JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Analysis of Moods in a Multilingual Context using Deep Learning: Comparative Analysis

Dr. Deepak Kumar,

Asst. Professor, Shri Shankaracharya Professional University, Bhilai, CG, India

Abstract

This research paper investigates the efficacy of various machine learning models, including deep learning and hybrid models, for text classification in the English and Hindi languages. The study focuses on sentiment analysis of comments from a popular Hindi e-commerce site, "DARAZ," which comprises both Hindi and translated English reviews. Our study is a comparative analysis of various models and evaluates its effectiveness in the field Mood analysis. The research method includes implementations of seven learning machines, deep learning models and models such as long short-term memory (LSTM), LSTM Bipartite (bi-MSTN), Slue 1d (conv1d) and Conv1d-MSTM combination. Pre-treatment method applies to installed modification text to increase model accuracy. The vector machine support model shows superior performance compared to other models, 82.56%, 86.43% accuracy for analysis of English text mood Analyze Hindi's text mood using Stemming Porter Algorithm. Furthermore, bi-The LSTM-based model shows the best performance of any deep learning model. It reaches 78.10% in English text and 83.72% in Hindi text. This study implies significant advances in natural language therapy research; In particular, in the case of Hindi, the improved model improvement and text classification methodology. The results of this study make a significant contribution to the field of mood analysis and sentences. Study is a precious idea for future research and practical applications.

Keyword: Sentiment analysis, Hindi text sentiment, Data preprocessing, Tokenization.

1. Introduction

The e-commerce platform is a major option adopted by businesses, offering a diverse range of lines online Possibility of purchase and sale. These platforms allow consumers to shop without having to visit physical stores. A regular website that is often used to collect information [1]. One of these e-commerce websites, Daraz, is popular online South Asian shopping markets including India, Sri Lanka, Pakistan, Myanmar and Nepal. This will provide a wide range of sets Products with over 60 million items in various categories, including home appliances, home needs, beauty Products, fashionable items, products, sports goods. The Gmail has caused noticeable growth on the Internet Purchase, how the state used the resident's directives. As a result of the widespread closure of retail trade and fears about the transmission of the Gmail, online purchases have become a dominant method for consumers to meet their needs for consumption [2]. Hindi is a wide speaking language in India and in many other countries of the world. As a result, Hindi natural language processing (BNLP) has garnered considerable attention in the field of NLP [3]. Text classification stands as one of the foundational challenges in the field of NLP [4]. However, despite the large number of e-commerce sites with comments sections by allowing the expression of opinions in Hindi, little research has been conducted on the analysis of Hindi sentiment. I'm worried. Meanwhile, English is widely used in a variety of industries, including search engines, social media, and customers service. This makes it an important language for NLP support. It is also the main language of many popular online platforms such as Google. Facebook and Wikipedia generate a huge amount of data in English. What's more, it's a common language adopted in science. Publications have become the focus of NLP applications in scientific research and knowledge extraction [5,6]. Mood analysis is an important factor in assessing opinions on subjects such as product policy, sports, finance, and products. How People are subjective and opinions are valuable. For example, the commerce site is filled with a variety of product and manufacturing perspectives. It is essential to detect which statements are positive or negative [7]. Using machines and deep learning algorithms, NLP techniques, it is now possible to identify cases of cyberbullying and distinguish between intimidation and non-control Instructions [8]. Automated learning has recently become an effective approach to processing data and calculations. It provides intelligent features in a variety of applications [9]. Algorithm statistics, probability theories, and Optimization techniques that draw data conclusions and identify models on large, unstructured data sets [10]. The potential uses of these algorithms are numerous, including automatic text categorization [11], machine learning for enhanced data in dermatological image recognition [13], machine learning for speech processing [14], improving medical diagnosis accuracy with causal machine learning [15], statistical arbitrage in crypto currency markets using machine learning [16], and classifying fake news [17], among others. Moreover, deep learning algorithms are becoming increasingly significant in different research areas, including but not limited to binary classification-supported multi-class skin lesion classification [18], and analysis of cellular images [19]. The content generated by users (UGC) has become a precious source to understand consumer emotions and Experiences concerning online retail services. Text extraction methods can be used to detect various facets in online content, Including information on products, retailers' promotions, delivery services, payment procedures, communication, return / reimbursement policies, and price details [20]. Deep learning technology, ongoing advances in convolutional neural networks (CNNs) Long-term long-term memory (LSTM) networks are increasing as two of the most influential neural network architectures [21]. With the use of The CNN model, which is the adjustment and salience map of hyper parameters, can be constructed directly from the infrared spectrum [22]. Text classification It can be performed using automated and detailed learning methods such as the BI-MSTM model and the CNN model [23]. CNN-MSTM-TTTNION (CCLA) model [24] and regional structure tree model CNN-MSTM [25] for text mood Mood classification and analysis. Therefore, these methods can help you analyze and understand user-generated content in consumer perception and experience.

Based on our literature review, it is obvious that studies on automatic learning and in -depth learning algorithms for the text The analysis was mainly in English. However, there is a notable gap in a study linked to Hindi. This gap in knowledge considerably prevents achievements in fields such as mood analysis, NLP and text Hindi classification. Consequently, it is extremely important to expand research efforts to analyze Hindi text data [26-28] and study the potential of using automatic learning and in -depth learning methodologies in this context. To fill this gap, our article aims to make the following contributions.

1 and 1 Dataset Development: Collected a complete dataset for analyzing Hindi mood and collected Hindi magazines and their proper English translation. Additionally, we performed cleaning and pretreatment data to ensure a set of data Analysis was accepted.

