UNVEILING EMOTIONAL NUANCES: LEVERAGING FINE-TUNED TRANSFORMER MODELS FOR SENTIMENT ANALYSIS OF HINGLISH TEXTS ON MENTAL HEALTH IN NLP

Shruti mathur[1], Dr. Gourav Shrivastava[2]
[1]PhD Scholar, Sanjeev Agrawal Global Educational University, Bhopal, Madhya Pradesh, India
[2]Associate Professor, Sanjeev Agrawal Global Educational University, Bhopal, Madhya Pradesh, India

Abstract: Understanding emotions in Hinglish, a mix of Hindi and English commonly used in India, especially when it comes to mental health discussions, is a tough nut to crack for us, folks in the field of natural language processing (NLP). It's like trying to decipher a secret code where emotions are hidden. Few work exists for NLP to navigate these rough waters of mixed languages and sensitive topics like mental well-being. In our quest, we decided to take a deep dive and test three powerful deep learning models: HingRoBERTa, HingBERT, and RoBERTa. We meticulously curated a treasure trove of 6000 Hinglish samples from social media platforms like Twitter and Reddit. Imagine it as a collection of real voices and experiences! We then fine-tuned these models with great care and evaluated them using standard tools like accuracy, precision, recall, and F1-score. The results were exciting! HingRoBERTa emerged as the champion, demonstrating an impressive 88.5% accuracy in navigating the complexities of Hinglish. It was like watching a master linguist effortlessly untangle a coded message. HingBERT also impressed, showing a strong ability (86% accuracy) to distinguish between neutral and emotional Hinglish text. Even RoBERTa, though not built specifically for Hinglish, surprised us with decent results. These findings highlight the critical need for NLP tools designed for diverse languages, especially in regions like India where Hinglish and other mixed languages are commonplace.

Keywords: Code-mixed, NLP, Deep Learning, Mental Health, BERT.

1. INTRODUCTION

Understanding Codemixed Language is crucial in the field of Natural Language Processing. The field of Natural Language Processing encompasses various techniques and methodologies to analyze and interpret human languages used in codemixed texts.

The area of natural language processing (NLP) integrates rule-based human language modelling with computational linguistics, statistical, and machine learning models. The purpose of this combination is to help computers and other digital devices identify, understand, and produce text or speech. Developing programming software that can accurately interpret data from text or voice seems like a difficult task due to human language. The most challenging elements include vagueness, homonyms, as well as ambiguities such as sarcasm, idioms, and metaphors.

Increasing the amount of transcribed training data can lead to remarkable improvements in Automatic speech recognition(ASR) performance for large or world languages as it is one of the more effective ways. In the case of low-resource languages and also languages in the developing world where there is a high cost associated with transcription, the massive amounts of transcribed training data are unavailable to all but a few languages. Consequently, there is a growing need for methods that will allow ASR models to be trained on a short supply of training data without transcription [1].

Moreover, the differences between grammar, usage, and sentence structure can also be considered as difficulty factors. In this sense, the natural language systems that software developers must include in their applications are alien to the end-users who will not be able to operate them without some additional effort.

Code-mixed expressions with Hinglish and regional languages blend in India with English, creating a harmonious mix of diverse languages [2]. Understanding mental health within this linguistic quilt requires state-of-the-art NLP models that can appreciate the subtle nuances brought on by the multilingual variables, as they need to translate cultural expressions, approach stigma sensitively, and much more. The development of such models helps foster better communication channels, inclusive reach, and personalized support for individuals addressing their own mental well-being across India's vast linguistic landscape.

India is a country where diverse languages are spoken without any differences, for instance, code-mixed language such as Hinglish combines with regional languages and English. To catch mental health in this language mosaic, we need NLP tools that are attentive to details. To decode the entire conversation correctly, they must be able to identify cultural references in different tongues or tackle negative perceptions attached to it [3]. In terms of the richness of Indian languages, the model creation provides the ability to communicate well, accessibility, and support for overcoming personal obstacles associated with mental illness.

