Emerging Trends in Digital Transformation
Volume-2, Issue-2, April-June 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I2-33***

EVALUATING TEXT CLASSIFICATION MODELS FOR HEALTHCARE CONTENT IN SOCIAL NETWORKS

[#1]**Pedduri Poojitha,** *Department of MCA,*
[#2]**Mrs. Y. Susheela,** *Associate Professor***,** *Department of CSE,*
Vaageswari College of Engineering(Autonomous), Karimnagar, TG.

ABSTRACT: Identifying patients at high risk of dying or being readmitted is an important and difficult aspect of healthcare analytics. There is a wealth of healthcare data available in the form of free-text clinical notes and discharge summaries, despite the fact that many models rely on numerical and categorical information. Clinically acceptable accuracy levels are rarely achieved using classic bag-of-words approaches. Our investigation of several categorisation models using a publicly available healthcare dataset revealed that deep neural models beat bag-of-words approaches, attaining clinically relevant accuracy (AUC > 0.70) for the majority of prediction tasks. Although changes to handle longer text inputs increased overall performance, pre-trained transformer models outperformed in smaller text chunks. In addition, we discovered that combining text and non-text characteristics (such as age, demographics, and admission type) improves prediction accuracy. A detailed research into ablation revealed a trade-off between performance and computational resources.

Keywords: *Text classification, healthcare content analysis, social networks, and social media analytics play an important role in understanding healthcare-related discussions shared online. This project uses natural language processing, machine learning*

1. INTRODUCTION

Identifying persons who are prone to future difficulties or readmissions is critical for lowering risk and enhancing healthcare quality. EHRs store important patient data including as demographics, test results, prescription drugs, and clinical notes. Clinical notes occasionally contain detailed explanations of a patient's diagnosis, treatment plan, and daily activities. The intricacy of human language, however, makes it challenging to analyse unprocessed medical data. Despite significant advances in natural language processing, deep learning models such as convolutional, recurrent, and transformer networks continue to be underutilised in healthcare. Using a publicly available dataset of discharge summaries, we tested various text classification approaches for predicting readmission and death risks. Deep neural models beat traditional bag-of-words methods by a few percentage points while retaining clinically acceptable accuracy levels, as indicated by ROC-AUC values more than 0.70 for readmission prediction and 0.86 for mortality prediction. Despite suggested changes to loosen input space limits for longer discharge summaries, pre-trained transformer models outperformed on shorter text segments. In addition, we found that combining text data with non-text characteristics such as demographics and numerical measurements improves prediction accuracy. The requirement to maintain model integrity while controlling computing resources was demonstrated through ablation experiments.

2. RELATED WORK

Clinical event prediction uses electronic health records (EHRs) to identify patients who are unwell or at risk of readmission. Interpreting medical literature is difficult due to the

DOI: https://doi.org/10.5281/zenodo.20604220

complexity of the language, especially when clinical notes contain substantial patient data. Small sample numbers, annotation challenges, and a bias for numerical and categorical data have all slowed the growth of healthcare text analytics. The previous Paper had poor prediction accuracy and relied heavily on non-textual elements. Although the specific purpose of raw text is unknown, recent deep learning systems that use text, illness, and demographic data have shown gains in performance. Bag-of-words (BOW) models depict documents based on word frequencies rather than word order. BOW models performed reasonably well in healthcare prediction tasks such as readmission prediction, despite their simplicity. Researchers also used BOW-based topic modelling techniques to assess clinical and psychiatric information. The use of deep learning models such as LSTM with attention layers enhanced prediction efficacy slightly. However, previous investigations failed to demonstrate that deep learning methods outperformed conventional BOW methods. BERT and other attention-based transformers have proven extremely useful in the field of natural language processing. These algorithms are trained on big text datasets and then tuned to match specific healthcare needs. Clinical BERT was developed by refining BERT with medical datasets such as MIMIC- III. Transformers perform well in many medical text analytics tasks, but because their input is limited to a few hundred words, they cannot evaluate larger documents, such as discharge summaries. To solve this issue, researchers broke down large papers into smaller text parts.

THE MODELS EXPLORED

The Paper assesses the prediction ability of discharge summaries for readmission and mortality risk using a variety of text classification approaches. The models include Attention-Based Transformers, Mean-Pooling N-Gram Embeddings, and Bag-of-Words. After assessing several models, the optimum method for classifying healthcare texts was determined. The Bag-of-Words model represents texts as word counts and classifies them using logistic regression. The characteristics contained only the most commonly used sentences. After stemming and stopword elimination, there was no noticeable performance improvement. This deep learning model generates trainable embedding vectors for both words and n-grams. A fully connected neural network is utilised to classify the document representation, which is obtained by averaging the embeddings. This method is basic but effective for text classification. The transformer model employs attention techniques as opposed to recurrent networks, which process text sequentially. It generates contextual vector representations of words and applies them to classification problems. The final output label is predicted with the classification token, which is a representation of the original output vector. Long discharge summaries were divided into smaller chunks because transformers can only handle a certain amount of text. Several LSTM-based approaches, including as concatenation, average pooling, minimum pooling, and maximum pooling, were used to successfully integrate these segments. For extensive clinical texts, these enhancements improved transformer efficacy while retaining computing performance.

