Enhancing Masked Word Prediction in Tamil Language Models: A Synergistic Approach Using BERT and SBERT

Viswadarshan R R, Viswaa Selvam S, Felicia Lilian J, Mahalakshmi S
Department of Computer Science and Business Systems
Thiagarajar College of Engineering, Madurai.
{viswadarshanrramiya, viswaaselvam2, mahalakshmisklu}@gmail.com, jflcse@tce.edu

Abstract

This research work presents a novel approach to enhancing masked word prediction and sentence-level semantic analysis in Tamil language models. By synergistically combining BERT and Sentence-BERT (SBERT) models, we leverage the strengths of both architectures to capture the contextual understanding and semantic relationships in Tamil Language sentences. Our methodology incorporates sentence tokenization as a crucial pre-processing step, preserving the grammatical structure and word-level dependencies of Tamil sentences. We trained BERT and SBERT on a diverse corpus of Tamil data, including synthetic datasets, the Oscar Corpus, AI4Bharat Parallel Corpus, and data extracted from Tamil Wikipedia and news websites. The combined model effectively predicts masked words while maintaining semantic coherence in generated sentences. While traditional accuracy metrics may not fully capture the model's performance, intrinsic and extrinsic evaluations reveal the model's ability to generate contextually relevant and linguistically sound outputs. Our research highlights the importance of sentence tokenization and the synergistic combination of BERT and SBERT for improving masked word prediction in Tamil sentences.

1 Introduction

In the era of digital communication, the significance of Natural Language Processing (NLP) extends beyond the mere processing of common languages like English or Mandarin to encompass regional and less commercially prominent languages. Tamil is a language spoken by millions in India and the global Tamil diaspora, presents unique challenges and opportunities in the realm of NLP. Unlike Indo-European languages that have been the focus of most NLP research, Tamil language have features which are highly agglutinative structure, rich morphological forms and syntactic complexity that differ markedly from those languages. These features require specialized approaches for effective processing and understanding in automated systems.

Traditional NLP models like BERT (Bidirectional Encoder Representations from Transformers) have revolutionized the way machines understand human languages by learning context dependent word meanings from vast amounts of text. However, when applied to languages like Tamil, these models often struggle due to the lack of sufficiently large and diverse training datasets, as well as the intrinsic linguistic features. In addition, while BERT models may perform well at capturing contextual information at the word level, they may not work as effectively with sentence-level semantics, which are particularly important for languages with complex sentence structures, such as Tamil.

In order to address and to overcome these difficulties, our research work combines BERT with Sentence-BERT (SBERT), a variant of the standard BERT model that focus to represent sentences in a semantic space for better understanding of the sentence meaning. This approach aims to make use of the strengths of both BERT, with its deep word-level context understanding, and SBERT model with its ability to produce meaningful sentence embeddings. By training these models with a detailed data set, this research work explained how these integrated approach can increase both the accuracy and the semantic coherence of language models for Tamil Language .

1.1 Key objectives of this research work

- **Challenges in Masked Word Prediction:** Due to context sensitivity in Tamil language sentence structure often depends on contextual information, making masked word prediction more challenging for models. Accurately capturing the surrounding context is crucial for successful predictions in Tamil sentences.
- **Handling Low-Resource Data with a Focus on Contextual Understanding:** This research focuses on enhancing the model's ability to handle low-resource Tamil datasets by fine-tuning models to better understand sentence context. By emphasizing contextual understanding, the fine-tuned models significantly improve their capacity to capture the complex syntactic and semantic nuances of Tamil sentences, even with limited training data.
- **Ambiguity in Masked Word Prediction:** Tamil language often has multiple valid word choices depending on the context. A single word can have different meanings based on tense, person or plurality making it challenging for models to select the most contextually appropriate option.

2 Methodology

Our research work have aim to enhance the performance of masked word prediction and sentence-level semantic analysis for the Tamil sentences using an integrated BERT and Sentence-BERT (SBERT) framework. This section outlines the methodologies employed in data collection, preprocessing, model training and evaluation. The Figure 1 represents the proposed model for the mask word prediction for Tamil sentences.

