BERT models to specialized domains presents limitations, as they often fail to capture domain-specific terminology and contextual nuances.

Since BERT is pre-trained on general corpora such as Wikipedia and BookCorpus, it tends to underperform when applied to highly specialized fields like insurance, legal, and scientific texts [8]. A potential solution to this challenge is domain-adaptive pre-training, in which BERT models are further pre-trained on domain-specific corpora to improve their performance.

This study examines this strategy in the insurance domain, where specialised corpora contain extensive domain-specific terminology, distinct vocabulary, and unique formal semantics and syntax compared to general-domain texts. The distributional differences between insurance and general corpora make it difficult for general-purpose BERT models to generalize effectively, necessitating the development of domain-specific BERT variants.

This research contributes to a deeper understanding of BERT adaptation in specialized domains by proposing RiskBERT, a domain-specific language model created by further pre-training LegalBERT [9] on insurance corpora. Given LegalBERT's strong performance in legal NLP tasks, adapting it for insurance-related text analysis could offer substantial benefits. The key findings indicate that RiskBERT significantly outperforms the general BERT model and the baseline LegalBERT on insurance clause and provision classification tasks. The model's effectiveness is evaluated on two benchmark datasets [1], demonstrating its potential for insurance document analysis and classification [2].

The paper is structured as follows: Section 2 provides a literature review of domain-adaptive

Manuscript received on 19 April 2025 | First Revised Manuscript received on 27 April 2025 | Second Revised Manuscript received on 16 May 2025 | Manuscript Accepted on 15 June 2025 | Manuscript published on 30 June 2025.
*Correspondence Author(s)
Rida Ghafoor Hussain*, Researcher, Department of Information Engineering, University Of Florence, Siena Italy. Email ID: rida.ghafoor@unifi.it, ORCID ID: 0000-0002-5646-5910

BERT models. Section 3 details the proposed methodology and model pre-training process. Section 4 presents the experimental setup and results, while Section 5 concludes the study with key findings and future research directions.

II. RELATED WORK

Most prior research on BERT domain adaptation has focused on fields such as biomedical, scientific, and legal domains. In [8], the BIOBERT model was introduced by further pre-training BERT-BASE on biomedical corpora. While this model showed notable improvements over the baseline, increasing the number of pre-training steps did not yield further enhancements. Similarly, Alsentzer et al. [10] developed Clinical BERT and Clinical BIOBERT by further pre-training BERT-BASE and BIOBERT on clinical notes, demonstrating superior performance compared to the base models.

SCIBERT [11], another specialized BERT variant, was designed for the scientific domain, particularly biomedical literature. It was obtained either by further pre-training BERT-BASE or training it from scratch (SC) on biomedical scientific corpora, yielding improvements in downstream tasks. Likewise, Sung et al. [12] enhanced short answer evaluations for intelligent tutoring systems by further pre-training BERT-BASE on textbook data.

In the legal domain, Moens et al. [13] explored supervised classifiers for extracting arguments from legal documents, while other studies have tackled claim detection [14], citation of legal principles and facts in judicial decisions [15], and judicial desion prediction [16]. Fabian et al. [17] examined machine learning techniques for analyzing private policies, whereas Ashley and Walker [18] conducted a case study on constructing legal arguments for vaccine injury compensation using NLP tools.

In the healthcare sector, Laila et al. [19] proposed Med-BERT, a contextualized embedding model adapted for Electronic Health Records (EHR). Tested on disease prediction tasks, Med-BERT outperformed larger baseline models. The study also demonstrated the feasibility of structured EHR data and the effectiveness of contextualized embeddings in capturing EHR semantics.

Other research has explored BERT adaptation in various languages and specialized domains. Hend Zouari [11] developed French AXA Insurance BERT, comparing its performance with BERT and Camembert for insurance domain adaptation. The results showed that Camembert was more adaptable to small insurance datasets, while BERT performed better without fine-tuning. Thomas et al. [20] adapted Swedish BERT by de-identifying clinical corpora in Swedish, demonstrating improved performance across six clinical downstream tasks.

In historical document analysis, Iiro et al. [21] used BERT to predict the publication year of digitized texts from the Eighteenth Century Collections Online (ECCO) dataset. The model achieved an average absolute error of fewer than seven years, outperforming linear models and human judgment.

