International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 4, 2025

Improving Data Entry Accuracy Using Distil-BERT: An Efficient Extension to BERT-Based NLP Models

E. Monisree[1], Department of AIML, MJR College of Engineering and Technology, Piler, India

Mr.R. Althaf [2], Assistant Professor, Department of CSE, MJR College of Engineering and Technology, Piler,India

Abstract: *In corporate applications, correct data input is essential for optimizing process efficiency and facilitating sound decision-making. The current systems mostly use old-fashioned machine learning methods like TF-IDF+SVM and Word2Vec+SVM, which don't give a lot of contexts, therefore the accuracy of data validation jobs is only modest. To get over these problems, the suggested system uses modern NLP and deep learning methods, especially the BERT model, to automatically check the quality of data submission. BERT's bidirectional transformer design captures deep semantic and contextual linkages, which makes it much easier to find missing or wrong data. Also, an improved version called Distil-BERT is developed to make things more efficient by lowering model complexity and computation time while keeping or enhancing classification accuracy. This model is light yet powerful, which makes data validation faster, more scalable, and less resource-intensive. It also works better and is more useful for real-time business applications.*

Index Terms— NLP, BERT, classification, data validation, risk management.

1. INTRODUCTION

In the digital age, correct data entry is essential for corporate application performance. Organisations increasingly use ERP systems to streamline operations, resources, and workflows. The quality of user data strongly impacts the reliability of these systems. Data input errors can delay procedures, misinterpret data, and impair decision-making, lowering company productivity. Maintaining corporate integrity and efficiency requires high-quality data input.

Traditional data validation systems use rule-based or traditional machine learning methods like TF-IDF+SVM or Word2Vec+SVM, which cannot grasp text contextual and semantic links. These models treat text as tokens, failing to understand meaning across sentences or phrases, limiting their performance in recognizing incomplete or incorrect submissions. These solutions are less scalable and responsive to changing data patterns in business contexts due to human feature engineering.

Advanced NLP and deep learning models are used to provide an intelligent data validation framework to overcome these restrictions. The system extracts bidirectional context and semantic meaning from textual entries using BERT. This method automatically identifies missing or wrong data with improved accuracy, recall, and F1-scores than older methods.

Distil-BERT, an expanded model, improves computing efficiency without losing accuracy. For large-scale corporate applications, Distil-BERT provides quicker processing and reduced resource usage than BERT due to its fewer layers. With rich contextual awareness and optimum performance, the suggested model offers more accurate, quicker, and scalable ERP data validation, improving decision-making and data quality across organizational systems.

2. LITERATURE SURVEY

2.1 Novel AI Framework Using NLP + LLM for Automated Chart Review

Uses rule-based NLP + GPT-4 Turbo (LLM) to automate extraction of spinal surgical data from EHRs. Accurately extracts surgery type, levels operated, number of disks removed, and durotomy occurrence. Achieves high performance (accuracy, precision, F1) with major improvements in time & cost efficiency

2.2 Resumate: NLP-Based Resume Parsing for Recruitment

Converts uploaded resumes to structured format using NLP + ML. XGBoost performs best (0.65–0.73) for resume classification after k-fold validation.

ISSN 2347–3657

Volume 13, Issue 4, 2025

Automates extraction, categorization, and improved processing of resumes

2.3 Adaptable Framework for Medication Adherence

Systematic review of 102 conceptual frameworks linked to WHO's five adherence dimensions. Identifies eight patient categories with different adherence-impacting factors. Helps clinicians identify barriers and design targeted interventions

2.4 Enhancing REST API Testing with NLP

NLP to REST extracts meaningful rules from natural-language API descriptions to enhance test cases. Validates rules to reduce false positives and inconsistencies. Improves performance of eight major REST API testing tools

2.5 Smarter People Analytics Using NLP

Demonstrates text-based prediction for HRM using classic (BoW) & advanced (Doc2Vec) NLP models. Uses text to forecast outcomes like organizational ratings and personality traits. Provides beginner-friendly instructions for practical NLP adoption

3. METHODOLOGY

The suggested method improves data entry quality utilizing NLP and deep learning. Model training and assessment employ a labeled dataset with 'Complete' or 'Incomplete' values. Preprocessing comprises removing stop words, URLs, special characters, and stemming and lemmatizing to normalize text. TF-IDF and Word2Vec for conventional machine learning models and contextual embeddings for BERT and Distil-BERT deep learning models extract features. Masked language modeling and next sentence prediction train the bidirectional transformer-based BERT model to capture semantic and syntactic links in text. Distil-BERT, a lightweight BERT extension, is used for quicker training and inference with good accuracy. Based on criteria like accuracy, precision, recall, and F1-score, Distil-BERT surpasses classical and deep learning models, providing efficient and accurate data entry quality categorization in corporate applications.