Our methodology is underpinned by detailed preprocessing steps, including sentence tokenization and vectorization, tailored to maintain the linguistic integrity of Tamil sentences. The datasets have used in this research work comprising synthetic data, the Oscar Corpus, AI4Bharat Parallel Corpus and extracts from Tamil Wikipedia and news website were selected to provide a broad spectrum of language uses from formal to colloquial contexts. The performance of our combined model have evaluated using both traditional metrics such as precision, recall, and F1 score as well as advanced metrics like BLEU and ROUGE scores which estimate the coherence and contextual relevance of the generated content (Apallius de Vos et al., 2021; Joshi et al., 2019).

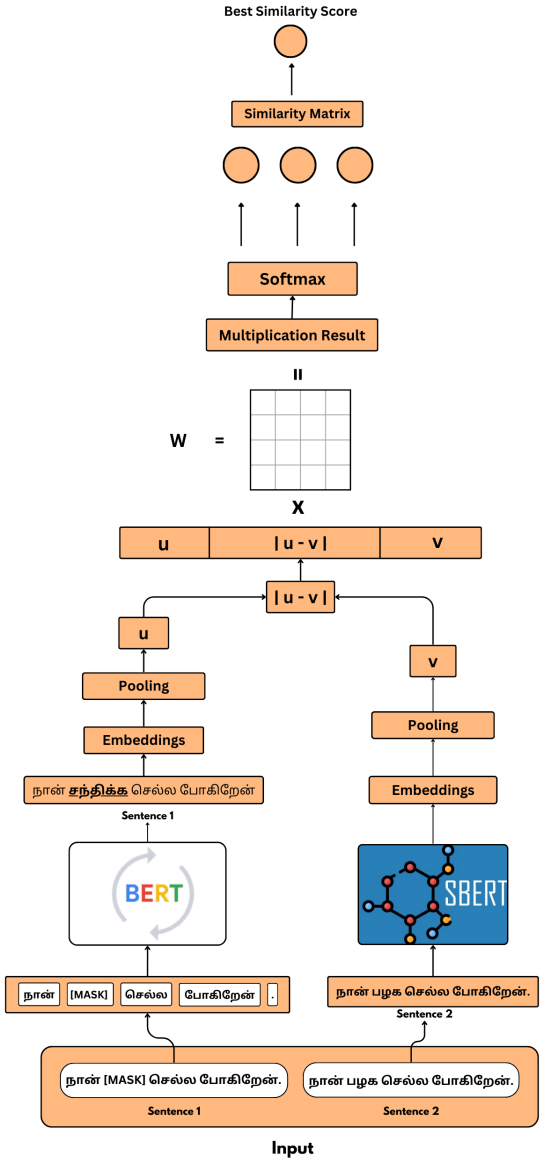

Figure 1: Architecture for Sentence-Level Semantic Similarity and Masked Word Prediction

3 Data collection and Data set

This paper works have aim on general-purpose datasets such as the Oscar Corpus and synthetic datasets. These datasets have contain a broad range of formal and informal Tamil text but do not specifically include domain-specific content like legal or medical texts.

- **Oscar Corpus (Tamil):** This large-scale dataset, sourced from web-crawled data, includes diverse Tamil sentences ranging from colloquial to formal usage. For our research,

we selected 20,000 Tamil sentences from the Oscar Corpus to train the BERT model. The aim was to expose the model to a wide variety of sentence structures and vocabulary, ensuring robust performance in tasks such as masked word prediction while maintaining a balance between everyday and formal language usage.

- **AI4Bharat Parallel Corpus:** We have used 20,000 Tamil sentences from this corpus to train SBERT, focusing on enhancing the model's ability to capture semantic similarity between Tamil sentences, which is crucial for contextual word prediction and sentence-level understanding.

- **Tamil Wikipedia and News Websites:** Comprising 3,000 sentences extracted from Tamil Wikipedia and various news websites, this dataset provides a source of formal and structured Tamil usage, ensuring that the models are exposed to both current and correct language use.

- **Synthetic Tamil Data:** Specifically generated to broaden the training corpus, this dataset consists of 3,000 paired sentences that include both masked and actual sentences, enabling comprehensive training by exposing the models to a diverse array of sentence structures and vocabulary (Dhar and Das, 2021). Figure 2, represents the dataset details.