Pre-trained language models have also been extensively analyzed for their general capabilities. Hang et al. [22] highlighted BERT's remarkable performance as a universal NLP tool, emphasizing its ability to process domain-specific content effectively. In another study, Rukhma et al. [23] applied transfer learning for classifying COVID-19 fake news and extremist content, evaluating the model using various metrics.

Mikhail et al. [24] conducted a comprehensive survey of BERT applications, summarizing key areas of implementation and comparing BERT to other NLP models. Ather et al. [25] reviewed different BERT models and embedding techniques, focusing on their numerical vector representations for word analysis. In [26], researchers explored BERT-based contextualized language models (LMs) for automatic speech recognition (ASR), using N-best hypothesis reranking to minimize word error rates (WER). Their approach outperformed **RNNs** and BERT models that relied on pseudo-log-likelihood (PLL) scores.

The insurance domain has received comparatively less attention. Wang et al. [27] proposed a BERT-based knowledge extraction method to extract key knowledge points from unstructured insurance clauses, reducing manual labor in knowledge graph construction. Unlike conventional template- and rule-based approaches, this model generates question-answer pairs from domain knowledge and contextualizes responses.

In the financial sector, FILBERT [28] was developed to address legal and risk aspects beyond standard financial language models. This model was designed to enhance knowledge process automation (KPA) solutions for firms, improving productivity and deployment.

This study presents a framework for integrating advanced analytics in insurance, leveraging machine learning and real-time data for improved risk assessment and claims processing. The approach enhances underwriting accuracy and operational efficiency by incorporating diverse data sources, such as IoT sensors and telematics. The study also addresses challenges in data quality, regulatory compliance, and integration, offering strategic recommendations for insurers seeking to adopt digital transformation.

This study [29] explores the impact of big data analytics on the insurance industry, focusing on risk classification, privacy concerns, and economic implications. It proposes a framework to analyze the trade-off between classification accuracy and policyholder incentives for data sharing. The findings highlight that using private data for risk classification does not always drive insurers to create more granular risk classes.

Another paper [30] worked on vehicle insurance fraud detection using advanced machine learning models, including a stacked ensemble approach integrated with Explainable AI (XAI) techniques. Feature selection algorithms identified key features that influence fraud detection, thereby improving model interpretability and accuracy. The study highlights future research directions, such as behavioural biometrics, blockchain security, and deep reinforcement learning for adaptive fraud prevention.

Big data analytics is transforming the insurance industry by enhancing predictive accuracy, reducing fraudulent claims by 28.9%, and improving the efficiency of risk assessment. AI-driven systems streamline claim processing with 94.7% accuracy, leading to a 34.7-point increase in customer satisfaction. Emerging technologies, such as edge AI and quantum computing, promise further advancements, reducing processing latency and enhancing prediction accuracy. [31]

The Loss Dirichlet Multinomial Mixture (LDMM) model [32] enhances individual claims reserving by integrating Bayesian probabilistic methods with textual claim descriptions. This approach effectively captures multimodality, skewness, and

heavy-tailed claim characteristics, improving predictive accuracy.

The model's ability to leverage unstructured data supports digital transformation in the insurance sector, enabling more informed decision-making.

For insurance-related NLP tasks, Anuj et al. [33] introduced IBLMs (Insurance-Based Language Models) by further pre-training domain-specific language models (ULMFiT and BERT) on insurance claim notes with an expanded vocabulary. These models achieved superior results in binary classification tasks compared to traditional approaches, offering practical benefits for insurance claims analysis. Similarly, Shan et al. [34] employed BERT for intent classification on insurance data, significantly reducing error rates compared to alternative models such as TextCNN, HAN, and ELMo**.**

Despite BERT's dominance as a state-of-the-art NLP model, its adaptation to the insurance domain remains largely unexplored. In specialized fields, the use of excessively large and computationally expensive models is often unnecessary, as domain-specific terminology and syntax are more structured and standardized. As artificial intelligence, particularly natural language processing**,** gains traction in insurance document analysis and classification, further research is needed to develop optimized models for this domain.

III. METHODOLOGY

A. RiskBERT: Adaptive Pre-training on Insurance Corpora

I introduce RiskBERT, a domain-adapted language model designed explicitly for insurance-related NLP tasks in this research. This approach involves further pre-training a LegalBERT model [9] on an insurance-specific corpus to enhance its performance on downstream insurance text classification tasks.

