Pinealai_StressIdent_LT-EDI@EACL2024: Minimal configurations for Stress Identification in Tamil and Telugu

Anvi Alex Eponon[1]
Ildar Batyrshin[1]
Grigori Sidorov[1]
[1]Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC),
Mexico City, Mexico
{epononanvialex@gmail.com, Ibatyr1@cic.ipn.mx, sidorov@cic.ipn.mx}

Abstract

This paper introduces an approach to stress identification in Tamil and Telugu, leveraging traditional machine learning models—Fasttext for Tamil and Naive Bayes for Telugu—yielding commendable results. The study highlights the scarcity of annotated data and recognizes limitations in phonetic features relevant to these languages, impacting precise information extraction. Our models achieved a macro F1 score of 0.77 for Tamil and 0.72 for Telugu with Fasttext and Naive Bayes, respectively. While the Telugu model secured the second rank in shared tasks, ongoing research is crucial to unlocking the full potential of stress identification in these languages, necessitating the exploration of additional features and advanced techniques specified in the discussions and limitations section.

1 Introduction

In Natural Language Processing (NLP), a diverse range of tasks is undertaken to comprehend and process human language. These tasks encompass the intricate understanding of emotional nuances, from identifying hate speech Yigezu et al. (2023); Shahiki-Tash et al. (2023a) to recognizing hope speech Shahiki-Tash et al. (2023b) and stress detection. This broad spectrum of tasks includes text classification, named entity recognition, machine translation, text generation, question answering, text summarization, and part-of-speech tagging. The evolution of NLP models has transitioned from traditional methods such as rule-based systems and statistical models to sophisticated deep learning architectures. Notable examples of these architectures include Recurrent Neural Networks (RNNs), and Long Short-Term Memory Networks (LSTMs)Tonja et al. (2022).

Stress, simply characterized as a negative emotional response stemming from external factors, notably societal pressures, represents a significant facet of the human experience. Zaydman (2017) underscores the increasing trend of individuals sharing their feelings on the internet. While substantial research has explored the identification of such emotions Tash et al. (2023), particularly in languages like English ("Joshi et al. (2005); Nagle and Sharma ("2018"), there exists a noticeable gap in studies focused on Dravidian languages such as Tamil and Telugu.

Our investigation delves into emotion detection, building upon existing studies and laying a foundational framework for future directions in this domain. The insights gleaned from our research contribute to a growing body of knowledge, providing valuable groundwork for forthcoming explorations and advancements in understanding emotions in linguistic contexts.

In the upcoming sections, we will explore the latest studies about emotion and stress identification in Dravidian languages. We will share our approach, configuration, and methodology, followed by a presentation of the results and a concise analysis with future orientations.

2 Literature Review

India, renowned for its rich history and culture, is witnessing a growing prevalence of stress across its diverse population. From the young to the adults, stress permeates various aspects of life, encompassing academic pressures and professional challenges.

In education, the pressure on students to achieve high standards has led to prolonged stress, impacting mental health and, at times, resulting in severe consequences, as noted by "Joshi et al. (2005) in their research study. Research by Nagle and Sharma ("2018") showcases the role of societal expectations and family pressures in exacerbating student stress.

The advent of social media platforms has provided an avenue for individuals to express

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 152–156
March 21, 2024 ©2024 Association for Computational Linguistics

their emotions and state of mind. Zaydman (2017) fetched 2.3 million mental health-relevant Tweets, shedding light on stress and suicide tendencies(corpora in English).

In response to this societal landscape, research on stress identification in textual corpora has flourished.Nijhawan et al. (2022) explored stress detection in social interactions, achieving high accuracy using models like Bert and Random Forest. Similarly, Inamdar et al. (2023) employed Elmo embeddings, Bag of Words, and Bert models to detect mental stress in Reddit Posts, achieving an F1-score of 0.76 (in English).