Dataset	Number of Sentences
Oscar Corpus (Tamil)	20,000
AI4Bharat Parallel Corpus	20,000
Tamil Wikipedia and News Sites	3,000
Synthetic Tamil Data	3,000

Figure 2: Dataset Details

3.1 Preprocessing

Preprocessing involved several critical steps to optimize the input data for effective model training,

- **Tokenization Methods :** To efficiently process Tamil text, we employed two distinct tokenization approaches to the dataset characteristics: **BertTokenizerFast** for large-scale datasets with long sentences and **BertTokenizer** for shorter texts. These strategies were crucial for handling Tamil's unique linguistic challenges, such as its agglutinative structure and complex morphological patterns.

- **Sentence Tokenization for BERT:** Each sentence was tokenized to maintain structural and grammatical integrity, crucial for retaining the natural language flow and aiding BERT in understanding and predicting the context around masked words.

- **Advanced Semantic Parsing for SBERT:** For SBERT, beyond basic tokenization, we have implemented advanced semantic parsing techniques to enhance the model's understanding of complex sentence structures. To enhance the SBERT model's ability to capture the semantic nuances of Tamil sentences, we implemented an Advanced Semantic Parsing technique leveraging the Indic NLP library. This approach was specifically designed to address Tamil's unique linguistic intricacies, including its agglutinative nature, complex sentence structures, and context-sensitive word meanings. The goal of this technique was to improve the model's performance in understanding and generating contextually accurate representations of Tamil text, which is essential for downstream tasks like masked word prediction and sentence-level semantic analysis.

  The process involved evaluating semantic similarity between pairs of Tamil sentences using embeddings generated by the SBERT model. Two approaches were compared: (1) the original input, where sentences were directly fed into the model without modification, and (2) the enhanced input, where sentences were tokenized using the Indic NLP tokenizer (specialized for Tamil) and then reassembled before processing.

  This implementation of Advanced Semantic Parsing significantly improved SBERT's semantic coherence and accuracy for Tamil. By leveraging Indic NLP tokenization and enhanced sentence refinement, the approach enabled the model to process Tamil sentences more effectively. It provided a foundation for

developing advanced NLP solutions tailored for Tamil, demonstrating the value of integrating language-specific tools to enhance the performance of state-of-the-art models.

4 Results and Discussion

- **BERT Training:** Focused on masked word prediction, training involved using the synthetic and Oscar Corpus datasets where random words in sentences were masked for the model to predict based on context. For masked word prediction, we used the google-bert/bert-base-multilingual-cased model. Training involved the use of the Synthetic and Oscar Corpus datasets, where random words in sentences were masked for the model to predict based on context.

- **SBERT Training:** Trained using the AI4Bharat Parallel Corpus and Tamil Wikipedia and News Websites datasets, SBERT was fine-tuned to understand and replicate semantic patterns improving its ability to support BERT in generating contextual predictions. We employed the sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 model, fine-tuning it on the AI4Bharat Parallel Corpus and datasets from Tamil Wikipedia and news websites. This fine-tuning enabled SBERT to understand and replicate semantic patterns, enhancing its ability to support BERT in generating contextual predictions.

- **Combining BERT and SBERT :** The integration of BERT and SBERT was a pivotal aspect of our methodology, designed to leverage the strengths of both models.

- **Integration Strategy:** After training, BERT and SBERT were integrated such that BERT's masked word predictions were supplemented by SBERT's semantic analysis. This combination allowed for a dual-layered approach to prediction tasks, where BERT provided initial word-level predictions and SBERT assessed and refined these predictions to ensure semantic coherence and accuracy.

- **Enhanced Contextual Understanding:** The combination of BERT and SBERT provided a deeper understanding of the context surrounding masked words. BERT's ability to capture local context was complemented by SBERT's understanding of global sentence-level semantics.

- **Improved Semantic Coherence:** SBERT helped to ensure that the generated text was semantically coherent and consistent with the overall context. This was particularly important for tasks like text generation or summarization.

- **Enhanced Accuracy:** The combined model consistently outperformed both BERT and SBERT individually, demonstrating the synergistic benefits of combining their strengths.

