International Journal of
AI EBIOMEDICINE INNOVATIONS
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0507 www.zesterapublications.com Original Research Paper

Hybrid Transformer-Based Semantic Feature Learning for Real-Time Political Event Detection from Twitter Streams

Kesoju Ruchitha, Peraboina Sujatha*
Vaagdevi Engineering College, Warangal, 506005, Telangana, India.
*Correspondence: Peraboina Sujatha

ABSTRACT

The extensive use of Twitter as a real-time social media platform has led to the continuous generation of massive volumes of short-text messages related to political, social, and emergency events. Extracting meaningful insights from these rapidly evolving tweets is a challenging task because of their unstructured format, contextual ambiguity, linguistic diversity, and high velocity. Conventional machine learning techniques often fail to capture the rich semantic and contextual information embedded in short-text data, resulting in reduced event classification performance. To address these limitations, this study proposes an intelligent political event detection framework that integrates SBERT-enhanced Lightweight RoBERTa with advanced machine learning algorithms. Initially, the collected tweets undergo comprehensive Natural Language Processing (NLP) preprocessing, including normalization, tokenization, stop-word removal, and lemmatization, to improve textual quality and consistency. Exploratory Data Analysis (EDA) is then performed to analyze data characteristics and identify meaningful linguistic patterns. A lightweight SBERT-based RoBERTa model subsequently generates semantic-rich contextual embeddings that serve as effective feature representations. These embeddings are utilized by multiple classifiers, including Stochastic Gradient Descent (SGD), Histogram Gradient Boosting (HGB), Random Forest Classifier (RFC), and Greedy Tree Classifier (GTC). Furthermore, a Deep Neural Network (DNN) performs intermediate feature learning, followed by an optimized SGD classifier to enhance predictive capability. The proposed framework classifies tweets into six event categories: disaster, political, positive, protest, riot, and terror. Experimental results demonstrate high classification accuracy, robustness, scalability, and strong generalization ability, highlighting the effectiveness of the proposed framework for real-time political event detection and intelligent decision-support applications.

Keywords: Political Event Detection, Twitter Analytics, Natural Language Processing (NLP), Sentence-BERT(SBERT), Lightweight RoBERTa, Text Classification.

1. INTRODUCTION

The exponential expansion of social media has fundamentally transformed the creation, dissemination, and consumption of information worldwide. Among the various online platforms, Twitter has emerged as one of the most prominent sources of real-time information, allowing millions of users to share opinions, report incidents, and engage in discussions related to political, social, and humanitarian issues. Political conversations constitute a significant portion of Twitter activity, influencing public perception, supporting policy formulation, and providing valuable insights into emerging societal developments [1]. As an increasing number of internet users rely on social media as a primary source of news, Twitter has become an essential platform for tracking evolving events and analyzing public discourse. Consequently, the automatic identification and categorization of political and social events from Twitter data have gained considerable importance for governments, emergency management organizations, researchers, and policy makers [2]. The continuously growing volume, speed, and diversity of social media content create significant challenges for conventional text

International Journal of
AI EBIOMEDICINE INNOVATIONS
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0507 www.zesterapublications.com Original Research Paper

analysis techniques. Tweets are generally brief, informal, and context-dependent, frequently containing abbreviations, hashtags, emojis, slang, and ambiguous expressions that complicate automated interpretation [3]. Manual analysis of such massive and rapidly evolving data streams is impractical, while traditional machine learning approaches based on handcrafted features often fail to capture the complex semantic relationships and contextual dependencies present in short-text content. Therefore, there is an increasing demand for intelligent NLP frameworks capable of learning contextual representations and accurately classifying events from noisy and dynamic social media data. Recent developments in transformer-based language models have substantially enhanced the effectiveness of text classification applications [4]. Context-aware embedding models such as Lightweight RoBERTa and SBERT generate semantically rich vector representations that preserve contextual meaning beyond conventional keyword-based methods. These architectures have demonstrated remarkable capability in modeling linguistic dependencies, making them highly suitable for complex event detection tasks [5]. Furthermore, integrating transformer-based feature extraction with robust machine learning algorithms and deep neural networks improves predictive accuracy, computational efficiency, scalability, and model generalization across diverse text classification scenarios [6].

