IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 June 2025, accepted 16 July 2025, date of publication 21 July 2025, date of current version 25 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3590983

RESEARCH ARTICLE

Improving Data Entry Quality in Enterprise Applications With NLP Methods: A Model Proposal Based on BERT and Deep Learning

H. CANLI
Department of Software Engineering, İstanbul Gedik University, 34876 İstanbul, Türkiye
e-mail: hikmet.canli@gedik.edu.tr

ABSTRACT In digital transformation, which is one of the most important keywords of our time, the completeness and accuracy of the data that users enter into applications directly affects the quality of the process, the accuracy of decision-making systems, and the speed at which data turns into information. Incorrect or incomplete data causes many problems such as prolonged approval processes, decreased trust in data, and negative impact on analysis capabilities. In this study, a data validation system was developed to improve the accuracy of risk management data collected from an ERP application and to minimize data entry errors. In order to prevent users from incorrectly entering or confusing important data such as Potential Risk, Internal Control, Control and Impact of the Risk during data entry, it is aimed to ensure accurate data entry by using NLP methods. Within the scope of the study, training was conducted on historical data and errors in user data entry were detected with various classification methods. Different methods such as BERT, RoBERTa, GPT-2, TFIDF+SVM, Word2Vec+SVM, Embedding GRU and Embedding LSTM were used to prevent these errors. The results show that the BERT model achieves the highest success rate with 94% accuracy. The strong language modelling capabilities of BERT gave it a significant advantage over other methods in detecting errors in data input.

INDEX TERMS NLP, BERT, classification, data validation, risk management.

I. INTRODUCTION

In the era of digital transformation, the accuracy of the data used by organizations is critical to the effectiveness and efficiency of business processes [18], [19], [20]. In particular, digital platforms such as Enterprise Resource Planning (ERP) systems allow for accurate and fast management of organizational processes [21]. However, errors made by users during data entry both negatively affect the efficiency of processes and cause reliability problems. Incorrect or incomplete data entry leads to prolonged approval processes, inaccurate analysis results and ultimately weakens decision-making mechanisms [22], [23]. Therefore, ensuring data accuracy is an important factor that directly affects the success of organizations in the digital transformation process.

In this study, a data validation system was developed to improve the accuracy of risk management data entry in an ERP application. The accuracy of the Potential Risk, Internal Control, Control and Impact of the Risk fields, where users make the most mistakes, enter incorrectly or incompletely while entering data, were checked. In the study, natural language processing (NLP) methods were utilized to prevent these errors, ensure data accuracy and speed up the processes. NLP is a set of techniques that enable language to be understood by computers and plays an important role in improving the accuracy of data [24], [25]. This study aims to detect the errors made on this data using different classification methods and help the user to make decisions during input.

Within the scope of the study, training was performed on risk management data using historical data and the data validation process was tested with five different

The associate editor coordinating the review of this manuscript and approving it for publication was Yilun Shang.

NLP-based classification methods that have proven themselves as natural language processing models in the literature: Bidirectional Encoder Representations from Transformers (BERT), Robustly Optimized BERT Pretraining Approach (RoBERTa), Generative Pre-trained Transformer 2 (GPT-2), Term Frequency – Inverse Document Frequency + Support Vector Machine (TF-IDF+SVM), Word2Vec+SVM, Embedding Gated Recurrent Units (GRU) and Embedding Long Short-Term Memory (LSTM). The results show that the BERT model achieved 94% accuracy, which is higher than the other methods. The models with the best accuracy are GPT-2 with %91 accuracy, RoBERTa with %90 accuracy, TF-IDF+SVM with 88% accuracy, Embedding LSTM with 87% accuracy, Embedding GRU with 81% accuracy and Word2Vec+ SVM with 73% accuracy. In addition, Precision, Recall and F1 score values of each model were examined in the experiments and details are given in the proposed model section. BERT's language modelling capabilities provided a significant advantage, especially in understanding complex input data and detecting erroneous entries, and much more successful results were obtained compared to other classification methods.

In summary, the contributions of this study are as follows:

a) An application-oriented method is proposed using BERT for ERP risk data validation, compared with other NLP-based classifiers.

b) By minimizing data entry errors, organizations have been enabled to save time (Error detection and prevention).

c) Contributed to the decision-making process with more reliable data.

d) The performance of all different classification methods in the literature are compared and the capabilities of the BERT model in language processing are demonstrated (RoBERTa, GPT-2, BERT, TF-IDF+SVM, Word2Vec+SVM, Embedding GRU, Embedding LSTM).

In the second part of this study, the studies on data checking and validation are analyzed in detail. In Section III, the dataset and NLP models used and the model evaluation parameters are explained in detail. In Section IV, the findings of the proposed models are discussed in detail. In the last section, the success of the proposed method and its contribution to the literature are explained and suggestions for future work are made.

II. RELATED WORK

In recent years, NLP and machine learning techniques have played an important role in the development of data validation and classification applications in different fields. Especially in the analysis of large datasets, providing accurate and reliable data directly affects the efficiency of business processes. Many studies have focused on the use of NLP-based methods to automate data validation processes.

Mert Marcel Dağlıet al. aimed to develop and validate a NLP algorithm integrated with the Big Language Model (LLM; GPT4-Turbo) to automate the extraction of spine surgery data from electronic health records (EHRs). They achieved 95% accuracy rate with their proposed method [1]. Asha Rajbhoj et al. introduce RClassify, a system that automatically extracts business rules from requirement documents written in natural language and classifies them into eight different classes. RClassify combines NLP and machine learning techniques to provide accurate and efficient classification of business rules in large and complex software products [2]. The study in [3] aims to extract meaningful information from unstructured data by using NLP and machine learning techniques to automatically analyze and categorize resumes. The study in [6] emphasizes that both simple and advanced NLP methods are effective in organizational analysis in human resources management and provides guidance for beginners. In [4], a NLP based system was developed to detect hidden social determinants of health (SDOH) in clinical notes. The pilot program aims to identify patients in the emergency department who need SDOH intervention and communicate them to social workers so that unmet social needs can be addressed in a timely manner. Another study presents an automated approach called NLPtoREST that extracts additional test information from natural language descriptions to improve REST API testing [5]. Emon et al. present a comparative analysis of three different transformative models for cyberbullying detection in Bangla language social media. On a dataset of 44,001 Bangla comments, the best result was obtained with the XLM-RoBERTa model with 85% accuracy and 86% F1 score [7]. In [8], a new NLP and logic-based framework called I-SNACC is introduced to automate structure code conformance checking. Tested on the International Building Codes, the system showed high accuracy with 95.2% precision and 100% recall, providing an efficient and near fully automated solution that can replace manual checks. In [9], data quality is studied from a different perspective. It provides a comprehensive taxonomy for assessing data quality in NLP, including linguistic, semantic and diversity dimensions. Moreover, by developing a new metric to measure the difficulty level of datasets, it is shown that this approach provides effective and holistic insights on datasets with different tasks [9].