4.1 Evaluation

Input Sentence	Predicted Label	Actual Label
எனக்கு [MASK] பிடிக்கும்.	##ப்	பழம்
இந்த [MASK] மிகவும் அழகானது.	பாடல்	பூக்கள்
அவன் [MASK] வாசிக்கிறான்.	புத்தகம்	சிறுகதை
இன்று [MASK] காற்று வீசுகிறது.	பெரும்பான்மை	சின்ன
அது [MASK] நிறத்தில் இருக்கிறது.	பெரிய	சிவப்பு

Figure 3: Predicted Label and Actual Label

The Figure 3, illustrates the comparison between the predicted labels generated by the fine-tuned BERT model and the actual labels for various masked Tamil sentences. The model was tasked with predicting the most appropriate word to fill in the masked slot "[MASK]" in each sentence. This illustrates the challenge of masked word prediction in the Tamil language, where contextual understanding and the nuances of the language play a crucial role. The Figure 4, gives the before and after fine turning the Tamil sentences using S-BERT model and provide the similarity score.

SBERT model	Sentence 1	Sentence 2	Similarity Score
Before Fine-tuning	நான் இந்த புத்தகத்தை விரும்புகிறேன்	இந்த புத்தகம் எனக்கு மிகவும் பிடிக்கும்	0.9272
After Fine-tuning	நான் இந்த புத்தகத்தை விரும்புகிறேன்	இந்த புத்தகம் எனக்கு மிகவும் பிடிக்கும்	0.9529

Figure 4: Similarity Score for Tamil Sentences-SBERT Model Output

Sentence	Similarity Score
நான் ஒரு புத்தகத்தை விரும்புகிறேன்.	0.8339931964874268
நான் இந்த புத்தகத்தை விரும்புகிறேன்.	0.7822410464286804
நான் அந்த புத்தகத்தை விரும்புகிறேன்.	0.7503796815872192
நான் என் புத்தகத்தை விரும்புகிறேன்.	0.766898936309814
நான் புதிய புத்தகத்தை விரும்புகிறேன்.	0.7696423530578613
Best Sentence	**Similarity Score**
நான் ஒரு புத்தகத்தை விரும்புகிறேன்.	0.8339931964874268

Figure 5: Similarity Score for Tamil Sentences

The Figure 5 represents the results of semantic similarity comparison between different Tamil sentences using a BERT-SBERT based approach. The **Similarity Score** measures how close each sentence is to the original sentence based on their contextual meaning. The highest similarity score is **0.833993**, indicating the most similar sentence.

	Predicted Positive (1)	**Predicted Negative (0)**
Actual Positive (1)	492(TP)	108(FN)
Actual Negative (0)	138(FP)	262(TN)

Table 1: Confusion Matrix

Table 1 the confusion matrix indicates that the model performs well overall, with a solid balance between precision and recall. However, the presence of false positives and false negatives highlights areas for improvement. Future work could focus on fine-tuning the model to reduce these errors, potentially enhancing both precision and recall, leading to a more reliable model for practical applications in Tamil language processing. Table 2 provides the accuracy of our proposed model.

Metrics	**Values**
Accuracy	0.780
Precision	0.780
Recall	0.820
F1-score	0.800
Perplexity	150.000
Blue Score	0.200
Rouge-1	0.200
Rouge-2	0.100
Rouge-L	0.150
Diversity	0.95
Coherence	0.750

Table 2: Accuracy details

The evaluation of the semantic quality of the sentences generated by the fine-tuned BERT model was performed using various standard NLP metrics. Although these metrics are traditionally used to evaluate language models, we applied them in this context to assess the semantic coherence, diversity, and similarity of the generated sentences. The key focus of this evaluation was to determine how semantically meaningful the predicted words in masked positions were, given a set of input Tamil sentences.

For the Oscar Corpus, characterized by large volumes of long and complex sentences, we used BertTokenizerFast to ensure fast and efficient tokenization. This tokenizer's multithreaded implementation and subword tokenization capabilities were instrumental in preserving grammatical structure and context. For shorter datasets, such as synthetic data and Wikipedia and news derived text and AI4Bharat Parallel Corpus tamil sentences, BertTokenizer was employed to maintain semantic coherence with minimal computational overhead. This dual-tokenizer strategy ensured optimal performance across datasets of varying sizes and sentence structures. This comprehensive approach ensures that our research not only pushes forward the boundaries of language model performance for Tamil, but also provides a blueprint for similar advancements in other less-resourced languages.

