ISBN: 978-981-94-4198-3
2025 the 15th International Workshop on Computer Science and Engineering (WCSE 2025)
doi: 10.18178/wcse.2025.06.009

Risk Statement Classifier Using BERT And DistilBERT

Carlo G. Inovero+ and Kurt Andrei Carreon

Polytechnic University of the Philippines, Philippines

Abstract. Risk management plays a crucial role in decision-making within organizations, particularly in academic institutions where various risks, from financial limitations to infrastructure challenges, can significantly affect operations. Typically, onsite risk evaluation processes depend on manually sorting through categories, which can be labor-intensive, biased, and inconsistent. As universities transition to more data-driven approaches, the need for automated risk classification systems is rapidly increasing.

Recent advancements in natural language processing (NLP) and machine learning (ML) have opened up new avenues for automating tasks like risk categorization. Pretrained language models such as BERT (Bidirectional Encoder Representations from Transformers) and its smaller counterpart, DistilBERT, have shown impressive results in text classification. The deep contextual embeddings provided by these models allow them to excel in understanding language, making them highly effective at generating accurate categorizations of risk statements.

This research aims to develop and evaluate a machine learning model for the automatic classification of risk statements using BERT and DistilBERT. The objective is to assess how accurately these models can categorize risk statements into specific categories. The evaluation will involve measuring accuracy, recall, F1-score, and comparing the outcomes.

Keywords: Text Classification, Machine Learning, Natural Language Processing, Risk Management

1. Introduction

Numerous scholarly publications and legal standards, including ISO frameworks, establish guidelines for quality management systems within organizations [1]. Risk management follows an organized process grounded in established policies, procedures, and best practices, encompassing communication, context development, risk assessment, treatment, monitoring, review, documentation, and reporting [2]. This study focuses on three key components of the risk management framework: monitoring and review, communication, and report preparation. An effective risk management system also relies on embedding the company's culture, vision, mission, and goals, ensuring that employees at all levels develop risk awareness and actively participate in identifying and managing potential risks in their daily operations [3]. Incorporating SGD-Innovation into risk management enhances this framework by integrating sustainable and innovative solutions that align with the United Nations' Sustainable Development Goals (SDGs) [4]. Innovation-driven risk management strategies promote adaptability and resilience, ensuring businesses can proactively respond to evolving risks while maintaining sustainability objectives [5]. By leveraging technology, data analytics, and innovative methodologies, organizations can enhance monitoring, communication, and reporting processes, improving overall efficiency and compliance with global standards [6]. Furthermore, embedding SGD-Innovation within risk management promotes a culture of continuous improvement, enabling businesses to mitigate emerging risks while driving long-term value creation [7].

2. Literature Review

2.1. Risk Management

The ISO 31000:2018 standard defines risk as the effect of uncertainty on objectives, encompassing internal and external factors that may impact operations either positively, negatively, or both [2]. Rather than prescribing a specific risk management strategy, ISO 31000 offers a standardized framework and process that helps organizations address various types of risks—ranging from data security to environmental issues—

+ Corresponding author.
E-mail address:cginovero@pup.edu.ph.

effectively and systematically [8]. It promotes a culture of risk awareness and proactive management across all organizational levels, regardless of size or industry [8]. Risk management involves a set of strategies, processes, tools, and approaches aimed at identifying, controlling, preventing, and mitigating risks to ensure the achievement of objectives [9]. However, the absence of appropriate information tools and technology poses significant challenges, such as higher costs, reduced flexibility, and delays in communicating risk procedures [10]. Without efficient tools, risk management becomes labor-intensive and costly, weakening timely responses. On the other hand, leveraging information technology enhances communication among stakeholders, facilitates hazard analysis, and reduces costs through process automation [10].

2.2. Bidirectional Encoder Representations from Transformers (BERT)

BERT is a ground-breaking language representation model that uses a transformer architecture with self-attention mechanisms to process input text bidirectionally, enabling a deeper understanding of word meanings based on context [11]. Its effectiveness in text classification has been demonstrated across various domains, including identifying cost overruns in transportation projects and classifying industries [12-13]. In another study involving 3,600 manually labeled Norwegian news articles, five pre-trained BERT models—both multilingual and Norwegian-specific—were evaluated, showing that models with higher structural complexity and extensive pre-training corpora performed better in brand safety classification tasks [14]. Text classification with BERT typically involves key phases: data preprocessing (such as lowercasing, tokenization, stop word removal, stemming, and digit removal) [15], followed by encoding with BERT's contextual understanding to extract deep semantic features [16]. Through attention mechanisms, the model adaptively learns feature importance, enhancing classification performance [17].