While research on stress identification is prolific in languages like English, limited attention has been given to Dravidian languages such as Tamil and Telugu. Noteworthy studies by S et al. (2022) on analyzing emotions in Tamil and Gokhale et al. (2022) using diverse deep learning models shed light on this underexplored area. However, They presented a lexicon-based approach that led to an F1-score grounding at 0.0300. In the context of a shared task, García-Díaz et al. (2022) secured first place with a neural network trained on linguistic features and various sentence embeddings achieving only an F1-score of 0.15. However, no stress emotions have been included in the research.

Through our experiment, we aspire to provide the community with a dependable method for predicting stress in Tamil and Telugu. We aim to establish a foundational benchmark for subsequent research endeavors in the field.

3 Task Description

The primary goal of this shared task is to develop a system capable of discerning between textual corpora that exhibit signs of stress and those that do not in Dravidian languages, specifically Tamil and Telugu. Stress, a multifaceted emotion stemming from various life factors, often prompts individuals to share their feelings online. Constructing such a system holds the promise of aiding and supporting individuals displaying stress characteristics on social media, ultimately contributing to the reduction of depression rates within a broad community.

Despite its noble objectives, this shared task poses unique challenges. Addressing the abstract concept of emotions, particularly stress, is inherently complex. The task's focus on Tamil and Telugu languages introduces additional difficulties, such as the limited resources available for processing and the lack of standardization observed in these languages. Decrypting the nuances of emotional variations in these languages amplifies the complexity of the problem. However, the absence of tonal characteristics shared by languages like Mandarin Chinese renders the task more approachable.

This undertaking not only underscores the significance of emotional analysis but also holds the potential to make a meaningful impact on mental health outcomes, emphasizing the importance of computational linguistics in addressing real-world challenges.

4 Approach

To address the challenge at hand, we systematically evaluated various options to arrive at a logical and explainable solution. Our decision-making process involved a thorough analysis of the syntactical structures in both languages, leveraging the support of the IndicNlp project from Kunchukuttan (2020). Upon examining the IndicNLP library Tokenization process and our dataset, we concluded that our corpora required no further preprocessing to align with our objectives.

During the feature extraction stage, we segmented the corpus into sentences and employed Term-Frequency Inverse Document Frequency (TF-IDF) as the sole feature type, alongside consideration for Bigrams. These choices were made with a focus on their compatibility with the dataset characteristics and the task requirements.

4.1 Model Selection

The model selection process was conducted transparently, guided by our current knowledge of machine learning model performances. To align with the datasets provided by the Organizers and based on our experimentation philosophy, we excluded deep learning models as the size of the dataset fits best for traditional and shallow machine learning models. So, we focused on traditional machine learning (ML) Tash et al. (2022) models and, at best, shallow models specifically Fasttext developed by Meta Joulin et al. (2016, 2017) because of its ability to not easily overfit over the training phase.

For the Tamil language, we experimented with five ML models: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Naive Bayes, and Fasttext. The trained Fasttext model demonstrated stability and robustness, leading to its selection for submission.

In the case of Telugu, similar experiments were conducted with the same models. However, we opted to submit the trained Naive Bayes model for Telugu, guided by its performance also to diversify the set of our models tackling the task of identifying stress in Dravidian languages.

This approach, grounded in a thoughtful and systematic evaluation, positions our system for effective stress identification in both Tamil and Telugu languages.

5 Experimental Setup

To conduct our experiments, we relied primarily on the Scikit-learn Pedregosa et al. (2011), and Fasttext packages due to their versatility and effectiveness in implementing various machine learning models. Scikit-learn provided a robust set of tools for traditional machine learning models, while Fasttext, with its efficient text classification capabilities, complemented our exploration.

Concerning the hardware, our experiments were executed on a computer running Ubuntu 22.04, equipped with 64 GiB of Random Access Memory (RAM), and powered by an AMD Ryzen 7000 Series 7 processor running at 3.3 GHz. This configuration was chosen for its capability to handle the computational demands of traditional machine learning models, even with a limited amount of data.

