CUET_DUO@StressIdent_LT-EDI@EACL2024: Stress Identification Using Tamil-Telugu BERT

Abu Bakkar Siddique Raihan, Tanzim Rahman, Md. Tanvir Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1804004, u1804015, u1804002, u1704039, u1704057}@student.cuet.ac.bd
{avishek, moshiul_240}@cuet.ac.bd

Abstract

The pervasive impact of stress on individuals necessitates proactive identification and intervention measures, especially in social media interaction. This research paper addresses the imperative need for proactive identification and intervention concerning the widespread influence of stress on individuals. This study focuses on the shared task, "Stress Identification in Dravidian Languages," specifically emphasizing Tamil and Telugu code-mixed languages. The primary objective of the task is to classify social media messages into two categories: stressed and non stressed. We employed various methodologies, from traditional machine-learning techniques to state-of-the-art transformer-based models. Notably, the Tamil-BERT and Telugu-BERT models exhibited exceptional performance, achieving a noteworthy macro F1-score of **0.71** and **0.72**, respectively, and securing the 15^{th} position in Tamil code-mixed language and the 9^{th} position in the Telugu code-mixed language. These findings underscore the effectiveness of these models in recognizing stress signals within social media content composed in Tamil and Telugu.

1 Introduction

Along with the hectic pace of contemporary life, stress has become an unavoidable force impacting the mental well-being of humans. It is a complicated emotional state produced by multiple events that might inspire displeasure, rage, or worry. Recognizing and resolving stress in its early stages is crucial since persistent stress may lead to devastating diseases, including depression (Masood et al., 2012). Recent surveys indicate that 48% of Gen Z individuals experience depression symptoms, often triggered by the pervasive impact of social media. Issues like the fear of missing out heightened concerns about judgment, and increased insecurity further contribute to stress levels (Milyavskaya et al., 2018). This highlights the need for efficient stress detection and support methods within online platforms. Global stress statistics emphasize the importance of proper stress management, impacting various aspects of people's lives, from businesses and educational institutions to family contexts (Mahmud et al., 2021). Automatic stress detection provides an effective solution to address this global health crisis, offering help and resources to individuals dealing with stress-related challenges.

This research addresses the problem of stress identification in Tamil and Telugu code-mixed languages. This proposed study consists of the following key contributions:

- Investigate various machine learning (ML), deep learning, and transformer-based models for stress identification from code-mixed Tamil and Telugu texts.
- Fine-tuned Tamil-BERT and Telugu-BERT models on respective datasets to enhance stress identification performance from code-mixed data.

2 Related Work

While various studies have studied stress detection in English and other high-resource languages, attention to low-resource languages like Tamil and Telugu has been sparse (Hegde et al., 2022). Chauhan et al. (2017) conducted a study using electrocardiogram data to analyze mental stress. They employed discrete wavelet transform for preprocessing and feature extraction techniques. Nijhawan et al. (2022) used the application of Unsupervised Topic Modeling using Latent Dirichlet Allocation has facilitated the identification of emotions in online user data. This approach has proven effective in analyzing stress or depression, which achieved a high detection rate. Another study (Jadhav et al., 2019) focused on social media stress detection using textual data, highlighting the effectiveness of combining BiLSTM with an attention

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 265–270

mechanism. Dreaddit, a corpus of 190K Reddit posts with 3.5K labeled for stress identification, was introduced by (Elsbeth, 2019). Few studies (Li and Liu (2020), Oryngozha et al. (2023)) demonstrated high accuracy rates in stress identification through the application of conventional and neural supervised learning techniques on the Dreaddit dataset. Ahuja and Banga (2019) focused on exam pressure and recruitment stress frequently ignored factors and aimed to determine the extent of stress experienced by college students. The researchers utilized four classification algorithms (LR, NB, RF, and SVM) with a dataset comprising 206 student records from the Jaypee Institute of Information Technology. Their study yielded the highest accuracy for SVM. In another study conducted by (Lin et al., 2017), the relationship between users' stress states and their friends on social media was investigated using a large-scale real-world social platform dataset.