2. Comparative analysis: we have carried out a comparative analysis, which includes traditional automatic learning models, such as logistics Regression (LR), a solutions solution (DT), a random forest (RF), naive bays (NB), a neighbor K-Grand (KNN), a vector of support Classifier (SVM), stochastic gradient (SGD) gradient, as well as models based on deep learning. We evaluate each performance This model analyzed a variety of indicators, including accuracy, accuracy, review, F1 indicators, and losses from the confusion matrix. Additionally, we studied the maximum, minimum and average lengths of Hindi and English

text comments. This analysis Provides ideas for the effectiveness of traditional and deep teaching methods for classifying texts in Hindi and English comment

3 Neuron Network Architecture: We designed four unique neural network architectures: Short-term memory (LSTM), bi-directional LSTM (BI-MSTM), a convolutional neuron network with single-dimensional convolutional layer (CONV1D) and hybrid an architecture that combines CONV1D with the LSTM layer (CONV1D-MSTM). These architectures were developed to explore them the capacity of modeling sequential data, especially in the context of Hindi text analysis.

2. Literature Review

Many universities and commercial researchers are currently studying and investigating the analysis of emotions [29–32]. In this literary classification of mood using a structure based on BI-MSTM, which includes the attention mechanism and a layer of bundle. In the same way, in article [34], a model based on CNN was presented for the classification of moods. In paper [35], two-layer bidirectional LSTM A network was proposed in combination with a complex sentiment analysis unit. In contrast, Du et al. [36] suggested that CNN is feasible A model that extracts textual attention and leads a certain number of experiences. Kim et al. [37] Complete number completed Experiments on single-layer convolutional neural networks. Chatterjee et al. [38] Two components have been applied - Content and general polarity Mood content - to accurately represent the emotions seen in textual data. In the financial field of Nelson et al. [39] Recommended Using LSTM networks with technical analytic indicators to predict action courses. In paper [40], the polarity of emotions is Important signs of customer well-being, helping businesses to better understand their customers. From the perspective of classification approach, Zhou et al. [41] provided a standard approach for classifying mixed texts for CNN and LSTM. Alhawarat et al. [42] developed a TF-DF extraction of Arabic text classification system and features using CNNS reaches 98.89% accuracy. By learning a language, Chowdhury et al. [43] Mood analysis of 4000 films manually translated in bangles, reaching 82.42% of accuracy with LSTM. Chakraborty et al. [44] predicts text and diagnostic capabilities using automated learning, and Zhang et al. [45] We presented a CNN for text classification in terms of characters with significantly improved accuracy. Paper [46] is a unique way The use of discriminatory confidentiality has been proposed to improve the predictive capabilities of LSTM models of actions. The paper [47] is complete Comparative analysis of mood classification in Hindi News comments using traditional SVM and detailed learning (LSTM and and) CNN) Algorithm. The paper [48] compared the productivity of neural networks as a function of the inverse distribution of text classification. Comparing with other controlled automated learning models. In general, these studies provide valuable information about development, using various models based on neural networks (NNs) for text mood analysis and classification.

3. Data and Methodology

The methodology utilized in this research comprises multiple phases, including data collection, data preprocessing, model selection, statistical assessment, and implementation, all of which are depicted in Fig. 1

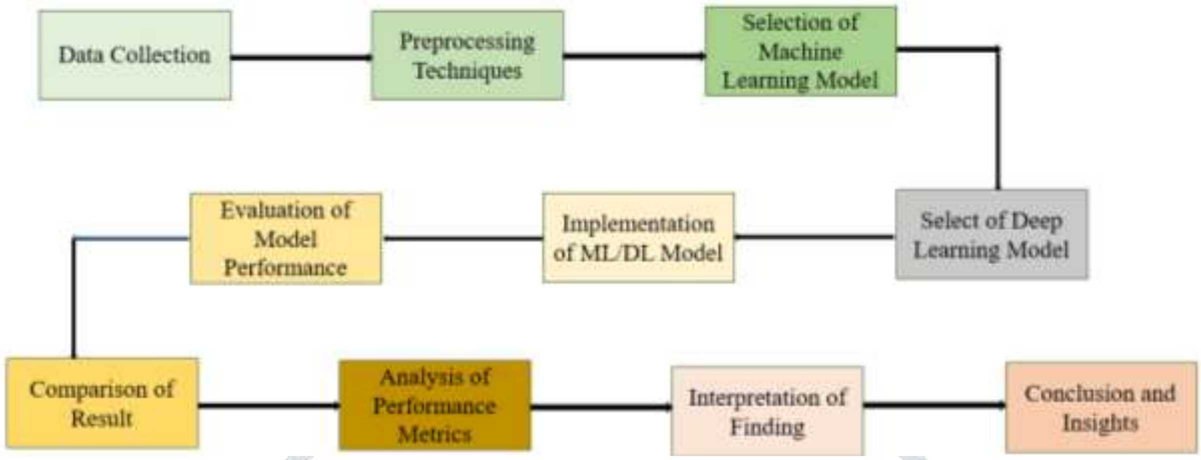

Fig 1: Proposed text classification methodology

3.1. Dataset preparation

The preparation of a dataset for Natural Language Processing (NLP) typically involves several stages, including data collection and curation, classification into distinct categories, pre-processing and cleaning, as well as splitting the data into training, annotation, balancing, validation, and test sets. This process entails organizing and cleaning the data to make it suitable for use in NLP tasks. The next step is to select an appropriate algorithm for the task to be accomplished. Model performance is evaluated using statistics The analysis, and the final step consists in deploying the model and monitoring its performances in parameters of the real world. It should be noted, data set the preparation process includes two key steps, which are data collection and preliminary data processing.

3.2. Data Collection

We have collected a set of data made up of 2577 journals in Hindi, as well as their corresponding English translations and class Labels. The dataset was assembled from the Daraz E-Commerce website using the network scratch method. Opinions regarding dataset coverage Positive and negative comments were carefully supervised for the purposes of this study. Table 1 is shown Examples of texts collected to analyze moods in both ball and English. After data collection, the dataset was subject to manual annotation to ensure class accuracy and consistency. after that, Data is pre-converted to remove information that is not linked to the case, fix errors, and ensure compatibility with nature Speech Therapy Model (NLP). This stage of preprocessing is required to improve and improve the quality of the dataset Productivity of NLP models under training and evaluation. As shown in Fig. 2, a set of data for English and the text of the Bange contained A total of 1439 positive comments and 1138 negative comments.

