

# Guest Editorial: Special Issue on Generating Human Readable Explanations in NLP

**I**N recent years, the quality of state-of-the-art models has significantly improved, but this progress has come at the cost of reduced interpretability. Developing explainable systems is a critical challenge in natural language processing (NLP), as most machine learning models do not provide explanations for their predictions. Explainability in deep learning (DL) is an emerging field addressing how DL methods make decisions. It involves techniques that produce human-comprehensible solutions, enhancing prediction accuracy, decision understanding, and traceability of actions. Explainable DL aims to improve human understanding, justify machine-made decisions, introduce trust, and reduce bias. Building explainable methods is crucial in NLP because existing models often lack transparency. Most current explainability methods focus on interpreting the outputs or the connections between inputs and outputs. However, they often overlook fine-grained information, making it difficult for humans to understand the generated explanations. A transparent, interpretable, and explainable system is necessary to provide human-understandable information, better equipping us to address the challenges and mitigations of NLP vulnerabilities. This special issues, focus on recent development on interpretable, and explainable system in human understandable form. In total, we have received 67 submission, and 19 articles have been accepted which represent the most recent research on generating explanations.

In the legal domain, ensuring fairness and reasonableness in every judge's decision is crucial. Unjust or incorrect judgments can undermine public trust in the national judicial system. Therefore, it is important for assistance systems aiding judges in decision-making to provide interpretable results, enhancing transparency and confidence in the legal process. Most of the existing methods on charge prediction often treats it as a multiclass classification task, resulting in poor interpretation due to weak semantic links between fact descriptions and charge labels. To address this, Li et al. propose a generative evidence mining method based on prompt learning [A1]. During training, charge labels are reformulated into a prompt template to enhance semantic correlation with fact descriptions. In testing, charge labels are generated using prompt learning, and the sentence with the highest attention score from the transformer's multihead self-attention is selected as evidence.

Traditional methods focus on quantitative measures, such as citation counts, and ignore the qualitative context, such as whether citations are positive, negative, or neutral. Daud et al.

[A2] study introduces a new metric, the context-based article impact factor (CBAIF), which evaluates scientific articles by considering the sentiment of citations and potential conflicts of interest between authors i.e. CBAIF not only considers the positive, negative, or neutral context of the citations but also involves the citing and cited author's conflict-of-interest relationship for the evaluation of their scientific impact. Experimental results demonstrate that CBAIF provides more accurate and fair rankings of articles compared to traditional article impact factors (AIFs) by incorporating the context of citations. Khan et al. explore the pervasive issue of fake news on social media and the various countermeasures, primarily involving DL and NLP [A3]. The model uses user comments as well as their news contents to mutually apprehend top-k explainable check-worthy user comments and sentences for detecting fake news. It highlights the importance of explainability in fake news detection systems, noting that understanding why news is identified as fake is as crucial as the detection itself and presented a new model that uses a co-attention subnetwork to analyze user comments and news content, improving the detection of fake news. Experiments show that model surpassed the existing methods in accuracy and explainability, achieving higher scores in F1, normalized cumulative gain (NCG), and precision in identifying top user comments that explain why an article might be fake. Text steganography, used for covert communication, faces challenges in balancing capacity and imperceptibility. Traditional text-selection methods have low hidden capacity and are impractical, while text-generation methods, though higher in capacity, often produce semantically incoherent long texts. To address these issues, Cao et al. introduce novel method based on generating long, readable texts [A4]. It uses the plug and play language model (PPLM) to ensure topic relevance and semantic coherence. Secret messages are embedded by choosing appropriate words from an embeddable candidate word pool (ECWP), which prevents low-quality or grammatically incorrect outputs by avoiding low-probability word choices during text generation. Muhuri et al. [A5] evaluate the role and impact of women only universities and institutions in India's research domain, analyzing their collaborative efforts within a social network of educational institutions. Using centrality metrics such as closeness, betweenness, and eigenvector, the influence of individual institutes is assessed. An overlapping community detection method is introduced to evaluate the position of gender-biased universities within this network. Results

demonstrate that this approach outperforms existing methods on real-life datasets, providing a new framework for strategic analysis in higher education. In another work [A6], Wang et al. introduce a novel deep tensor evidence fusion (DTEF) network for multimodal sentiment classification. The approach includes a common view evaluation network using LSTM and tensor-based neural networks to extract intermodal and intramodal information, and a unique time cue evaluation network leveraging temporal granularity. It incorporates uncertainty through a trusted fusion layer to enhance accuracy and robustness. Validated on the CMU-MOSEI and CMU-MOSI datasets, the experimental results show that the proposed network outperforms state-of-the-art methods in accuracy. Fake news can impact our society in different ways.