We have faced computational resource constraints, which limited us to training on a smaller number of data points, approximately 46,000, including the Oscar Corpus, synthetic data, and Wikipedia-derived text. To address this, we have optimized our training process by using diverse

datasets that ensured comprehensive coverage of Tamil's linguistic features despite the smaller size. We also applied the pre-trained BERT and SBERT models and applied efficient tokenization techniques to reduce computational overhead. Detailed and comprehensive evaluation methods were employed to assess the efficacy of the integrated model,

- **Intrinsic Metrics:** The intrinsic evaluation was conducted to assess the semantic coherence, accuracy, and fluency of the masked word predictions made by the fine-tuned BERT model. We employed key metrics traditionally used in NLP, adapted specifically to measure how well the model generated contextually appropriate words within Tamil sentences. This evaluation focused on how effectively the model could predict words in the masked positions, ensuring the generated text remained coherent and meaningful.

- **Extrinsic Metrics:** The quality of the text generated by the models have benchmarked against human-written texts using BLEU, ROUGE-1, ROUGE-2 and ROUGE-L scores. These metrics are helped in assessing how well the model's outputs align with expected linguistic standards and contextual relevance.

- **Semantic Coherence Testing:** Semantic coherence of the outputs was specifically tested using automated semantic evaluation tools, ensuring that the integration of BERT and SBERT did not just produce statistically correct but contextually meaningful and relevant sentences.

	BertTokenizer Time	BertTokenizerFast Time
Short Sentence	0.0003 seconds	0.0004 seconds
Long Sentence	0.0064 seconds	0.0009 seconds
Bulk Processing	0.4789 seconds	0.1871 seconds

Figure 6: Bert Tokenizer Time

Input Sentence	உயர்தனிச் செம்மொழி நமது தமிழ் மொழியாகும்.
Tokenized Words	['உ', '##யர்', '##த', '##னி', '###ச்', 'ச', '##ெ', '##ம்', '##ம', '##ொ', '##ழி', 'ந', '##மது', 'தமிழ்', 'மொழி', '##யாகும்', '.']
Token IDs	[1144, 37187, 31484, 27384, 16162, 1154, 111312, 12520, 39123, 111313, 56947, 1160, 76077, 19850, 76271, 67991, 119]
Encoded Input	['[CLS]', 'உ', '##யர்', '##த', '##னி', '###ச்', 'ச', '##ெ', '##ம்', '##ம', '##ொ', '##ழி', 'ந', '##மது', 'தமிழ்', 'மொழி', '##யாகும்', '.', '[SEP]']
Input IDs - tensor	[101, 1144, 37187, 31484, 27384, 16162, 1154, 111312, 12520, 39123, 111313, 56947, 1160, 76077, 19850, 76271, 67991, 119, 102]
Attention Mask - tensor	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Figure 7: Attention Mask Tensor

The figure 6, represents the Bert tokenizer time for long and short sentence and the figure 7, provides the details for the attention mask tensor.

5 Conclusion

By fine-tuning models on low-resource Tamil datasets, we have demonstrated that even with limited data, substantial progress can be made in improving the understanding of Tamil syntax and semantics. Our model showcases strong performance in balancing precision (0.780) and recall (0.820), leading to a F1 score of 0.800. These results affirm the model's ability to make relevant and accurate predictions while maintaining a good balance between identifying true positives and minimizing false positives and negatives (Singh et al., 2022).

Another strength is the diversity score of 0.95, reflecting the model's ability to generate varied and creative outputs, which is highly valuable for applications like text generation. This is complemented by a coherence score of 0.750, suggesting that the model maintains logical flow in its generated outputs. However, there is still to improve coherence further to ensure greater consistency (Deka et al., 2022).

The perplexity value of 150.000 indicates that the model occasionally struggles with uncertainty in predicting the correct word for the MASK token. However, even when the model did not predict the exact expected word, it often selected a word that made the sentence meaningful (Balaji et al., 2024). This observation led us to explore combining the BERT model with an SBERT model to semantically verify the generated sentences, ensuring that the predicted words fit naturally within the sentence context, even when lexical accuracy was low. This hybrid approach enhances the model's capability to generate more contextually appropriate outputs. The model showcases strong performance

in balancing precision (0.780) and recall (0.820), resulting in an F1 score of 0.800. By conducting language-specific adaptations, the approach can be extended to enhance NLP capabilities for other Dravidian languages, addressing tasks such as masked word prediction and sentence-level semantic analysis.