3.3. Preprocessing data

Data preprocessing plays a key role in data preparation for natural language processing tasks (NLP). NLP uses preliminary data processing Delete irritability information, remove correct errors, or use text to words, sentences, or Supplements, stems, and lemmatization are the process of breaking words towards the simplest components, identifying and deleting Data points without data presentation and standardization to ensure consistency. These steps help improve precision and efficiency NLP algorithms. In this study, we present our contribution to the pre -treatment of data on the comments of English and Hindi, as illustrated in the figure. 3. The pre -treatment steps involved in our study include cleaning the text to eliminate relevant information and errors, tokenization

The text in words and sentences, by jumping the words to their basic form, by identifying and deleting all non - representative data points and normalize the data to ensure consistency.

3.3.1. Transform

The normalization of the text is a general rhythm of preliminary treatment in the analysis of moods for the English text, in which all the characters are converted into tiny. However, for Hindi's text, this process is not necessary. In our study, we have transformed all words into the lower register of English Language in order to simplify data and facilitate the processing and analysis of moods using automatic learning models. Capitalization options C Tunneling data analysis sets can prevent precise identification of mood using algorithms. Lower hut eliminates such variations and generates a more coherent data set. In addition, the reduction in the number of unique tokens through this process leads to an increase Efficiency of calculating the analysis and formation of models, resulting in a more effective analysis of feelings. Reducing variations in Pre-processing steps such as text and improvements in data consistency and low. Identify and classify moods accurately.

3.3.2. Delete double lines

Deleting a double line is an important step in preprocessing data for NLP tasks. Because it can cause inaccuracies in the model Increased processing time. In this study, we removed replicas from lines in English and Hindi datasets. This process for example, double line decision by choosing a strategy for deletion, save the first copy or with the best copy Evaluating confidence and strategy implementation using language language or tools. By eliminating the double line, we tried Improves the accuracy of the NLP model and reduces treatment time. This step contributes to the quality and overall dataset Efficacy of emotional analysis. Therefore, to ensure reliability, we removed the English dataset and line duplication Results obtained from the analysis.

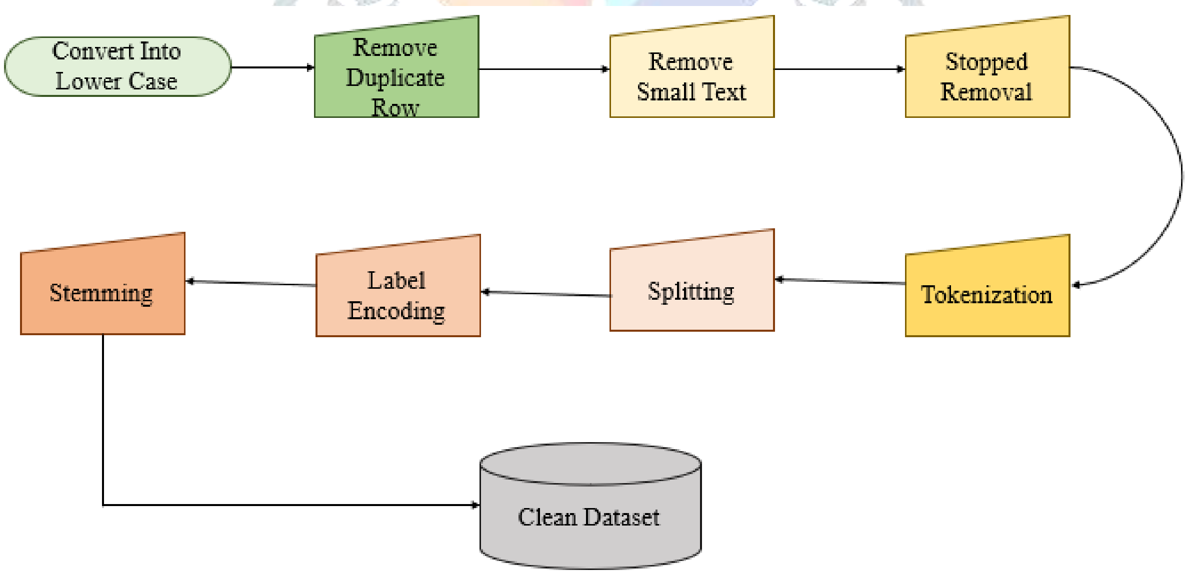

Fig 2: Data Preprocessing Setup

3.3.3.3. Delete small text

Filtering small text or text less than a certain length is an important step in data before processing an NLP task. Such text, Individual words and short sentences can be removed because there is not enough information to improve quality Dataset. This process involves determining the minimum length of text stored in the dataset and identifying text below the threshold. Using programming languages or software tools for removal. This approach can improve accuracy It can improve the quality of NLP models and datasets. In our study, we determined the duration of two English sentences. and Hindi. We then provided the minimum length of text stored in the dataset. After cleaning the small text, I deleted the small text Conversations for English offers and small conversations for Hindi proposals.

3.3.4. Retracting the term stop

Deleting a stop is often used by preprocessing techniques for NLP tasks. stop words are words such as etc. It does not allow specific meanings and can be removed to simplify datasets and accelerate processing. This process includes stop identification Words that determine your deletion strategy, such as removing everything or using programming, the most common one's A language or software tool for implementing your strategy. This makes a set of data less and process more quickly, potentially improve the accuracy of NLP models. For all English data, we have erased words such as - "o", "not", "through", "after", "still", "yours", "I", "I", "Not"}, etc. For Hindi datasets, I erased the word "how" - {" यह", "इकान", "उद्देश्य", "आपका", ", "ईज़", "मेन", "किर "} etc.