In [10], we present a novel approach that combines NLP and domain verification to detect disposable email addresses. Using various machine learning algorithms (e.g. SVC, XGBoost, Random Forest), this method effectively detects and classifies new disposable email addresses with 97% accuracy compared to traditional blacklist-based methods. In [11], a comprehensive review of datasets, data validation techniques and prediction approaches used in software defect prediction is presented. The literature review shows that existing datasets contain incomplete labels and insufficient details and provides futuristic proposals for software defect prediction, emphasizing the importance of statistical validation techniques for more comprehensive dataset development and multi-label classification. In [12], he formalizes the definition of data validation and examines

some of the properties that can be derived from this definition. In particular, it shows how a formal definition of data validation can be used to classify and rank data quality requirements at increasing levels of complexity and some of the subtleties that arise in this process. In [13], he develops and validates a data quality assurance framework, ML-DQA, using real-world data (RWD) from clinical trials and machine learning applications. The framework includes methods such as redundant data elements, automated utilities, and rules-based transformations that improve data quality in clinical projects and provide a common quality control approach. In [14], in order to improve the data validation process for machine learning (ML) models on production lines, we introduce the Partition Summarization (PS) approach, providing an automated, accurate and scalable system for detecting corrupted data. By comparing each summarized partition with data quality metrics, the PS method provides a significant increase in precision over previous methods by reducing false positives and improving validation accuracy. In [15], he empirically examines the impact of dirty data on training and test data by relating the performance of 19 popular machine learning algorithms to six data quality dimensions. The results show that errors in training and test data have significant negative impacts on model accuracy and that high quality data is critical for reliable AI applications. In [16], he addresses data collection, quality management and fairness measures for data-centric artificial intelligence (AI) and deep learning applications. Given the imperfections and biases of real-world datasets, he emphasizes the importance of applying robust data management techniques and unfairness mitigation methods in the model training process. In [17], it emphasizes the importance of data validation for the reliability of ML algorithms and examines the challenges associated with data validation in ML systems. The paper discusses the strengths and weaknesses of data validation solutions, noting that these solutions have not been sufficiently adopted in industrial applications and that data validation shortcomings, together with cultural, ethical and legal factors, limit the effectiveness of ML systems. In [52], CNN was used for text classification and sentiment analysis, while CNN-BERT was used for text classification and sentiment analysis. In [53], challenges such as incompatibility in data structures, quality issues and security vulnerabilities are discussed. In particular, it is emphasized that while the widespread use of electronic health records (EHR) has led to significant improvements in patient care, data entry errors can jeopardize patient safety. In this context, it was argued that systematic tools such as the high reliability organizations (HRO) approach and root cause analysis (RCA) should also be used in health data. In [54], Wand and Wang's 1996 paper was revisited, which had a great impact on the information systems literature on understanding and evaluating data quality (DQ). Using representation theory, inadequate mappings between the real world and information systems are analyzed and a new intrinsic DQ classification is proposed.

TABLE 1. RISK.

Parameters	Descriptions
Potential Risk	This is the text data entry field where the definition of the risk is entered.
Impact of the Risk	This is the text data entry field where the scenarios that will occur when the risk occurs are entered.
Internal Control	This is the text data entry field where the measures taken to prevent the risk from occurring and how the relevant risk is currently under control within the department are entered.
Control	This is the text data entry field where the measures to be taken in the opinion of the auditor are entered.

A review of the studies shows that NLP and machine learning techniques are effectively used in data validation, classification and analysis processes. Each study focuses on developing NLP-based solutions to accurately process and analyze data in different areas, especially on large datasets. Most of the studies aim to improve data accuracy, improve efficiency by automating business processes, and extract meaningful information from various types of data. The overall goal has been to improve data quality, minimize errors and provide robust and reliable data to make the right decisions. In the light of these studies, a significant contribution to the literature has been made with the proposed method and study.

III. PROPOSED METHOD

In this section, the data, methods and implementation process used in the study will be explained in detail. In line with the purpose of the study, information about the preparation of the dataset, the characteristics of the models used and the evaluation methods will be provided.

A. RISK DATASET

The data used in the study was taken from the Risk Management module of an ERP Software used in the automotive sector. It consists of a total of 3148 risk records and 11 different features. The risk notification form is the screen where every user in the company is authorized and can enter data when necessary. When users are not experienced enough in risk while entering data, it is very likely that they may enter incorrect or incomplete data. In this study, we focused on the 4 most confusing and incomplete features when entering risk notifications. Since the other fields are not based on multiple choice and interpretation, the rate of users making mistakes is much lower. Table 1 shows the risk dataset.

Figure 1 shows the statistics of the dataset. While there is the most data in the potential risk category, the shortest entered data contains 3 characters and the longest entered data contains 10179 characters. Potential Risk, Impact of the Risk, Internal Control and Control fields are open to interpretation as they contain long texts and words. This significantly affects

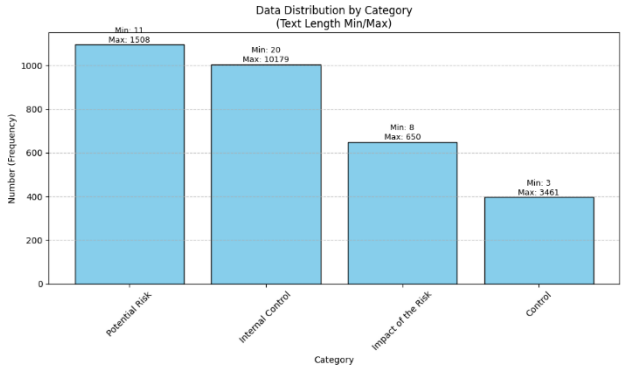

FIGURE 1. Dataset statistics.

TABLE 2. PRE-processing operations.

No	Pre-processing Operations
1	Special characters, URLs, and email addresses were removed.
2	Numbers were removed.
3	White spaces were removed.
4	Punctuations were removed.
5	Tokenization was performed.
6	All text data were converted to lowercase for eliminating to eliminate variations in word forms.
7	Stop words were removed.

the quality of the data entered and the quality of the data entered.

B. DATA PRE-PROCESSING

Text preprocessing is the process of preparing texts for easier processing. It includes processes such as removing stop words and special characters. It is used to prepare data before the NLP classification stage.

Only the most important words are retrieved or existing words are cleaned of their unnecessary tags and all are converted into normalized form to achieve better and more reliable classification results. Table 2 shows the preprocessing steps performed.

C. NLP MODELS FOR DATA VALIDATION

In this study, natural language processing models are used to prevent incorrect or incomplete entries of risk notification data. In the following, a description of the models used for error detection will be given. These models classify the data entered in the past and determine the appropriateness of the newly entered data.

D. CLASSIFICATION BY Word2Vec+SVM

Word2Vec represents words in a vector space. This expresses the semantic meaning of each word with a vector. This technique helps to understand the relationships and similarities between words [26]. SVM is an algorithm frequently used in classification and regression problems [27]. In this study, the word vectors transformed with Word2Vec are given as input to the SVM model. SVM classifies risk texts using these vectors.