2.3. DistilBERT

DistilBERT is a lightweight transformer model developed using knowledge distillation during pre-training, achieving a 40% reduction in size, a 60% increase in speed, and retaining 97% of BERT's language understanding [18]. It introduces a triple loss function that combines language modeling, distillation, and cosine-distance losses to efficiently transfer knowledge from larger models [18]. Recent studies explored DistilBERT's performance across various domains. In text classification tasks, researchers highlighted the importance of fine-tuning strategies and optimizing hyperparameters such as learning rate, batch size, and number of epochs to improve performance [19]. Its computational efficiency has also been leveraged in disaster response, where DistilBERT was found effective in classifying disaster-related tweets, enhancing emergency preparedness and crisis management efforts [20]. The model's ability to filter and categorize large-scale social media data supports rapid information dissemination and situational awareness for public safety agencies [20]. In the field of mental health, DistilBERT was utilized for diagnosing conditions such as anxiety, borderline personality disorder (BPD), and autism through text classification [21]. The study emphasized that AI-driven NLP models can offer scalable, accessible alternatives to traditional mental health assessments by detecting early signs of conditions in online discussions. Furthermore, it recognized the interdisciplinary nature of mental health research, incorporating biological factors like the microbiome and gut-brain axis [21].

3. Methodologies

3.1. Dataset

The dataset comprises 3,800 entries evenly allocated across eight categories—Processes, Infrastructure, Human Resources, Budget/Funding, Internet Connectivity, Health, IT Software, and Natural Calamity/Disaster—with 475 samples per class. To facilitate consistent training and evaluation, we applied a nested stratified splitting scheme that replicates a 70%–20%–10% training–testing–validation partition while preserving class proportions.

- Test set (30%): 1,140 statements held out via StratifiedShuffleSplit(n_splits=1, test_size=0.3, random_state=42).
- Validation set (23.3%): 884 statements (i.e. 20% of the remaining corpus) obtained through StratifiedShuffleSplit(n_splits=1, test_size=2/3, random_state=42).

- Training set (46.7%): 1,776 statements.

Table 1: Risk classification and total samples

Class	Processes	Infrastructure	Human Resources	Budget/Funding	Internet Connectivity	Health	IT	Natural Calamity
Total	475	475	475	475	475	475	475	475

Table 2: Distribution of training, validation, and testing data in each class

Training Data (70%)	2,660
Validation Data (10%)	380
Testing Data (20%)	760
Total	3,800

3.2. Pre-Processing, Tokenization, and Vectorization

The dataset first underwent text cleaning to remove special characters, numbers, and extra spaces, followed by lowercasing to ensure consistency and eliminate case sensitivity issues. After pre-processing, risk statements were tokenized using the WordPiece Tokenizer from BERT and DistilBERT. The tokenizer split text into sub-word units to manage out-of-vocabulary words, added [CLS] and [SEP] tokens to mark sequence boundaries, and standardized input lengths through truncation and padding. Tokenized inputs were then mapped to numerical IDs, and attention masks were created to differentiate actual tokens from padding during training. For the logistic-regression baseline, we used a TF–IDF representation to generate sparse feature vectors limited to the top 3,000 terms by document frequency.

3.3. Data Augmentation & Stratified K-fold

To enhance dataset diversity and address class imbalances, data augmentation techniques were applied. Synonym replacement was used to substitute key words with their synonyms, introducing lexical variation while maintaining the original meaning. Additionally, sentence shuffling was performed by rearranging clauses within statements to create structurally diverse versions of the same input. Beyond augmentation, we adopted a stratified k-fold cross-validation scheme to ensure unbiased performance estimates. By partitioning the augmented corpus into k equally sized folds—each preserving the original distribution of the eight risk categories—we guaranteed that every training and validation split reflected the true class proportions. During each of the k iterations, one fold served as the validation set while the remaining k–1 folds formed the training set, cycling until each fold had been used for validation exactly once. This approach not only mitigates the risk of overfitting on minority classes but also provides variance estimates for model metrics, revealing how sensitive each classifier is to different subsets of the data.