A. Proposed Work:

The suggested work builds on the current BERT-based data validation system by adding Distil-BERT, a lighter and more efficient variant of BERT, to make it more efficient and scalable. BERT is good at finding incomplete or wrong data inputs by understanding their meaning and context. However, its enormous size and high computing requirements make it hard to use in real time in businesses. To solve these problems, Distil-BERT is used with fewer layers and better training goals, which leads to the same or better accuracy with quicker inference time. The proposed system uses basic NLP preprocessing approaches to handle ERP-based text data. It then uses the Distil-BERT model to sort the entries into "Complete" or "Incomplete." This extension lets you check the data that users provide in real time, uses fewer resources, and makes sure that the predictions are accurate. The solution is built into a Flask-based interface for practical use. This lets businesses check the correctness of their data right away and make sure that their enterprise apps perform more reliably.

B. System Architecture:

In three primary steps, the system architecture is built: Preparation for Model, Creating a Model, and Training and Output. In the first stage, the risk dataset is gathered and cleaned, tokenized, and converted into integers. Then, the dataset is divided into training and testing subsets to make sure that the model is evaluated fairly. In the second step, the BERT-based model is built and trained to sort and forecast four important outputs: Potential Risk, Control, Impact of the Risk, and Internal Control. BERT's tokenizer turns text into contextual embeddings that show how words are related to each other. In the last step, the model is trained with parameters like epochs, batch size, and learning rate. Then it is tested with data it has never seen before. Standard metrics like accuracy, precision, recall, and F1-score are used to measure how well the system works. This makes sure that corporate applications can reliably and accurately find missing or wrong data inputs.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