Researchers enhanced transformer-based models, including BERT and MentalBERT, by incorporating extra-linguistic data for depression and stress detection in social media (Ilias et al., 2023). Their approach involved a multimodal adaptation gate for combined embeddings, inputting data into a BERT (or MentalBERT) model, and model calibration through label smoothing (Aspillaga et al., 2020). The study highlighted the robustness of transformer-based models like RoBERTa, XLNet, and BERT in stress tests but also identified fragility and unexpected behaviors, suggesting potential directions for further advancements in the field.

3 Task & Dataset Descriptions

The task organizers curated a standardized dataset for identifying stress-related statements in Tamil and Telugu code-mixed social media texts. This effort aims to develop a system that proficiently recognizes stress expressions within a given social media text. The dataset is derived from the organizers' corpus (S et al., 2022), categorized into *Stressed (St)* and *Non Stressed (NSt)*. Table 1 displays the dataset distribution summary for Stress Identification Dataset in Tamil, including details on the train, test, and validation datasets, along with the total word count for each class. The same information is presented in Table 2 for Stress Identification Dataset in Telugu.

Class	Train	Validation	Test	W_T
St	1784	439	370	238434
NSt	3720	939	650	30876
Total	**5504**	**1378**	**1020**	269310

Table 1: Summary of SID in Tamil where W_T denotes total words

Class	Train	Validation	Test	W_T
St	1783	440	400	267320
NSt	3314	799	650	26663
Total	**5097**	**1239**	**1050**	293983

Table 2: Summary of SID in Telugu where W_T denotes total words

4 Methodology

The suggested methodology encompasses assessing diverse feature extraction techniques, integrating ML and DNN, and exploring various transformer-based architectures. The comprehensive approach aims to explore the effectiveness of different strategies in addressing the challenge of stress identification in the specified linguistic context. Figure 1 illustrates an overall outline of the stress identification technique in Tamil and Telugu code-mixed texts.

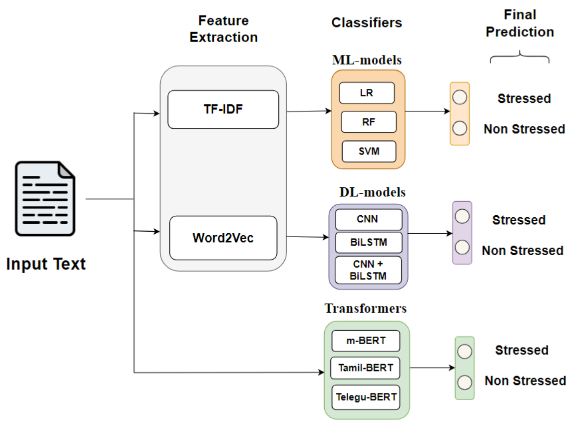

Figure 1: Schematic process of Stress Identification

4.1 Textual Feature Extraction

This study adopted several feature extraction methods to facilitate the training of classifier models for stress identification. We have employed TF-IDF (Sundaram et al., 2021) for ML models and Word2Vec embeddings (Rashid et al., 2020) for

DL models. The Keras embedding layer is vital in generating 100-dimensional embedding vectors, which encode the semantic meaning of words in the document.

4.2 ML Approaches

Various ML-based approaches (including LR, DT, and NB) are explored in developing a robust stress recognition system. Meticulous parameterization was applied to optimize each algorithm's efficiency. For instance, logistic regression underwent fine-tuning with a regularization parameter of 0.01, and the decision tree was configured with a maximum depth of 10. Naive Bayes incorporated an RBF kernel with a gamma value of 0.001, enhancing algorithm effectiveness in stress pattern recognition.

4.3 DL Approaches

A hybrid CNN and LSTM architecture (Wu et al., 2020) is employed, featuring seven layers. The model starts with a 200-length sequence vector input into the embedding layer, followed by two convolution layers with 'relu' activation and downsampling via a max-pooling layer. The Bidirectional LSTM layer, with 128 units, addresses complex patterns, and a dropout rate of 0.5 mitigates overfitting. The final layer uses a sigmoid activation function for binary classification. Pre-trained word vectors are explored, and training spans 20 epochs with a batch size of 64, achieving a balance between performance and computational efficiency in stress identification.