Code mixing occurs when an individual employs a combination of two or more languages in a single sentence, conversation, or even phrase. It is common in societies where people speak more than one language fluently [4]. This can include many aspects such as grammar syntax, as well as the usage of words in the vocabulary. Hinglish, spoken in India, is one example of a code-mixed language that reflects the cultural heritage and the impacts of globalization experienced by this country. Code-mixing is one of the hybrid data types, and any computational analysis of such mixed data types must not ignore this factor. In India, Hinglish stands as a good example of code-mixed language where Hindi and English intermingle, demonstrating its multicultural nature and the impact of globalization. Computational analysis of such mixed data types requires special attention given that code mixing falls into the category of hybridity which means it cannot be understood by usual methods of calculations.

2. Background

2.1 Challenges faced by NLP in India

In Indian circumstances, NLP usage adapting to mental health problems must factor in-country health-related culture, language, and environment. Here are some specific challenges:

Culture and stigma: Mental health in India is often stigmatized, leading to lack of awareness and denial. ask for help. It is difficult to develop NLP techniques that respect cultural differences, eliminate stigma, and encourage help-seeking behavior.

Linguistic Diversity: India is linguistically diverse with hundreds of languages and dialects spoken across the country. Developing NLP models that can understand and translate mental health messages in multiple languages is critical to accessibility and efficiency.

Data Accessibility and Quality: Despite India's large population, access to mental health data, especially digital data, can be limited. It is difficult to ensure the availability and quality of data used to train NLP models while complying with privacy laws such as the Indian Data Protection Act.

Low Digital Literacy: A large portion of India's population has low levels of digital literacy, which may impact the availability and effectiveness of NLP-based psychotherapy tools. It is important to create user-friendly interfaces and provide easy access for users with different levels of digital knowledge.

Speech, incorrect speech, different speech, stuttering, etc. It will become difficult for machines to understand. However, these problems can be reduced as the database grows and the smart assistant is trained by a single user.

2.2 Challenges in code-mixed languages

Challenges in code-mixed languages in code-mixed language processing, various challenges are

Syntactic and Semantic Variability: For example, in Hinglish, the sentence structure may follow Hindi syntax while using English vocabulary, or vice versa, leading to complex parsing requirements.

Limited Resources and Datasets: The lack of extensive, well-annotated datasets in languages like Spanglish (Spanish-English) or Taglish (Tagalog-English) limits the development of effective models.

Word matching and tokenization: In a language like Tanglish (Tamil-English), it is very difficult to match Tamil words written in Romani with English words for proper tokenization.

Contextual Ambiguity: A word like "set" in Manglish (Malayalam-English) might have different meanings based on whether it's used in an English or Malayalam context.

Cultural and Regional Variations: In Franglais (French-English), Canadians and European French speakers may have different pronunciations, which can affect the translation mix.

2.3 Addressing mental health in India.

Healthcare in India faces many challenges, especially in a mixed language context.

Cultural Boundaries: Mental illnesses are often stigmatized, leading to underreporting of the problem. For example, discussing depression can be taboo in some communities.

Language Diversity: India's linguistic diversity makes it difficult to provide resources in all languages. Sentiment analysis tools will have difficulty accurately translating the regional language into British English.

Access to Care: Adequate mental health services are not available in many areas, especially in rural areas. This makes it difficult for people to get the help they need. There are many apps to consider mental health in Code Mix Language:

Digital Health Platforms: Apps and websites offering counseling in English Hindi are available to further assist with Health and wellness.

Prevention Research: Sentiment analysis of Indian English social media posts may help early detection of mental health problems.

Awareness campaigns: Campaigns in Indian English can reach a wider audience, raise awareness about mental health and reduce stigma.

3. Related work

Recent research on code-switching has led to some interesting findings, including a comprehensive review of emotional evaluation methods, the benefits and challenges of emotional evaluation, and a comparison of the advantages and disadvantages of various methods. This article also discusses the general acceptance of emotional evaluation and its importance in various fields (Wankhede et al., 2022) [5]. Sharma et al. The 2022 Assessment believes that needs assessment is a topic that has been researched for many years using different methods depending on the type and size of the data set [6]. Machine learning and Transformer-based methods have had a significant impact on the analysis of emotions, leading to the development of accuracy and state-of-the-art models. Due to its richness and presence, Twitter has become the main point of opinion analysis used in many areas such as business, political predictions, line analysis, disease, data security, disaster response.