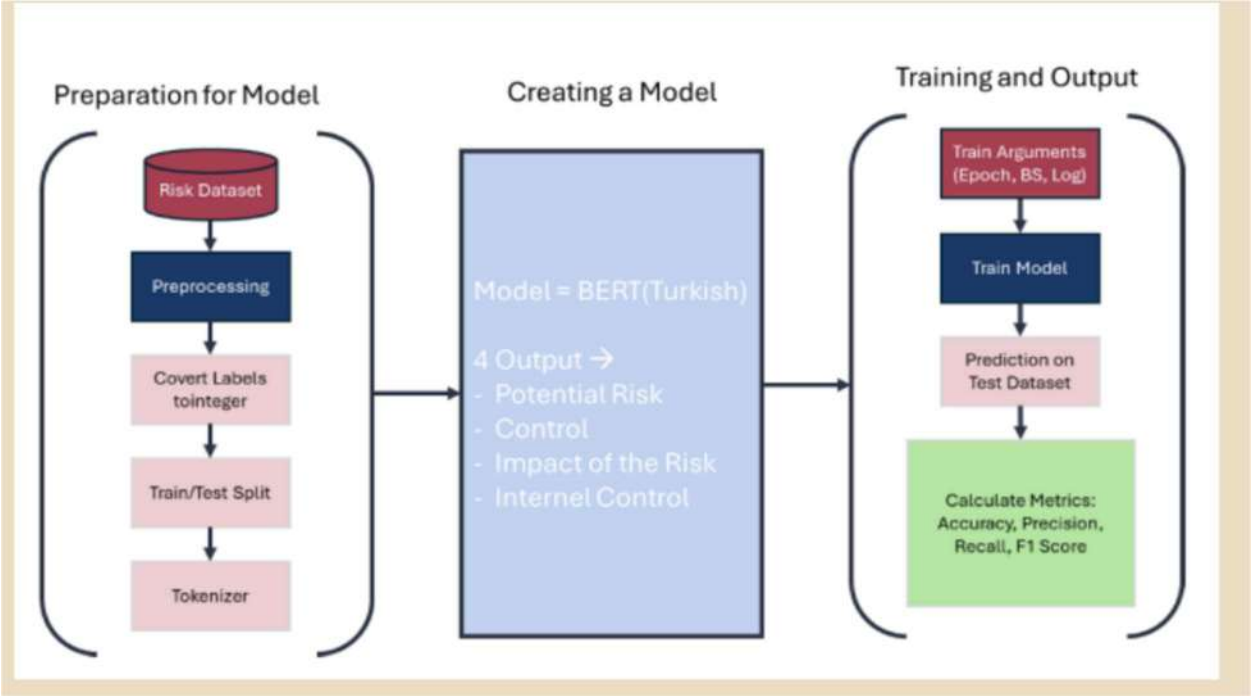

Fig proposed architecture

C. MODULES:

a) Data Collection and Preprocessing

The risk-related dataset is collected from enterprise systems and undergoes cleaning processes such as removal of stop words, special symbols, and URLs. Text normalization, tokenization, stemming, and lemmatization are applied to prepare the data for model training.

b) Data Labeling and Splitting

Each data entry is labeled as 'Complete' or 'Incomplete,' and labels are converted into integer form. The dataset is then divided into training and testing sets to ensure balanced learning and evaluation.

c) Model Creation using BERT/Distil-BERT

A transformer-based deep learning model (BERT or Distil-BERT) is built to analyze the textual data. The model captures contextual meaning and semantic relationships within data entries to identify incomplete or erroneous inputs effectively.

d) Model Training and Validation

The model is trained using defined hyperparameters such as epochs, batch size, and learning rate. During this phase, it learns from the training data and is validated on the test set to fine-tune performance.

e) Performance Evaluation

The trained model is evaluated using metrics such as accuracy, precision, recall, and F1-score. These metrics help in assessing the effectiveness of the proposed model in detecting and classifying data entry quality.

f) Deployment through Flask Interface

The validated model is deployed in a Flask-based web interface, allowing users to input sentences and instantly receive classification outputs ('Complete' or 'Incomplete'), demonstrating the real-time applicability of the system.

D. Algorithms:

a) TF-IDF + SVM (Support Vector Machine):

This method combines the TF-IDF feature extraction technique with a Support Vector Machine classifier to categorize data entries. TF-IDF converts text into numerical representations by assigning weights based on how important a word is to a document relative to the corpus. SVM then constructs an optimal

hyperplane that separates the 'Complete' and 'Incomplete' data entry classes. Although this approach provides a good baseline for classification, it struggles to capture deeper semantic meaning and contextual relationships within text. It performs well on small datasets but has limitations with complex enterprise-level data.

b) Word2Vec + SVM:

Word2Vec transforms words into dense vector embeddings based on their contextual similarity, allowing words with similar meanings to have similar vector representations. These embeddings are then fed into an SVM classifier for training and prediction. The model effectively captures linear relationships between words but lacks the ability to understand sentence-level dependencies or long-range context. While it improves over basic bag-of-words models, its accuracy remains moderate for nuanced text validation tasks where context plays a vital role.

c) BERT (Bidirectional Encoder Representations from Transformers):

BERT is a transformer-based deep learning architecture that reads text bidirectionally—both left-to-right and right-to-left—enabling it to understand the full context of each word within a sentence. It uses Masked Language Modeling (MLM) to predict missing words and Next Sentence Prediction (NSP) to learn relationships between sentences. This makes BERT highly effective in identifying subtle inconsistencies or incomplete entries in enterprise data. Its deep contextual understanding significantly improves classification accuracy and reduces human error in data entry validation.

d) Distil-BERT (Extension Model):

Distil-BERT is a compact and optimized version of BERT that retains around 97% of BERT's accuracy while being 40% smaller and 60% faster. It achieves this by applying a knowledge distillation process, where a smaller model learns from a larger one without losing core representational power. Distil-BERT is ideal for real-time data validation in enterprise systems, as it reduces computational cost and inference time. Despite its reduced size, it maintains superior accuracy, effectively handling large-scale textual data and providing a perfect balance between performance, speed, and resource efficiency.