The selected hardware configuration proved adequate for our experimental goals, ensuring efficient execution and allowing us to gain meaningful insights into stress identification in Dravidian languages.

5.1 Hyperparameters for Tamil language Training

To train the selected model in Tamil language(Fasttext), we randomized the datasets, ran successively several training with different parameters, and finally applied the following hyperparameters which gave the best results:

Lr	Epochs	N-grams	Bucket	Em-Dims	Loss
0.5	10	2	200	30	ova

Table 1: Fasttext Hyperparameters Configurations

Finally, for the testing phase, we set the **threshold at 0.5**.

5.2 Hyperparameters for Telugu language Training

We split the dataset with a random state at 42, and applied the following hyperparameter configurations:

Loss	Penalty	Max-iter
hinge	L2	5

Table 2: Naive Bayes Hyperparameters Configurations

6 Results

We present the performance metrics of our stress identification model for Tamil and Telugu languages. The evaluation metrics include Accuracy, Macro F1-score, Macro-Recall, and Weighted Precision.

6.1 Tamil Language

For the Tamil language, our model achieved the following results:

Metrics	Score
Accuracy	0.724
Macro F1-score	0.723
Macro-Recall	0.775
Weighted Precision	0.822

The bar chart in Figure 1 provides a visual representation of the metrics. Notably, the Macro F1-score is highlighted in red for emphasis.

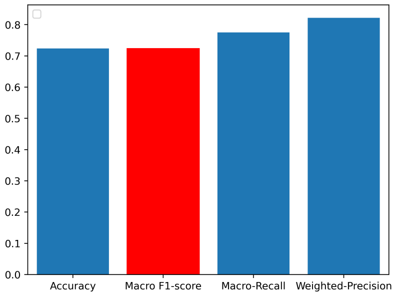

Figure 1: Tamil Stress Identification Results

6.2 Telugu Language

Similarly, for the Telugu language, our model's performance is summarized below:

Metrics	Score
Accuracy	0.729
Macro F1-score	0.727
Macro-Recall	0.756
Weighted Precision	0.779

The bar chart in Figure 2 visually presents the Telugu language metrics, with Macro F1-score highlighted in red.

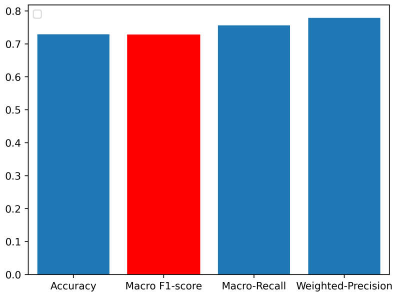

Figure 2: Telugu Stress Identification Results - Naive Bayes

These results demonstrate the effectiveness of our stress identification model in both Tamil and Telugu even with strict and minimal preprocessing and hyperparameter tuning, providing a good foundation and baseline for further exploration as most previous works were focused on general emotion detection in Tamil or Telugu or the targeted language were mostly English.

7 Discussions and limitations

Our model for the Telugu language exhibited commendable performance by securing the second rank in the shared tasks. We can observe that both our models did not overfit and remained stable on the test set. This achievement underscores the effectiveness of the chosen approach and the potential for accurate stress identification in Dravidian languages.

It is noteworthy that sufficient data and additional features related to the two languages Tamil and Telugu can help our models become more reliable in the identification of emotional stress in textual corpora in both Tamil and Telugu. This observation emphasizes the importance of data abundance in enhancing model performance, paving the way for more accurate and robust stress identification systems.

However, it's essential to acknowledge certain limitations in our current approach. One notable limitation is we did not make use of phonology or phonetic features which are important factors in extracting meaningful information from Tamil or Telugu. Despite the capabilities of IndicNlp for phonetic feature extraction, this aspect was not explored in our experiments. Future investigations could delve into extracting phonetic features, potentially enriching the model's understanding of stress patterns in the spoken language.