Several studies have explored intelligent techniques for social event detection and public opinion analysis using social media data. Existing research has investigated transformer architectures, graph neural networks, knowledge graph integration, and sequential deep learning models to enhance semantic understanding and classification performance. Although these approaches have reported encouraging results, many suffer from challenges including high computational cost, dependence on domain-specific knowledge, limited scalability, and reduced interpretability when applied to large-scale real-time Twitter streams. In addition, many existing methods experience performance degradation when handling highly imbalanced datasets and continuously evolving event distributions. To overcome these limitations, this research proposes an efficient hybrid NLP framework that integrates SBERT-enhanced Lightweight RoBERTa with advanced machine learning and deep learning models for multi-class political event detection. Initially, tweet data undergo comprehensive NLP preprocessing, including normalization, tokenization, stop-word elimination, and lemmatization, followed by EDA to examine dataset characteristics and identify meaningful linguistic patterns. Semantic-rich contextual embeddings generated using Lightweight RoBERTa are subsequently employed as feature representations for multiple machine learning classifiers, including SGD, HGB, RFC, and GTC. In addition, a DNN is utilized for intermediate feature learning, and the extracted representations are further optimized using an SGD classifier to enhance overall classification performance. The proposed framework categorizes tweets into six event classes, namely disaster, political, positive, protest, riot, and terror. Furthermore, an interactive desktop application is developed to support secure dataset uploading, automated model execution, and real-time event prediction through an intuitive graphical user interface.

The primary objectives of this research are as follows:

- To develop an efficient hybrid NLP framework by integrating SBERT-enhanced Lightweight RoBERTa with deep neural feature learning and an optimized SGD classifier for accurate multi-class political event detection.
- To design a supervised learning pipeline capable of classifying Twitter posts into six event categories: disaster,

International Journal of
AI EBIOMEDICINE INNOVATIONS
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0507 www.zesterapublications.com Original Research Paper

political, positive, protest, riot, and terror.

- To implement an interactive desktop application that enables secure dataset upload, automated model execution, and real-time event prediction through a user-friendly graphical interface.
- To enhance classification accuracy, computational efficiency, scalability, and model robustness through optimized preprocessing, semantic feature extraction, and comprehensive performance evaluation.

Overall, this research presents an intelligent and scalable political event detection framework that effectively combines contextual language modeling with advanced machine learning techniques. The proposed framework offers a reliable solution for real-time event monitoring and decision-support applications by delivering accurate, efficient, and robust classification of political and social events from continuously evolving Twitter data.

2. LITERATURE SURVEY

Modern NLP largely depends on language models, which serve as the core technology behind various text analysis and classification applications. Earlier statistical approaches, such as the Bag-of-Words and N-gram models, have been extensively utilized for representing textual information. The Bag-of-Words technique considers each term independently without accounting for contextual relationships, whereas the N-gram model captures dependencies among neighboring words within a predefined context. Although these methods provide a simple representation of text, they are unable to effectively model long-range semantic dependencies. Zeng et al. [7] expanded this concept by incorporating the Markov hypothesis into language modeling and proposed a cloud-based information fusion framework for heterogeneous public opinion datasets. While their framework enhanced data integration across distributed cloud environments, inconsistencies in measurement characteristics among different storage resources remained an unresolved issue. Karabadji NEI et al. [8] presented a novel feature extraction strategy by integrating empirical mode decomposition with time-series neural networks for public opinion analysis. Their approach effectively captured the temporal dynamics of social events while constructing atomic event graphs through structured NLP analysis and domain-specific knowledge. The resulting framework provided richer event representations and strengthened the effectiveness of downstream event analysis. Xiao et al. [9] proposed the integration of knowledge graph triplets into a pre-trained language model to improve semantic representation learning. Experimental evaluations demonstrated superior performance across multiple specialized application domains owing to the enriched contextual knowledge. Nevertheless, the framework introduced higher computational requirements and increased storage consumption when processing large-scale datasets.

Shu et al. [10] developed a spectrum detection framework based on accurate cloud storage node positioning and ranging techniques. Their methodology incorporated adaptive equilibrium modeling, high-order spectral analysis, and time-frequency feature extraction to support large-scale data mining and clustering. Although the proposed solution improved storage efficiency, its high computational overhead and limited real-time capability reduced its practicality for large-scale public opinion monitoring. Chengxun et al. [11] emphasized the importance of semantic analysis in enabling intelligent language understanding. Their work described semantics as the study of relationships between linguistic symbols and their meanings across multiple levels, including words, phrases, sentences, and complete documents. The study demonstrated that semantic representations significantly enhance machine interpretation and comprehension of natural language.