Mabokela et al. 2022 [7] provides a comprehensive guide to multilingualism research, surveys the latest trends in MSA approaches to non-verbal languages, and provides popular literature on MSA (multilingual study) research. It focuses on multilingualism, emphasizing the role of deep learning and machine translation. This study identifies inconsistencies in current analytical thinking regarding negative language use, highlighting the need for more robust models and data. This work highlights the challenges and possibilities of advancing emotional analysis in different languages and changing environments and argues for a variety of NLP models and adaptations. Shanmugavadivel et al. 2022 studied sentiment analysis and negative language recognition using the Adapter-BERT model and showed an accuracy rate of 65% and a negative language recognition rate of 79%. The article also highlights the serious problem of trolling, hostility and derogatory messages on social media

Ondrej Klejch et al. 2021[8] said they compared two sets of examples: the first used a dictionary containing the pronunciation of the Bengali/Hindi language and the phonemes of the English language, while the second used a dictionary called the Bengali/Hindi language. balance. Initially, they were trained in different languages based on Bengali and Hindi teaching materials for each group. The training data is then refined, improving performance in both languages (absolute improvement of over 3% and 6% for Bengali and Hindi respectively). They further improved the results by reclassifying the test data using WebRTC VAD and achieved word error reduction (WER) of 0.8% and 1.8% for Bengali and Hindi, respectively. Then, we use Web LM to analyze the test data and use RNN LM to improve the results. Interestingly, for Bengali, rescoring with RNN LM causes performance degradation. Through this step, we reduced the WER of the phoneme model from 28.5% to 23.3% for Bengali and from 27.2% to 18.1% for Hindi. We also attempted single-level transfer learning from the different languages studied in Track 1, but did not achieve any improvements over those previously reported.

Sengupta et al. 2021[9] proposed HIT (Hierarchically attentive Transformer) framework significantly improves performance in representing code-mixed texts for multiple languages and NLP tasks. HIT demonstrates adaptability in a transfer learning setup, indicating its potential for capturing task-invariant linguistic features. The paper presents a novel hierarchical process tracing transformation for learning word representations of code-mixed texts for six non-English languages. Nayak et al. 2022[10] proposed L3Cube-HingCorpus, the significant improvements of the models pre-trained on real code-mixed data, and the effectiveness of pretraining on real code-mixed data. The intervention is the pre-training of BERT models (HingBERT, HingMBERT, HingRoBERTa) and GPT model (HingGPT) on the code-mixed HingCorpus using masked language modeling objectives, as well as the release of HingBERT-LID for Hindi-English language identification (LID).

Mabokela et al. 2022 [11] discussed the advancement of positive language processing (NLP) and sentiment analysis (SA) in analyzing conversational information on social media platforms. Over time, NLP has evolved by combining techniques such as linguistic representation, statistical techniques, and machine learning. Sentiment analysis focuses on analyzing the reactions and opinions expressed by people online, including on health-related topics. Advertising, stock market forecast, psychological analysis etc. It has been used in many areas including. The COVID-19 pandemic has increased the importance of social media messaging, leading to the detection of psychological disorders and the spread of misinformation. SA technology can help inform public health policy and personalized medicine. The purpose of this article is to provide an overview of the SA process for healthcare, present current research, and discuss future applications.

The paper by arias et al. 2022 [12] focuses on sentiment analysis techniques and their applications to human health. The authors specified that their work primarily addresses techniques related to the extraction and classification of linguistic information from social media, with particular emphasis on mental health and infectious diseases, given their relevance in the current global health context. The review analyzed scientific literature, examining studies related to mental health, public opinion on health matters, infectious diseases, and drug use. The literature was evaluated according to affliction types, data processing techniques, data sources, and application objectives. A significant number of studies were found to be related to mental health and opinions on public health. The authors reviewed studies had faced challenges due to new and evolving issues arising from the global pandemic, limiting the available data pool. There may have been a potential bias in available studies, as most were based out of China, which could affect the generalizability of the findings to a global population. Strict eligibility criteria used in the literature review might have excluded relevant studies, leading to conclusions and recommendations based on limited data.