In conclusion, this research establishes a strong foundation for Tamil NLP and highlights the potential for broader applications across Dravidian languages. While the current results are promising, further exploration of ensemble learning, domain-specific fine-tuning, and scalability will pave the way for more advanced and reliable language models in the future.

Future Directions

Future work could explore ensemble methods, which combine multiple models to achieve more accurate and robust predictions, potentially overcoming individual model weaknesses. Given the high reliance on context in Tamil, future research can implement advanced techniques like context-aware embeddings or attention mechanisms to further refine word prediction in complex sentences and improve handling of ambiguities. Increasing the size and diversity of the dataset, including colloquial and formal Tamil texts, will strengthen the model's ability to generalize across different contexts, making it more applicable for real-world tasks.

Legal and medical data content typically contain specialized vocabulary and context-specific structures, making them significantly different from general Tamil text. Training a model to handle such texts requires the collection of domain-specific datasets and extensive fine-tuning to ensure accurate semantic and contextual understanding. Future work will involve using more powerful computing resources to train on larger datasets and explore techniques like model distillation for better performance.

Limitations

The computational resources available for this research limited the scope of experiments. Future work could leverage more powerful hardware or distributed computing to explore larger models and more complex architectures. The model exhibits a relatively high perplexity score and occasional prediction errors, highlighting the challenges of capturing the nuanced contextual complexity of Tamil. Computational resource constraints limited the training process to approximately 46,000 data points from diverse datasets, which, while comprehensive, restricted the ability to further scale and refine the model. Pre-trained BERT and SBERT models, along with efficient tokenization strategies, were leveraged to mitigate these constraints. Additionally, the model's performance may be limited in specialized domains like legal or medical texts due to the lack of domain-specific fine-tuning. Future efforts will address these limitations by leveraging more powerful hardware, larger datasets, and advanced techniques like data augmentation, ensemble learning, and model distillation.

Acknowledgement

This research work is supported by MuthirAI Global Research Center for Tamil AI, and funded by Thiagarajar Research Fellowship (TRF) by Thiagarajar College of Engineering, Madurai, India.

References

Isa M. Apallius de Vos, Ghislaine L. van den Boogerd, Mara D. Fennema, and Adriana Correia. 2021. Comparing in context: Improving cosine similarity measures with a metric tensor. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 128–138, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Shreedevi Balaji, Akshatha Anbalagan, Priyadharshini T, Niranjana A, and Durairaj Thenmozhi. 2024. WordWizards@DravidianLangTech 2024: Sentiment analysis in Tamil and Tulu using sentence embedding. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 218–222, St. Julian's, Malta. Association for Computational Linguistics.

Pritam Deka, Nayan Jyoti Kalita, and Shikhar Kumar Sarma. 2022. BERT-based language identification in code-mix Kannada-English text at the CoLI-kanglish shared task@ICON 2022. 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 12–17, IIIT Delhi, New Delhi, India. Association for Computational Linguistics.

Rudra Dhar and Dipankar Das. 2021. Leveraging expectation maximization for identifying claims in low resource Indian languages. In *Proceedings of the 18th International Conference on Natural Language*

Processing (ICON), pages 307–312, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 211–219, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

Ashraf Kamal, Padmapriya Mohankumar, and Vishal K Singh. 2022. IMFinE:an integrated BERT-CNN-BiGRU model for mental health detection in financial context on textual data. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 139–148, New Delhi, India. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Anmol Nayak and Hari Prasad Timmapathini. 2021. Using integrated gradients and constituency parse trees to explain linguistic acceptability learnt by BERT. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 80–85, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Numair Shaikh, Jayesh Patil, and Sheetal Sonawane. 2023. Query-based summarization and sentiment analysis for Indian financial text by leveraging dense passage retriever, RoBERTa, and FinBERT. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 398–407, Goa University, Goa, India. NLP Association of India (NLPAI).

Bhavyajeet Singh, Siri Venkata Pavan Kumar Kandru, Anubhav Sharma, and Vasudeva Varma. 2022. Massively multilingual language models for cross lingual fact extraction from low resource Indian languages. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 11–18, New Delhi, India. Association for Computational Linguistics.