Sufi and Khalil [A7] develop a comprehensive system to analyze social media feeds related to disasters in 110 languages using AI and NLP techniques such as sentiment analysis, named entity recognition (NER), anomaly detection, regression, and Getis Ord Gi algorithms. The system was deployed and tested on live Twitter feeds from 28 September 2021 to 6 October 2021, processing tweets in 39 different languages. The framework successfully extracted 9727 location entities with over 70% confidence from the live Twitter feed, providing disaster intelligence. The system achieved high precision, recall, and F1-score rates of 0.93, 0.88, and 0.90 respectively, resulting in an overall accuracy of 97%. This study is the first to incorporate location intelligence with NER, sentiment analysis, regression, and anomaly detection on social media messages related to disasters across a wide range of languages.

To categorize, modify, and expand training instances, Ahmed and Lin employ a multifaceted approach [A8]. Initially, hate speech lexicons and online forums were used to train embeddings through transfer learning, followed by synonym expansion to enhance semantic vectors. Active learning was then applied, training the model with result-label pairs, using entropy-based selection and visualization to identify unlabeled text for each cycle. This iterative process increased the number of training instances and improved model accuracy. As a result, the semantic embedding and lexicon expansion improved the model's ROC. Johnson et al. [A9] present ML-based social media thematic campaign classification (TCC) framework utilizes a novel campaign network attribute (CNA) concept to encode network data features for classifying campaign types. This CNA-based analysis was trained and validated with Twitter and Instagram data, demonstrating scalability. The goal is for CNA-based neural networks to identify and classify thematic campaigns from general social media data, which is valuable for governmental and commercial security services analyzing large datasets. The key contributions of this work are proposing a new TCC framework for social media networks and introducing a novel CNA model for encoding social media features and their thematic classification patterns. Shahid et al. [A10] systematically serve the existing state-of-the-art approaches designed to detect and mitigate the dissemination of fake news, and based on the analysis, authors discuss several key challenges and present a potential future research agenda, especially incorporating AI explainable fake news credibility system.

More than 800 000 people die each year from suicide globally, which makes it a severe public health problem. Most of existing methods for identifying suicide risk on social media have shown promise but struggle to capture both low- and high-level features within the complex structure of social media data and lack explanatory capability. To address this, the article [A11] introduces a novel hybrid text representation method combining word and document-level representations to explain suicide risk identification. This method is integrated into a transformer-based encoder with ordinal classification. Results demonstrate that the proposed method outperforms current benchmarks, achieving an F-score of 0.79 (a 15% increase) on a public suicide dataset. The study suggests that explainable models can match the performance of nonexplainable ones while offering advantages for translation into clinical and public health practices. Jain et al. present co-learning-based approach to enhance explainability in NLP-based multimodal sentiment analysis [A12]. It addresses issues of noisy or missing data during training or testing by determining modality dominance and extracting both local and global model explanations. Validated with LIME and SHAP methods, the approach provides insights into modality contributions and interactions, ensuring trust and robustness. These explanations help system designers and developers understand complex model dynamics, which is particularly challenging in multimodal applications.

Salim et al. present differentially privacy blockchain-based explainable federated learning (DP-BFL) framework to develop privacy-aware ML models within a federated learning ecosystem [A13]. This framework leverages SM 3.0 networks to enable decentralized learning from Internet-enabled devices while preserving privacy. Participants upload differentially private local updates to blockchain miners, who evaluate and reward them. Experimental results on real-world datasets, SM 3.0 and MNIST, demonstrate high utility, enhanced privacy, and improved efficiency. DP-BFL achieves performance improvements with high privacy and comparable utility to standard FL and centralized learning approaches. It also enhances recognition of user preferences and image class prediction while mitigating the impact of malicious entities' poisoned updates. DP-BFL ensures differential privacy on local model updates, making it equivalent to standard FL with additional privacy preservation on uploaded model updates.