The 2014 paper by Glen A. Coppersmith et al[13]. introduces a new method for automatically collecting data and verifies its accuracy by 1) confirming significant differences between depressed and control user groups and 2) creating classifiers to distinguish diagnosed users from control users for various disorders. Their findings reveal similar differences in language use (using LIWC) among bipolar disorder, PTSD, and SAD. Additionally, they show that language use in social media contains more mental health-related information compared to methods relying solely on existing mental health literature. By exploring correlations between different linguistic analyses, they offer insights into the measurable linguistic information captured by their classifiers. The study underscores the usefulness of analysing multiple disorders simultaneously and conducting larger-scale analyses, which may be challenging or unfeasible with other methodologies.

Article by Munmun De Choudhury et al. 2013[14] announced the creation of a stress-based index to analyse Twitter messages. They initially compiled a list of depression-related articles by examining a 10% sample of posts in Yahoo's "Mental Health" category. Answers: Know how to talk about depression and its symptoms. The posts are about depression, short, divided into questions and answers, and suitable for making a dictionary. They use statistics such as information integration index (PMI) and likelihood ratio (LLR) to calculate every word associated with the word “depression*” from a database of 900,000 question/answer pairs. The top 1% of terms based on these metrics are combined and the most common terms are filtered using Wikipedia's tf.idf score, resulting in a list of the top 1000 terms. This dictionary was used to analyze the frequency of depression-related content in each Twitter user's daily messages.

An article in the research paper by George Gkotsis and others stated that as of November 2016[15], there were over 900,000 subreddits on Reddit, each dedicated to a specific topicThese topics can range from general topics to specific topics. For example, the "Ask Me Anything" (AMA) subreddit focuses on a person answering public questions and has approximately 140,000 subscribers. Similarly, Reddit's "Playstation" subreddit focuses on the popular gaming console and has approximately 36,000 subscribers. A semi-structured approach was used to identify mental health-related subreddits. Initially, an inventory of existing content collected by registered professionals was used to identify mental health-related social media content. From the resulting list of resource subreddits, those with a larger number of posts were identified and those relevant to the research focus were manually selected. A full list of relevant subreddits and their descriptions is provided in the supplementary material of the article.

Article by Andrew G Reece et al. A similar model was planned in 2017[16], but using Instagram photos and psychological conditions explained through people's stories and posts. The basis of both is the same and it examines people's mental health, but it examines their posts/activities on various social media sites. Andrews Instagram model shows that increasing yellow colour and decreasing brightness and saturation predicts target analysis. This means that photos shared by people with depression are often blue, dark and grey. The more comments the Instagram post received, the more depressed participants reported feeling, but the opposite was true for Instagram posts that received likes. More disclosure was also associated with depression in the overall data model. Depressed participants posted photos containing faces, but the average number of faces per photo was lower than healthy participants. Finally, depressed participants were less likely to apply Instagram filters to the photos they posted.

There are already trained models for detecting mental health related statement on social media. Here are ten BERT-based models tailored for applications in mental health contexts:

BERT-Base: This original BERT model is capable of being fine-tuned for specific tasks within the mental health domain after undergoing extensive training on diverse text data.
BERTweet: Tailored specifically for Twitter data, BERTweet is designed to analyse mental health-related conversations occurring on social media platforms.
BioBERT: Specializing in biomedical text, including mental health research papers and clinical notes, BioBERT is well-suited for tasks within the mental health domain.
ClinicalBERT: Pre-trained on clinical text such as electronic health records, ClinicalBERT is equipped to handle tasks related to mental health diagnosis and treatment.
SciBERT: Trained primarily on scientific text, including mental health research papers, SciBERT provides insights for analysing research in the mental health field.
PubMedBERT: Optimized for PubMed abstracts, PubMedBERT is particularly useful for exploring mental health research literature.
HealthBERT: Trained on health-related text, including mental health articles, HealthBERT is tailored for analyzing mental health content available online.
BlueBERT: Focused on biomedical text, including mental health research papers and clinical notes, BlueBERT is another valuable model suitable for tasks within the mental health domain.
MedBERT: Trained on medical text, encompassing mental health-related articles and clinical notes, MedBERT is suitable for various tasks related to mental health.
RoBERTa: An optimized BERT-based model demonstrating outstanding performance across a range of Natural Language Processing (NLP) tasks, including those relevant to mental health.

