www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 15(1), Jan-April 2024, 32-38. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2024.v15i1.430

Stress Recognition in Code-Mixed Social Media Texts using Machine Learning

Tewodros Achamaleh [1], *Lemlem Eyob*[2], *Muhammad Tayyab*[2], *Grigori Sidorov*[2], *Ildar Batyrshin*[2]
[1] Department of Technology and Engineering, Rift Valley University.
[2] Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN).

E-mails: teddymas97@gmail.com

Abstract. Stress, being a complex emotional state caused by a variety of multiple sources, has the potential for serious effects if left untreated. The primary goal of this research is to select and consider AI models that effectively recognize stress within the complicated domain of social media posts. The significance of this study is not only the categorization of stress but also the interpretation of the sophisticated methods that serve as the basis for these emotional responses. Among the traditional machine learning models, Random Forest, K-Nearest Neighbor, Logistic Regression, Decision Tree, and Support Vector Machine are used. The deep learning model's LSTM, BiLSTM, and transformer-based models m-BERT, AL-BERT, XLM-RoBERTa, IndicBERT, and Distil-BERT were used. Of those models, LSTM proved to be the best-performing model, with an F1-score of 0.75.
Keywords: stress, Tamil, code-mixed, transformers, non-stressed

Article Info
Received Feb 26, 2024
Accepted April 16, 2024

1 Introduction

Understanding mental health concerns like depression, stress, fear, and sadness through social media analysis aids in early identification and intervention strategies (Borah, & Kumar, 2022). This collaborative project is an important addition to mental health awareness by using the capabilities of natural language processing and AI technologies.

Today, social media platforms have experienced a transformational transition in the ever-evolving digital sphere. It assists in the early detection of stress [2], due to the expansion and significance of social media in communications. Stress detection of social media code-mixed text has been an interesting subject of study in recent years (Tonja, et al., 2022). Measuring the degree of depression in social media texts is essential to treat them and prevent any negative effects.

Research on natural language processing (NLP) is currently using code-mixed (Tonja, et al., 2022). This research is focused on the Tamil language, which is often code-mixed with English. The term "code- mixing" describes the use of several different languages in a single document or statement (Tash, 2023). Tamil is predominantly from the Indian state of Tamil Nadu and the northern and eastern regions of Sri Lanka. It is one of the oldest living languages still in use today, as it is over 5000 years old. With its first grammar book having made its first appearance in 300 BC.

Despite its rich linguistic heritage, Tamil, an ancient Dravidian language, faces challenges in NLP due to limited resources compared to major languages such as English, Chinese, Spanish, etc.

We conduct a study that intends to contribute to the area by utilizing a variety of traditional machine learning models and deep models to improve the performance of stress identification in the Tamil–English code-mixed language, which is a low-resource language.

2 Statement of the Problem

Stress management can help a person deal with challenges in a healthier way. It is natural and normal to be stressed sometimes, but long-term stress can cause physical symptoms, emotional symptoms, and unhealthy behaviors. Stress can be managed and

prevented by using different strategies. The strategy of stress monitoring systems requires an accurate stress classification technique. To resolve the above issue, we present a study on the task of stress detection in Tamil–English code-mixed text.

Analyzing these textual traces not only provides a detailed picture of the user's mental health but also provides an important chance for insightful investigation. The importance of early stress identification cannot be overemphasized, given the potential development of chronic stress into more serious mental health problems such as depression. The goal is to be able to detect stress and non-stressed texts within the huge panorama of social media postings.

Some researchers detect various steps in stress detection and identification. A few algorithmic approaches have been proposed for automatic stress detection, but the proposed classifiers have some drawbacks. We propose twelve different models, including state-of-the-art methods like transformers, and evaluate their performance for automatic stress detection. In the proposed method, we use a Tamil–English code-mixed data collected from social media sources.

3 Related Work

Prior studies have explored various approaches to understanding mental health through texts from social media. Every single study turns out distinct insights, and cumulatively, such knowledge represents a complete picture of the developmental trajectory this discipline is heading down. Below are summarized related works for this study.

The work (Thannickal, Sanmati, Rajalakshmi & Deborah, 2023) aspires to extract potentially depressive signs among English language comments on social media. The three methods were provided, namely BERT, Word2Vec, and SVC, and TFIDF-LinearSVC, and BERT model showed the highest approval in the empirical study with a F1 macro score of 0.407. This study suggests BERT model is an optimal tool for distinguishing the voices of depression among social media texts.