3.3.5. Tokenization

NLP tokenization is an important process of involving splitting of text in a clear word or sentence called tokens. The tokenization process is a critical step in preprocessing data for NLP tasks such as analysis of sentiment, classification text, and machine translation.

Tokenization helps normalize text data by breaking it down into small units that are easy to process and analyze. With us I used the Kera's talk Naser class to tokenize the code, statement. The main stages of the token are as follows: Define the number of functions (words) to consider, a copy of the tokenizer class, and the maximum number of words that need to be taken into consideration, an entry that shows a separation symbol in the form of space, adjusts the Tinkerer to text data, converts the text statement into a sequence of integers Sequence to ensure uniform length. These steps allowed us to convert text into digital ideas. This is handled by our detailed learning model.

3.3.6. Special deletion of punctuation, characters and numbers

The dataset is: It has been converted to remove punctuation marks such as: ,? ! Special characters such as @, #, $, %, ^, and *. So are numbers that are not important for mood analysis.

3.3.7. Split

Splitting is a fundamental method of NLP that involves splitting text into small, meaningful units for analysis. That's important Preprocessing steps for NLP tasks to enable processing of text data with machine learning models. Text splitting can be performed at various levels, such as the word, sentence, or paragraph level. At the word level, the text is divided into individual words, also known as tokens, by a process called tokenization.

3.3.8. Categorical coding

Categorical coding is a process in NLP where categorical variables are converted into digital values which can be used as entries for automatic learning models. The categorical variables are those which have a limited number of possible values, such as words in a vocabulary or elements on a menu. Categorical coding enters two forms: coding and embedding the label. Coding tags is a way The transformation of categorical data into a numerical format by prescribing integer values of each unique category. In our study, the coding of the label Provides categorical data transformation that cannot be processed directly, in a digital form. The brand's coding is assigned Each category is a single whole value. Overall values are assigned unclassified or ordered between categories. Meanwhile, it integrates methods used in NLP and auto-learning to convert data into density and low vectors. For NLP, words or phrases are transformed into word embeddings, which are high-dimensional vectors that depict their semantic significance and the relationships between words. Word embedding methods rely on the principle that words with similar meanings tend to be used in similar contexts. For example, in a set of English comments, "good" and "best" are probably similar So there is a context, and therefore a similar value. For example, in our dataset, comments such as Hindi, and It is found in the appropriate context and therefore has an appropriate value. The word incorporation of these words is similar, The relationship between them. Integration improves the performance of the NLP model by encapsulating words and word meanings in words Density and low-dimensional form.

3.3.9. Overall

STEM is a widely used technique in NLP, simplifying basic shape words. Its main purpose is to simplify the text A data set that makes analysis easier. In our study, we applied Lancaster algorithms and bring to our models. The purpose of the following to convert the words to their general basic form. For example, in English texts, words such as \ "Recommended, \" \"recommended, \" and \"recommended via stem in \"recommended\". Similarly, in Hindi texts. It involves extracting words, applying the chosen algorithm, and storing the stemmed words for analysis. Stemming improves NLP model accuracy but can sacrifice information and interpretability. We looked into this compromise when using stemming algorithms in our models.

3.5.2. Vectorization or distribution of data

The Vectizer count, part of the handy Python Scikit-Learn toolbox, takes into account the frequency of each word in the text Converts a statement to a vector. The size of the n-grams used can be specified using the ngram_range parameter. As an illustration, the installation of the value of 1, 1 will be carried out by unigrams (n-grams made up of a word), while the value of 1-3 would give an N-Gram One to three words.

LANGUAGES.

• Unigram: Setting n = 1 in the n-grams function generates unigrams or 1-g, allowing for word frequency calculations.

• Bigram: Configuring n = 2 in the n-grams function generates bigrams or 2-g, facilitating word frequency calculations.

• Trigram: indication n = 3 in the N-grams functions generates trigrams or 3-G, which allows calculations of the frequency of words.

3.6 DNN -based models

DNN, which means a deep neural network, is a category of artificial neural networks characterized by several levels Including hidden layers, in addition to the layers for input and output. Networks with these concealed layers enable the learning and representation of intricate data patterns and relationships, rendering DNNs valuable for tasks like high-dimensional data involving image and speech recognition, natural language processing, and more. In the current study, we designed and implemented four unique NN models, namely LSTM, Bi-LSTM, Convolutional Neural Network with one-dimensional convolutional layers (Conv1D), and a hybrid architecture comprising Conv1D and LSTM layers (Conv1D-LSTM). The experimental installations of four depths Neural Network Model (DNN).

3.6.1. LSTM (Long Short-Term Memory)

The proposed LSTM model is a sequential model, commonly used for sequential data processing such as time series, speech, and text. It consists of the following layers: Build - Layer: This layer converts input functions into fixed, dense vectors Allows the size (constructed with -in Dim) and models to study ideas specific to the input data. Input length parameters are set.It consists of neurons and abandonment rate of 0.2, normalizes the network and prevents overregulation. Repeated dropout the parameter is set to 0.4, indicating that abandonment applies to recurring connections in LSTM cells. Dark layer: Completely Connected layers with 256 units and Softmax activation functions. This layer helps to convert LSTM output signals to format Suitable for the final classification task. Output Layer: Another dense layer with two units and a sigmoid activation function. This layer It produces the ultimate probability of conclusions for binary classification tasks. Models are edited with cross-entropy binary loss Metric optimizer for features, Adam, and precision. The model summary shows a total of 82,178 parameters Model All this is taught.