For pre-training, I collected insurance-related textual data from the EDGAR database, which includes a diverse set of documents such as insurance policies, contracts, treaties, and schedules. A pre-trained tokeniser, consistent with the vocabulary used in the original LegalBERT training, was employed to tokenise all textual data. I used Masked Language Modelling (MLM) as the training objective to effectively adapt the model. MLM randomly masks specific tokens in the input text, and the model is trained to predict these masked tokens. To optimize this process, the tokenized text was concatenated and divided into chunks to maintain a consistent sequence length. Additionally, a data collator was employed for dynamic random masking, ensuring variation in masked tokens across different training epochs. This approach prevents the model from overfitting to fixed masked patterns and enables it to learn more generalized representations, ultimately enhancing its adaptability and performance.

B. Fine-Tuning for Downstream Classification Tasks

After pre-training, RiskBERT was fine-tuned for a multi-class clause/provision classification task using two benchmark datasets: Ledgar [1], which consists of legal contract provisions, and Unfair ToS [2], containing clauses extracted from Terms of Service agreements. To assess the effectiveness of RiskBERT, a comparative performance analysis was conducted against two baseline models—BERT-Base, a general-domain model, and LegalBERT, which is pre-trained on legal texts. This evaluation aimed to demonstrate the benefits of domain-specific pre-training in improving classification accuracy for specialized insurance-related texts.

IV. EXPERIMENTS AND RESULTS

This section presents the experimental setup and results obtained in adapting BERT for the insurance domain. The primary objective of the experiments is to evaluate the performance of RiskBERT, a model derived by further pre-training LegalBERT on insurance-specific corpora, and to compare its classification effectiveness with that of state-of-the-art models. The evaluation focuses on clause/provision classification using the LEDGAR and Unfair Terms of Service (ToS) datasets.

A. Data Preprocessing and Experimental Setup

The Ledgar dataset, initially available in JSON format, and the Unfair ToS dataset, provided in XML format, were first converted into CSV format using dataframes to enhance readability and facilitate structured analysis. The datasets were then randomly split into training (80%), validation (10%), and test (10%) sets to ensure an unbiased evaluation. For fine-tuning, the Adam optimizer was employed, with categorical cross-entropy as the loss function, given the multi-class classification nature of the task. After training, the model's performance was evaluated on the test set using standard metrics, including accuracy, F1-score, precision, and recall. This systematic methodology ensures that RiskBERT is effectively adapted to the insurance domain, enhancing its capability for specialized text classification tasks.

B. Pre-training of RiskBERT

To construct RiskBERT, I performed additional pre-training on an insurance-specific corpus. The LegalBERT model was used as the base model due to its prior adaptation to legal text. The training corpus comprises insurance-related documents collected from EDGAR, including policies, contracts, treaties, and schedules. Pre-training was performed using the masked language modelling (MLM) objective, where random tokens were masked, and the model was trained to predict them based on context. The tokenization process utilized a pretrained tokenizer to maintain consistency with the original LegalBERT vocabulary.

For pre-training, the model was configured with the following hyperparameters: a batch size of 25, a maximum sequence length of 128 tokens using Sentence Piece tokenization, a learning rate of 2e-5, and a weight decay of 0.01. A masked language modelling (MLM) probability of 0.15 was also applied to facilitate effective contextual learning. To assess the effectiveness of pre-training, perplexity was used as a key evaluation metric, indicating how well the model learned from the insurance corpus. A lower perplexity score signifies improved language modelling capability, demonstrating the model's ability to better understand and represent domain-specific insurance texts.

C. Fine-tuning for Clause/Provision Classification

After pre-training, RiskBERT was fine-tuned on downstream classification tasks to evaluate its effectiveness in clause and provision classification. The fine-tuning process was conducted using two benchmark datasets: LEDGAR, a large-scale dataset comprising contractual provisions extracted from legal agreements, and Unfair ToS, which contains 50 online consumer contracts annotated for unfair terms of service. To optimize

performance, dataset-specific hyperparameters were applied during training. For LEDGAR, the model was trained using a learning rate 1e-5 for five epochs, with a batch size of 64 and a maximum sequence length of 512 tokens, while maintaining a fixed dropout rate of 0.5. Similarly, for the Unfair ToS dataset, the model was trained for 20 epochs with a learning rate of 1e-6, while keeping other hyperparameters consistent with those used in LEDGAR fine-tuning. To ensure robust performance evaluation, both datasets were split into training (80%), validation (10%), and testing (10%) sets, allowing for a comprehensive assessment of RiskBERT's classification capabilities in domain-specific legal and insurance text processing.