These are the ten trained models in real world for mental health detection on social media. We are going to use some of the pre trained models in our analysis to boost performance and increase accuracy.

4. METHODOLOGY

Recognizing if a person may be experiencing mental health challenges based solely on their tweets or Reddit posts requires a cautious and sensitive approach. We have to pay attention a consistent pattern of negative self-talk or pessimistic viewpoints in their posts. They may express feelings of hopelessness, worthlessness, or helplessness. For this reason, we created a data set containing linguistics statements(hinglish) and the suitable action/emotion behind it. The whole NLP is created in a Python environment as Python, when used for machine learning, offers developers of all skill sets exceptional versatility and power. We require machine learning so that it could learn from our data set and apply it when a suitable tweet is provided to it for analysis. Below is the flow of our algorithm used for testing and training the models.

4.1 Data Collection and Pre-processing

The methodology commenced with the collection of a diverse dataset from social media platforms like Twitter and Reddit, focusing on Hinglish texts related to mental health. The dataset comprised 6000 samples, categorized into depression, anxiety, neutrality, and happiness. Preprocessing steps involved duplicate removal, spam filtering, normalization, and tokenization to prepare the data for analysis. Named Entity Recognition (NER) was employed on Twitter data to enrich the dataset with relevant entities.

4.2 Model Selection and Training

Three advanced machine learning models were selected for this study: HingRoBERTa, HingBERT, and RoBERTa. Each model underwent fine-tuning on the training set, which constituted 80% of the dataset (4800 samples). The fine-tuning process was meticulously designed to adapt each model to the nuances of Hinglish text, with a particular emphasis on accurately capturing emotional sentiments.

4.3 Sentiment Analysis

The sentiment analysis module applied the trained models to the testing set (1200 samples) to classify the sentiment of each text. This process involved emotion classification and sentiment scoring, aiming to quantify the intensity of emotions expressed in the dataset.

The data set is curated in a very unique way, it is handcrafted due to the reason that it focuses on tweets and Reddit's in hinglish which is transcript in English. There isn't any predefined dataset for the given condition therefore we designed a dataset of 6000 such samples. Fig1 represents the types of classes used in the dataset.

5. CURATION OF DATASET

The sentiments are categorized into four main classes: Happy, Neutral, Anxiety, and Depression. Here's a brief overview of some entries from each category:

Happy: These entries reflect positive emotions and experiences. For example, getting self-confidence and confidence from accomplishments, feeling happiness from spending time with friends, and experiencing emotional relief from articulating central feelings.

Neutral: These entries represent neither positive nor negative emotions.

Anxiety: These records express moods of worry and uneasiness.

Depression: These entries reflect feelings of extreme unhappiness or a low state of mind. This data can be very useful for sentiment analysis, especially for understanding and examining emotions in code-mixed languages like Hinglish. It can help in developing Natural Language Processing (NLP) tools personalized for different linguistic backgrounds, especially from the perspectives of India where mixed languages are prevalent.