3.6.2. BI-MSTM (Short-term Control Double Memory)

The BI-MSTM model uses two LSTM models to process neuronal repetitive networks, front input sequences and the other was delayed to capture the last and future context shown in the proposed Bi-ALSTM model is its coherent model It contains the following layers: Layer Build - In: Converts input functions into dense vectors of sizes built for learning Specific to the input data. Dicerium LSTM Level: Process both contiguous data using double detained LSTM cells Front and back direction. 64 units/neurons, dropout rate for normalization is 0.2, recurrence dropout rate is 0.4 For dropouts for recurring connections. Dark Layer: Fully connected layer with 256 units and Softmax activation, Output of the bidirectional LSTM layer of classification tasks. Output layer: another dense layer with 2 units and sigmoid activation, Creating the final output probabilities for binary classification. The model is compiled with binary loss of cross entropy, Adam Optimizer and accuracy metric. There are a total of 131,586 preparation parameters.

3.6.3. CNN (Neural Network Business)

Neuronal Networks (CNNs) are a type of neural network commonly used in image recognition and computer vision. task. They consist of several layers of interconnected knots; each knot operates the package with sagging Input data. The proposed CNN model is shown in the diagram. 10 consists of the following layers: layer built -in: this layer converts the input The characteristics of dense size vectors are built -In DIM to capture significant representations specific to input data Conv1D layer: this

The level performs the operation of the input data using 128 filters, each with a size of 5. The activation function used is Relu, which introduces nonlinearity into the network. GlobalMaxpooling1d -layer: This layer gets the maximum value via a temporary value Measure curled functions, reduce data dimensions. Dark Layer: 256 units of fully connected layers and Activation Soft Max. Turns the join functions into a format suitable for the ultimate classification task. Layout: Another dense layer the layer with two units and sigmoid activation creates the ultimate probability of output data for binary classification. The model is compiled Cross-entropy, binary losses for Adam Optimizer and Precision Metrics. There are a total of 106,626 preparation parameters.

3.6.4. con1d-mstm hybrid

CNNs and LSTMs are often used in conjunction with other DNN types to increase productivity for a variety of tasks. A The architecture shown in the diagram. 11 is commonly employed in tasks that involve analyzing sequential data, such as natural language processing and speech recognition. The proposed Hybrid Conv1D-LSTM model combines Conv1D and LSTM layers to process sequential data. The architecture is as follows: Embedding Layer: Converts input features into dense vectors of size embed dim to capture meaningful representations specific to the input data. Conv1D Layer: Performs a 1D convolution operation on the input sequence using filters and a kernel size. The reread activation function. Maxpooling1d layer: reduces the dimension Functions fulfilled, taking the maximum value in the temporary dimension. LSTM layer: treats the functions of Conv1D A layer using LSTM cells with 64 units. It includes the regulation of the decline at speed of 0.2 and a recurring murder speed 0.4.

Thick Layer: Fully connected layers with 256 units and SoftMax activation, convert LSTM output signals to format Suitable for classification problems. Output Layer: Another dense layer with 2 units and sigmoid activation, the final output is created Probability of binary classification. This model summarizes entropy binary loss, Adam Optimizer, and precision metrics. He There is a total of 139,650 training parameters.

3.7 Experimental Installation

This experiment was conducted using the Google Cola platform with the aim of teaching training models. This platform provides unlimited access to high-performance graphics units (graphic processors) with minimal configuration requirements. For evaluation Model performance, train test separation was used. In particular, splits are used in 80-20, highlighting 80% of the data Test the diagram of the diagram for the sample for model formation and the remaining 20% sample. 12. This division provides an unbiased assessment Capabilities of models. During the education process, key hyperparameters were set up as follows: The number of ERAs is set to 50 years. Lot 32 size for effective optimization. The level of the word was set by 1 to give an idea about the progression of learning. Training updates and display performance indicators.

4. Results and Discussion

The "Results and Discussion" section provides a complete overview of the empirical results and their corresponding interpretations. This section presents a detailed analysis of the received data, covering the appropriate assessment of the effectiveness Such metrics as accuracy, accuracy, review and indicator F1. In addition, the study of the influence carried out by various parameters and hyperparameters on the performance of the model are undertaken. It should be noted that a comparative evaluation is carried out, contrasting the results obtained from traditional automatic learning models against those who carry out deep training models. Contain a comparison of various models used for natural language treatment tasks. It includes the model names, stemming algorithm used (Lancaster or Porter), and the performance metrics: accuracy, precision, recall, and F1 score for English and Hindi text respectively. In the field of English text analysis, a comparison was made between different models using a dataset. The model that demonstrated the highest accuracy was the SVM (vector machine support) algorithm combined with the carrier. The algorithm reaches an impressive accuracy of 82.56%. SVM works by identifying the best hyperplane to separate data into Make the classes separate, while maximizing the margins between classes. Meanwhile, the KNN model (k-nearest des violins) The Lancaster Stemming algorithm shows the lowest performance with an accuracy of 73.97%. KNN gives a name to the course label Copy based on most votes from my recent neighbors. A similar evaluation was performed for the analysis of Hindi texts. Of the models tested, SVM stood out as the most efficient using Porter and Lancaster. Stemming algorithm. He reached an impressive accuracy of 86.43%. He reached an impressive precision rate of 86.43%. Meanwhile, whatever the rod algorithm used (Porter or Lancaster), the KNN model lags behind in terms of performance, reaching 77.91% accuracy. Table 7 shows performance assessments of different learning models to analyze English texts. Among the models evaluated, there were great artists and models with the worst performances. Best model from a general perspective Performance was a model-based model with the stemming porter algorithm. This model has exceptional results and demonstrated Accuracy: 78.10%.