D. Comparative Analysis and Performance Evaluation

RiskBERT was compared against LegalBERT and the general BERT model to provide a comprehensive evaluation. Each model was fine-tuned on the same downstream tasks and datasets, using identical experimental settings, to ensure a fair comparative analysis.

While experimental settings were consistent across models, minor dataset-specific variations in hyperparameters were introduced to enhance performance. This approach ensures that each dataset is optimally leveraged while maintaining the validity of comparisons across models. The results from these experiments demonstrate the effectiveness of RiskBERT in insurance text classification, highlighting the impact of domain-specific pre-training in improving the accuracy and robustness of NLP models in specialized domains.

E. Pre-Training and Fine-Tuning Results

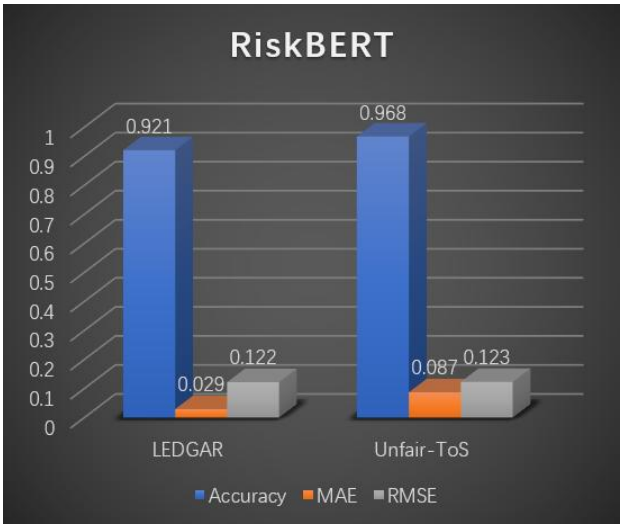
In this section, I present the results of pre-training and fine-tuning the RiskBERT model on insurance-specific corpora. The results demonstrate the effectiveness of RiskBERT in the context of domain-specific insurance text classification tasks, as well as its performance comparison with LegalBERT and BERT-Base on downstream tasks.

i. Pre-training Performance

The RiskBERT model, created by adapting LegalBERT to an insurance-specific corpus, demonstrates fast and effective adaptation to the domain. The perplexity of RiskBERT after pre-training on the insurance corpus is 1.32, indicating that the model has effectively learned the domain-specific language. A lower perplexity suggests that the model can generate meaningful representations of the insurance text, subsequently improving performance on downstream tasks.

ii. Fine-Tuning and Comparative Analysis

The fine-tuning of RiskBERT was conducted across both the LEDGAR and Unfair Terms of Service (ToS) datasets. Results from the fine-tuning experiments on these datasets indicate that RiskBERT consistently outperforms the baseline LegalBERT and BERT-Base models on all evaluation metrics. This suggests that RiskBERT's domain-specific pre-training offers substantial performance advantages, particularly for insurance-related text classification tasks.



[Fig.1: Fine-Tuning Result Performance of RiskBERT Over both Datasets. Accuracy, Mean Square Error (MAE), Root Mean Square Error (RMSE) are Plotted]

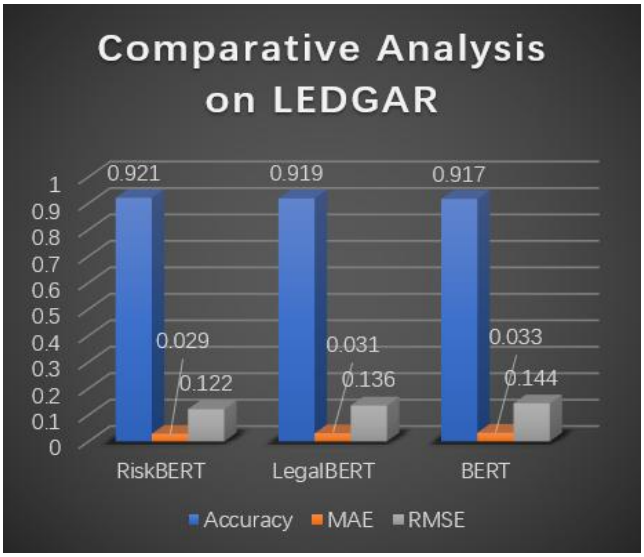
Figure 1 illustrates Riskbert's performance across the development data of both datasets, showcasing its superior adaptability to insurance corpora compared to the other models. On the test data, RiskBERT consistently produces better results in terms of accuracy, root mean square error (RMSE), and mean absolute error (MAE) when compared to the baseline models (LegalBERT and BERT-Base).