Online education and distance learning are now widely adopted. Numerous platforms and institutions, such as Coursera, edX, Udemy, Virtual University, DigiSkills, Open Courseware, and Khan Academy, offer online courses. These courses typically feature audio and video lectures, assignments, quizzes, exams, and grading. MOOC platforms offer discussion forums where students can share their thoughts and problems about the course. Instructors need to monitor these forums to identify student difficulties and improve their teaching methods. With the goal to help instructors enhance their teaching strategies and improve student understanding. Amjad et al. [A14] present a framework to categorize discussion threads into topics and subtopics using topic modeling, followed by sentiment analysis to determine the sentiment of the comments.


Part-of-speech (POS) taggers are essential for NLP. Traditional POS taggers and libraries are typically expert-created or static and focus on literary text. This limitation affects the performance of other NLP tasks such as polarity detection, sentiment analysis, and opinion mining. Samantra et al. [A15] first critically evaluate the shortcomings of the conventional POS taggers and then present a preliminary study on the suitability of neural taggers over static or manual taggers, supported by the accuracy of 97% achieved on Hamlet. NER is challenging due to the need to identify entity boundaries and handle hierarchically nested entities. Most existing work focuses on either Flat NER or Nested NER, but few methods effectively determine entity positions and use text grammar. Zhou et al. introduce PANNER, a POS-aware Nested NER model, to address these challenges [A16]. It constructs a heterogeneous graph using POS information and employs a dilated random walk (DRW) algorithm to sample neighbors for each node based on grammatical paths. An attention mechanism aggregates messages from various neighbors, and a bidirectional decoding module identifies and categorizes flat and nested entities layer by layer. Rodrigues et al. present reproducible framework to create a forensic pipeline for extracting information from texts using NLP [A17]. The findings demonstrate that it is feasible to develop NER and RE models in any language, fine-tune their hyperparameters to achieve state-of-the-art performance, and construct comprehensive knowledge graphs, thereby reducing the need for human supervision and review. Additionally, the research suggests that addressing this task in stages can further enhance performance.


Malik et al. [A18] present eFRVFL model which incorporates three types of features: original, supervised randomized, and unsupervised randomized, which help capture nonlinear relationships in the dataset. Despite its advanced features, eFRVFL is an unstable classifier. To enhance performance, an ensemble learning approach is proposed, resulting in the ensemble of extended feature RVFL (en-eFRVFL) model. This ensemble model trains each base model on different feature spaces, leading to more accurate and diverse base models, offering greater stability, accuracy, and generalization compared to single models. IFRVFL assigns each sample an intuitionistic fuzzy number (IFN) based on its membership and nonmembership scores. These scores are determined by the sample's distance from the centroid of its class and its neighborhood. The model effectively mitigates the influence of outliers. To assess its efficiency, Malik et al. apply IFRVFL to diagnose Alzheimer's disease (AD) [A19]. Experimental results indicate that IFRVFL outperforms in distinguishing between mild cognitive impairment (MCI) and AD cases, suggesting its potential for early AD diagnosis in clinical settings.


#### ACKNOWLEDGMENT

This special issue would not have been possible without the support of Bin Hu, the Editor-in-Chief of IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS; colleagues from IEEE Systems, Man, and Cybernetics Society and IEEE Computer Society; and all reviewers involved the review process.


IMRAN RAZZAK , *Guest Editor*  
University of New South Wales  
Sydney, NSW 1466, Australia  
e-mail: imran.razzak@unsw.edu.au

REDA BOUADJENEK , *Guest Editor*  
Deakin University  
Geelong, VIC 3216, Australia  
e-mail: reda.bouadjenek@deakin.edu.au

AAMIR CHEEMA , *Guest Editor*  
Monash University  
Melbourne, VIC 3800, Australia  
e-mail: aamir.cheema@monash.edu

IBRAHIM A. HAMEED , *Guest Editor*  
Norwegian University of Science and Technology  
7491 Gjøvik, Norway  
e-mail: ibib@ntnu.no