E. CLASSIFICATION BY EMBEDDING + LSTM/GRU

Embedding is a technique that places words, sentences or documents into a low-dimensional, continuous vector space [32]. This technique tries to capture semantic relationships between words. That is, words with similar meanings are located close to each other in the embedding vectors [31].

LSTM network is an extension of the recurrent neural network (RNN) network used in deep learning [36], [37]. Unlike standard feed-forward networks, LSTMs have feedback connections [38]. It has three gates: Input, Output and Forget gate. The input gate controls the flow of input activations to the memory cell. Output gate controls the output flow of cell activation. The forget gate filters the information in the input and the previous output and decides which to remember or forget [39].

GRU is another variant of RNN architecture that addresses the short-term memory problem in deep learning models and offers a simpler structure compared to LSTM [34]. GRU combines the input gate and the forget gate of LSTM into a single update gate, leading to a more streamlined design. Unlike LSTM, GRU does not contain a separate cell state [35].

In this study, the text data is first converted into a vector representation through the embedding layer. Then, these vectors are passed to the GRU and LSTM layer. The GRU and LSTM models learn important features by processing sequential and dependent information in the text. In the final step, classification is performed using these features.

F. CLASSIFICATION BY TF-IDF+SVM

TF is a method for calculating the weights of terms in a document and is expressed by Equation (1) [29]. IDF, on the other hand, analyzes words that appear in multiple documents and tries to determine whether the word is really an important term (stop words). For this purpose, the absolute value of the logarithm of the number of documents in which the term occurs is divided by the total number of documents. This process is represented by Equation (2) [28], [30], [33].

TF-IDF score form term i in document j=TF (i, j) * IDF (i)

$$TF\ (i,j) = \frac{Term\ i\ frequencyin\ document\ j}{Total\ words\ in\ document\ j} \tag{1}$$

$$IDF(i) = \log\left(\frac{Total\ documents}{documents\ with\ term\ i}\right)$$

$$t = \text{Term}, j = \text{Document} \tag{2}$$

In this study, the input data generated over the TF-IDF model is classified with SVM vectors.

G. RoBERTa

RoBERTa (Robustly optimized BERT approach) is an improved version of the BERT model developed by Facebook

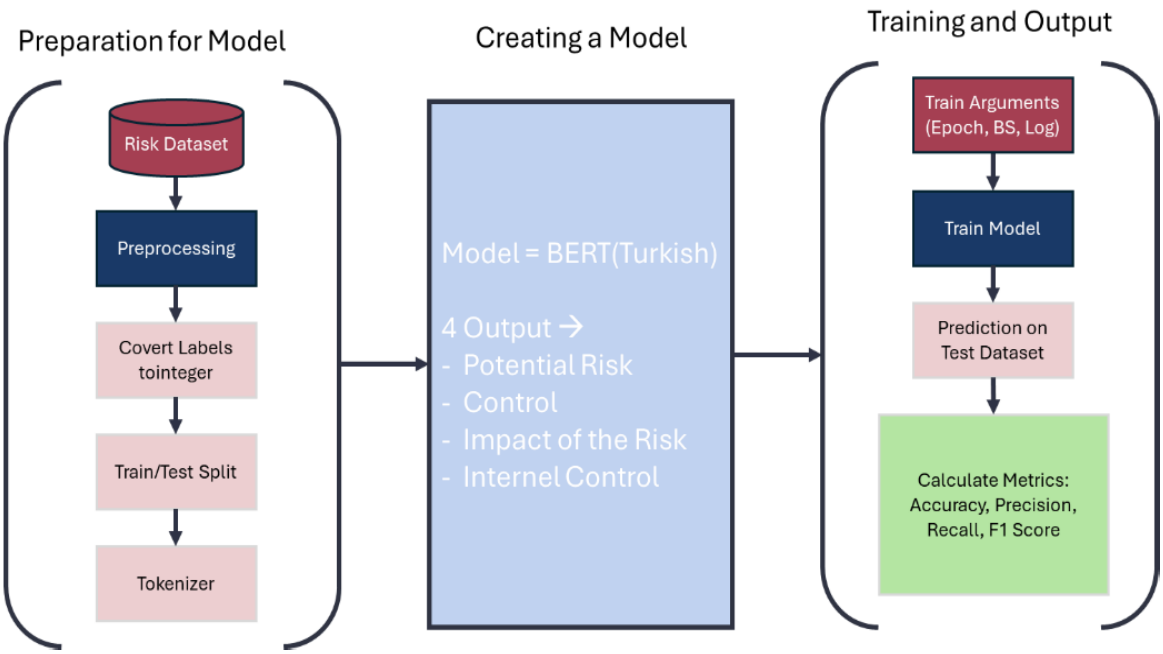

FIGURE 2. The proposed BERT model.

AI in 2019. RoBERTa aims to achieve higher performance by increasing the amount of data used in the training process of BERT, extending training times and changing some hyper-parameters [50]. In particular, training on larger datasets and using a dynamic masking strategy have enabled RoBERTa to provide better results in natural language processing tasks compared to BERT. Thanks to these developments, RoBERTa shows state-of-the-art performance in many areas such as text classification, semantic similarity and language modeling.

H. GPT-2

GPT-2 (Generative Pre-trained Transformer 2) is a deep learning model developed by OpenAI that makes significant contributions to scalable language modeling in natural language processing [51]. Built on the Transformer architecture, the model is trained on large text collections with a one-way pre-training method to predict the next word based only on the previous context. With 1.5 billion parameters, GPT-2 demonstrated strong overall performance on a wide range of tasks (e.g. text generation, translation, summarization) without the need for fine-tuning. The model's capacity for such efficient language generation has led to significant technical advances, but also to debates about the ethical and security implications of large language models.

I. CLASSIFICATION BERT (PROPOSED MODEL)

BERT is a very powerful model, especially in NLP tasks. Classification is one of these tasks. BERT uses a Cloze task-inspired ''masked language model'' (MLM) pre-training target, eliminating the one-way context limitation of traditional language models [41]. By randomly masking (hiding) some words in the input, the masked language model aims to predict the original lexical identity (ID) of these masked words based on context alone [42]. Unlike left-to-right language model pre-training, MLM allows target representations to combine information from both left and right contexts. This allows us to pre-train a deep bidirectional Transformer. In addition to the masked language model, we jointly pre-train text pair representations using the ''next sentence prediction'' (NSP) task [43], [44], [45]. Figure 2 shows the BERT model proposed in this work in detail.

J. EVALUATION METRIC

In this study, the classification algorithms were used to evaluate the models developed with the confusion matrix [40]. According to Table 3, the performance evaluation criteria for the classification algorithms are given below.

