2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC) | 979-8-3503-1653-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/IMCEC59810.2024.10575693

Data-Mining of Social Media Users with Embedding Techniques and Neural Network

Hengshan Cao[1], Mengjiao Yin[1*], Ying Xi[1]
1. College of Business, Taihu University of Wuxi, Jiangsu, China
caohs@wxu.edu.cn, yinmj@wxu.edu.cn, wennie2085@163.com
Corresponding Author: Mengjiao Yin Email: yinmj@wxu.edu.cn

***Abstract*—In this study of ten thousand representative users from the popular Chinese social media platform bilibili.com, we find that usernames−ordinarily considered mere indexing elements rather than useful modeling features − alongside personalized signatures and other natural language data, should and must be integrated into neural network modeling via embedding techniques, leveraging pretrained large language models such as BERT. This work challenges conventional wisdom and emphasizes the importance of harnessing all potentially informative user-generated content, even those previously deemed extraneous. The successful application of embedding methods to these unconventional data sources underscores their versatility in extracting meaning from diverse text types, extending the scope of text-based feature engineering. This discovery holds promise for uncovering subtle user traits, enriching user profiles, and guiding practical applications such as user engagement optimization, content personalization, and targeted marketing efforts. Overall, our study advances the understanding of the value of usernames and signatures as valuable modeling inputs, contributing to the refinement of social media analytics and user profiling.(*Abstract*)**

Keywords—data-mining, social media users, NLP, embedding, LSTM, neural network (key words)

I. INTRODUCTION

Data mining of social media users is a growing field that utilizes advanced computational techniques to extract valuable insights from the vast amount of data generated by users on various social media platforms. The rise of social media data mining has been fueled by factors such as the increasing availability of user data online, the reduced costs associated with data collection and processing, and the expansion of social media platforms [1].

Researchers have explored various applications of data mining in social media, including sentiment analysis, community detection, identification of adverse drug reactions, and prediction of mental illness [2,3]. Through data mining techniques, researchers have successfully extracted valuable information from social media data streams, enabling tasks such as personalized travel recommendations and the detection of irregular user behaviors in online social networks [4].

Furthermore, data mining in social media has played a crucial role in pharmacovigilance, where innovative text and data mining methods have been developed to analyze social media data for adverse drug reactions[5]. Additionally, the application of data mining in social media has expanded to humanitarian efforts, with projects focusing on utilizing social media data for disaster relief and community assistance [6].

While challenges exist in mining knowledge from social media data, recent advancements in data mining techniques have allowed researchers to uncover hidden patterns, sentiments, and user behaviors that were previously challenging to discern, moving beyond earlier keyword-based methods for public health tasks [7].

Thus, data mining of social media users offers a vast landscape for exploration, providing opportunities to extract valuable insights, predict trends, and enhance decision-making processes across various domains. By harnessing data mining techniques, researchers can unlock the potential of social media data to drive innovation and address real-world challenges.

In the realm of data mining of social media users, there is a tendency to overlook natural language data while emphasizing structured data features. Despite the wealth of information embedded in the unstructured text found in social media posts, researchers and practitioners often prioritize structured data characteristics for analysis. This inclination towards structured data features can be attributed to the challenges associated with processing and extracting insights from the vast amounts of unstructured text data available on social media platforms [8,9].

While structured data features offer clear patterns and easily quantifiable metrics, such as likes, shares, and user demographics, natural language data presents complexities in terms of sentiment analysis, context understanding, and noise reduction. The structured data features provide a more straightforward approach to data analysis, allowing for easier classification, clustering, and prediction tasks compared to the nuanced and context-dependent nature of natural language data [10,11].

However, the oversight of natural language data in social media mining can limit the depth of insights that can be derived from user-generated content. Natural

language data holds rich information regarding user sentiments, opinions, and intentions that may not be fully captured by structured data features alone. By incorporating advanced natural language processing techniques, such as sentiment analysis, topic modeling, and text mining, researchers can unlock valuable insights from social media data that go beyond what structured data features can offer [12,13].

In conclusion, while structured data features play a crucial role in data mining of social media users due to their ease of analysis and quantifiability, it is essential to recognize the significance of natural language data in capturing the nuanced aspects of user-generated content. By striking a balance between structured data features and natural language data analysis, researchers can harness the full potential of social media data for gaining comprehensive insights into user behaviors, sentiments, and trends.