4.4 Transformer-based Approaches

This research exploited three pre-trained transformer models, namely M-BERT (Devlin et al., 2018), Tamil-BERT (Joshi, 2022), and Telugu-BERT (Joshi, 2022). These models, sourced from the Hugging Face[1] transformers library, underwent fine-tuning using the Ktrain (Maiya, 2022) package. Pre-trained versions of the transformer-based models are used with a maximum sequence length of 100 and a batch size of 16. The training spanned three epochs with a learning rate of $1e^{-4}$, enhancing their effectiveness for the specific task of stress identification.

5 Results and Analysis

Table 3 demonstrates the performance of the employed techniques for stress identification on the test set for Tamil code-mixed language and Table 4 for Telugu code-mixed language. The macro F1-score (F) was employed as a significant metric to determine model dominance, while we also evaluated the models on accuracy (A), precision (P), and recall (R) scores.

Method	Classifier	P	R	F	A
ML	LR	0.72	0.57	0.64	0.76
	DT	0.58	0.94	0.71	0.73
	NB	0.52	0.99	0.68	0.67
DL	CNN	0.61	0.82	0.68	0.68
	BiLSTM	0.59	0.88	0.65	0.67
	CNN+BiLSTM	0.54	0.99	0.70	0.72
Transformers	m-BERT	0.77	0.75	0.68	0.68
	Tamil-BERT	0.78	0.77	**0.71**	0.71

Table 3: Performance for stress identification for Tamil code-mixed language

Method	Classifier	P	R	F	A
ML	LR	0.66	0.17	0.27	0.65
	DT	0.58	0.91	0.70	0.72
	NB	0.56	0.97	0.70	0.70
DL	CNN	0.52	0.90	0.69	0.70
	BiLSTM	0.60	0.92	0.71	0.71
	CNN+BiLSTM	0.58	0.96	0.71	0.72
Transformers	m-BERT	0.72	0.73	0.70	0.70
	Telugu-BERT	0.78	0.76	**0.72**	0.72

Table 4: Performance for stress identification in Telugu code-mixed language

The LR displays competitive performance across ML models, reaching an accuracy of 0.72, a balanced recall of 0.57, and a macro F1-score of 0.64 for the Tamil dataset. DT excels in recall (0.94), resulting in a higher macro F1-score (0.71), whereas NB displays high recall (0.99) but poorer accuracy, generating a macro F1-score of 0.68. The DL model gets a competitive macro F1-score of 0.70. Among Transformers, m-BERT and Tamil-BERT demonstrate comparable performance, with macro F1-scores of 0.68 and 0.71, respectively.

For Telugu code-mixed language, LR obtains a moderate accuracy of 0.66, paired with a reduced recall, resulting in a macro F1-score of 0.27. Decision Tree stands out with solid recall (0.91) and a large macro F1-score of 0.70. Naive Bayes displays excellent recall (0.97) but poorer accuracy, pro-

[1] https://huggingface.co/

viding a macro F1-score of 0.70. CNN+BiLSTM delivers balanced accuracy, recall, and the greatest macro F1-score (0.71). Transformer models, m-BERT and Telugu-BERT, demonstrate decent performance, with Telugu-BERT (0.72) marginally beating m-BERT (0.70) in the macro F1-score.

The comparison analysis underlines the different performance of models in stress identification tasks for Tamil and Telugu code-mixed languages. LR and DT demonstrate different strengths in Tamil, whereas in Telugu, DT and CNN+BiLSTM do exceptionally well. The transformer models exhibit competitive performance but with variances in efficacy throughout the two languages.

6 Error Analysis

The stress detection performance of the BERT model in both Tamil and Telugu code-mixed languages demonstrates excellent accuracy in recognizing stressed situations, with a large true positive count in both datasets. However, a substantial difficulty develops in the form of false positives, indicating examples incorrectly categorized as stressed, particularly within an environment of class imbalance when non-stressed instances outweigh stressed ones. Figures 2 and 3 illustrate the performance of Tamil-BERT and Telugu-BERT models concerning the confusion matrix for the stress identification task.