The poorest performance was the Conv1D based model with the Lancaster stemming algorithm. This model demonstrated lower scores compared to the other models, with an accuracy of 72.86%. Table 8 presents a complete overview of the performance scores obtained by various in -depth learning models when applied to Hindi text analysis, using different rod algorithms. In particular, the model based on BI-MSTM combined with the bearer entirely the algorithm appeared to be the most efficient among the DL models. This model showed an impressive accuracy of 83.72%, which indicates its Strong skill in accurate classification and predicting Hindi text data. On the contrary, the Cron1D -based model using Lancaster the STEMMMING algorithm has demonstrated the poorest in assessment. With a precision of 71.70%, the precision of 71.66%, Review of 59.03% and 64.73% F1, this model is faced with effective classification problems of Hindi text data. Therefore, further Improvement and alternative methods may be necessary to improve its performance. In Fig. 13(a -h), Graphical comparison of four different models, i.e., models based on LSTM, models based on bi -mstm, conv1d Models based on conv1d-mstm are expressed in English and Hindi text classifications wearing both Lancaster algorithm. These comparisons focus on the accuracy curves of the most efficient model for each stem algorithm. Relationship between the accuracy and number of trainings for each model for training and validation datasets the times are shown in the diagram. The results show that the accuracy of the model generally increases with the number of trainings This indicates that additional ERAs do not significantly improve the accuracy of the model until ERAS reaches the tray.

Furthermore, the drawing shows that the increase in accuracy differs between models and languages. Compared to all training data, it highlights the general trends for lowest accuracy in all validation data. Overflow of educational data models. Provides valuable information on the performance of various models of neural networks Classification of Hindi and English texts. The results underline the importance of carefully selecting appropriate models and balanced the number of trainings wishes to achieve optimal precision on training and validation data sets. In the figure. 14 (a -h), graphic illustrations are presented, number of training times and lose four different models from neural networks, as classifying English and Hindi texts. Models considered in this comparison Wear using an algorithm and use a Lancaster. The purpose of this analysis is to the loss curve of the most efficient model. About each rod algorithm. This diagram shows how the loss of the model changes with the desired number of models being formed process. More specifically, the figure indicates that the loss of the models generally decreases with the number of trainings wishes until it Reached a tray, indicating that additional eras do not significantly improve the loss of models. In addition, the figure reveals that the loss of models on the validation data set is generally higher than that of the training data set, which suggests the presence of over-adjustment. This observation highlights the importance of using appropriate validation strategies to assess the ability to generalize Models. The pseudo presented in Table 9 defines a step-by-step procedure for the implementation of SVM in the tasks to predict moods on the external text of the Banlla and the English. The inclusion of SVMs in the table is based on its incredible accuracy and outweighs everything else A model evaluated in the study. It serves as a guide for developers and researchers interested in using SVM for emotional analysis. It utilizes its high accuracy compared to other models investigated in the study.

Table 10 demonstrates the external performance of the SVM algorithm using the STEMMMING PORTER technique. This table the previously mentioned pseudo is complemented in table 9, which sets out the stages of SVM implementation to predict mood on the outer text of the banlla and English. Among the studies listed in Table 11, our study is distinguished as the most interpreted in terms of accuracy of both mood predictions. Hindi and English textbooks. Previous work reached accuracy of 44.20% to 79.3%, but our study reached impressive accuracy of 86.43% in Hindi and 82.56% in English text using the stemming porter algorithm using SVM. compared to other approaches such as sent wordnet, SVM, LSTM, VAE, and various combinations of models and techniques, our research has been in a consistent way. Duplicate them in terms of accuracy. The accuracy scores affected by our study were above the scores of previous studies margins.

5. Recommendations and Policy Makers Consequences

Based on the conclusions and deposits of this study on mood analysis in a multilingual context, the following recommendations You can then propose the results of the policy.

I. Dataset Enlargement: Future research should focus on increasing reliability and generalizing models that predict mood When increasing the size of the dataset and including a wider range of sources. This provides more complete performance Data mood model and English text data.

II. Research of alternative methods: while Porter Algorithm and SVM model demonstrated effective performance In this study, researchers should study the alternative STEMMING algorithms, machine learning models and deep training architectures. Comparative studies between different methods will give an idea of their strengths and Weak sides, providing the choice of the most suitable approach for analysis of moods.

III. Specific analysis of the field: extend the scope of the analysis of feelings beyond magazines to different fields, such as social media, Press articles or customer comments would be beneficial.

Investigating the performance of different emotions the text genre provides valuable information to the analysis of field-specific emotions, enabling a tailor-made approach.

IV. Aspect-based mood analysis: This study focused on predicting mood predictions in general. However, future research should be done Focus on fine grain analysis of mood and mood analysis according to aspects. Mood analysis at a more detailed level, Special definitions of moods regarding a particular aspect or function provide a detailed understanding of textual data; A more convenient way to analyze mood.

V. V. Practical Applications: The results of this study will affect practical applications in industries such as e-commerce. Manage customer comments and social media surveillance. Organizations can leverage sentiment analysis models to extract valuable insights from customer reviews and improve their products, services, and customer satisfaction.

VI. Development of politics: politicians can use mood analysis methods to assess public moods for a specific policy, Social initiatives or problems.

By analyzing sentiment in multilingual contexts, policymakers can make informed decisions, address concerns, and improve governance strategies. Study limitations and scope for future research While our study on sentiment prediction for Hindi and English text using the Porter stemming algorithm and SVM achieved impressive accuracy results, to acknowledge specific limitations and pinpoint potential avenues for future research is of paramount importance. The scope of our study was limited to a specific dataset comprising 2577 reviews in both Hindi and English. This is true though the dataset provided valuable insights and produced promising results. It is present in all Hindi and English text data. Future research can increase the size of the dataset and include a wider range of sources Increases the robustness and generalizability of emotional prediction models. Our study was mainly focused on using the porter Stemming and SVM algorithm. Although these methods have been effective in our study, there are other Stemming algorithms,

Machine training models and deep education architecture, which can be studied for greater precision of mood forecasts. Comparative studies between various algorithms and models will give valuable information about their strengths and Weak sides. In addition, our study primarily considered the prediction of moods in the context of reviews. Future research can study Analysis of moods in different fields, such as social networks, press articles or customer reviews, to understand how a model Productivity varies in different kinds of text. This study was concentrated exclusively on the prediction of moods, without diving into such aspects as aspect of mood analysis or classification of moods with fine grain. Investigating these aspects would enable a more nuanced understanding of sentiment in text data and could lead to more specialized sentiment analysis techniques.