Our research aims to delve into the natural language text data left by users on social media platforms, employing feature embedding techniques alongside neural network modeling methodologies to investigate latent, intricate relationships among variables.

II. METHODS

A. Source of Data

Bilibili is a significant Chinese video-sharing platform that has had a notable impact on Chinese youth culture and social interactions. Through qualitative data analysis and digital-field surveys, Bilibili has been identified as a platform that not only entertains but also influences the cohort identity and aesthetic choices of Generation Z Peng [14]. The platform's unique features, such as Danmu comments, have facilitated a sense of community among users, promoting social contact and interaction through creative linguistic and semiotic resources [15]. Furthermore, Bilibili's dynamic community cultures and innovative content have played a role in its growth, increased social recognition, and enhanced commercial value, challenging previous perceptions of being a niche or fragmented platform [16].

For the above reasons, it is a multifaceted platform that not only entertains but also shapes cultural trends, encourages community engagement, and provides unique opportunities for content creators and users. Consequently, it constitutes an excellent proving ground for data mining endeavors targeting social media users.

Due to computational resource constraints, our dataset was sampled from Bilibili, drawing upon a subset of the original open-source dataset consisting of over 800,000 entries available on the Kaggle platform[17]. Utilizing the .sample() method of pandas DataFrame, we randomly extracted a total of 10,000 user records, thereby assembling a small-scale yet sufficiently representative dataset capable of ensuring accurate predictions by neural network models.

TABLE I. FEATURE LIST

Data Type	Feature	Explanation
Index	uid	Identifier, int 1,2...
Image	avatar	Profile Picture
Text(NLP)	name	User name
	sign	User signature, a short text shows their personality
Structured	level	User level, reflects user stickiness
	sex	Users stated gender, male/female/secret
	vip type	Current subscribe plan, int 0/1/2, means non-subscriber/monthly/yearly
	vip_status	Current Membership Status, bool, 1 for yes, 0 for no
	vip_role	Rank of subscriber,int, 0/3/1/7/15
	archive	Number of Submissions as a creator
	fans	Number of followers, reflects popularity
	friend	Number of mutual followers
	like_num	Number of recieved “like”
	is_senior	Senior member or not, bool, 1for yes, 0 for no

Table 1 is a feature list presenting the data types, names, and respective explanations for each attribute. Observation of the dataset reveals that the feature columns within the structured data share certain commonalities, as they quantitatively and precisely capture aspects of the sample's engagement with the social media platform, such as popularity, level of creative enthusiasm, and loyalty to platform usage. Consequently, these structured data metrics can serve dual roles as both features for model construction and as targets for model prediction. In our study, we will accordingly employ “is_senior” and “level” successively as target labels.

B. Embedding technique

Word embeddings are a technique that maps words or phrases from a vocabulary to low-dimensional real-valued vector spaces, thereby imbuing them with a mathematical notion of similarity. The mathematical expression for word embeddings can be written as:

$$w_i \mapsto \mathbf{v}_i = (v_{i1}, v_{i2}, \ldots, v_{id})^\top \in \mathbb{R}^d$$

In the above formula: W_i denotes the i-th word in the vocabulary; V_i is the corresponding word vector for W_i, which is a d-dimensional column vector; V_{ij} represents the component of word vector V_i in the j-th dimension; T denotes the transpose operation; R_d represents the d-dimensional real vector space.

Using embedding techniques before inputting text data into neural networks has been shown to significantly enhance accuracy. A comparative study was conducted and found that pretrained embeddings, when fine-tuned with neural-topic models, enhanced interpretability and performance on short-text corpora [18]. Another study highlighted the use of word embedding and external resource-based embedding techniques for sentiment analysis, showcasing the effectiveness of embeddings in enhancing text analysis tasks [19].

While prior researches have extensively investigated text-based features, distinctive feature such as username, conventionally regarded as unique identifiers and often relegated to index-type data, have received relatively scant attention. This study innovatively transforms usernames into tensors of shape (10000, 768) through the employment of embedding techniques, with the pre-trained large language model BERT serving as the foundation for word embeddings.

BERT (Bidirectional Encoder Representations from Transformers) embedding has emerged as a powerful technique for enhancing various natural language processing tasks. The effectiveness of BERT embeddings has been demonstrated in different applications, such as sentiment analysis and semantic similarity search. A modification of BERT that generates semantically meaningful sentence embeddings suitable for tasks like clustering and similarity comparison [20]. Furthermore, BERT embeddings have been successfully applied in news recommendation systems [21], forensic email author attribution [22], and humor detection [23], showcasing their versatility and effectiveness across diverse domains.