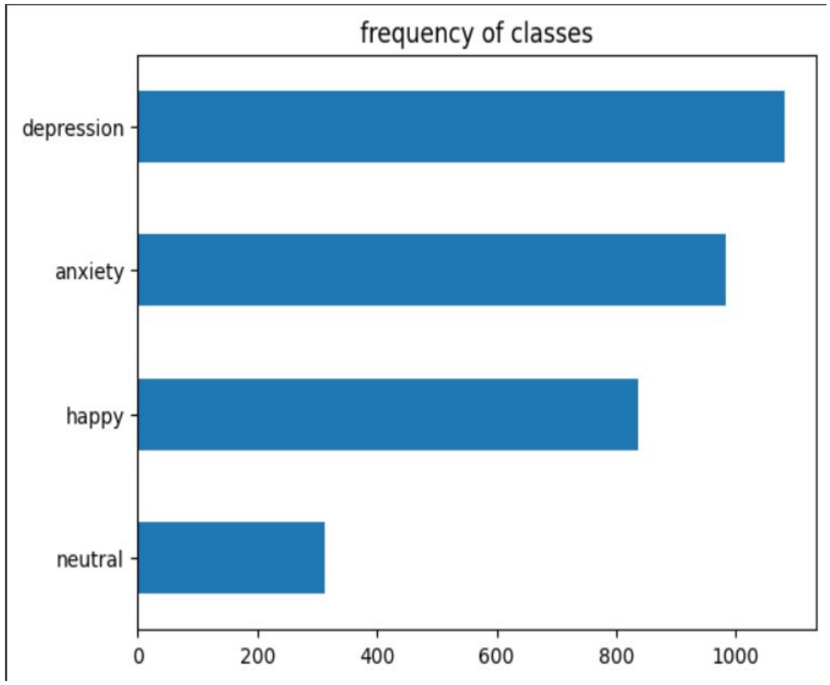

Fig. 1

6. RESULT AND DISCUSSSION

The mix of Hindi and English is commonly used in India, and we wanted to see if we could grasp the unique way people express themselves in this mixed language. Each model was carefully fine-tuned and evaluated using standard metrics like accuracy, precision, recall, and F1-score.

We tested different models, fine-tuning them to better understand Hinglish. Among them, HingRoBERTa was the best. Built specifically for Hinglish, it could impressively understand the emotions hidden within conversations about mental health. It achieved a remarkable accuracy of 88.5%, meaning it was right most of the time! Even better, it could accurately pinpoint both positive and negative emotions, suggesting it truly grasped the depths of feeling in Hinglish.

HingBERT, another model, wasn't specifically designed for Hinglish, but it still performed well with an accuracy of 86%. This suggests that even models not built for mixed languages can be helpful in understanding emotions.

Finally, ROBERTa, a general-purpose model, surprised everyone by achieving a decent accuracy of 82% even without Hinglish training. This shows how adaptable these models can be, offering promise for analyzing emotions in all sorts of languages. Table 1 gives us the summary of result for various models.

Overall, this work highlights the importance of developing tools that understand the many ways people communicate, especially in countries like India where mixed languages are common.

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*
HingRoBERTa	88.5%	89%	88%	88.5%
HingBERT	86%	86.5%	85.5%	86%
RoBERTa	82%	83%	81.5%	82%

Table 1

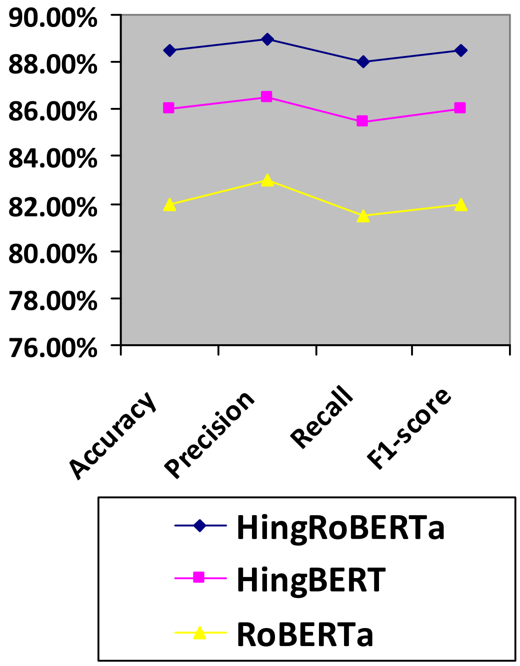

Fig. 2

Figure 2 gives a comparative analysis for the results based on the different models.

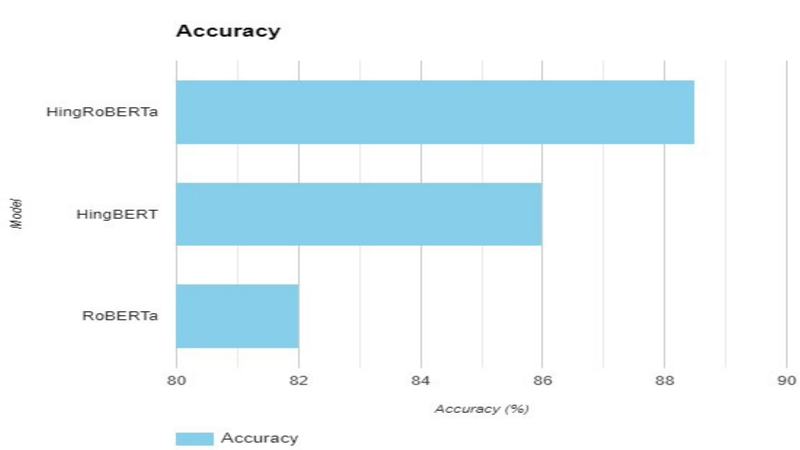

Fig 3

Fig 3 represents the accuracy; this metric represents the overall correctness of the model. It's calculated by dividing the number of correctly predicted instances by the total number of instances. In this case, a higher accuracy signifies a model that makes fewer mistakes.