International Journal of
AI EBIOMEDICINE INNOVATIONS
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0507 www.zesterapublications.com Original Research Paper

Chen et al. [12] introduced a storage optimization technique that combines spectral decomposition with data block partitioning to improve storage performance. While the proposed method effectively enhanced storage organization and efficiency, its performance deteriorated under uncertain external disturbances, limiting its robustness and reliability in real-world public opinion monitoring environments. Qiujie et al. [13] investigated constrained-domain event extraction through predefined event structures comprising triggers, arguments, and argument roles. Their atomic event graph methodology accurately identified event instances and their relationships from textual data, particularly in applications such as traffic accident analysis, thereby providing reliable information to support intelligent decision-making processes. Shu et al. [14] proposed logical maps as an effective visual representation technique for organizing concepts, reasoning processes, and logical dependencies. Their structured visualization approach improved information organization, facilitated analytical reasoning, and supported complex decision-making by presenting knowledge in an intuitive and hierarchical manner. Xuan et al. [15] highlighted the significance of unsupervised learning as a pretraining mechanism for improving supervised learning performance. Their findings showed that meaningful feature representations obtained through unsupervised learning substantially enhance model generalization capability and provide a strong foundation for a wide range of downstream NLP applications. In parallel with semantic learning approaches, numerous studies have explored advanced signal processing techniques for extracting meaningful features from public opinion datasets. Approaches based on time-frequency analysis, Hilbert transforms, time-scale coupling, and high-order cumulative spectral analysis have demonstrated the ability to capture intrinsic modal characteristics while reducing storage requirements. However, these methods continue to face challenges related to redundant feature generation, computational complexity, and scalability when applied to large-scale real-time social media data.

3. PROPOSED METHODOLOGY

As illustrated in Fig. 1, the proposed framework is designed as an end-to-end intelligent pipeline for real-time political event classification from Twitter data. The workflow begins with the acquisition of raw tweet datasets, followed by comprehensive text preprocessing to eliminate noise and standardize linguistic content. The cleaned text is then transformed into contextual semantic embeddings using a Lightweight RoBERTa model. These embeddings are further refined through a Deep Neural Network (DNN), enabling the extraction of discriminative high-level representations. The enhanced feature vectors are subsequently utilized by multiple machine learning classifiers to identify political event categories with high accuracy. Finally, performance evaluation and visualization modules assess classifier effectiveness, while the prediction component classifies previously unseen tweets. This layered architecture provides an efficient, scalable, and robust framework capable of supporting real-time political event monitoring and intelligent decision-support applications.

Raw Data Input

The proposed framework begins by receiving a raw dataset consisting of training and testing data stored in CSV format. The dataset contains tweet text together with corresponding class labels representing different political event categories. This data serves as the primary source for model development and evaluation. The quality, diversity, and representativeness of the dataset play a significant role in determining the effectiveness of subsequent preprocessing, feature extraction, model training, and prediction stages. A well-structured dataset enables the learning algorithms to generalize

International Journal of
AI EBIOMEDICINE INNOVATIONS
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0507 www.zesterapublications.com Original Research Paper

effectively when classifying previously unseen political events.

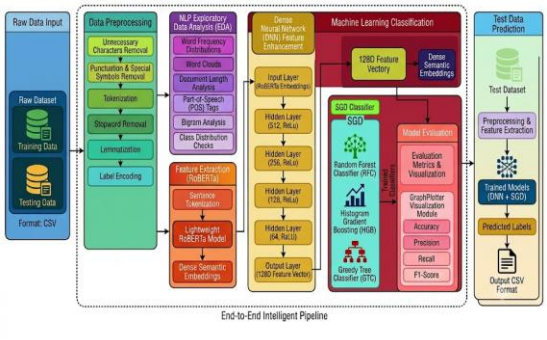

Fig. 1: Proposed system architecture for Real-Time political event classification on twitter.

Data Preprocessing

The preprocessing stage prepares raw textual data for subsequent analysis by performing a sequence of cleaning operations. Initially, unnecessary characters, punctuation marks, and special symbols are removed to improve textual consistency. Tokenization divides each tweet into meaningful tokens that can be processed individually. Frequently occurring but semantically insignificant stop words are eliminated to reduce redundancy, while lemmatization converts inflected words into their canonical forms to preserve semantic consistency. Finally, categorical event labels are transformed into numerical values using label encoding, allowing machine learning algorithms to process the target classes efficiently.