GUANDONG XU , *Guest Editor*  
University of Technology, Sydney  
Sydney, NSW 2007, Australia  
e-mail: guandong.xu@uts.edu.au

AMIN BEHESHTI , *Guest Editor*  
Macquarie University  
Sydney, NSW 2109, Australia  
e-mail: amin.beheshti@mq.edu.au

#### APPENDIX RELATED ARTICLES

- [A1] L. Li, D. Liu, L. Zhao, J. Zhang, and J. Liu, "Evidence mining for interpretable charge prediction via prompt learning," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4556–4566, Aug. 2024.
- [A2] A. Daud, S. Ghaffar, and T. Amjad, "Citation count is not enough: Citation's context-based scientific impact evaluation," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4567–4573, Aug. 2024.
- [A3] F. Khan, R. Alturki, G. Srivastava, F. Gazzawe, S. T. U. Shah, and S. Mastorakis, "Explainable detection of fake news on social media using pyramidal co-attention network," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4574–4583, Aug. 2024.
- [A4] Y. Cao et al., "Generative steganography based on long readable text generation," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4584–4594, Aug. 2024.
- [A5] S. Muhuri, S. Kumari, S. Namasudra, and S. Kadry, "Analysis of the pertinence of Indian women's institutions in collaborative research," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4595–4604, Aug. 2024.
- [A6] Z. Wang, G. Xu, X. Zhou, J. Y. Kim, H. Zhu, and L. Deng, "Deep tensor evidence fusion network for sentiment classification," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4605–4613, Aug. 2024.
- [A7] F. K. Sufi and I. Khalil, "Automated disaster monitoring from social media posts using AI-based location intelligence and sentiment analysis," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4614–4624, Aug. 2024.
- [A8] U. Ahmed and J. C.-W. Lin, "Deep explainable hate speech active learning on social-media data," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4625–4635, Aug. 2024.
- [A9] N. Johnson, B. Turnbull, M. Reisslein, and N. Moustafa, "CNA-TCC: Campaign network attribute based thematic campaign classification," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4636–4648, Aug. 2024.
- [A10] W. Shahid et al., "Detecting and mitigating the dissemination of fake news: Challenges and future research opportunities," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4649–4662, Aug. 2024.
- [A11] U. Naseem, M. Khushi, J. Kim, and A. G. Dunn, "Hybrid text representation for explainable suicide risk identification on social

- media," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4663–4672, Aug. 2024.
- [A12] D. K. Jain, A. Rahate, G. Joshi, R. Walambe, and K. Kotecha, "Employing co-learning to evaluate the explainability of multimodal sentiment analysis," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4673–4680, Aug. 2024.
- [A13] S. Salim, B. Turnbull, and N. Moustafa, "A blockchain-enabled explainable federated learning for securing Internet-of-Things-based social media 3.0 networks," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4681–4697, Aug. 2024.
- [A14] T. Amjad, Z. Shaheen, and A. Daud, "Advanced learning analytics: Aspect based course feedback analysis of MOOC forums to facilitate instructors," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4698–4706, Aug. 2024.
- [A15] A. Samantra, P. K. Sa, T. N. Nguyen, A. K. Sangaiah, and S. Bakshi, "On the usage of neural POS taggers for Shakespearean literature in social systems," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4707–4717, Aug. 2024.
- [A16] L. Zhou, J. Li, Z. Gu, J. Qiu, B. B. Gupta, and Z. Tian, "PANNER: POS-aware nested named entity recognition through heterogeneous graph neural network," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4718–4726, Aug. 2024.
- [A17] F. B. Rodrigues, W. F. Giazza, R. de Oliveira Albuquerque, and L. J. García Villalba, "Natural language processing applied to forensics information extraction with transformers and graph visualization," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4727–4743, Aug. 2024.
- [A18] A. K. Malik, M. A. Ganaie, M. Tanveer, and P. N. Suganthan, "Extended features based random vector functional link network for classification problem," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4744–4753, Aug. 2024.
- [A19] A. K. Malik, M. A. Ganaie, M. Tanveer, P. N. Suganthan, and A. D. N. I. Initiative, "Alzheimer's disease diagnosis via intuitionistic fuzzy random vector functional link network," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4754–4765, Aug. 2024.