An impressive advantage of RiskBERT is its computational efficiency. Despite its enhanced domain-specific performance, RiskBERT is comparable to LegalBERT and BERT-Base in terms of resource usage and can be trained on standard GPU configurations. This makes RiskBERT a viable option for researchers with limited computational resources, ensuring its practicality for broader adoption in insurance text analysis.

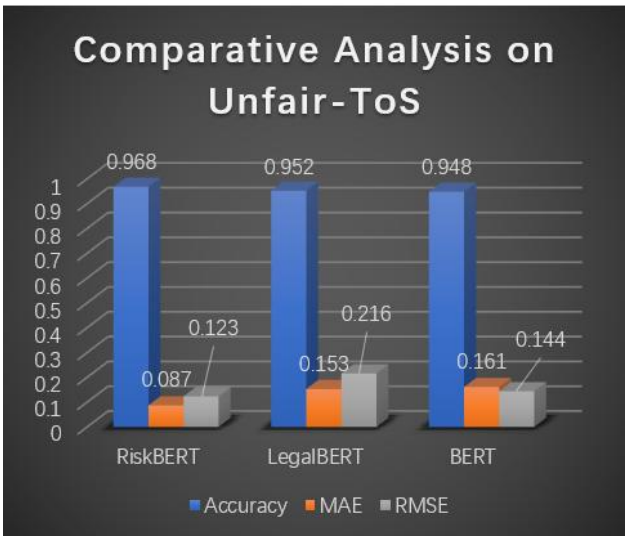
Table-I: Test Results on Datasets: Unfair-ToS and LEDGAR. Accuracy, Mean Square Error (MAE), Root Mean Square Error (RMSE) are Reported in the Comparative Analysis of Each Model. The Best Scores are in Bold, which were Obtained from the RiskBERT

Models	Accuracy		MAE		RMSE	
	Unfair -ToS	Ledgar	Unfair- ToS	Ledgar	Unfair -ToS	Ledgar
RiskBERT	0.968	0.921	0.087	0.029	0.123	0.122
LegalBERT	0.952	0.919	0.153	0.031	0.216	0.136
BERT	0.948	0.917	0.161	0.033	0.227	0.144

Table 1 summarizes Riskbert's performance improvements over LegalBERT and BERT-Base on both datasets. I report accuracy, RMSE, and MAE values, which highlight RiskBERT's superior performance compared to the baseline models across both datasets.



[Fig.2: Comparative Analysis on LEDGAR for Three Models. Accuracy, Mean Square Error (MAE), Root Mean Square Error (RMSE) are plotted for Each Model (RiskBERT, LegalBERT, BERT) Calculated During Fine-Tuning on the downstream task on Ledgar]



[Fig. 3: Comparative Analysis on Unfair-ToS for Three Models. Accuracy, Mean Square Error (MAE), Root Mean Square Error (RMSE) are Plotted for Each Model (RiskBERT, LegalBERT, BERT) Calculated During Fine-tuning on the Down-Stream Task on Unfair-ToS]

Figures 2 and 3 display the individual performance of each model on the LEDGAR and Unfair ToS datasets, respectively, providing a clear visual comparison of RiskBERT against the baseline models for each dataset. These figures demonstrate that RiskBERT outperforms in accuracy and exhibits lower RMSE and MAE values across both datasets.

The results confirm that RiskBERT outperforms baseline models in terms of classification accuracy and error metrics, demonstrating the feasibility of adapting BERT models for specialised domains such as insurance. These findings open the door to new research directions focused on further domain-specific adaptations of BERT models, enabling improved performance across specialized NLP tasks in other domains as well. The approach of fine-tuning pre-trained models with domain-specific corpora can significantly enhance the capabilities of transformer-based models in various industry sectors, including insurance, legal, and healthcare.

V. CONCLUSION AND FUTURE WORK

This study examined the development and application of the RiskBERT model for domain adaptation in the insurance sector, highlighting its potential to improve industry-specific natural language processing tasks. By further pre-training LegalBERT on insurance-specific corpora, I introduced RiskBERT, a model specifically tailored for insurance-related text classification tasks, such as clause and provision classification. The experiments demonstrated that RiskBERT outperforms the baseline models, including LegalBERT and BERT-Base, across multiple evaluation metrics, including accuracy and mean absolute error (MAE). The significant performance improvements highlight RiskBERT's potential to enhance classification tasks within the insurance domain, with minimal task-specific modifications required.