4. EXPERIMENTAL RESULTS

The experimental study used a labeled dataset of 'Complete' and 'Incomplete' enterprise data items. After preprocessing, TF-IDF+SVM, Word2Vec+SVM, BERT, and Distil-BERT were trained on the dataset. Due to their inability to incorporate contextual semantics, traditional models like TF-IDF+SVM and Word2Vec+SVM have middling accuracies of 65% and 63%. However, the BERT model improved to 94% accuracy with increased precision and F1-score, indicating its ability to interpret complicated textual input. Extending the Distil-BERT model improved results to 98.08% accuracy, beating all other approaches and lowering computing time. Accuracy, precision, recall, F1-score, and ROC curves showed Distil-BERT's greater ability to categorize incomplete or incorrect inputs. The system proved that rich contextual embeddings and optimal transformer topologies improve corporate data entry quality.

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy = TP + TN TP + TN + FP + FN.

$$Accuracy = \frac{(TN + TP)}{T}$$

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 4, 2025

relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{(FN + TP)}$$

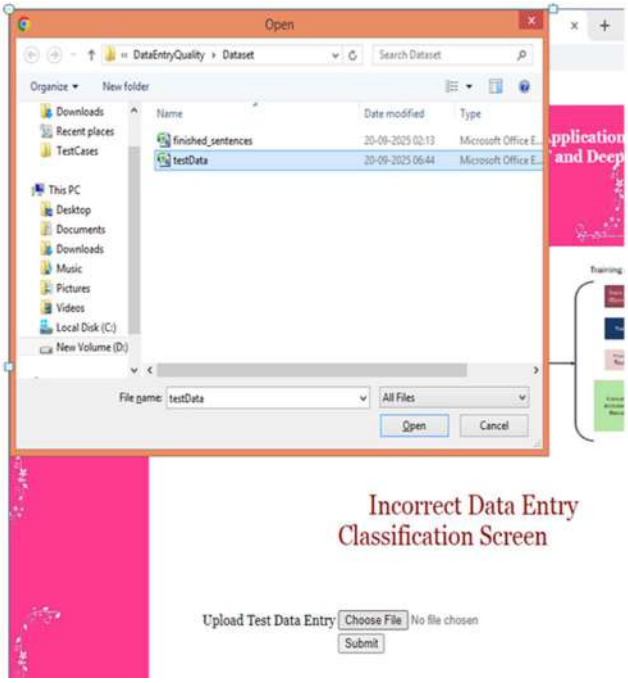

Fig 2 uploading testData.csv code file

Input Data Entry	Classification Status
Technical foul: Chinese traders in online sneaker market punish NBA after HK controversy	Complete
Fed's Powell still has a chance to save the economy before it's too late, Cramer says	Complete
Amazon shareholders reject facial recognition ban as concern grows	Incomplete
Canopy can generate $1 billion in revenue this fiscal year,	Incomplete
'It's a mistake' to write off FANG even as the stocks could still go lower, Cramer says	Complete
Boeing 737 MAX groundings plague U.S. airlines, 'frustrated' Southwest exits Newark	Complete
Trump presses China to reverse stance on structural reform if it	Incomplete
Exclusive: Ride-hailing firm Grab plans major investment in Vietnam	Incomplete

Fig 3 Results

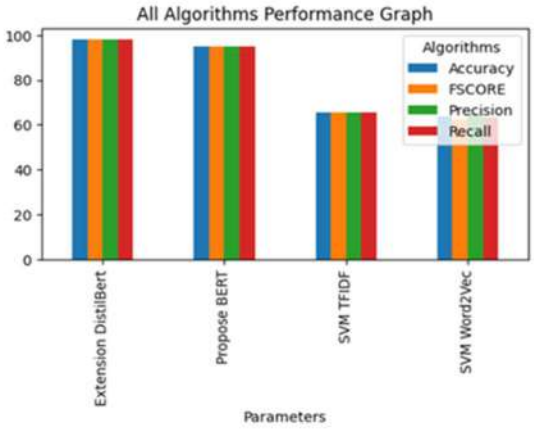

Fig 4 Accuracy graph

Algorithm Name	Accuracy	Precision	Recall	F-Score
SVM TFIDF	65.385	65.374	65.382	65.374
SVM Word2Vec	63.654	65.151	63.360	62.426
Propose BERT	94.808	94.940	94.757	94.798
Extension DistilBERT	98.077	98.073	98.083	98.077