The accuracy value is shown in Eq. (3).

$$Accuracy\ (ACC) = \frac{T_P + T_N}{M} \tag{3}$$

The recall value is shown in Eq. (4).

$$Recall\ (Rcc) = \frac{T_P}{T_P + F_N} \tag{4}$$

Precision value is shown in Eq. (5).

$$Precision\ (PPV) = \frac{T_P}{TPos.} = \frac{T_P}{T_P + F_P} \tag{5}$$

F-measure value is shown in Eq. (6).

$$F - measure\ (F) = \frac{2{*}Precision\ {*}Sensitivity}{Precision + Sensitivity} \tag{6}$$

TABLE 3. Confusion matrix.

		Actual Result		
		Yes	**No**	**Total**
Prediction Result	**Yes**	True Positive (TP)	False Positive (FP)	TPos
	No	False Negative (FN)	True Negative (TN)	TNeg
	Total	Pos	Neg	M

TABLE 4. Parameters used in the model.

Parameters	Values
Batch Size	32,**64**,128,256,512,1024
Epoch	5,**10**,20,30
Activation Function	**Relu**, softmax, tanh
Optimizer	**Adam**, RMSprop, SGD
Number of units and Number of Conv filters	64,100,**128**,200,300
Kernel	2,3,**4**,5,6,7
Dropout	**0.3**, 0.4, 0.5, 0.6

K. HYPERPARAMETERS SETTING

In order to provide better results for the models used in this study, the ReducePlateau [46] method and various callback functions in the Keras library [47] as well as the Early Stopping mechanism [48] were utilized. During model training, if the verification accuracy does not improve for a certain period of time, the training process is terminated before completion. This early stopping is realized through the EarlyStopping function of Keras. In the study, training was stopped when the verification accuracy did not improve in 5 consecutive steps.

Furthermore, in cases where no improvement in the verification loss was observed, the learning rate was reduced by a certain factor to enable the model to learn more precisely. In this context, if the verification loss did not improve in 5 consecutive steps, the learning rate was reduced by a factor of 0.1 and the model continued to learn in smaller steps.

During the training of the model, performance monitoring and optimization were performed using the TensorBoard tool. TensorBoard is a powerful tool running on the TensorFlow infrastructure that allows monitoring metrics such as accuracy and loss, visualizing the architecture of the model, and efficiently experimenting with a large number of parameter combinations [49]. In this study, the hyperparameters in the first column of Table 4 were run in nested loops with different values in the second column to determine the optimal parameter values. The best parameters are highlighted in bold in Table 4.

Table 5 presents all summary information of the model used in the study, as determined by TensorBoard.

The study was carried out on a computer with a "13th Gen Intel(R) Core(TM) i9-13950HX 2.20 GHz" processor, "64.0 GB" RAM and "Windows 11 Pro" operating system.

TABLE 5. Model parameter summary information.

Parameters	Values
Train size	70%, 80%
Test size	30%, 20%
Validation split	10%
Optimizer	Adam
Number of Epochs	10
Recurrent Dropout	0.3
Batch Size	64
Filter	128
Learning Rate	0.001
Max Length	200
Metrics	Accuracy, Precision, Recall, F1 Score
Activation Function	Relu
Loss Function	Binary Cross Entropy

TABLE 6. Model calculate metrics.

Model	Accuracy	Precision	Recall	F1 Score
BERT	**0.9301**	**0.9338**	**0.9301**	**0.9310**
GPT-2	0.9116	0.9132	0.9111	0.9116
RoBERTa	0.8968	0.8998	0.8968	0.8976
TF-IDF + SVM	0.8857	0.8871	0.8857	0.8855
Embedding LSTM	0.8666	0.8757	0.8666	0.8672
Embedding GRU	0.8158	0.8309	0.8158	0.8164
Word2Vec + SVM	0.7365	0.7379	0.7365	0.7305

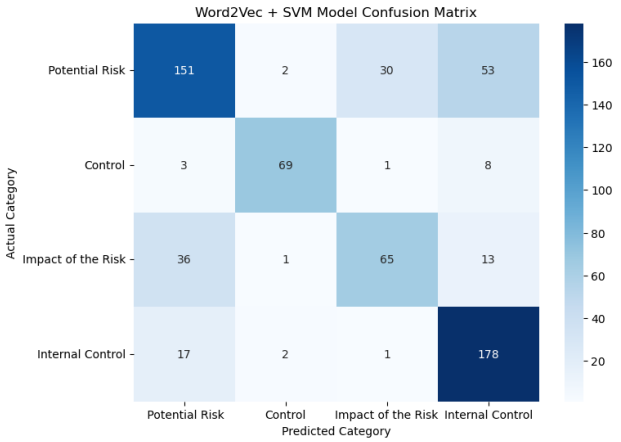

FIGURE 3. Word2Vec + SVM model confusion matrix.

IV. FINDINGS AND DISCUSSION

In this section, the proposed and compared methods, the analysis performed and the results of the models used and the details of these results will be discussed.

The results presented in Table 6 show the performance of various NLP-based classification models in the

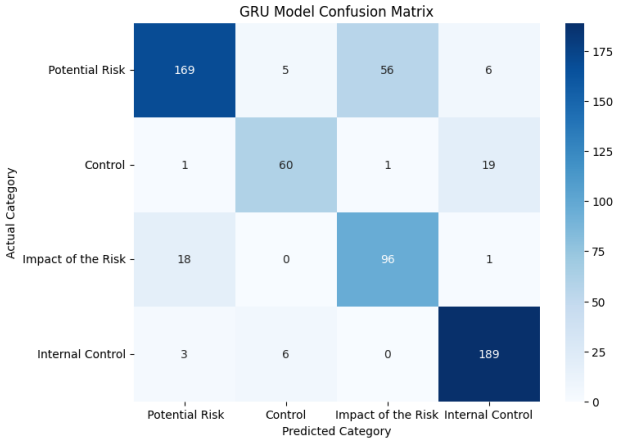

FIGURE 4. Embedding + GRU model confusion matrix.

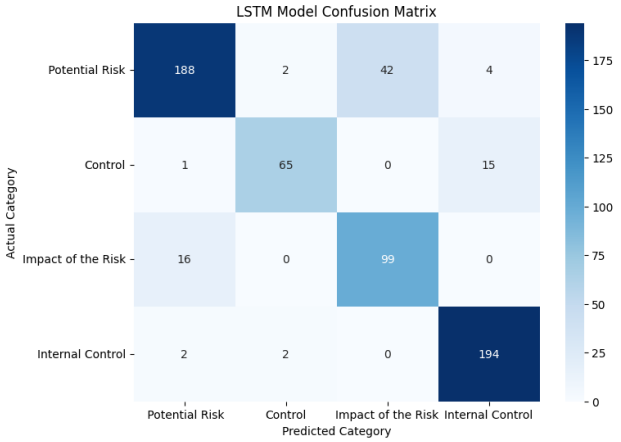

FIGURE 5. Embedding + LSTM model confusion matrix.