Figure 1 presents word clouds separately for male and female usernames, from which we can indeed discern several common patterns in user naming preferences. Notably, there is a propensity for adopting possessive structures such as "Something of Somebody," with the term "cat" emerging as the most frequently occurring animal within popular usernames. In summary, given the observation that usernames encompass a mix of characters from languages including Chinese, English, and Japanese, we have opted for the "bert-base-multilingual-cased" model as the base for training word embeddings.

Fig. 1.Multilingual Word Clouds of Username Texts

Figure 2 depicts the resulting word embeddings, where the initial textual features of dimensionality (10000, 1) have been transformed into floating-point features of size (10000, 768), rendering them amenable for computation within a neural network.

```
tensor([[ 0.3485, -0.5345,  0.2847,  ..., -0.0355, -0.0182, -0.3593],
        [ 0.3499, -0.5367,  0.2859,  ..., -0.0352, -0.0176, -0.3618],
        [ 0.3506, -0.5368,  0.2991,  ..., -0.0352, -0.0181, -0.3586],
        ...,
        [ 0.3489, -0.5365,  0.2862,  ..., -0.0348, -0.0172, -0.3615],
        [ 0.3488, -0.5357,  0.2850,  ..., -0.0354, -0.0170, -0.3620],
        [ 0.3486, -0.5362,  0.2875,  ..., -0.0379, -0.0176, -0.3616]])
```

Fig. 2.Result of embedding

Subsequently, all remaining structural data, excluding "level," were incorporated as dense features (below called dense 11 for it has 11 feature columns) into a custom-built neural network model, the computational graph of which is illustrated in Figure 3. Next, the length of the additional textual feature "sign" was extracted using the len() function, thereby generating a novel variable "sign_len" which was subsequently fed into the model for prediction purposes.

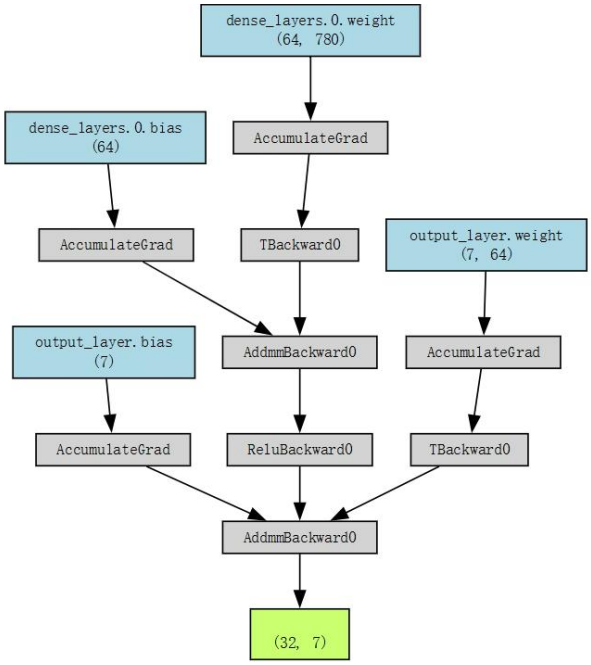

Fig. 3. Calculating Plot of Neural Network Model

C. LSTM Network

Unlike traditional neural networks, LSTM networks are designed to mitigate the vanishing and exploding gradient problems, making them well-suited for processing long sequences without losing important information [24]. The unique architecture of LSTM, with components like input gates, forget gates, and output gates, enables the network to store and retrieve information over extended time periods, facilitating the modeling of sequential data effectively [25,26].

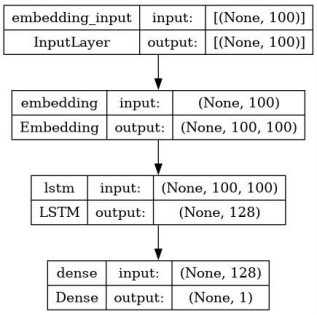

Fig. 4.LSTM model summary

Based on the aforementioned advantages, we decided to employ LSTM as our predictive model. Initially, we performed data cleaning, handling missing values in the "sign" column with the fillna('unknown') method. Subsequently, a dictionary of size=10000 was set up, utilizing the keras.preprocessing.text.Tokenizer method to tokenize the text and generate a dictionary index, assigning indices to all words. The sequences were then padded using the X_train_padded = pad_sequences(max_sequence_length = 100) method. Thereafter, we constructed the model depicted in Figure 4. Given our objective of predicting binary outcomes, we employed binary_crossentropy as the loss function and utilized the adam optimizer. We compared different learning rates to identify the optimal model parameters, ultimately completing the training of the LSTM.