Fig 5 accuracy table

5. CONCLUSION

The method shows a good way to use NLP and deep learning to automatically check the quality of data submission. We used a text-based dataset with the labels "Complete" or "Incomplete." We pre-processed the dataset using typical NLP approaches such removing stop words, URLs, special characters, stemming, and lemmatization. This made sure that the input for model training was clean and normalized. We used a lot of different techniques, such as SVM with TF-IDF, SVM with Word2Vec embeddings, BERT, and Distil-BERT. This made it possible to compare classic machine learning with more sophisticated deep learning approaches in a whole way. The SVM models did a fair job of classifying, getting 65% and 63% accuracy using TF-IDF and Word2Vec features, respectively. The

BERT model performed much better since it used a bidirectional transformer architecture and masked language modeling, reaching 94% accuracy. Distil-BERT improved categorization even more by improving layer depth and execution speed. It achieved the greatest accuracy of 98.08%, showing that it was better at capturing contextual semantic information. Using precision, recall, F1-score, confusion matrices, and ROC curves to test Distil-BERT's ability to predict data input quality showed that it was strong. This system offers a pragmatic solution for automatic data quality evaluation, facilitating precise digital record management.

6. FUTURE SCOPE

This system may be expanded to validate multilingual data, allowing companies in diverse locations to verify data quality. Real-time validation in ERP and CRM platforms may improve operational efficiency by detecting and rectifying data entry mistakes. Advanced transformer models like GPT-based designs or hybrid BERT variations can manage complicated data formats and increase contextual comprehension. The solution might be scaled and made accessible for large organizations in cloud and edge settings. Continuous learning from fresh data helps the model adapt to changing patterns, maintaining data validation dependability and accuracy.

REFERENCES

[1] H.A.M. Abdeljaber, S. Ahmad, A. Alharbi, S. Kumar, XAI-based reinforcement learning approach for text summarization of social IoT-based content, Secur. Commun. Netw. 2022 (2022) 1–12.

[2] A.Z.Z. Abidin, Y. Murdianingsih, U.T. Suryadi, D. Setiyadi, Text summarizing system of english subjects and text mining subjects for computer science students, J. Crit. Rev. 7 (05) (2020) 730–742.

[3] A.A. Abro, M.S.H. Talpur, A.K. Jumani, Natural language processing challenges and issues: a literature review, Gazi Univ. J. Sci. (2022), https://doi.org/ 10.35378/gujs.1032517.

[4] L. Abualigah, M.Q. Bashabsheh, H. Alabool, M. Shehab, Text summarization: a brief review, Stud. Comput. Intell. 874 (December 2019) (2020) 1–15.

[5] D. Adkins, H. Moulaison Sandy, Information behavior and ICT use of Latina immigrants to the U.S. Midwest, Inf. Process. Manag., 57 (3) (2020) 102072.

[6] M. Adnan, M. Ghazali, N.Z.S. Othman, E-participation within the context of e- government initiatives: a comprehensive systematic review, Telemat. Inform. Rep. 8 (2022) 100015.

[7] S.T. Al-Amin, C. Ordonez, Efficient machine learning on data science languages with parallel data summarization, Data Knowl. Eng. 136 (2021) 101930.

[8] H. Alam, A. Kumar, M. Nakamura, F. Rahman, Y. Tarnikova, Che Wilcox, Structured and unstructured document summarization:design of a commercial summarizer using Lexical chains, in: Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings 1, 2003, pp. 1147–1152.

[9] Z. Alami Merrouni, B. Frikh, B. Ouhbi, EXABSUM: a new text summarization approach for generating extractive and abstractive summaries, J. Big. Data 10 (1) (2023) 163.

[10] A.S. Albahri, A.M. Duhaim, M.A. Fadhel, A. Alnoor, N.S. Baqer, L. Alzubaidi, O. S. Albahri, A.H. Alamoodi, J. Bai, A. Salhi, J. Santamaría, C. Ouyang, A. Gupta, Y. Gu, M. Deveci, A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion, Inf. Fus. 96 (2023) 156–191.

[11] S. Alias, Unsupervised text feature extraction for academic chatbot using constrained FP-growth, ASM Sci. J. 14 (2021) 1–11.

[12] F. Amato, V. Moscato, A. Picariello, G. Sperlí, A. D'Acierno, A. Penta, Semantic summarization of web news, Encycl. Semant. Comput. Robot. Intell. 01 (01) (2017) 1630006.