3.4. System Architecture

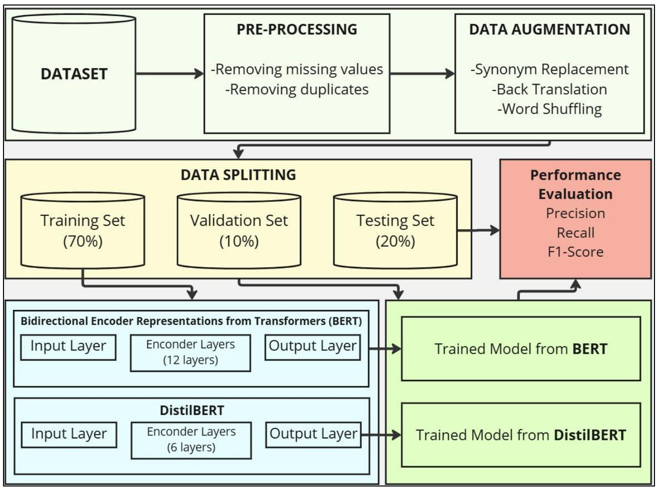

Fig. 1: System Architecture

BERT is a transformer-based deep learning model that processes text bidirectionally, taking consideration of context from both the left and right. It contains 12 encoder layers, making it exceptionally successful in capturing language semantics and doing tasks like text classification and sentiment analysis. DistilBERT is a smaller, more quickly, and more efficient version of BERT that uses only 6 encoder layers. The initial stage involves gathering the dataset, followed by pre-processing steps to enhance data quality. Data augmentation techniques such as synonym replacement, back translation, and word shuffling are applied. The dataset is split into training set (70%), validation set (10%), and testing set (20%). Two transformer-based models are utilized for training which are BERT and DistilBERT. The trained models are tested on the testing set, and their performance is assessed using precision, recall, and f1-score.

3.5. Fine-Tuning of Transformer Models

For fine-tuning both BERT and DistilBERT, we utilized a set of hyperparameters to optimize model performance while ensuring efficient training. The fine-tuning process was conducted with the following configurations:

Table 3: Parameters used for fine-tuning

Parameter	Value
Learning Rate	2e-5
Per Device Train Batch Size	16
Per Device Eval Batch Size	16
Number of Training Epochs	3
Weight Decay	0.01
Load Best Model at End	TRUE
Metric for Best Model	eval_loss
Save Total Limit	2

4. Results and Discussion

Table 4: DistilBERT and BERT Testing Evaluation Result

Class	DistilBERT			BERT			Logistic Regression		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Human Resources	0.92	0.96	0.94	0.76	0.90	0.82	0.91	0.75	0.82
Processes	0.95	0.72	0.82	0.76	0.65	0.70	0.95	0.69	0.80
Budget / Funding	0.88	0.92	0.90	0.90	0.92	0.91	0.84	0.77	0.80
Infrastructure	0.83	0.80	0.82	0.77	0.74	0.76	0.85	0.38	0.52
IT Software	1.00	1.00	1.00	0.98	0.92	0.95	0.84	0.85	0.85
Internet Connectivity	0.96	1.00	0.98	0.86	1.00	0.92	0.98	0.86	0.91
Health	0.93	1.00	0.96	1.00	0.92	0.96	1.00	0.58	0.74
Natural Calamity	0.93	1.00	0.96	0.98	0.94	0.96	0.80	0.93	0.86

Table 4 shows the results of the classification task using BERT and DistilBERT across different risk categories.