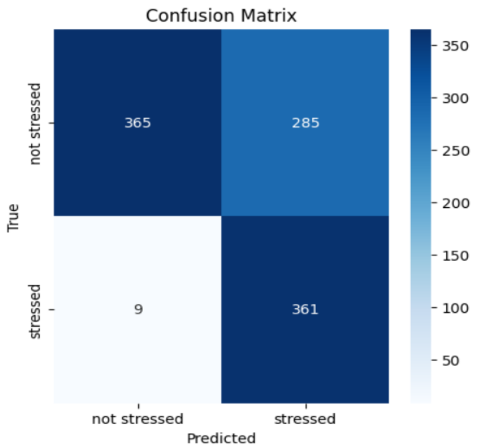

Figure 2: Confusion matrix of stress identification in Tamil using Tamil-BERT

Figure 2 illustrates the Tamil-BERT model's robust performance, accurately classifying 365 not stressed and 361 stressed samples out of 650 and 400, respectively. Despite this, precision is limited, with 285 not stressed samples misclassified as stressed and 9 stressed samples misclassified as not stressed, indicating susceptibility to false positives, particularly in identifying not stressed samples. In Figure 3, the Telugu-BERT model demonstrates strong performance, correctly tagging 371 not stressed and 378 stressed samples out of 650 and 400, respectively. However, precision is limited, with 279 not stressed samples misclassified as stressed and 22 stressed samples misclassified as not stressed. This highlights a vulnerability to false positives, especially in identifying not stressed samples.

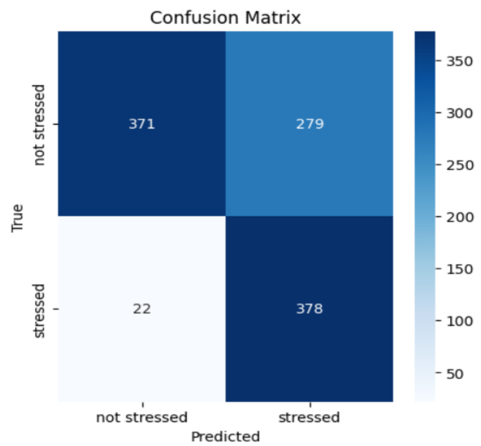

Figure 3: Confusion matrix of stress identification in Telugu using Telugu-BERT

Text Sample	Actual	Predicted
Sample1. நா யார் என்பதற்காக நா நேசிக்கப்பட விரும்புகிறேன், nA ènனவAgha இருப்பேன் என்று நா உறுதியளிக்கவில்லை: யென்ன nA adhai ஒருபோதும் கடைப்பிடிக்க மாட்டேன், நீ என்னை vèRughghiRIrghaL.	stressed	Non stressed
Sample2. One of my favourite song ஒரு பெண் தான் கொண்ட காதலை மிகவும் அழகாகவும், அமைதியாகவும் வெளிப்படுத்தும் விதமாகவும் அமைந்துள்ளது.	Non stressed	Non stressed
Sample3. அம்மாவின் அழகு இவர்களிடம்.	Non stressed	Non stressed

Figure 4: Few examples of predicted outputs by the best performing model (Tamil-BERT)

Text Sample	Actual	Predicted
Sample1. మరణం గురించి ఆందోళన : నేను తెల్లవారుజామున 3:30 గంటలకు ఆందోళన చెందడం ఇష్టపడతాను ఎందుకంటే నేను మరణం గురించి ఆలోచిస్తాను మరియు ఏమీ పట్టింపు లేదు మరియు ప్రతిదీ ఒక రోజు నల్లగా మారుతుంది. .	stressed	stressed
Sample2. Kajal enti ni voice agarbathila undiiiiKajal enti ni voice agarbathila undiiii	Non stressed	Non stressed
Sample3. Raktha Sambandam ఎంత కరుణా మూర్తివయ్యా ఎంత చల్లని తండ్రివయ్యా	Non stressed	Non stressed

Figure 5: Few examples of predicted outputs by the best performing model (Telugu-BERT)

Figures 4 and 5 illustrates some correct and incorrect predicted outcomes by the best-performed models (**Tamil-BERT and Telugu-BERT**).

Limitations

Several challenges were encountered in the stress identification task, primarily from using code-mixed language and an imbalanced dataset. The major limitations of the developed models are as follows:

- Incorporating multiple languages in code-mixed text introduces linguistic variations, making it intricate for models to discern stress-related patterns precisely.
- The dataset exhibits an imbalance, with a prevalence of non-stressed instances compared to stressed ones, potentially affecting the model's generalization capabilities. These factors collectively contribute to the task's intricacy, necessitating strategic approaches for enhanced model adaptability and accurate stress identification.