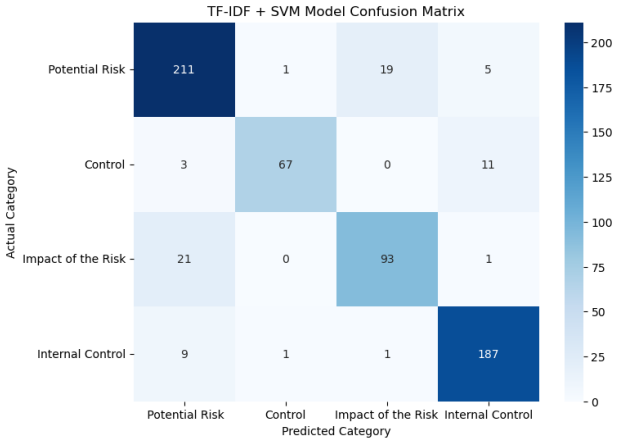

FIGURE 6. TF–IDF + SVM model confusion matrix.

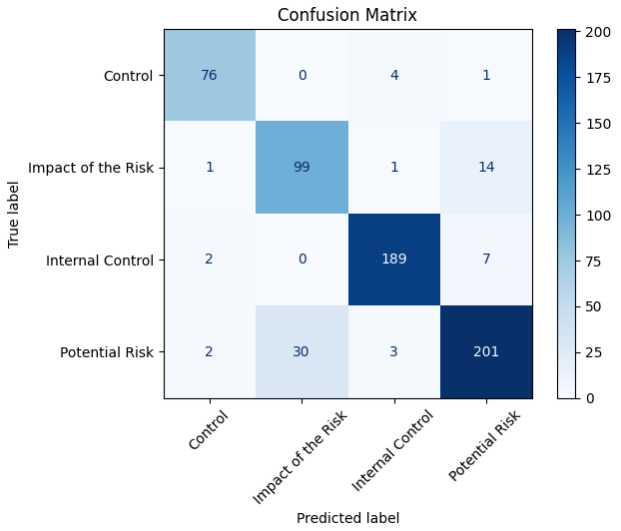

FIGURE 7. RoBERTa model confusion matrix.

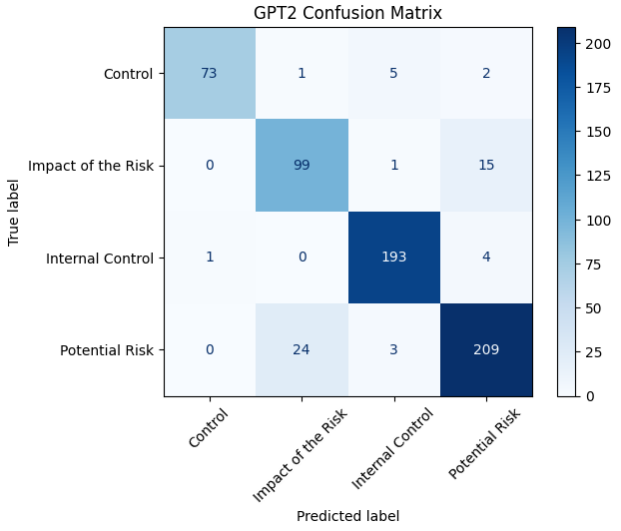

FIGURE 8. GPT-2 confusion matrix.

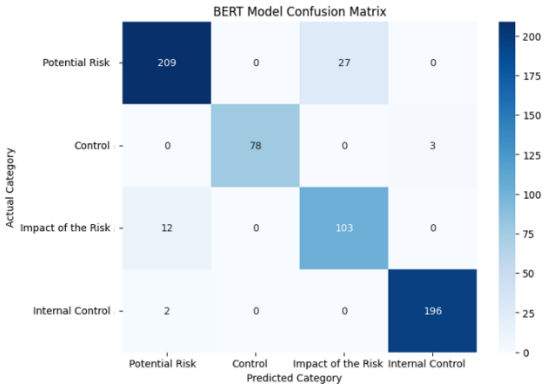

FIGURE 9. BERT model (proposed) confusion matrix.

category-based classification task of risk data. This evaluation includes both deep learning-based and traditional machine learning models. The BERT model showed the highest performance in all metrics (Accuracy: 0.9301, F1:0.9310). This demonstrates that BERT is able to represent the meaning of texts more accurately thanks to its ability to handle

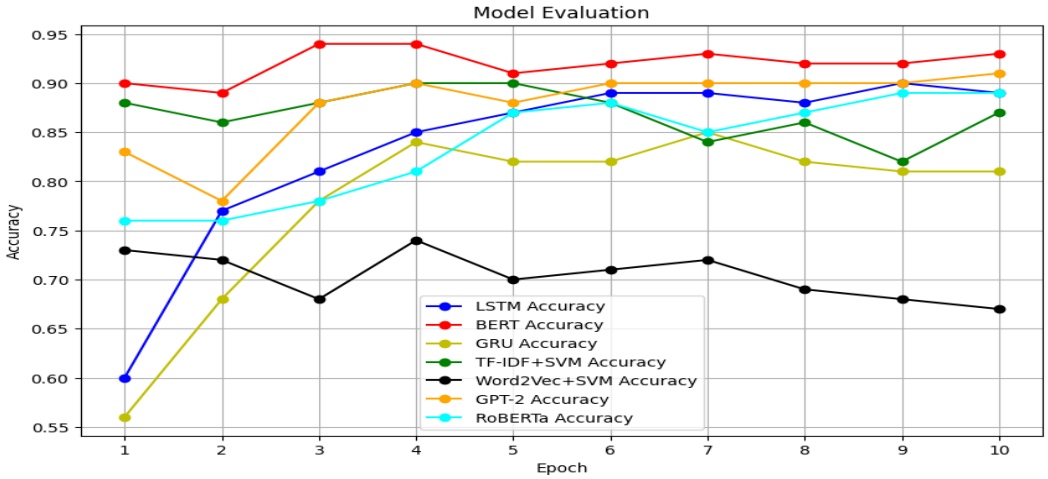

FIGURE 10. Model evaluation epoch/accuracy.

context bidirectionally. BERT's superior performance in text classification tasks involving complex language structures is also expected when reviewing other studies in the literature. The GPT-2 model follows closely behind BERT in terms of accuracy (0.9116) and F1 score (0.9116). Although GPT-2 was primarily designed as a generative language model, it has been shown to be effective in classification tasks through transfer learning. However, due to its unidirectional (left-to-right) structure, it can understand context more limitedly compared to BERT. RoBERTa, although a model based on BERT's optimizations, performed lower than BERT and GPT-2 in this study (F1:0.8976). This difference stems from the type and size of the training data used. The combination of TF-IDF vectorization and SVM, which are among the classical methods, yields highly competitive results (F1:0.8855). This shows that high-dimensional text representations can still be a powerful alternative when combined with the right classifiers. However, since this method does not consider context, it is limited in capturing semantic integrity compared to deep learning models. Results obtained using embedded vectors with recurrent neural networks such as LSTM and GRU showed lower performance in terms of accuracy and F1 score (F1 for LSTM: 0.8672, F1 for GRU: 0.8164). The lowest performance belongs to the combination of Word2Vec feature extraction and the SVM classifier (F1:0.7305). While Word2Vec can model semantic similarities at the word level, it may fail to capture context at the sentence or text level.