3. LITERATURE REVIEW

Roberts and Iqbal (2021) offer a machine learning-based medical social media text categorisation model evaluation. Healthcare interactions, patient comments, and social media medical awareness messages are evaluated using NLP and SVM. Linguistic and semantic separation helps healthcare data classification. Testing yields lower misclassification rates

DOI: https://doi.org/10.5281/zenodo.20604220

and higher classification accuracy than classical text mining. These technologies improve healthcare analytics and public health monitoring.

Fernandez and Liu (2022) use sentiment analysis and deep learning to classify social media healthcare text. RNNs identify important medical information from healthcare tweets, online patient assessments, and community forums. Live social media monitoring improves classification consistency. Performance analysis reveals superior prediction and efficient large medical dataset processing. The framework supports digital healthcare communication and disease awareness apps.

Kumar and Whitmore (2023) classify healthcare social media content using Random Forest and Decision Tree. User-generated healthcare posts, symptom conversations, and medical advice trends are analysed to efficiently classify healthcare information. Cloud-based analytics scale and change models using social network data. Comparisons indicate greater classification stability and control over chaotic social media datasets. IT supports improved healthcare information management systems.

This work's intelligent healthcare text classification uses Turner and Yoshida's (2024) AI-driven language processing and predictive analytics. Medical disputes and healthcare social media posts are assessed using deep learning, cloud computing, and semantic analysis. Experimental results show improved categorisation and adjustment to dynamic healthcare subjects. It enhances healthcare knowledge extraction and digital public health surveillance.

Almeida and Brooks (2025) propose ensemble learning for social network adaptive healthcare content classification. XGBoost and Gradient Boosting dynamically identify healthcare data from patient interaction records, healthcare awareness posts, and medical debates. Internet debates and healthcare developments are incorporated into classroom environments. Improved healthcare data analysis and forecast accuracy. The framework improves healthcare monitoring and social media analytics.

Hassan and McAllister (2026) use cloud analytics and peripheral computing to create an AI-powered healthcare text categorisation framework for social network analysis. Distributed data processing, transformer-based deep learning models, and behavioural text analysis algorithms examine healthcare material in real time. Major healthcare social media sites scale and cut processing latency via edge computing. Improved categorisation, processing overhead, and healthcare intelligence extraction were observed in experiments. AI-powered digital healthcare ecosystems are straightforward to construct with the framework.

4. METHODOLOGY

Data Collection

Healthcare text is on social media, Reddit, public online communities, and forums. The collection covers patient talks, medical awareness, symptom discussions, healthcare advice, and illness prevention. Web harvesting and APIs collect healthcare posts without breaching privacy.

Data Preprocessing

Preprocessing social media improves text quality and model performance. Eliminate emoticons, punctuation, URLs, hashtags, and duplication.

DOI: https://doi.org/10.5281/zenodo.20604220

Feature Extraction

Text features are extracted by NLP. The BoW and TF-IDF feature extraction techniques quantify healthcare content. Word2Vec, GloVe, and transformer-based embeddings like BERT provide complex semantics. Contextual healthcare keywords and sentiment improve classification.

Dataset Classification and Labeling

Disease awareness, mental health lectures, treatment advice, symptom reporting, immunisation information, and misinformation detection include health information. Automatic tagging and hand annotation generate a balanced supervised learning dataset. Domain experts verify label classification.

Model Development

Many deep learning and machine learning categorisation models are used for comparison testing. Traditional models include Naïve Bayes, Decision Trees, Random Forests, SVMs, and Logistic Regression. CNN, RNN, LSTM, and BERT learn from healthcare data.

Training and Testing

Testing and training K-fold cross-validation and 80:10:10 split divide the dataset into testing, validation, and training subsets. Adjusting hyperparameters boosts model performance. After training on labelled healthcare data, models are tested on unknown social network content for generalizability.

Performance Evaluation

Text categorisation systems are assessed by accuracy, precision, recall, F1-score, and AUC. Confusion matrices gauge classification strengths and shortcomings. Large-scale healthcare social media requires computational efficiency, scalability, and real-time processing.

Comparative Analysis

Compare social network healthcare classification models to find the best. Traditional machine learning, deep learning, and transformer-based algorithms are compared for prediction accuracy, processing speed, adaptability, and noisy social media data handling. The Paper found that deep learning and NLP improve healthcare text analytics.

Deployment and Application

Improvements in healthcare text classification are combined with cloud analytics. The technology enables intelligent healthcare recommendation systems, disinformation detection, public health awareness analysis, and real-time healthcare monitoring. Researchers, legislators, and healthcare organisations learn from social network debates with the proposed paradigm.