6. Conclusion

This research work has made a significant contribution to the analysis of moods in multilingual contexts, in particular, concentrating in English and Hindi. The study successfully developed a comprehensive dataset for sentiment analysis in Hindi by collecting reviews in Hindi and their corresponding English translations. This dataset contributes to the availability of resources for sentiment analysis in Hindi. Comparative analysis was conducted to evaluate seven traditional models: learning-based machine learning and deep learning. A model implemented by four unique architectures of neural networks. Specially designed to model Hindi continuous data Text analysis. These architectures include LSTM, BI-LSTM, CONV1D, and Hybrid Conv1D-LSTM. The findings showed that Support Vector Machine (SVM) models exhibited superior performance compared to other models, achieving high accuracies for sentiment analysis in both English and Hindi text using the porter stemming algorithm. The Bi-LSTM Based Model demonstrated the best performance among the deep learning models. The findings offer valuable insights for researchers and practitioners in sentiment analysis, particularly in the context of multilingual sentiment analysis.

Reference

[1] S. Kaur Chatratet, G.S. Batra, Yu. Chaba, Handling consumer vulnerability in e-commerce product images using machine learning, Heliyon 8 (9) (2022), https:// doi.org/10.1016/j.heliyon.2022.e10743.

[2] V. Balakrishnan, et al., In -depth learning approach in the prediction of the mood ratings of the products: comparative analysis, J. SuperComput. 78（5）（2021）7206、https：// doi.org/10.1007/S11227-021-04169-6。 –7226。
[3] N.R. Blowmik, M. Arifuzazaman, M.R. Mondal, Analysis of moods on the Bangla text using an extended vocabulary and deep learning algorithms, array 13 (2022), 100123, https://doi.org/10.1016/j.array.2021.100123.

[4] T. Alam A. Khan, F. Alam, Classification of text of a ban using Transformers, 2020. https://arxiv.org/abs/2011.04446.

[5] S. Dreisbach, T.A. Koleck, P.E. Bourne, S. Bakken, A systematic review of natural language processing and text mining of symptoms from electronic patientauthored text data, Int. J. With。Inf. 125 (2019) 37–46, https://doi.org/10.1016/j.ijmedinf.2019.02.008.

[6] R. Ferreira-Mello, M. Andru, A. Pinheiro, E. Costa, C. Romero, Text Mining in Education, Data Wire. org / 10.1002 / spectrum 1332。

[7] M. Rahman, S. Haque、Z。 J. Comput. Appl。 176 (2020）13–17、https://doi.org/10.5120/IJCA2020920119。

[8] MD T. Ahmed、M。 Rahman, S. Nuru, A. Islam, D. Deploying automatic learning and detailed learning algorithms in the detection of Das, bale, and bale in a bale confusion Rhomized Text Bangla: A Comparative Study, International Conference on 2021 Outcomes in the Fields of Power, Computation, Communication and Stability Technologies, Icaect), 2021, https://doi.org/10.1109/icaet49130.2021.9392608.

[9] I.H. Sarker, M.KH. Furhad, R. Nowrozy, Cybersecurity controlled by AI: Review, Modeling and Research on Intelligence, SN Computer Science 2 (3) (2021), https://doi.org/10.1007/S42979-021-00557-0.

[10] T. Mitchell, Machine Learning, vol. 1, McGraw-Hill, New York, 1997. No. 9.

[11] F. Sebastiani, Machine Learning in Auto-Kegorisation of Texts, ACM Comput. Survivor 34（1）（2002）1–47、https：//doi.org/10.1145/505282.505283。

[12] C. Ming、V。 Viassolo, N. Probst-Hensch, P.O. Chappuis, I.D. Dina, M.S. Patapod, a machine learning method for personalized prediction of breast cancer risk: Comparison with BCRAT and BOADIATEA models, breast cancer resolution. 21(1) (2019), https://doi.org/10.1186/S13058-019-1158-4.

[13] Increased data when recognizing dermatological images using St Lt Aggarwal, Autolearning, Skin Res. Technological 25（6）（2019）815–820、https：//doi.org/ 10.1111/srt.12726。 [14] Y.H. Jung et al., Flexibles Piezoelectric Acoustic Sensors and Machine Learning for Voice Processing, Adv. maternal 32 (35) (2019), 1904020, https://doi.org/ 10.1002/adma.201904020.

[14] Y.H. Jung and colleagues have developed flexible piezoelectric acoustic sensors integrated with machine learning techniques for the purpose of speech processing, as detailed in their publication in Advanced Materials, volume 32, issue 35, article number 1904020, published in 2019. The DOI for this work is https://doi.org/10.1002/adma.201904020.

[15] J.G. Richens, C.M. Lee, and S. Johri have contributed to the field of medical diagnostics by enhancing accuracy through the application of causal machine learning, as reported in Nature Communications, volume 11, issue 1, in 2020. The DOI for their research is https://doi.org/10.1038/s41467-020-17419-7.

[16] T. Fischer, C. Krauss, and A. Deinert have investigated statistical arbitrage strategies within cryptocurrency markets, which is documented in the Journal of Risk Financial Management, volume 12, issue 1, article 31, published in 2019. The DOI for this study is https://doi.org/10.3390/jrfm12010031.

[17]. S. Hakak, M. Alazab, S. Khan, T.R. Gadekallu, P.K. Maddikunta, and W.Z. Khan have proposed an ensemble machine learning methodology that employs effective feature extraction techniques to classify fake news, as presented in Future Generation Computer Systems, volume 117, pages 47–58, in 2021. The DOI for their article is https://doi.org/10.1016/j.future.2020.11.022.

[18] B. Harangi, A. Baran, and A. Hajdu have established an assisted deep learning framework aimed at multi-class skin lesion classification, which also incorporates a binary classification support system. This work is published in Biomedical Signal Processing and Control, volume 62, article 102041, in 2020. The DOI for this research is https://doi.org/10.1016/j.bspc.2020.102041.