Figure 3 shows the confusion matrix of the model trained with Word2Vec + SVM. When the model is examined, although it has a high accuracy rate especially in the Internal Control class, it made more misclassifications in the Impact of the Risk and Potential Risk classes.

Figure 4 shows the confusion matrix of the model trained with Embedding + GRU. In general, the model achieved very good accuracy in the Internal Control class, while it also performed strongly in the Impact of the Risk and Control classes.

Figure 5 shows the confusion matrix of the model trained with Embedding + LSTM. While the Embedding + LSTM model performs quite strongly in the Internal Control and Impact of the Risk classes, some misclassifications were observed in the Potential Risk and Control classes.

Figure 6 shows the confusion matrix of the model trained with TF - IDF + SVM. It is observed that the overall performance of the TF-IDF + SVM model is quite high (89% accuracy) and it makes strong correct predictions in Potential Risk and Impact of the Risk classes.

Figures 7 and 8 show the confusion matrices of the models trained with RoBERTa and GPT-2, respectively. Both models made successful predictions in almost all categories. Accuracy rates of 90% and 91% support their inclusion among the preferred models.

Figure 9 shows the confusion matrix of the model trained with BERT (proposed method). The BERT model has a high rate of correct classifications in all classes of Potential Risk, Control, Impact of the Risk, and Internal Control. 94% accuracy rate is a very successful result compared to other methods.

TF - IDF + SVM, Word2Vec + SVM, RoBERTa, GPT-2 and BERT provided much higher accuracy rates in the Potential Risk metric compared to other deep learning methods. This is because the Potential Risk parameter has more context on the text than the other parameters and these models work in this context. LSTM and GRU, on the other hand, were more successful in the Impact of the Risk parameter, which has less textual context.

Figure 10 shows that the BERT model is the most robust model in terms of accuracy rates, and also shows a stable structure with high accuracy values at each epoch. The TF-IDF + SVM model, on the other hand, exhibited a stable accuracy rate like the BERT model from the beginning and

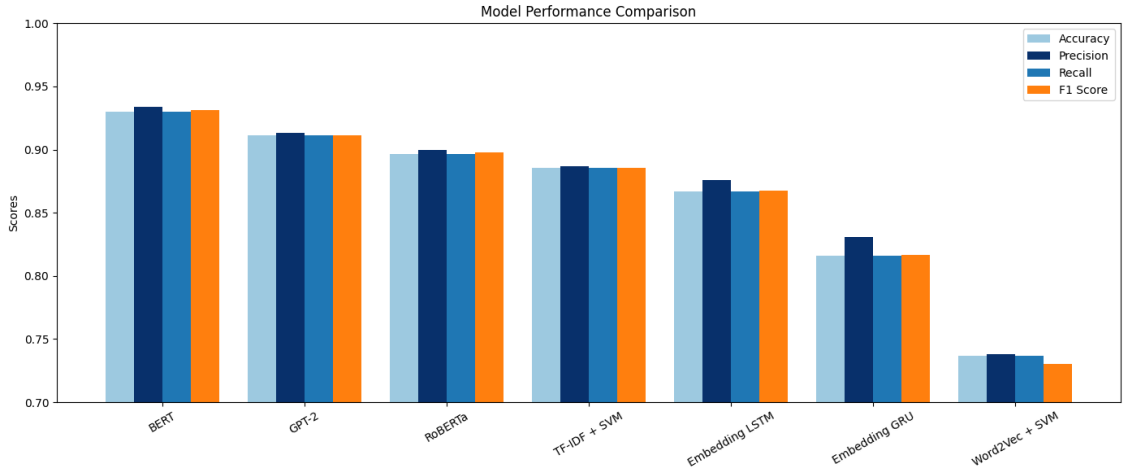

FIGURE 11. Model evaluation scores/metric.

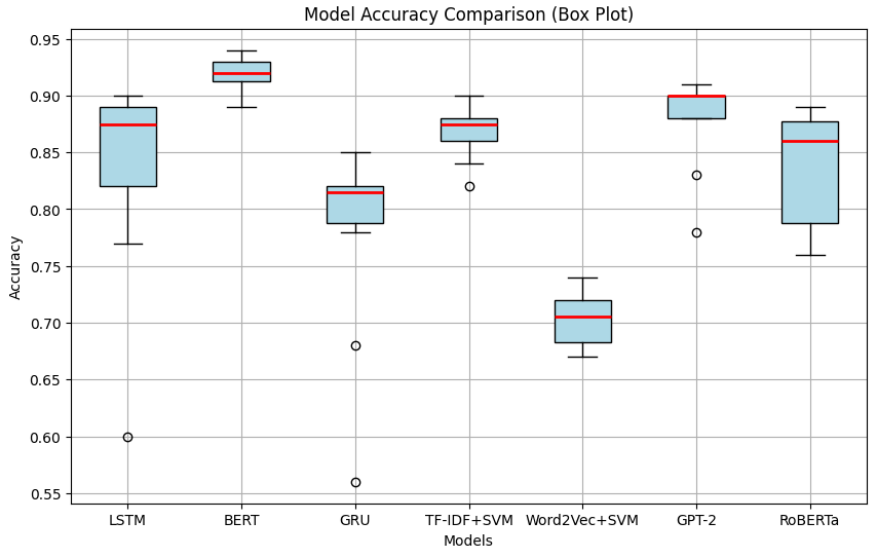

FIGURE 12. Model accuracy comparison with box plot.

gave the closest results to the proposed method. The accuracy values of the Word2Vec + SVM model, on the other hand, are low compared to the other models, indicating that the model is less successful than the others in understanding the text representation or in the classification task.

Figure 11 shows the performance metrics for all models. The proposed model performs better than other models in all metrics. In addition, models other than the Word2Vec + SVM model have shown successful performance in the prediction and classification of risk datasets.

Figure 12 Model Accuracy Comparison with Box Plot is given. The figure shows that the BERT model exhibits the highest and most stable accuracies. In addition, the accuracy values are generally above 0.90 and fall within a narrow range. LSTM and GRU, on the other hand, start with low accuracies at the beginning but improve over time and reach around 0.85, but their accuracies are distributed in a wider range. The TF-IDF + SVM model provides stable but average accuracies, while Word2Vec + SVM has the lowest accuracies, generally between 0.67 and 0.74. These results show that BERT provides stronger and more consistent performances than the other models, while Word2Vec + SVM is weaker in the classification task.

V. CONCLUSION

In this study, a data validation system was developed using natural language processing (NLP) techniques to detect and prevent data entry errors encountered in risk management processes. In the analysis of 3,148 risk notification data obtained from ERP systems, it was found that user errors were common especially in the "Potential Risk", "Impact of the Risk", "Internal Control" and "Control" headings, which

are open to interpretation and text-based fields. Accordingly, seven different NLP-based classification models were applied to improve data quality and support accurate data entry: Word2Vec + SVM, Embedding + GRU, Embedding + LSTM, TF-IDF + SVM, RoBERTa, GPT- 2 and BERT.