In the research, the authors (Ilias, Mouzakitis & Askounis, 2023) designed the study to calibrate transformer-based models, namely BERT and MentalBERT models, for identifying depression and stress in social media postings from Reddit.

The research (Zulqarnain, et al., 2023) is focused on attention-aware deep learning approaches for an effective stress classification domain. The E-LSTM, which is an enhancement on traditional LSTM and combines pre-attention, resulted in superior performance when compared to other classification methods, including Naïve Bayesian, SVM, deep belief network, and standard LSTM. This research has been evaluated using a selected dataset accessed from the sixth Korea National Health and Nutrition Examination Survey conducted from 2013 to 2015 (KNHANES VI) to analyze health-related stress data.

The authors in the study (Siam, Gamel, & Talaat, 2023) proposed a way of using bio signals to detect stress in car drivers using physiological data recorded continuously during real-world driving tasks in Boston. In this study, six different traditional machine learning models (KNN, SVM, DT, LR, RF, and MLP) have been used. The Random Forest classifier series yields the best result of 98.2% of classifying accuracy and out- performs other techniques.

The authors in (Ding, Zhang, & Ding, 2023) provide the outcome model, which was a hybrid of boosting machines (GBM) and random forests (RF), and was perfect as it gave 100% accuracy. Using stress detection dataset from Kaggle (Rachakonda, Bapatla, Mohanty & Kougianos, 2020).

Additionally, in the work of (Wongkoblap, Vadillo & Curcin, 2017) the authors discuss current takes on mental health predictive analytics limitations and ethics. They used data mining techniques to analyze the literature to find terms related to computer science and medical journal articles.

The authors in (Islam, Kamal, Sultana, Islam, Moni, & Ulhaq, 2018) use machine learning approaches k-nearest neighbors classification technique for detecting individuals suffering from depression from Facebook user comments.

To conclude, the article reveals the objective of several researches that are designed to detect stress. This analysis is to discover the present stage of the knowledge and summarize the research findings and methodologies used in the discipline.

4 Dataset Analysis

The main point of this research lies in the Tamil-English code-mixed dataset, which is collected from social media (Kayalvizhi, Durairaj, Chakravarthi, & Mahibha, 2022). Each entry is extensively annotated for stress or non-stressed content. Table 1 shows the distribution of labels for training, evaluation, and test data.

Code-mixing is used in nearly all social media networks where individuals speak many languages. The dataset is tab-separated valued data that has been distributed into three splits, namely the training set, the evaluation set, and the test set. The size of the training set is 5,504, of which 2,621 are Tamil, 1,801 are Tamil-English code mixed, and 1,082 are English, while the size of the validation set is 1,318, and the size of the test set is 1,020. Table 1 display samples for the dataset.

Table 1. Distribution of labels for training, evaluation, and test data

	Stress	Non-stressed
Training	32.4%	67.6%
Evaluation	31.9%	68.1%
Test	36.3%	63.7%

```
Label: Non stressed
Text: Appa nice
Label: Non stressed
Text: முதுன மூஞ்சி வரவு மட்டும் சொல்றியே
Label: Non stressed
Text: இல்ல பா கஜானா தான் ரொம்ப பெருசு
Label: Non stressed
Text: என்னன்ன சொல்றான் பாருங்க.... கம்பி
Label: Non stressed
Text: முழு கதைக்கு அந்த கதைநாயகிதான் தே
```

Figure 1. Example for the Tamil-English dataset

```
stressed
Bro dog ah pathukonga bro, I know how it feels.
stressed
So sweet both of you god bless them!omsairam
stressed
Pepper beef bread omlette try panni parunga bro ...
Really Awesome try this recipe .....
Beef pidikathavanga eduthuka vena
stressed
Super dialogue. Issue on demand. Regional language
stressed
Shiva Bro Always Rocking
```

Figure 2. Example for the stress dataset

```
Non stressed
Me too
Non stressed
Intha mathiri amma iruntha depression agathu
Non stressed
Watched more than 20 times
Non stressed
Any one after his daughter death .....
Non stressed
 @nemesiscricketvlogs8357  it's mentioned in
```

Figure 3. Example for the non-stressed dataset

5 Methodology

A To find the best-performing model, we ran several experiments. The methodology employed in this paper consists of three approaches.

In the first approach, we implement various traditional machine learning algorithms provided by the "scikit-learn" library, which are Random Forest, K-Nearest Neighbors, Logistic Regression, Decision Tree, and Support Vector Machine, on a Tamil–English code-mixed training, evaluation, and test data set. The dataset has been preprocessed into lowercase, and we remove special characters and digits from the text.