[13] F. Ansari, Knowledge management 4.0: theoretical and practical considerations in cyber physical production systems, IFAC-PapersOnLine 52 (13) (2019) 1597–1602.

[14] L. Anthopoulos, V. Kazantzi, Urban energy efficiency assessment models from an AI and big data perspective: tools for policy makers, Sustain. Cities. Soc. 76 (2022) 103492.

[15] A. Arora, A. Gupta, M. Siwach, P. Dadheech, K. Kommuri, M. Altuwairiqi, B. Tiwari, Web-based

news straining and summarization using machine learning enabled communication techniques for large-scale 5G networks, Wirel. Commun. Mobile Comput. 2022 (2022).

[16] L. B, P Venkata, An overview of text summarization, Int. J. Comput. Appl. 171 (10) (2017) 1–17.

[17] Baralis, E., Cagliero, L., Jabeen, S., Fiori, A., & Shah, S. (2014). Combining semantics and social knowledge for news article summarization (pp. 209–230).

[18] R. Baumgartner, P. Arora, C. Bath, D. Burljaev, K. Ciereszko, B. Custers, J. Ding, W. Ernst, E. Fosch-Villaronga, V. Galanos, T. Gremsl, T. Hendl, C. Kropp, C. Lenk, P. Martin, S. Mbelu, S. Morais dos Santos Bruss, K. Napiwodzka, E., …. Nowak, R. Williams, Fair and equitable AI in biomedical research and healthcare: social science perspectives, Artif. Intell. Med. 144 (2023) 102658, https://doi.org/10.1016/j.artmed.2023.102658.

[19] K. Benharrak, F. Lehmann, H. Dang, D. Buschek, SummaryLens – A smartphone app for exploring interactive use of automated text summarization in everyday life, in: 27th International Conference on Intelligent User Interfaces, 2022, pp. 93–96.

[20] A. Bhaskar, A. Fabbri, G. Durrett, Prompted opinion summarization with GPT-3.5, Find. Assoc. Comput. Linguistics: ACL 2023 (2023) 9282–9300.

[21] A. Bhola, J. Mullapudi, S. Kollipara, T. Sanaka, Text summarization based on ranking techniques, in: 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), 2022, pp. 1463–1467.

[22] D.S. Bitterman, E. Goldner, S. Finan, D. Harris, E.B. Durbin, H. Hochheiser, J. L. Warner, R.H. Mak, T. Miller, G.K. Savova, An end-to-end natural language processing system for automatically extracting radiation therapy events from clinical texts, Int. J. Radiat. Oncol.*Biol. (Basel)*Phys. (College Park Md) 117 (1) (2023) 262–273.

[23] J. Burton, Algorithmic extremism? The securitization of artificial intelligence (AI) and its impact on radicalism, polarization and political violence, Technol. Soc. 75 (2023) 102262.

[24] L. Cao, J. Fu, Improving efficiency and accuracy in english translation learning: investigating a semantic analysis correction algorithm, Appl. Artif. Intell. 37 (1) (2023).

[25] N.V. Chandran, V.S. Anoop, S. Asharaf, TopicStriKer: a topic kernels-powered approach for text classification, Results Eng. 17 (2023) 100949.

[26] H. Chen, Z. Zhang, S. Huang, J. Hu, W. Ni, J. Liu, TextCNN-based ensemble learning model for Japanese Text Multi-classification, Comput. Electr. Eng. 109 (2023) 108751.

[27] Z. Chen, H. Lin, Improving named entity correctness of abstractive summarization by generative negative sampling, Comput. Speech. Lang. 81 (2023) 101504.

[28] W. Cheng, P. Hu, S. Wei, R. Mo, Keyword-guided abstractive code summarization via incorporating structural and contextual information, Inf. Softw. Technol. 150 (2022) 106987.

[29] A. Chikhi, S.S. Mohammadi Ziabari, J.-W. van Essen, A comparative study of traditional, ensemble and neural network-based natural language processing algorithms, J. Risk. Financ. Manage 16 (7) (2023) 327.

[30] T. Cui, S. Li, System movement space and system mapping theory for reliability of IoT, Fut. Gener. Comput. Syst. 107 (2020) 70–81.