According to the results, the BERT model showed the highest performance with 94% accuracy. The depth of BERT's linguistic understanding and its ability to model context robustly enabled it to more effectively distinguish semantic differences and errors between texts. Compared to other methods, especially the Word2Vec+SVM model had the lowest performance with 73% accuracy, demonstrating the inability of classical methods to capture complex language structures. With the high accuracy rate of the BERT model, very positive results were obtained in terms of automating manual data control processes in organizations, improving data quality and saving time. In addition, the proposed system can be configured to not only detect erroneous data, but also to guide the user on correct data entry. In this way, it is possible to perform more accurate analyses, create more reliable decision support systems and increase the efficiency of ERP systems in general.

In future studies, the scope of the proposed system can be expanded by adapting it to different sectors, adding multilingual support and integrating real-time warning mechanisms. Furthermore, the integration of artificial intelligence-supported personalized recommendation systems that analyze user behaviour can further improve data entry accuracy.

REFERENCES

[1] M. M. Dagli, Y. Ghenbot, H. S. Ahmad, D. Chauhan, R. Turlip, P. Wang, W. C. Welch, A. K. Ozturk, and J. W. Yoon, "Development and validation of a novel AI framework using NLP with LLM integration for relevant clinical data extraction through automated chart review," *Sci. Rep.*, vol. 14, no. 1, Nov. 2024, doi: 10.1038/s41598-024-77535-y.

[2] A. Rajbhoj, P. Nistala, A. Pathan, P. Kulkarni, and V. Kulkarni, "RClassify: Combining NLP and ML to classify rules from requirements specifications documents," in *Proc. IEEE 31st Int. Requirements Eng. Conf. (RE)*, Hannover, Germany, Sep. 2023, pp. 180–189, doi: 10.1109/re57278.2023.00026.

[3] S. Mohanty, A. Behera, S. Mishra, A. Alkhayyat, D. Gupta, and V. Sharma, "Resumate: A prototype to enhance recruitment process with NLP based resume parsing," in *Proc. 4th Int. Conf. Intell. Eng. Manage. (ICIEM)*, London, U.K., May 2023, pp. 1–6, doi: 10.1109/ICIEM59379.2023.10166169.

[4] U. Ravichandran, D. Jungst, and E. Kwan, "Implementing an NLP tool to address SDOH needs," in *Proc. IEEE 11th Int. Conf. Healthcare Informat. (ICHI)*, Houston, TX, USA, Jun. 2023, pp. 522–524, doi: 10.1109/ichi57859.2023.00091.

[5] M. Kim, D. Corradini, S. Sinha, A. Orso, M. Pasqua, R. Tzoref-Brill, and M. Ceccato, "Enhancing REST API testing with NLP techniques," in *Proc. 32nd ACM SIGSOFT Int. Symp. Softw. Test. Anal.*, New York, NY, USA, Jul. 2023, pp. 1232–1243, doi: 10.1145/3597926.3598131.

[6] F. Guo, C. M. Gallagher, T. Sun, S. Tavoosi, and H. Min, "Smarter people analytics with organizational text data: Demonstrations using classic and advanced NLP models," *Human Resource Manage. J.*, vol. 34, no. 1, pp. 39–54, Jan. 2024, doi: 10.1111/1748-8583.12426.

[7] M. I. H. Emon, K. N. Iqbal, M. H. K. Mehedi, M. J. A. Mahbub, and A. A. Rasel, "Detection of Bangla hate comments and cyberbullying in social media using NLP and transformer models," in *Advances in Computing and Data Sciences. ICACDS 2022. Communications in Computer and Information Science*. Cham, Switzerland: Springer, 2022, pp. 86–96, doi: 10.1007/978-3-031-12638-3_8.

[8] J. Wu, X. Xue, and J. Zhang, "Invariant signature, logic reasoning, and semantic natural language processing (NLP)-based automated building code compliance checking (I-SNACC) framework," in *Proc. ITcon Special Issue Eastman Symp.*, vol. 28, pp. 1–18, doi: 10.36680/j.itcon.2023.001.

[9] V. M. H. Dang and R. M. Verma, "Data quality in NLP: Metrics and a comprehensive taxonomy," in *Advances in Intelligent Data Analysis XXII*, 2024, pp. 217–229, doi: 10.1007/978-3-031-58547-0_18.

[10] R. Alanazi and S. Alanazi, "A hybrid NLP and domain validation technique for disposable email detection," *Alexandria Eng. J.*, vol. 102, pp. 200–210, Sep. 2024, doi: 10.1016/j.aej.2024.05.068.

[11] J. Pachouly, S. Ahirrao, K. Kotecha, G. Selvachandran, and A. Abraham, "A systematic literature review on software defect prediction using artificial intelligence: Datasets, data validation methods, approaches, and tools," *Eng. Appl. Artif. Intell.*, vol. 111, May 2022, Art. no. 104773, doi: 10.1016/j.engappai.2022.104773.

[12] M. P. J. Van der Loo and E. de Jonge, "Data validation," in *Wiley StatsRef: Statistics Reference Online*, 2025, doi: 10.1002/9781118445112.stat08255.

[13] M. Sendak, "Development and validation of ML-DQA—A machine learning data quality assurance framework for healthcare," in *Proc. 7th Machine Learn. Healthcare Conf.*, vol. 182, 2022, pp. 741–759. [Online]. Available: https://proceedings.mlr.press/v182/sendak22a.html

[14] S. Shankar, L. Fawaz, K. Gyllstrom, and A. Parameswaran, "Automatic and precise data validation for machine learning," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manage.*, New York, NY, USA, Oct. 2023, pp. 2198–2207, doi: 10.1145/3583780.3614786.

[15] S. Mohammed, L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, and H. Harmouch, "The effects of data quality on machine learning performance on tabular data," 2022, *arXiv:2207.14529*.

[16] S. E. Whang, Y. Roh, H. Song, and J.-G. Lee, "Data collection and quality challenges in deep learning: A data-centric AI perspective," *VLDB J.*, vol. 32, no. 4, pp. 791–813, Jul. 2023, doi: 10.1007/s00778-022-00775-9.

[17] F. Biessmann, J. Golebiowski, T. Rukat, D. Lange, and P. Schmidt. *Automated Data Validation in Machine Learning Systems*. [Online]. Available: https://www.amazon.science/publications/automated-data-validation-in-machine-learning-systems

[18] A. Bozkurt, N. B. Hamutoglu, A. L. Kaban, G. Tasçi, and M. Aykul, "Dijital Bilgi çagi: Dijital toplum, dijital dönüsüm, dijital egitim ve dijital yeterlilikler," *Açikögretim Uygulamalari ve Arastirmalari Dergisi*, vol. 7, no. 2, pp. 35–63, Apr. 2021, doi: 10.51948/auad.911584.

[19] A. Ailtürk, "Isletmelerde dijital Dönüsüm Yönetiminde nihai hedef: Dijital olgunluk ultimate goal of digital transformation management in businesses: Digital maturity," *Alanya Akademik Bakis*, vol. 5, no. 2, pp. 647–669, May 2021, doi: 10.29023/alanyaakademik.859300.