The results suggest that RiskBERT offers a promising solution for processing insurance documents, providing a more efficient and accurate approach than existing models. The findings highlight the potential for further pre-training models on domain-specific corpora to enhance their effectiveness in specialised fields.

For future work, I plan to evaluate the efficiency and scalability of RiskBERT on a broader range of insurance datasets and tasks. Additionally, I aim to explore the impact of further pre-training on sub-domains within the insurance sector to assess how domain-specific nuances can further improve model performance. i also intend to expand the research by testing other BERT-based models and their adaptations for domain-specific tasks. This may contribute to the development of more robust solutions for insurance document classification and other similar applications.

DECLARATION STATEMENT

I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** No organisation or agency has funded this article. This independence ensures that the research is conducted objectively and without external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** This article's authorship is solely contributed by the author.

REFERENCES

1. Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 1235–1241, Marseille, France. European Language Resources Association. https://aclanthology.org/2020.lrec-1.155.pdf
2. Lippi, M., Pałka, P., Contissa, G. et al. CLAUDETTE: an automated detector of potentially unfair

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

clauses in online terms of service. Artif Intell Law 27, 117–139 (2019). DOI: https://doi.org/10.1007/s10506-019-09243-2
3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need in *31th Annual Conference on Neural Information Processing Systems*, Long Beach, CA, USA. DOI: https://doi.org/10.48550/arXiv.1706.03762
4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, DOI: https://doi.org/10.48550/arXiv.1810.04805
5. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR, DOI: https://doi.org/10.48550/arXiv.1907.11692
6. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. CoRR, DOI: https://doi.org/10.48550/arXiv.1906.08237
7. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. CoRR, DOI: https://doi.org/10.48550/arXiv.1909.11942
8. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. In CoRR. DOI: https://doi.org/10.48550/arXiv.1901.08746
9. Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2898–2904, Online. Association for Computational Linguistics. https://aclanthology.org/2020.findings-emnlp.261.pdf
10. Emily Alsentzer, John Murphy, William Boag, WeiHung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 72–78, Minneapolis, Minnesota, USA. DOI: https://doi.org/10.48550/arXiv.1904.03323
11. Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3606– 3611, Hong Kong, China. https://aclanthology.org/D19-1371.pdf
12. Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. 2019. Pre-training BERT on domain resources for short answer grading. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP), pages 6071–6075, Hong Kong, China. Association for Computational Linguistics. https://aclanthology.org/D19-1628.pdf
13. Moens MF, Boiy E, Palau RM, Reed C (2007). Automatic detection of arguments in legal texts. In: Proceedings of the 11th International Conference on Artificial Intelligence and Law, ACM, pp 225 -230. DOI: https://doi.org/10.1145/1276318.1276362
14. Lippi M, Torroni P (2016a) Argumentation mining: State of the art and emerging trends. ACM Trans Internet Technol 16(2):10:1{10:25. DOI: https://doi.org/10.1145/2850417
15. Shulayeva O, Siddharthan A, Wyner A (2017) Recognizing cited facts and principles in legal judgements. Artificial Intelligence and Law 25(1):107{126. https://link.springer.com/article/10.1007/s10506-017-9197-6
16. Aletras N, Tsarapatsanis D, Preoiuc-Pietro D, Lampos V (2016) Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. PeerJ Computer Science 2:e93. https://peerj.com/articles/cs-93/
17. Fabian B, Ermakova T, Lentz T (2017) Large-scale readability analysis of privacy policies. In: Proceedings of the International Conference on Web Intelligence, ACM, pp 18- 25. DOI: https://doi.org/10.1145/3106426.3106427
18. Ashley KD, Walker VR (2013) Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law, ACM, pp 176-180. https://sites.hofstra.edu/vern-walker/wp-content/uploads/sites/69/2019/12/Ashley_Walker-Toward_Evidence-Based_Legal_Arguments-ICAIL2013.pdf
19. Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, Degui Zhi "Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction in Digital Medicine, 2021. https://www.nature.com/articles/s41746-021-00455-y
20. Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, Hercules Dalianis. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4245–4252, Marseille, 20-25 June 2022. https://aclanthology.org/2022.lrec-1.451/
21. Iiro Rastas, Yann Ciarán Ryan, Iiro Tiihonen, Mohammadreza Qaraei, Liina Repo, Rohit Babbar, Eetu Mäkelä, Mikko Tolonen, and Filip Ginter. 2022. Explainable Publication Year Prediction of Eighteenth Century Texts with the BERT Model. In Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 68–77, Dublin, Ireland. Association for Computational Linguistics. https://aclanthology.org/2022.lchange-1.7.pdf
22. Hang Li. Language Models: Past, Present, and Future. Communications of the ACM, July 2022, Vol. 65 No. 7, Pages 56-63, https://doi.org/10.1145/3490443
23. Rukhma Qasim, Waqas Haider Bangyal, Mohammed A. Alqarni, Abdulwahab Ali Almazroi, "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification", Journal of Healthcare Engineering, vol. 2022, Article ID 3498123, 17 pages, 2022. DOI: https://doi.org/10.1155/2022/3498123
24. Koroteev, Mikhail. (2021). BERT: A Review of Applications in Natural Language Processing and Understanding. DOI: https://doi.org/10.48550/arXiv.2103.11943
25. Athar Hussein Mohammed and Ali H. Ali 2021 *J. Phys.: Conf. Ser.*. 1963 012173. https://iopscience.iop.org/article/10.1088/1742-6596/1963/1/012173/pdf
26. Chiu, Shih-Hsuan & Chen, Berlin. (2021). Innovative Bert-Based Reranking Language Models for Speech Recognition. 266-271. DOI: http://doi.org/10.1109/SLT48900.2021.9383557
27. Wang, Z., Li, Y., & Zhu, Z. (2021). BERT-based knowledge extraction method for unstructured domain text. DOI: https://doi.org/10.48550/arXiv.2103.00728
28. DeepSee.ai. (2021). Meet FILBERT: Google's BERT trained on finance, insurance, legal. https://deepsee.ai/knowledge-center/meet-filbert-googles-bert-trained-on-finance-insurance-legal/
29. Eling, M., Gemmo, I., Guxha, D., & Schmeiser, H. (2024). Big data, risk classification, and privacy in insurance markets. *The Geneva Risk and Insurance Review.* DOI: https://doi.org/10.1057/s10713-024-00098-5
30. Khan, Z., Olivia, D., & Shetty, S. (2025). A Machine Learning and Explainable Artificial Intelligence Approach for Insurance Fraud Classification. *Inteligencia Artificial*, *28*(75), 140-169. DOI: https://doi.org/10.4114/intartif.vol28iss75
31. Dhanekulla, P. (2024). Real-Time Risk Assessment in Insurance: A Deep Learning Approach to Predictive Modelling. *International Journal For Multidisciplinary Research*. https://www.ijfmr.com/papers/2024/6/31710.pdf
32. Hou, Y., Xia, X., & Gao, G. (2024). Combining structural and unstructured data: A topic-based finite mixture model for insurance claim prediction. *arXiv preprint arXiv:2410.04684.* DOI: https://doi.org/10.48550/arXiv.2410.04684
33. A. Dimri, S. Yerramilli, P. Lee, S. Afra and A. Jakubowski, "Enhancing Claims Handling Processes with Insurance Based Language Models," *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 2019, pp. 1750-1755, https://ieeexplore.ieee.org/document/8999288
34. Tang, S., Liu, Q., Tan, Wa. (2019). Intention Classification Based on Transfer Learning: A Case Study on Insurance Data. In: Milošević, D., Tang, Y., Zu, Q. (eds) Human Centered Computing. HCC 2019. Lecture Notes in Computer Science(), vol 11956. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-37429-7_36

AUTHOR'S PROFILE

Rida Ghafoor Hussain (PhD, AI Researcher) completed her PhD from the University of Florence, Italy. She is actively working as an Artificial Intelligence Researcher. Her research interests include Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing, and Software Engineering, with a focus on Graph Optimisations, Software Quality Assurance, Software Testing,

Recommender Systems, and domain-specific NLP optimisations. She also has teaching experience at the National University of Computer and Emerging Sciences in Pakistan, spanning over 5 years with the Faculty of Computer Science and Software Engineering. She has collaborated with prominent institutions and organisations, including Nokia, Siena Artificial Intelligence Lab, GISEL Lab, Expert.ai and the University of East London.