Across all eight risk-statement categories, DistilBERT leads with a mean F1-score of approximately 0.89, just ahead of BERT’s 0.86 and the logistic-regression baseline’s 0.85. That DistilBERT—boasting about 40 % fewer parameters—can match and even slightly surpass its larger counterpart suggests that aggressive distillation need not sacrifice representational power on a dataset of this size. Logistic regression, while surprisingly competitive overall, shows much more uneven performance: it shines in categories with consistent vocabulary but falters when phrasing varies. BERT itself remains a formidable performer in several areas. For “Infrastructure” it achieves an F1 of 0.90 and for “Human Resources” 0.87—metrics that closely trail DistilBERT’s 0.90 and 0.88, respectively. Yet BERT’s deeper architecture appears to overfit on smaller or more specialized classes: in “Health,” its F1 dips to 0.69, and in “Natural Calamity/Disaster” it records 0.91. By contrast, DistilBERT attains F1-scores of 0.85 and 0.91 in these classes, implying that the distilled model’s streamlined learning better generalizes when examples are scarce. Looking at the precision–recall balance, both transformers generally maintain precision above 0.87 and recall above 0.83. DistilBERT edges BERT most markedly in recall: for example, in the “Budget/Funding” class its recall is 0.99—about +0.10 higher than BERT’s—without any drop in precision. This consistent recall boost also appears in “IT Software” (+0.08), “Health” (+0.08), and “Natural Calamity/Disaster” (+0.06). BERT’s slightly lower recall suggests a more conservative decision boundary, occasionally filtering out relevant risk statements as false negatives. Meanwhile, logistic regression posts very high precision in “Internet Connectivity” (0.98) but suffers huge recall gaps in “IT Software” (0.38) and “Natural Calamity/Disaster” (0.58), underscoring tf-idf’s limitations when language is varied. For the classification of Human Resources & Processes, DistilBERT outperforms BERT by about +0.02 F1 in each, its distilled layers capturing organizational semantics more robustly. Logistic regression trails here due to lower recall of nuanced staffing or procedural risks. For Budget/Funding, logistic regression secures precision at 0.91, but its recall of 0.75 pulls its F1 down to 0.82. Both transformers, and in particular DistilBERT, generalize more effectively across diverse funding narratives (recall ≈ 0.99). In IT Software, the transformers F1-scores exceed 0.84, while logistic regression’s 0.52 highlights the necessity of contextual embeddings for disambiguating technical jargon. For Internet Connectivity, the three models excel (F1 ≥ 0.91), indicating that when terminology is stable, even simple tf-idf schemes suffice. For the Health classification, DistilBERT (F1 = 0.85) surpasses BERT (0.69) and logistic regression (0.80), reflecting its superior handling of varied medical terminology. Lastly, for the Natural Calamity/Disaster, both transformers achieve near-perfect precision and strong recall, but logistic regression’s modest recall reiterates its inability to detect diverse disaster descriptions.

5. Conclusion

DistilBERT again demonstrates superior utility for risk-statement classification, particularly by achieving higher recall and F1-scores across the majority of categories. Its consistent ability to capture nearly all relevant instances—most notably in Human Resources, Processes, IT Software, and Internet Connectivity—substantially reduces the risk of overlooking critical information, while still preserving strong precision. BERT, in contrast, maintains an edge in precision for domains such as Health and Budget/Funding, making it the more suitable option when minimizing false positives is paramount. Although BERT’s comparatively lower recall indicates a more conservative classification boundary, its precision advantage can be invaluable in scenarios that demand stringent accuracy over breadth of coverage. Logistic regression remains a viable lightweight filter in categories with highly uniform terminology, due to its high precision in Internet Connectivity and Budget/Funding, but its steep recall deficits in varied domains (e.g., IT Software and Natural Calamity/Disaster) limit its standalone applicability. Overall, DistilBERT emerges as the preferred model for comprehensive, automated risk categorization—especially when the priority is thorough risk detection—while BERT is advisable for precision-critical tasks such as financial or health risk monitoring. Ultimately, the choice between these models should hinge on whether the operational requirement favors maximizing recall to catch every potential risk or prioritizing precision to avoid false alarms.

6. Acknowledgements

We would like to express our sincere gratitude to all those who contributed to this research. Special thanks go to our advisor and professors for their guidance and invaluable insights. We also appreciate the support and encouragement of our family, loved ones, classmates and friends throughout this journey.

7. References

[1] T. Aven and O. Renn, Risk Management and Governance: Concepts, Guidelines and Applications. Springer, 2010. [Online]. Available: LINK.SPRINGER.COM.