[20] H. USLU, "Dijital Dönüşüm ve kamu hizmetleri Yönetimde Yenilikçi yaklaşımlar ve zorluklar innovative approaches and challenges in digital transformation and public services management," *Int. J. Political Stud.*, pp. 15–31, Oct. 2023, doi: 10.25272/icps.1354693.

[21] M. Keskinkýlýç and M. Ýpkin, "Ýşletmelerde ERP Uygulamalarýnýn Dijital Dönüşüm Sürecine Katkýlarý," *Aurum Sosyal Bilimler Dergisi*, vol. 8, no. 1, pp. 49–74, 2023.

[22] K. A. Barchard and L. A. Pace, "Preventing human error: The impact of data entry methods on data accuracy and statistical results," *Comput. Hum. Behav.*, vol. 27, no. 5, pp. 1834–1839, Sep. 2011, doi: 10.1016/j.chb.2011.04.004.

[23] R. R. Panko, "Thinking is bad: Implications of human error research for spreadsheet research and practice," 2008, *arXiv:0801.3114*.

[24] Z. Xuan, T. Wang, C. Wang, and T. Li, "A tool for automatically identifying semantic conflicts in user stories by combining NLP and BERT model," in *Proc. IEEE 32nd Int. Requirements Eng. Conf. (RE)*, Reykjavik, Iceland, Jun. 2024, pp. 484–487, doi: 10.1109/re59067.2024.00057.

[25] M. S. Jahan, M. Oussalah, D. R. Beddia, J. K. Mim, and N. Arhab, "A comprehensive study on NLP data augmentation for hate speech detection: Legacy methods, BERT, and LLMs," 2024, *arXiv:2404.00303*.

[26] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013.

[27] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/bf00994018.

[28] S. N. Başa and M. S. Basarslan, "Sentiment analysis using machine learning techniques on IMDB dataset," in *Proc. 7th Int. Symp. Multidisciplinary Stud. Innov. Technol. (ISMSIT)*, Oct. 2023, pp. 1–5.

[29] A. Saha, O. Hassanzadeh, A. Gittens, J. Ni, K. Srinivas, and B. Yener, "Improving neural ranking models with traditional IR methods," 2023, *arXiv:2308.15027*.
[30] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022, *arXiv:2203.05794*.
[31] M. Xu, "Understanding graph embedding methods and their applications," *SIAM Rev.*, vol. 63, no. 4, pp. 825–853, Jan. 2021, doi: 10.1137/20m1386062.
[32] C. D. T. Barros, M. R. F. Mendonça, A. B. Vieira, and A. Ziviani, "A survey on embedding dynamic graphs," *ACM Comput. Surv.*, vol. 55, no. 1, pp. 1–37, Jan. 2023, doi: 10.1145/3483595.
[33] M. B. Çaký and M. S. Başarslan, "Classification of fake news using machine learning and deep learning," *J. Artif. Intell. Data Sci.*, vol. 4, no. 1, pp. 22–32, 2024.
[34] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU," 2023, *arXiv:2305.17473*.
[35] S. Abbaspour, F. Fotouhi, A. Sedaghatbaf, H. Fotouhi, M. Vahabi, and M. Linden, "A comparative analysis of hybrid deep learning models for human activity recognition," *Sensors*, vol. 20, no. 19, p. 5707, Oct. 2020, doi: 10.3390/s20195707.
[36] Z. Huang, F. Yang, F. Xu, X. Song, and K.-L. Tsui, "Convolutional gated recurrent unit–recurrent neural network for state-of-charge estimation of lithium-ion batteries," *IEEE Access*, vol. 7, pp. 93139–93149, 2019.
[37] H. Yan, Y. Qin, S. Xiang, Y. Wang, and H. Chen, "Long-term gear life prediction based on ordered neurons LSTM neural networks," *Measurement*, vol. 165, Dec. 2020, Art. no. 108205.
[38] H. Canli and S. Toklu, "Deep learning-based mobile application design for smart parking," *IEEE Access*, vol. 9, pp. 61171–61183, 2021, doi: 10.1109/ACCESS.2021.3074887.
[39] R. Fu, Z. Zhang, and L. Li, "Using LSTM and GRU neural network methods for traffic flow prediction," in *Proc. 31st Youth Academic Annu. Conf. Chin. Assoc. Autom. (YAC)*, Nov. 2016, pp. 324–328.
[40] P. Harrington, *Machine Learning in Action*. Shelter Island, NY, USA: Manning Publications Co, 2012.
[41] W. L. Taylor, "'Cloze procedur,': A new tool for measuring readability," *Journalism Quart.*, vol. 30, no. 4, pp. 415–433, Sep. 1953.
[42] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.
[43] Y. Sun, Y. Zheng, C. Hao, and H. Qiu, "NSP-BERT: A prompt-based few-shot learner through an original pre-training Task-Next sentence prediction," 2021, *arXiv:2109.03564*.
[44] S. Aroca-Ouellette and F. Rudzicz, "On losses for modern language models," 2020, *arXiv:2010.01694*.
[45] Ç. Aksoy, A. Ahmetoğlu, and T. Güngör, "Hierarchical multitask learning approach for BERT," 2020, *arXiv:2011.04451*.
[46] RedecuPlatue. *RedecuPlatue*. [Online]. Available: https://keras.io/api/callbacks/reduce_lr_on_plateau/
[47] Keras. *Keras*. [Online]. Available: https://www.keras.io
[48] *Early Stopping*. [Online]. Available: https://keras.io/api/callbacks/early_stopping/
[49] Google. (2021). *Tensorflow*. [Online]. Available: https://tensorflow.org
[50] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[51] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI, Tech. Rep., 2019.
[52] P. Nuthakki, M. Katamaneni, C. S. J. N., K. Gubbala, B. Domathoti, V. R. Maddumala, and K. R. Jetti, "Deep learning based multilingual speech synthesis using multi feature fusion methods," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, Sep. 2023.
[53] K. Merriman, E. Yu, A. Hawkins-Daarud, K. Adelson, K. Shaw, D. Shoenthal, and C. Chung, "Data events are safety events: High-reliability organization approach to improving data quality and safety," *JCO Clinical Cancer Inform.*, vol. 9, Apr. 2025, Art. no. e2400273.
[54] A. Haug, "Intrinsic data quality dimensions: Expanding on wand and wang's data quality model," *Ind. Manage. Data Syst.*, vol. 125, no. 1, pp. 238–261, 2025.

H. CANLI was born in Ordu, Türkiye, in 1992. He received the B.S. (Hons.) and M.S. degrees in computer engineering and the Ph.D. degree in computer engineering education from Duzce University, Düzce, Türkiye, in 2015, 2017, and 2022, respectively. His teaching areas include cloud computing, information and network security, and programming languages at the bachelor's degree level. His research interests include the Internet of Things, cybersecurity, computer networking, deep learning, machine learning, and data mining.