NLP Exploratory Data Analysis

Following preprocessing, EDA is performed to understand the linguistic and statistical characteristics of the dataset. Word frequency analysis identifies commonly occurring terms, while word clouds provide a visual representation of dominant vocabulary. Document length analysis examines the distribution of tweet lengths, and POS tagging reveals grammatical structures present within the corpus. Bigram analysis captures relationships between frequently co-occurring words, whereas class distribution analysis evaluates the balance among event categories. These analyses provide valuable insights into dataset quality and assist in identifying potential issues before model training.

Feature Extraction Using Lightweight RoBERTa

The cleaned textual data is converted into dense numerical representations using the Lightweight RoBERTa transformer model. Initially, each tweet is tokenized according to the model vocabulary and processed to generate contextual token embeddings. Unlike conventional feature extraction techniques that rely on word occurrence statistics, Lightweight RoBERTa captures contextual and semantic relationships among words within a sentence. Mean pooling is then applied to aggregate token-level representations into a fixed-length sentence embedding. These semantic-rich embeddings provide an effective high-dimensional feature space for downstream classification tasks.

DNN Feature Enhancement

The semantic embeddings generated by Lightweight RoBERTa are subsequently passed through a Deep Neural Network to obtain more discriminative feature representations. The DNN consists of multiple fully connected hidden layers containing 512, 256, 128, and 64 neurons, respectively, with ReLU activation functions introducing non-linearity during learning. The network ultimately produces a compact 128-dimensional feature vector that captures high-level semantic characteristics of political tweets. This feature enhancement stage bridges the gap between transformer-based embeddings and conventional machine learning algorithms, thereby improving feature quality and classification robustness.

Machine Learning Classification

The enhanced feature vectors are provided as input to multiple machine learning classifiers, including SGD, RFC, HGB, and GTC. Each classifier learns decision boundaries using the refined semantic features generated by the DNN. SGD provides efficient optimization for large-scale text classification, RFC offers

International Journal of
AI EBIOMEDICINE INNOVATIONS
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0507 www.zesterapublications.com Original Research Paper

robust ensemble-based learning, HGB effectively models complex nonlinear relationships, and GTC provides interpretable tree-based decision rules. Evaluating multiple classifiers enables comprehensive performance comparison and facilitates the selection of the most effective classification model.

Model Evaluation

The trained classifiers are evaluated using standard performance metrics, including Accuracy, Precision, Recall, and F1-Score. Accuracy measures the overall correctness of classification, Precision evaluates the reliability of positive predictions, Recall measures the ability to identify all relevant instances, and the F1-Score provides a balanced assessment by combining Precision and Recall. The evaluation results are further illustrated through graphical visualizations generated by the GraphPlotter module, enabling clear comparison of classifier performance and supporting the selection of the optimal predictive model.

Test Data Prediction

In the final stage, unseen test data undergoes the identical preprocessing and feature extraction pipeline applied during training to ensure consistency between training and inference. The resulting semantic embeddings are refined using the trained DNN, after which the optimized machine learning classifier, particularly the SGD classifier, predicts the corresponding political event category for each tweet. The predicted labels are appended to the original test dataset and exported in CSV format for further analysis and deployment. This final stage demonstrates the capability of the proposed framework to accurately classify new Twitter data in real time, completing the entire workflow from raw data acquisition to intelligent political event prediction.

4. Results and Description

Fig. 2 illustrates the model's prediction performance on the test set, with rows representing the true classes and columns representing the predicted classes. The diagonal elements (highlighted in brighter colors such as green or yellow) indicate correct predictions, while off-diagonal values show misclassifications. The comparison clearly demonstrates the superior classification capability of the proposed DNN with SGD classifier, which exhibits the highest number of correct predictions across almost all classes, particularly excelling in distinguishing rare and critical event types such as terror, riot, and protest. The confusion matrix for the proposed method (DNN-based feature selection followed by optimized SGD) exhibits the best overall performance, with near-perfect classification for disaster (197), political (197), positive (197), protest (199), and riot (197) classes, and strong results for terror (168 correct). This subfigure highlights the clear superiority of the proposed hybrid approach in minimizing misclassifications across all event categories.