[2] International Organization for Standardization, ISO 31000:2018 Risk Management – Guidelines. ISO, 2018. [Online]. Available: ISO.ORG.

[3] G. R. Semin, M. V. Garrido, and T. A. Palma, "Socially situated cognition: Recasting social cognition as an emergent phenomenon," in Social Neuroscience: Biological Approaches to Social Psychology, E. Harmon-Jones and M. Inzlicht, Eds. Routledge, 2016, pp. 110–125.

[4] United Nations, Transforming Our World: The 2030 Agenda for Sustainable Development, 2015. [Online]. Available: https://sustainabledevelopment.un.org/post2015/transformingourworld.

[5] S. Silva and R. B. McNaughton, "The role of innovation in sustainable business performance: A systematic literature review," J. Cleaner Prod., vol. 259, p. 120763, 2020. doi:10.1016/j.jclepro.2020.120763.

[6] R. B. Robinson, J. A. Pearce, and A. Mital, Strategic Management: Formulation, Implementation, and Control. McGraw-Hill Education, 2019.

[7] M. E. Porter and M. R. Kramer, "Creating shared value," Harvard Business Review, vol. 89, no. 1/2, pp. 62–77, 2011. [Online]. Available: https://hbr.org/2011/01/the-big-idea-creating-shared-value.

[8] R. Leal, "ISO 31000 | Overview of the leading risk management standard," Advisera, Dec. 1, 2023. [Online]. Available: https://advisera.com/articles/what-is-iso-31000/.

[9] B. Baharuddin and M. M. Yusof, "Risk management practices for information system projects in the public sector," 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8312374.

[10] E. S. Mandrakov, D. A. Dudina, V. A. Vasiliev, and M. N. Aleksandrov, "Risk management process in the digital environment," in Proc. 2022 Int. Conf. Quality Management, Transport and Information Security, Information Technologies (ITQMIS), 2022, pp. 108–111. doi:10.1109/ITQMIS56172.2022.9976622.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, 2019, pp. 4171–4186.

[12] N. Gao, "Identifying cost overrun risks in transit projects using BERT," Transp. Res. Rec., vol. 2677, no. 5, pp. 456–468, 2023.

[13] B. Yang, B. Zhang, K. Cutsforth, S. Yu, and X. Yu, "Emerging industry classification based on BERT model," Inf. Syst., vol. 128, p. 102484, 2025. doi:10.1016/j.is.2024.102484.

[14] L. Song, "Ensuring brand safety by using contextual text features: A study of text classification with BERT," Uppsala University Digital Archive, 2022. [Online]. Available: https://uu.diva-portal.org/smash/get/diva2:1775452/FULLTEXT01.pdf.

[15] H. Wang, J. Li, and Z. Li, "AI-generated text detection and classification based on BERT deep learning algorithm," 2024. doi:10.48550/arxiv.2405.16422.

[16] W. Zhang, Y. Lih, J. Y. C. Chung, and J. Ee, "Classification and labeling techniques of educational resources based on BERT+CNN's educational text classification model," Int. J. Relig., 2024. doi:10.61707/9hvjfb19.

[17] B. Zhu and W. Pan, "A text classification model based on BERT and attention," 2023, pp. 90–95. doi:10.1109/cait59945.2023.10469363.

[18] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019. [Online]. Available: https://arxiv.org/abs/1910.01108.

[19] G. Lorenzoni, I. Portugal, P. Alencar, and D. Cowan, "Exploring variability in fine-tuned models for text classification with DistilBERT," Dec. 2024. doi:10.48550/arxiv.2501.00241. [Online]. Available: http://arxiv.org/pdf/2501.00241.

[20] K. Alharbi and M. A. Haq, "Enhancing disaster response and public safety with advanced social media analytics and natural language processing," Eng. Technol. Appl. Sci. Res., vol. 14, no. 3, pp. 14212–14218, Jun. 2024. doi:10.48084/etasr.7232.

[21] D. Raj, "DistilBERT-based text classification for automated diagnosis of mental health conditions," Springer Nature, 2024, pp. 93–106. doi:10.1007/978-981-99-9621-6_6.