[19] E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen have explored the application of deep learning techniques for cellular image analysis, as detailed in Nature Methods, volume 16, issue 12, pages 1233–1246, in 2019.

[20]. J.-J. Wu and S.-T. Chang conducted a study titled "Exploring Customer Sentiment Regarding Online Retail Services: A Topic-Based Approach," published in the Journal of Retailing and Consumer Services, volume 55, in 2020, with the article number 102145. The work can be accessed via the following link: https://doi.org/10.1016/j.jretconser.2020.102145.

[21]. Y. Luan and S. Lin presented their research on text classification utilizing CNN and LSTM at the 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). The details of their work can be found at https://doi.org/10.1109/icaica.2019.8873454.

[22]. D.C.S.Z. Ribeiro et al. investigated the determination of lactose content in low-lactose milk through Fourier-transform infrared spectroscopy (FTIR) and a Convolutional Neural Network, published in Heliyon, volume 9, issue 1, in 2023. The article is available at https://doi.org/10.1016/j.heliyon.2023.e12898.

[23]. B. Jang and colleagues developed a BI-LSTM model aimed at enhancing accuracy in text classification by integrating word2vec, CNN, and an attention mechanism. Their findings were published in Applied Sciences, volume 10, issue 17, in 2020, article number 5841, accessible at https://doi.org/10.3390/app10175841.

[24]. Y. Zhang et al. proposed a text sentiment classification modeling method based on a coordinated CNN-LSTM-attention model, which appeared in the Chinese Journal of Electronics, volume 28, issue 1, in 2019, pages 120–126. The publication can be found at https://doi.org/10.1049/cje.2018.11.004.

[25]. J. Wang and co-authors introduced a Tree-Structured Regional CNN-LSTM model for dimensional sentiment analysis, published in the IEEE/ACM Transactions on Audio, Speech, and Language Processing, volume 28, in 2020, pages 581–591. The article is available at https://doi.org/10.1109/taslp.2019.2959251.

[26]. J.F. Ani and colleagues estimated gender based on Bengali conventional full names using various machine learning techniques, presented at the 2021 12th

[27] J.-J. Wu and S.-T. Chang conducted a study titled "Exploring Customer Sentiment Regarding Online Retail Services: A Topic-Based Approach," published in the Journal of Retailing and Consumer Services, volume 55, in 2020, with the article number 102145. The work can be accessed via the following link: https://doi.org/10.1016/j.jretconser.2020.102145.

[28]. Y. Luan and S. Lin presented their research on text classification utilizing CNN and LSTM at the 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). The details of their work can be found at https://doi.org/10.1109/icaica.2019.8873454.

[29]. D.C.S.Z. Ribeiro et al. investigated the determination of lactose content in low-lactose milk through Fourier-transform infrared spectroscopy (FTIR) and a Convolutional Neural Network, published in Heliyon, volume 9, issue 1, in 2023. The article is available at https://doi.org/10.1016/j.heliyon.2023.e12898.

[30]. B. Jang and colleagues developed a BI-LSTM model aimed at enhancing accuracy in text classification by integrating word2vec, CNN, and an attention mechanism. Their findings were published in Applied Sciences, volume 10, issue 17, in 2020, article number 5841, accessible at https://doi.org/10.3390/app10175841.

[31]. Y. Zhang et al. proposed a text sentiment classification modeling method based on a coordinated CNN-LSTM-attention model, which appeared in the Chinese Journal of Electronics, volume 28, issue 1, in 2019, pages 120–126. The publication can be found at https://doi.org/10.1049/cje.2018.11.004.

[32]. J. Wang and co-authors introduced a Tree-Structured Regional CNN-LSTM model for dimensional sentiment analysis, published in the IEEE/ACM Transactions on Audio, Speech, and Language Processing, volume 28, in 2020, pages 581–591. The article is available at https://doi.org/10.1109/taslp.2019.2959251. 7. J.F. Ani and colleagues estimated gender based on Bengali conventional full names using various machine learning techniques, presented at the 2021 12th

[33] Y.H. Jung and colleagues have developed flexible piezoelectric acoustic sensors integrated with machine learning techniques for the purpose of speech processing, as detailed in their publication in Advanced Materials, volume 32, issue 35, article number 1904020, published in 2019. The DOI for this work is https://doi.org/10.1002/adma.201904020.

[34] J.G. Richens, C.M. Lee, and S. Johri have contributed to the field of medical diagnostics by enhancing accuracy through the application of causal machine learning, as reported in Nature Communications, volume 11, issue 1, in 2020. The DOI for their research is https://doi.org/10.1038/s41467-020-17419-7.

[35]. T. Fischer, C. Krauss, and A. Deinert have investigated statistical arbitrage strategies within cryptocurrency markets, which is documented in the Journal of Risk Financial Management, volume 12, issue 1, article 31, published in 2019. The DOI for this study is https://doi.org/10.3390/jrfm12010031.

[36]. S. Hakak, M. Alazab, S. Khan, T.R. Gadekallu, P.K. Maddikunta, and W.Z. Khan have proposed an ensemble machine learning methodology that employs effective feature extraction techniques to classify fake news, as presented in Future Generation Computer Systems, volume 117, pages 47–58, in 2021. The DOI for their article is https://doi.org/10.1016/j.future.2020.11.022.

[37]. B. Harangi, A. Baran, and A. Hajdu have established an assisted deep learning framework aimed at multi-class skin lesion classification, which also incorporates a binary classification support system. This work is published in Biomedical Signal Processing and Control, volume 62, article 102041, in 2020. The DOI for this research is https://doi.org/10.1016/j.bspc.2020.102041.

[38]. E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen have explored the application of deep learning techniques for cellular image analysis, as detailed in Nature Methods, volume 16, issue 12, pages 1233–1246, in 2019. The DOI for their publication is