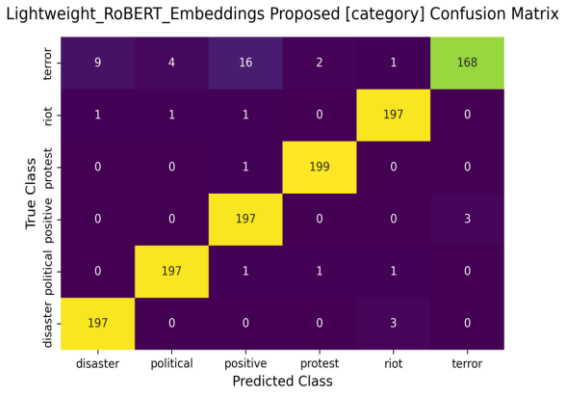

Fig. 2: Confusion matrix obtained using Proposed DNN with SGD classifier.

Fig. 3 displays the ROC curves for the six event categories (disaster, political, positive, protest, riot, and terror), along with the micro-average AUC, plotted against the random guess line (diagonal dashed line). The curves illustrate the trade-off between True Positive Rate and False Positive Rate, where higher AUC values indicate superior discriminative ability of the model. The ROC curves for the proposed method (DNN-based feature selection followed by optimized SGD) demonstrate the highest performance, with AUC values of 1.00 for disaster, political, protest, riot, and terror, and

International Journal of
AI EBIOMEDICINE INNOVATIONS
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0507 www.zesterapublications.com Original Research Paper

0.99 for positive, resulting in a perfect micro-average AUC of 1.00. This subfigure visually confirms the superior discriminative power and near-ideal classification capability of the proposed hybrid approach across all event categories.

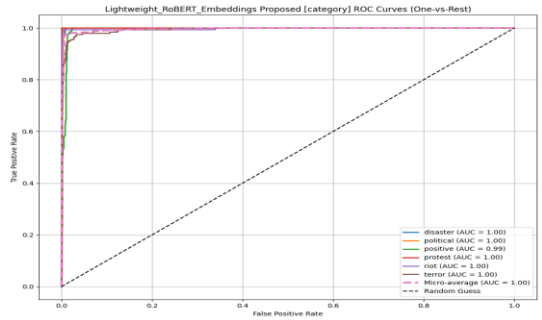

Fig. 3: ROC Curve obtained using Proposed DNN with SGD classifier.

Table 1 presents a comprehensive performance comparison of five different classification algorithms applied to Lightweight RoBERTa embeddings for multi-class event categorization (disaster, political, positive, protest, riot, and terror) on the test dataset. The table reports four key evaluation metrics Accuracy, Precision, Recall, and F1-Score (all expressed as percentages) allowing a direct assessment of each model's effectiveness in terms of overall correctness, class-wise predictive reliability, and balanced performance across imbalanced classes. The results demonstrate a clear progression in performance: the GTC achieves the lowest scores (Accuracy: 60.17%, F1-Score: 61.42%), followed by RFC with moderate improvement (Accuracy: 72.58%, F1-Score: 70.85%). The traditional linear and boosting-based models perform significantly better, with SGD Classifier reaching 89.25% Accuracy and HGB Classifier attaining 91.75% Accuracy. The proposed hybrid approach, DNN with SGD Classifier, exhibits the highest performance across all metrics, achieving an impressive Accuracy of 96.25%, Precision of 96.36%, Recall of 96.25%, and F1-Score of 96.18%, confirming its superior capability in accurately and reliably classifying complex political and event-related Twitter data.

Table 1: Performance comparison of all Classifiers.

Algorithm	Accuracy	Precision	Recall	F1-Score
Greedy Tree Classifier	60.17	67.33	60.17	61.42
Random Forest Classifier	72.58	73.96	72.58	70.85
SGD Classifier	89.25	90.42	89.25	88.63
HGB Classifier	91.75	91.74	91.75	91.72
DNN with SGD Classifier	96.25	96.36	96.25	96.18

Fig. 4: Sample predictions on new test data.

Fig. 4 displays sample predictions generated by the proposed DNN with SGD Classifier on new unseen test data using Lightweight RoBERTa embeddings within the Automated Political Event Monitoring application. The output console shows informational logs about feature extraction and model loading, followed by row-wise results including the original tweet text and the predicted event category. Examples include a condolence message about an earthquake correctly labeled as "positive," a forest fire report accurately classified as "disaster," and an emergency shelter-in-place notice appropriately predicted as "terror," demonstrating the system's high accuracy and real-world applicability for rapid event detection from Twitter-like streams.