III. RESULTS

In the context of training an LSTM model utilizing user signature features, we have conducted a series of comparative experiments of two prevailing learning rate values, specifically 1e-3 and 1e-4. Figure 5 presents a typical visualization of the model's training history from them for both configurations. The results indicate that a learning rate of 1e-4 constitutes a more reasonable choice, but the difference is very subtle at an average level of 2.4 percent.

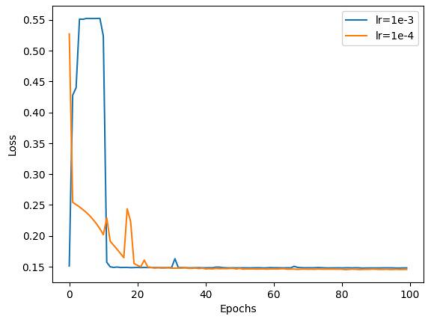

Fig. 5.Loss comparison between learning rates

Following over one hundred of exploratory experiments, we ultimately determined that a learning rate of 1e-3 yields the best prediction performance. Both increasing the learning rate by one order of magnitude or decreasing it by one order of magnitude significantly impairs the predictive efficacy (as shown in fig.6). We have selectively listed four critical, representative experimental outcomes in Table 2. These results indicate that, when the model structure is simple (as shown in fig.4) , increasing the number of hidden neurons leads to a substantial decline in prediction accuracy. Moreover, incorporating embeddings into the modeling process produces significantly superior outcomes compared to models relying solely on structured data (i.e., dense features), exhibiting an improvement exceeding 3 percentage points.

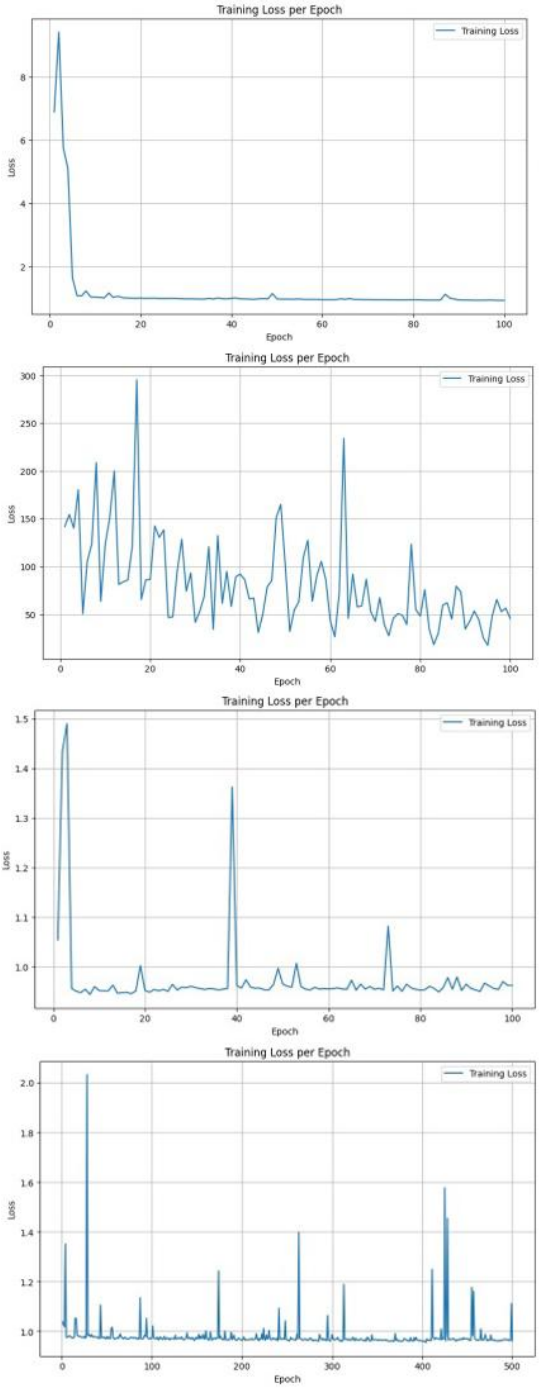

Fig. 6. Training Loss of Four Typical Experiments

TABLE II. COMPARATIVE RESULTS WITH/WITHOUT EMBEDDING

	Exp1	Exp2	Exp3	Exp4
L.r.	1.00E-03	1.00E-03	1.00E-03	1.00E-03
epoch	100	100	100	500
hidden units	64	64	128	128
feature sets	Embedding +len_sign+ dense 11	Dense11	Dense11	Embedding +len_sign+ dense 11
loss	0.9329	0.963	45.47	0.9629