7 Conclusion

This work presented a comprehensive study of stress detection within the code-mixed languages of Tamil and Telugu by exploiting various ML, DL, and transformer-based models. Remarkably, the transformer model Tamil-BERT emerges as a remarkable performer, achieving the most significant macro F1 score of 0.71 in the context of Tamil. Meanwhile, in the domain of Telugu, the leading model is Telugu-BERT, exhibiting a substantial macro F1 score of 0.72. Future endeavors may involve the integration of culturally sensitive features, thereby enhancing the effectiveness of stress detection in social media interactions within specific linguistic contexts.

References

Ravinder Ahuja and Alisha Banga. 2019. Mental stress detection in university students using machine learning algorithms. *Procedia Computer Science*, 152:349–353.

Carlos Aspillaga, Andrés Carvallo, and Vladimir Araujo. 2020. Stress test evaluation of transformer-based models in natural language understanding tasks. *arXiv preprint arXiv:2002.06261*.

Monika Chauhan, Shivani V Vora, and Dipak Dabhi. 2017. Effective stress detection using physiological parameters. In *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pages 1–6. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Turcan Elsbeth. 2019. Dreaddit: A reddit dataset for stress analysis in social media. In *Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis*.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40.

Loukas Ilias, Spiros Mouzakitis, and Dimitris Askounis. 2023. Calibration of transformer-based models for identifying stress and depression in social media. *IEEE Transactions on Computational Social Systems*.

Sachin Jadhav, Apoorva Machale, Pooja Mharnur, Pratik Munot, and Shruti Math. 2019. Text based stress detection techniques analysis using social media. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–5. IEEE.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Russell Li and Zhandong Liu. 2020. Stress detection using deep neural networks. *BMC Medical Informatics and Decision Making*, 20:1–10.

Huijie Lin, Jia Jia, Jiezhong Qiu, Yongfeng Zhang, Guangyao Shen, Lexing Xie, Jie Tang, Ling Feng, and Tat-Seng Chua. 2017. Detecting stress based on social interactions in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 29(9):1820–1833.

Sultan Mahmud, Sorif Hossain, Abdul Muyeed, Md Mynul Islam, and Md Mohsin. 2021. The global prevalence of depression, anxiety, stress, and, insomnia and its changes among health professionals during covid-19 pandemic: A rapid systematic review and meta-analysis. *Heliyon*, 7(7).

Arun S Maiya. 2022. ktrain: A low-code library for augmented machine learning. *The Journal of Machine Learning Research*, 23(1):7070–7075.

Khalid Masood, Beena Ahmed, Jongyong Choi, and Ricardo Gutierrez-Osuna. 2012. Consistency and validity of self-reporting scores in stress measurement surveys. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4895–4898. IEEE.

Marina Milyavskaya, Mark Saffran, Nora Hope, and Richard Koestner. 2018. Fear of missing out: prevalence, dynamics, and consequences of experiencing fomo. *Motivation and emotion*, 42(5):725–737.

Tanya Nijhawan, Girija Attigeri, and T Ananthakrishna. 2022. Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data*, 9(1):1–24.

Nazzere Oryngozha, Pakizar Shamoi, and Ayan Igali. 2023. Detection and analysis of stress-related posts in reddit acamedic communities. *arXiv preprint arXiv:2312.01050*.

Umar Rashid, Muhammad Waseem Iqbal, Muhammad Akmal Skiandar, Muhammad Qasim Raiz, Muhammad Raza Naqvi, and Syed Khuram Shahzad. 2020. Emotion detection of contextual text using deep learning. In *2020 4th International symposium on multidisciplinary studies and innovative technologies (ISMSIT)*, pages 1–5. IEEE.

Kayalvizhi S, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha 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*, pages 331–338, Dublin, Ireland. Association for Computational Linguistics.

Varun Sundaram, Saad Ahmed, Shaik Abdul Muqtadeer, and R Ravinder Reddy. 2021. Emotion analysis in text using tf-idf. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 292–297. IEEE.

Jheng-Long Wu, Yuanye He, Liang-Chih Yu, and K Robert Lai. 2020. Identifying emotion labels from psychiatric social texts using a bi-directional lstm-cnn model. *IEEE Access*, 8:66638–66646.