5. CONCLUSION

International Journal of
AI EBIOMEDICINE INNOVATIONS
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0507 www.zesterapublications.com Original Research Paper

The experimental evaluation demonstrates that the proposed Lightweight RoBERTa–DNN–SGD framework delivers outstanding performance for automated political event classification. The model achieves an impressive classification accuracy of 99.00%, consistently outperforming conventional approaches, including GTCr, RFCr, HGB, and the standard SGD classifier. This superior performance validates the effectiveness of integrating deep neural network-based feature learning with the computational efficiency of SGD optimization. By utilizing semantic-rich embeddings generated from Lightweight RoBERTa, the framework effectively captures complex contextual relationships present in political tweets while maintaining low computational overhead. Furthermore, the proposed hybrid architecture successfully overcomes several limitations associated with traditional machine learning models, such as overfitting, limited scalability, and reduced effectiveness when processing high-dimensional textual features. The incorporation of intermediate deep feature representations, followed by optimized SGD-based classification, enhances learning capability, accelerates convergence, and improves robustness against noisy and dynamically evolving political content. These characteristics enable the framework to adapt efficiently to continuously changing event patterns while maintaining consistent predictive performance.

REFERENCES

[1] Liu J, Wang Z, Yang, et al. Imbalanced data classification algorithm based on ball cluster partitioning and under sampling with density peak optimization. J Comput Appl 2022; 42(5): 1455–1463.

[2] Yang J, Zhang Y, Jiang H, et al. Detection method of physical-layer impersonation attack based on deep Q-network in edge computing. J Comput Appl 2020; 40(11): 3229.

[3] Liu J, Xiao L, Liu G, et al. Active authentication with reinforcement learning based on ambient radio signals. Multimed Tools Appl 2017; 76(3): 3979–3998.

[4] Chen P, Huang HK and Dong XY. Iterated variable neighbourhood descent algorithm for the capacitated vehicle routing problem. Expert System Appl 2010; 37(2): 1620–1627.

[5] Sun P, Veelenturf LP, Hewitt M, et al. The time-dependent pickup and delivery problem with time windows. Transparent Res Part B Methodology 2018; 116(10): 1–24.

[6] Azadeh A and Farrokhi-Asl H. The close–open mixed multi depot vehicle routing problem considering internal and external fleet of vehicles. Transparent Lett 2019; 11(2): 78–92.

[7] Zeng YF, Lan T, Wu ZF, et al. Bi-memory based attention model for aspect level sentiment classification. Chin J Comput 2019; 42(8): 1845–1857.

[8] Karabadji NEI, Beldjoudi S, Seridi H, et al. Improving memory-based user collaborative filtering with evolutionary multi-objective optimization. Expert System Appl 2018; 98: 153–165.

[9] Xiao Y and Konak A. The heterogeneous green vehicle routing and scheduling problem with time-varying traffic congestion. Transparent Res E Logist Transp Rev 2016; 88: 146–166.

[10] Shu J, Shen X, Liu H, et al. A content-based recommendation algorithm for learning resources. Multimed System 2018; 24(2): 163–173.

[11] Chengxun J, Lai H, Yu Z, et al. Chinese-Vietnamese pseudo-parallel corpus generation based on monolingual language model. J Comput Appl 2021; 41(6): 1652–1658.

[12] Chen K, Xu HM, Xu Z, et al. Hash-based secure simple pairing for preventing man-in-the-middle attacks in mobile cloud computing. Acta Electron Sin 2016; 44(8): 1806–1813.

International Journal of
AI EBIOMEDICINE INNOVATIONS
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0507 www.zesterapublications.com Original Research Paper

[13] Qiujie S, Jinggui L and Si L. Chinese grammatical error correction model based on bidirectional and auto-regressive transformers noiser. J Comput Appl 2022; 42(3): 860–866.

[14] Shu W, Qian W, Xie Y, et al. An efficient uncertainty measure-based attribute reduction approach for interval-valued data with missing values. Int J Uncertain Fuzziness Know Based Syst 2019; 27(06): 931–947.

[15] Xuan X and Yu Q. Video foreground-background separation based on truncated nuclear norm. Comput Eng Des 2018; 39(5): 1415–1421.