For feature extraction, the term frequency-inverse document frequency (TF-IDF) vectorization technique with considerations of each word's importance in the preprocessed text is used to convert the text data into TF-IDF feature vectors, which serve as input features for training the machine learning models and elaborating on them to learn patterns and make predictions based on the text data. Finally, model definition and training were performed, and a classification report was provided for each model Precision, recall, and F1-score.

The second approach involved the use of the deep learning model's LSTM and BiLSTM. We apply the following preprocessing techniques to clean the data, stripping HTML tags from the text, translating newlines, tabs, and converting the text to lowercase, removing accented characters, expanding contractions, removing special characters, and stripping leading and trailing white spaces.

The LSTM and BiLSTM models were tailored to the dataset, provided the training, evaluation, and test datasets were tokenized and converted into a sequence of integer sets. Which can feed into models for training. The model architecture consists of the embedding layer, which creates word embedding for the input sequence. The bidirectional LSTM layer processes the input sequence in both forward and backward directions, and two dense layers for classification finally make predictions on test data using a trained model and then evaluate the accuracy of the predictions. Based on the weighted average, the BiLSTM model achieves F1-score of 0.72, while the LSTM model outperforms it with F1-score of 0.75.

Table 2 shows the best-performing hyperparameter settings among the different parameters we experimented with for LSTM and BiLSTM.

Table 2. Parameter setting

Parameters	Values
vocab_size	3000
embedding_dim	100
max_length	200
padding_type	'post'
trunc_type	'post'
num_epochs	10

In the third approach, we implement various transformer-based algorithms. We use the ktrain library to build and train a text classification model based on four transformer-based models, namely m-BERT, AL-BERT, XLM-RoBERTa, IndicBERT, and Distil-BERT for this classification task.

The model passed with preprocessed training and evaluation data. We specify a maximum sequence length of 400 tokens (max Len = 400), and the batch size for training is specified as 'batch-size = 12'. After this setup, we proceed with training these models with a method that helps in finding an optimal learning rate, and the number of epochs for training is 3. After training, we evaluate the model's performance on the validation set to see how well it generalizes to the unseen data. Then predictions were performed in a test set. The general methodology pipeline for transformer-based models is shown in Figure 4

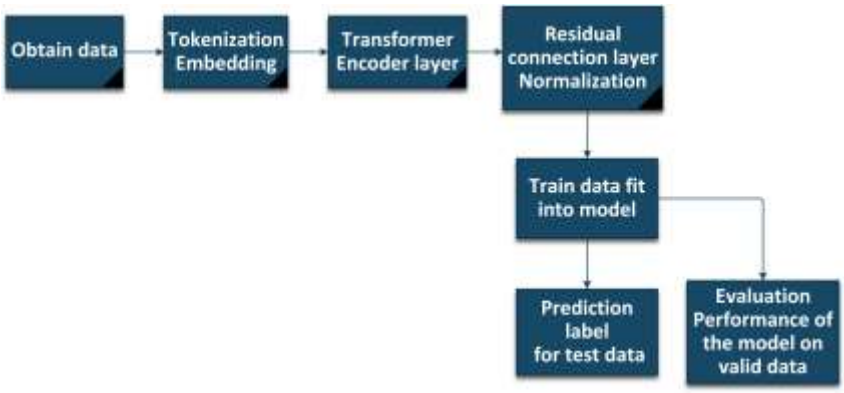

Figure 4: General methodology pipeline for transformer-based models

6 Results and Discussion

We evaluated the proposed stress detection method using twelve different traditional machine learning and deep learning models. Out of the above-mentioned models, the deep learning model LSTM gave the highest weighted average F1-score of 75% for the predicted labels on the test data and an accuracy of 74%. It is inferred from the results that we get the traditional machine learning models could not perform well when compared to the deep learning models.

While the models we use achieved promising results, we acknowledge the limits inherent in the above stated transformer-based models. We assume that the performance of the BERT model can be increased by training the model for a larger number of epochs, adding additional variables like sentiment analysis, and leveraging larger and more diverse datasets.

Precision, recall, accuracy rate, and F1-score have been considered to measure the performance of algorithms used. We show the evaluation of model performance in the Table 3.