IV. CONCLUSION

Through a sampling and modeling exercise involving a representative sample set of ten thousand users from the prominent Chinese social media platform bilibili.com, we conclude that username texts, typically regarded as indexing items rather than viable modeling features, along with other natural language data such as personalized signatures, should and must be incorporated into the nerual network modeling process via embedding techniques, with the help of pretrained large language models such as BERT.

Our finding carries significant implications for the field of social media analytics and user profiling.

Firstly, it challenges the prevailing notion that usernames, often treated as mere identifiers devoid of meaningful information content, are of negligible value in predictive modeling. By demonstrating that usernames and signatures can indeed contribute to enhancing model performance when properly embedded, our study underscores the importance of considering every potentially informative aspect of user-generated content, even those traditionally dismissed as peripheral.

Secondly, the successful application of embedding techniques to these unconventional sources of data highlights the versatility and adaptability of such methodologies in extracting latent semantic structures from diverse text types. This extends the scope of text-based feature engineering beyond conventional message contents, opening up new avenues for researchers and practitioners to harness unexplored or underutilized textual elements within social media platforms.

Moreover, this discovery contributes to the refinement of user profiling strategies by revealing additional layers of user characteristics that may not be evident from more overt forms of online behavior. The inclusion of embedded username and signature features in models can potentially provide deeper insights into users' personalities, interests, and social identities, enriching our understanding of user demographics and psychographics, and ultimately leading to more nuanced and accurate user segmentation.

Lastly, from a practical standpoint, our findings offer guidance to social media platforms, advertisers, and other stakeholders seeking to optimize user engagement, content personalization, or targeted marketing efforts. By integrating username and signature embeddings into their analytical frameworks, these entities can leverage previously overlooked data sources to refine their audience targeting, tailor content recommendations, and enhance overall user experience.

In summary, the recognition of usernames and signatures as valuable modeling inputs through embedding techniques not only expands the theoretical boundaries of social media analytics but also presents tangible opportunities for enhancing the precision, depth, and breadth of user profiling in practical applications.

ACKNOWLEDGMENT

This research is funded by 2023 University Philosophy and Social Science Research General Project of Jiangsu Province (No. 2023SJYB0948) and 2023 Science Association Soft Science Project of Wuxi City (No. KX-23-C006).

REFERENCES

[1] Kennedy, H., Elgesem, D., & Miguel, C. (2015). On fairness. Convergence the International Journal of Research Into New Media Technologies, 23(3), 270-288. https://doi.org/10.1177/1354856515592507

[2] Bulcock, A., Hassan, L., Giles, S., Sanders, C., Nenadić, G., Campbell, S., ⋯ & Dixon, W. (2021). Public perspectives of using social media data to improve adverse drug reaction reporting: a mixed-methods study. Drug Safety, 44(5), 553-564. https://doi.org/10.1007/s40264-021-01042-6

[3] Kamaruzaman, N. (2019). Towards a multimodal analysis to predict mental illness in twitter platform. International Journal of Advanced Trends in Computer Science and Engineering, 126-130. https://doi.org/10.30534/ijatcse/2019/1981.42019

[4] (2020). Discovering and expansion the irregular manners of users in online social networks using data mining techniques. Journal of Critical Reviews, 7(04). https://doi.org/10.31838/jcr.07.04.62

[5] Aswale, M. and Dharmadhikari, S. (2017). Survey on recommendation of personalized travel sequence. Ijarcce, 6(1), 108-113. https://doi.org/10.17148/ijarcce.2017.6122

[6] Stekelenborg, J., Ellenius, J., Maskell, S., Bergvall, T., Caster, O., Dasgupta, N., ⋯ & Pirmohamed, M. (2019). Recommendations for the use of social media in pharmacovigilance: lessons from imi web-radr. Drug Safety, 42(12), 1393-1407. https://doi.org/10.1007/s40264-019-00858-7