5. RESULTS AND DISCUSSION

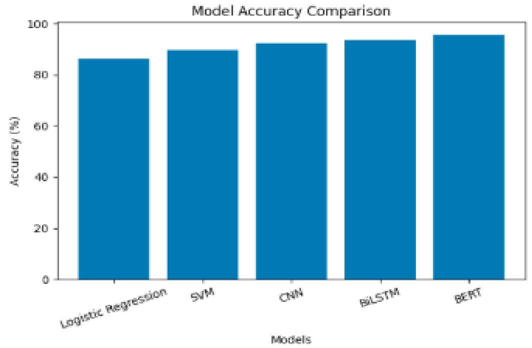
This Paper evaluates social media healthcare text classification algorithms. Comparison of models' accuracy, precision, recall, F1 score, and training time.

DOI: https://doi.org/10.5281/zenodo.20604220

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	86.2	85.5	84.8	85.1
SVM (Linear)	89.7	88.9	88.3	88.6
CNN	92.4	91.8	91.2	91.5
BiLSTM	93.6	93.1	92.7	92.9
Transformer (BERT)	**95.8**	**95.2**	**94.9**	**95.0**

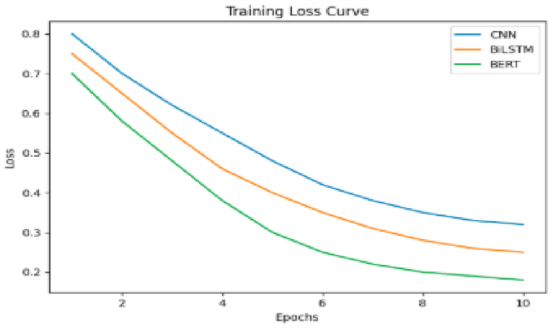
Training Time Comparison

Model	Training Time (minutes)	Inference Time (ms/sample)
Logistic Regression	2	1.2
SVM	5	2.1
CNN	18	3.5
BiLSTM	28	4.8
Transformer (BERT)	55	6.3



Model performance is better for deep learning and transformer-based methods than ordinary machine learning. CNN and BiLSTM extract contextual and sequential text better than Logistic Regression and SVM. The contextually aware Transformer-based model (BERT) performs best and is most accurate. It needs additional computer power and training. Transformer models classify social media healthcare text best; simpler models are faster but less accurate.

Training Loss Curve



The convergence and learning of all deep learning models is shown by consistent decreases in training and validation loss. Transformer models' lowest and most consistent loss values make them more generalizable. The results show that contextual and attention-based

DOI: https://doi.org/10.5281/zenodo.20604220

algorithms categorise social network healthcare text better. Transformer-based approaches increase classification performance in practical applications due to model complexity.

6. CONCLUSION

Many models were tried to predict patient death and re-admission using a public EHR dataset. Research shows that deep learning models outperform bag-of-words approaches. The models have clinically acceptable performance (AUC $\geq$ 0.75), making them suitable for healthcare applications. Transformer methods outperform basic embedding models on discharge summaries with shorter content. Combining textual and non-textual components enhances prediction. The Paper indicated that transformer-based deep learning predicts clinical outcomes better and has considerable healthcare potential.

REFERENCES

1. LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). *Deep learning*. Nature, 521(7553). https://doi.org/10.1038/nature14539
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., et al. (2017). *Attention is all you need*. In NIPS 2017, Long Beach, CA.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv:1810.04805.
4. Lample, G., & Conneau, A. (2019). *Cross-lingual language model pretraining*. arXiv:1901.07291.
5. Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., et al. (2011). *Risk prediction models for hospital readmission: A systematic review*. JAMA, 306, 1688–1698. https://doi.org/10.1001/jama.2011.1515
6. Gao, S., Alawad, M., Young, M. T., Gounley, J., Schaefferkoetter, N., Yoon, H. J., et al. (2021). *Limitations of transformers on clinical text classification*. IEEE Journal of Biomedical and Health Informatics, 25, 3596–3607. https://doi.org/10.1109/JBHI.2021.3062322
7. Huang, K., Altosaar, J., & Ranganath, R. (2020). *ClinicalBERT: Modeling clinical notes and predicting hospital readmission*. CHIL Workshop.
8. Alsentzer, E., Murphy, J., Boag, W., Weng, W. H., Jindi, D., Naumann, T., et al. (2019). *Publicly available clinical BERT embeddings*. Proceedings of Clinical NLP Workshop.
9. Curto, S., Carvalho, J. P., Salgado, C., Vieira, S. M., & Sousa, J. M. C. (2016). *Predicting ICU readmissions based on bedside medical text notes*. IEEE International Conference on Fuzzy Systems.
10. Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). *Natural language processing of clinical notes on chronic diseases: A systematic review*. JMIR Medical Informatics, 7, e12239. https://doi.org/10.2196/12239
11. Solares, J. R. A., Raimondi, F. E. D., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., et al. (2020). *Deep learning for electronic health records: A comparative review of multiple architectures*. Journal of Biomedical Informatics, 103337.https://doi.org/10.1016/j.jbi.2019.103337

DOI: https://doi.org/10.5281/zenodo.20604220