Table 3: Evaluation metrics table

Model	Weighted Avg.			Micro Avg.			Accuracy
	P	R	F1	P	R	F1	
LSTM	0.80	0.75	0.75	0.75	0.77	0.74	0.74
BiLSTM	0.84	0.72	0.72	0.78	0.78	0.72	0.72
XLM-RoBERTa	0.83	0.70	0.70	0.77	0.76	0.70	0.70
Distil-BERT	0.83	0.70	0.70	0.77	0.76	0.70	0.70
BERT	0.83	0.70	0.69	0.77	0.76	0.69	0.70
AL-BERT	0.83	0.70	0.69	0.77	0.76	0.70	0.70
IndicBERT	0.84	0.70	0.70	0.77	0.77	0.70	0.70
RF	0.82	0.72	0.73	0.77	0.77	0.72	0.72
LR	0.81	0.73	0.73	0.75	0.77	0.73	0.73
Decision tree	0.81	0.71	0.71	0.75	0.76	0.71	0.71
SVM	0.81	0.71	0.72	0.76	0.76	0.71	0.71
KNN	0.69	0.70	0.67	0.68	0.62	0.62	0.70

7 Conclusion

In this article, we present a variety of technologies from traditional machine learning algorithms to deep learning models for identifying stress in social media text including transformer-based models.

LSTM has outperformed all models used in testing, as shown by a weighted average F1-score of 75% and an accuracy of 74% for the test data. This implies the algorithm could be particularly acceptable for the task at hand and consequently, bring more accurate prediction in the real-life case. Therefore, focusing on fine-tuning and optimizing the LSTM model could lead to improved outcomes in stress recognition tasks.

Overall, the proposed approach has represented an exciting development in stress detection. So, this nuanced technique for stress recognition is a benchmark in AI models' learning environment, leading future efforts in mental health awareness and support.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Borah, T., & Kumar, S. G. (2022). Application of NLP and machine learning for mental health improvement. In *Proceedings of the International Conference on Innovative Computing and Communications: ICICC 2022, Volume 3* (pp. 219-228). Springer.

Ilias, L., Mouzakitis, S., & Askounis, D. (2023). Calibration of transformer-based models for identifying stress and depression in social media. *IEEE Transactions on Computational Social Systems*, 1-12. https://doi.org/10.1109/TCSS.2023.3283009

Tonja, A. L., Yigezu, M. G., Kolesnikova, O., Tash, M. S., Sidorov, G., & Gelbukh, A. (2022). Transformer-based model for word level language identification in code-mixed Kannada-English texts. *arXiv preprint arXiv:2211.14459*.

Tash, M., Armenta-Segura, J., Ahani, Z., Kolesnikova, O., Sidorov, G., & Gelbukh, A. (2023). Lidoma@Dravidianlangtech: Convolutional neural networks for studying correlation between lexical features and sentiment polarity in Tamil and Tulu languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages* (pp. 180-185).

Thannickal, K. E., Sanmati, P., Rajalakshmi, S., & Deborah, S. A. (2023). TechSSN1 at LT-EDI-2023: Depression detection and classification using Bert model for social media texts. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 149-154).

Zulqarnain, M., Shah, H., Ghazali, R., Alqahtani, O., Sheikh, R., & Asadullah, M. (2023). Attention aware deep learning approaches for an efficient stress classification model. *Brain Sciences*, 13(7), 994.

Siam, A. I., Gamel, S. A., & Talaat, F. M. (2023). Automatic stress detection in car drivers based on non-invasive physiological signals using machine learning techniques. *Neural Computing and Applications*, 35(17), 12891-12904.

Ding, C., Zhang, Y., & Ding, T. (2023). A systematic hybrid machine learning approach for stress prediction. *PeerJ Computer Science*, 9, e1154.

Rachakonda, L., Bapatla, A. K., Mohanty, S. P., & Kougianos, E. (2020). Sayopillow: Blockchain-integrated privacy-assured iomt framework for stress management considering sleeping habits. *IEEE Transactions on Consumer Electronics*, 67(1), 20-29.

Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2017). Researching mental health disorders in the era of social media: systematic review. *Journal of Medical Internet Research*, 19(6), e228.

Islam, M. R., Kamal, A. R. M., Sultana, N., Islam, R., Moni, M. A., & Ulhaq, A. (2018). Detecting depression using k-nearest neighbors (knn) classification technique. In *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)* (pp. 1-4). IEEE.

Kayalvizhi, S., Durairaj, T., Chakravarthi, B. R., & Mahibha, J. C. (2022). Findings of the shared task on detecting signs of depression from social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 331-338).