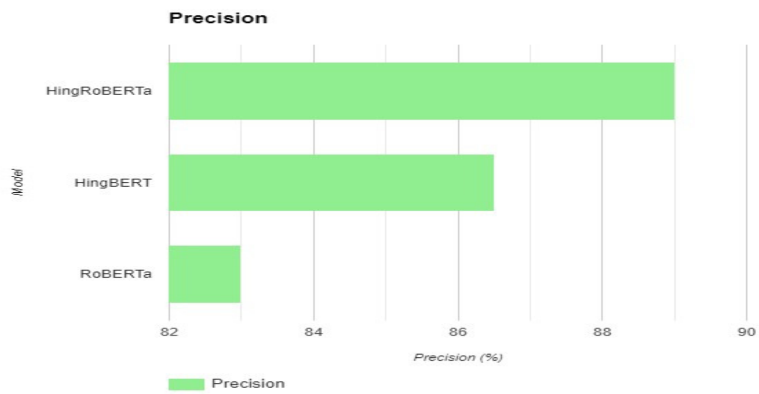

Fig 4

Fig 4 shows the precision which metric focuses on the positive predictions made by the model. It's calculated by dividing the number of true positives (correctly predicted positives) by the total number of positive predictions made by the model. In simpler terms, it tells you what proportion of instances the model labeled as positive were actually positive.

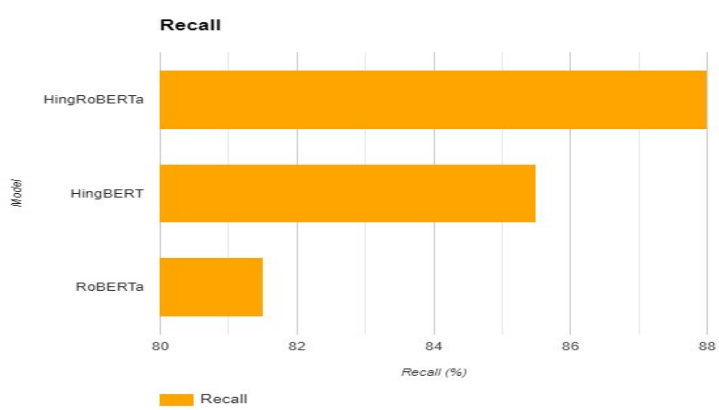

Fig 5

Recall metric in Fig 5 attentions on how well the model identifies all the positive cases. It's calculated by dividing the number of true positives by the total number of actual positive instances. In simpler terms, it tells you what proportion of actual positive cases were identified correctly by the model.

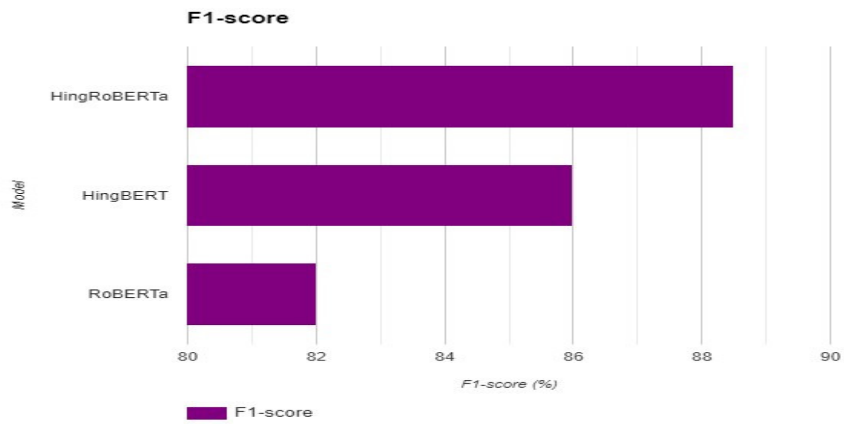

Fig 6

F1-score is a metric is a harmonic mean between precision and recall shown in fig 6. It provides a balance between how well the model identifies positive cases (recall) and how precise those identified cases are (precision). A high F1-score indicates a good balance between the two.