Moreover, there exists ample room for further experiments. Techniques such as embedding learning offer a promising avenue for feature extraction, potentially capturing nuanced linguistic representations. Alternatively, leveraging large language models can provide a comprehensive understanding of stress-related features in Dravidian languages since both languages make use of context.

8 Conclusion

In conclusion, while the success of our model in the Telugu language showcases the effectiveness of our approach, ongoing research, and experimentation are crucial to unlocking the full potential of stress identification in Tamil and Telugu. Exploring additional features, incorporating advanced techniques, and leveraging large-scale language models are avenues for future research, promising advancements in computational linguistics for emotion analysis in diverse linguistic contexts.

Ethics Statement

Our scientific work adheres to the ACL Ethics Policy[1], ensuring the responsible conduct of research and publication. We recognize the ethical implications of our work in stress identification in Dravidian languages and commit to upholding the highest standards throughout our research process.

As researchers, we remain vigilant about the ethical dimensions of our work and its impact on individuals and communities. We welcome open dialogue and scrutiny regarding the ethical considerations associated with our research and are dedicated to fostering responsible and ethical practices within the computational linguistics community.

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

[1] https://www.aclweb.org/portal/content/acl-code-ethics

the Laboratorio de Supercómputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

José García-Díaz, Miguel Ángel Rodríguez García, and Rafael Valencia-García. 2022. UMUTeam@TamilNLP-ACL2022: Emotional analysis in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 39–44, Dublin, Ireland. Association for Computational Linguistics.

Omkar Gokhale, Shantanu Patankar, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. Optimize$_{p}rime@DravidianLangTech - ACL2022 : EmotionAnalysisinTamil$. *arXiv (Cornell University)*.

Shaunak Inamdar, Rishikesh Chapekar, Shilpa Gite, and Biswajeet Pradhan. 2023. Machine learning driven mental stress detection on Reddit posts using natural language processing. *Human-Centric Intelligent Systems*, 3(2):80–91.

Aparna "Joshi, LV Gangolli, R Duggal, and A" Shukla. 2005. "mental health in india: review of current trends and directions for the future". *"Review of Health Care in India"*, pages "127–136".

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Y. K. Nagle and Usha Sharma. "2018". "academic stress and coping mechanism among students: An indian perspective". *Journal of child and adolescent psychiatry*, "2"(1).

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).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Varsini S, Kirthanna Rajan, Angel S, Rajalakshmi Sivanaiah, Sakaya Milton Rajendram, and Mirnalinee T T. 2022. Varsini_and_Kirthanna@DravidianLangTech-ACL2022-emotional analysis in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 165–169, Dublin, Ireland. Association for Computational Linguistics.

Moein Shahiki-Tash, Jesús Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023a. Lidoma at homomex2023@ iberlef: Hate speech detection towards the mexican spanish-speaking lgbt+ population. the importance of preprocessing before using bert-based models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*.

Moein Shahiki-Tash, Jesús Armenta-Segura, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023b. Lidoma at hope2023iberlef: Hope speech detection using lexical features and convolutional neural networks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), colocated with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS. org*.

M Shahiki Tash, Z Ahani, Al Tonja, M Gemeda, N Hussain, and O Kolesnikova. 2022. Word level language identification in code-mixed kannada-english texts using traditional machine learning algorithms. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 25–28.

Moein Tash, Jesus Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 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*, pages 180–185.

Atnafu Lambebo Tonja, Mesay Gemeda Yigezu, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbuk. 2022. Transformer-based model for word level language identification in code-mixed kannada-english texts. *arXiv preprint arXiv:2211.14459*.

Mesay Gemeda Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. Transformer-based hate speech detection for multi-class and multi-label classification.

Mikhail Zaydman. 2017. *Tweeting about mental health: Big Data text analysis of Twitter for public policy*.