[7] Gundecha, P. and Liu, H. (2012). Mining social media: a brief introduction., 1-17. https://doi.org/10.1287/educ.1120.0105

[8] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media. Acm Sigkdd Explorations Newsletter, 19(1), 22-36. https://doi.org/10.1145/3137597.3137600

[9] Dreisbach, C., Koleck, T., Bourne, P., & Bakken, S. (2019). A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. International Journal of Medical Informatics, 125, 37-46. https://doi.org/10.1016/j.ijmedinf.2019.02.008

[10] Nasralah, T., El-Gayar, O., & Wang, Y. (2020). Social media text mining framework for drug abuse: development and validation study with an opioid crisis case analysis. Journal of Medical Internet Research, 22(8), e18350. https://doi.org/10.2196/18350

[11] Alvarez, T. and Chen, H. (2021). Meta-analysis of social networking sites for the purpose of preventing privacy threats in the digital age. Journal of Applied Data Sciences, 2(3), 64-73. https://doi.org/10.47738/jads.v2i3.36

[12] Wang, X., Du, S., Feng, C., Zhang, X., & Zhang, X. (2018). Interpreting the fuzzy semantics of natural-language spatial relation terms with the fuzzy random forest algorithm. Isprs International Journal of Geo-Information, 7(2), 58. https://doi.org/10.3390/ijgi7020058

[13] Su, Q., Wan, M., Liu, X., & Huang, C. (2020). Motivations, methods and metrics of misinformation detection: an nlp perspective. Natural Language Processing Research, 1(1-2), 1. https://doi.org/10.2991/nlpr.d.200522.001

[14] Peng, H. (2023). Exploring symbolic effect of new media: the impact of bilibili on gen z's cohort identity and aesthetic choices in fashion., 176-187. https://doi.org/10.1007/978-3-031-38541-4_17

[15] Wang, R. (2022). Community-building on bilibili: the social impact of danmu comments. Media and Communication, 10(2), 54-65. https://doi.org/10.17645/mac.v10i2.4996

[16] Li, F. and Li, M. (2022). A study on the changing trends of popular videos on bilibili., 842-853. https://doi.org/10.2991/978-2-494069-89-3_99

[17] Datese url: https://www.kaggle.com/datasets/beats0/bilibili-user

[18] Murakami, R. and Chakraborty, B. (2022). Investigating the efficient use of word embedding with neural-topic models for interpretable topics from short texts. Sensors, 22(3), 852. https://doi.org/10.3390/s22030852

[19] Dovdon, E. and Batsuuri, S. (2021). text2plot: sentiment analysis by creating 2d plot representations of texts. Ieej Transactions on Electrical and Electronic Engineering, 16(6), 852-860. https://doi.org/10.1002/tee.23372

[20] Reimers, N. and Gurevych, I. (2019). Sentence-bert: sentence embeddings using siamese bert-networks.. https://doi.org/10.18653/v1/d19-1410

[21] Juarto, B. and Girsang, A. (2021). Neural collaborative with sentence bert for news recommender system. Joiv International Journal on Informatics Visualization, 5(4), 448. https://doi.org/10.30630/joiv.5.4.678

[22] Apoorva, K. and Sangeetha, S. (2021). Deep neural network and model-based clustering technique for forensic electronic mail author attribution. Sn Applied Sciences, 3(3). https://doi.org/10.1007/s42452-020-04127-6

[23] Miraj, R. and Aono, M. (2021). Integrating extracted information from bert and multiple embedding methods with the deep neural network for humour detection. International Journal on Natural Language Computing, 10(02), 11-21. https://doi.org/10.5121/ijnlc.2021.10202

[24] Mou, H. and Yu, J. (2021). Cnn-lstm prediction method for blood pressure based on pulse wave. Electronics, 10(14), 1664. https://doi.org/10.3390/electronics10141664

[25] Huang, B., Ji, Z., Zhai, R., Xiao, C., Yang, F., Yang, B., … & Wang, Y. (2021). Clock bias prediction algorithm for navigation satellites based on a supervised learning long short-term memory neural network. GPS Solutions, 25(2). https://doi.org/10.1007/s10291-021-01115-0

[26] Mamidala, K. and Sanampudi, S. (2022). Text summarization on telugu e-news based on long-short term memory with rectified adam optimizer. International Journal of Computing and Digital Systems, 11(1), 355-368. https://doi.org/10.12785/ijcds/110130