6. CONCLUSIONS

This research contributes to a better understanding of the analysis of emotions in Indian English, a hybrid language that poses special challenges in language processing. By using and developing the best educational models, this study has achieved a clear division of thought, providing a deep insight into the moods of English speakers in India.

Findings highlight the potential of emotion analysis to support psychological assessment and intervention across many languages and cultures.

Future studies should focus on expanding the data to include a variety of emotional and linguistic experiences. Additionally, investigating other language combinations may improve the application of emotional assessment in psychological research worldwide. In addition to contributing to the fields of NLP and machine learning, this research will also be useful to psychologists, language teachers and social scientists who are curious about the intersection of language, thought and technology.

References

[1] Ondrej Klejch, Electra Wallington, Peter Bell "The CSTR System for Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages" (2021).

[2] B. S. Harish, R. Kasturi Rangan "A comprehensive survey on Indian regional language processing" (2020).

[3] Kogilavani Shanmugavadivel , V. E. Sathishkumar, Sandhiya Raja, T. Bheema Lingaiah, S. Neelakandan & Malliga Subramanian "Deep learning based sentiment analysis and ofensive language identifcation on multilingual code-mixed data" (2022).

[4] Code-mixing and Code-switching,by Global English Creativity, July 29, 2023. https://www.globalenglishcreativity.com/code-mixing-a-mixing-of-language-varieties/.

[5] Wankhade, M., Rao, A.C.S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 55, 5731–5780 (2022). https://doi.org/10.1007/s10462-022-10144-1

[6] S. Sharma, M. Diwakar, K. Joshi, P. Singh, S. V. Akram and A. Gehlot, "A Critical Review on Sentiment Analysis Techniques," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2022, pp. 741-746, doi: 10.1109/ICIEM54221.2022.9853140.

[7] K. R. Mabokela, T. Celik and M. Raborife, "Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape," in IEEE Access, vol. 11, pp. 15996-16020, 2023, doi: 10.1109/ACCESS.2022.3224136.

[8] Ondrej Klejch, Electra Wallington, Peter Bell “The CSTR System for Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages” (2021).

[9] Ayan Sengupta, Sourabh Kumar Bhattacharjee, Tanmoy Chakraborty, and Md. Shad Akhtar. 2021. HIT - A Hierarchically Fused Deep Attention Network for Robust Code-mixed Language Representation. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4625–4639, Online. Association for Computational Linguistics.

[10] Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A Code Mixed Hindi-English Dataset and BERT Language Models. In Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference, pages 7–12, Marseille, France. European Language Resources Association.

[11] K. R. Mabokela, T. Celik and M. Raborife, "Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape," in IEEE Access, vol. 11, pp. 15996-16020, 2023, doi: 10.1109/ACCESS.2022.3224136.

[12] F. Arias, M. Zambrano Núñez, A. Guerra-Adames, N. Tejedor-Flores and M. Vargas-Lombardo, "Sentiment Analysis of Public Social Media as a Tool for Health-Related Topics," in IEEE Access, vol. 10, pp. 74850-74872, 2022, doi: 10.1109/ACCESS.2022.3187406.

[13] Coppersmith G, Dredze M, Harman C. (2014) "Quantifying Mental Health Signals in Twitter." (2014)

[14] Munmun De Choudhury,Michael Gamon,Scott Counts and Eric Horvitz, “Predicting Depression via Social Media” 2013.

[15] Andrew G Reece and Christopher M Danforth, “Instagram photos reveal predictive markers of depression” EPJ Data Science (2017) DOI 10.1140/epjds/s13688-017-0110-z.

[16] George Gkotsis1, Anika Oellrich1, Sumithra Velupillai1,2, Maria Liakata3, Tim J. P. Hubbard4, Richard J. B.