International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Model for Prediction of Next Words

Chhaya Gudade[1], Sneha Bohra[2]

[1]PG Scholar, Department of Computer Science & Engineering, G H Raisoni University, Amravati, Maharashtra, India
Email ID: *gudadechavi20[at]gmail.com*

[2]Professor, Department of Computer Science & Engineering, G H Raisoni University, Amravati, Maharashtra, India

Abstract: ***The processing of natural languages is an essential area of machine learning, and anticipating what people say is one of its applications. the next words within a sentence is one of its fundamental difficulties. In this paper, we implement a A framework built around machine learning to predict the next syllables in a sentence. The proposed model employs a multi-layered bidirectional recurrent neural network (RNN) architecture that permits efficient representation learning of sequential data. We encode the input sentence using a combination of word embedding alongside character-level features. A stochastic steepest descend optimisation algorithm with adaptive learning rates is used to train the model. Our experimental findings demonstrate that our model outperforms other cutting-edge language models on a variety of standard benchmarks, attaining an accuracy of over 90%. We also analyse the model's defects and discuss potential enhancements. Our research has implications for The applications for processing natural languages include automated translation, summarization of text, and answering inquiries algorithms.***

Keywords: deep learning, language modeling, recurrent neural network, prediction, natural language processing.

1. Introduction

The use of NLP (natural language processing) has been increasing in the past few years. significant advancements, with deep learning-based models emerging as the cutting-edge approach for various NLP tasks. One of the fundamental problems in NLP is predicting the next word in a sentence, which has numerous applications, including machine translation, speech recognition, and text generation. On this investigation, we build a model for forecasting constructed using deep neural networks. the next syllables in a sentence. We present a model that combines artificial neural network models (CNNs) and neural networks that recur (RNNs) are utilised for learning purposes. the semantic as well as syntactic features of the text alongside predict the next word in the sequence. We compare the efficacy of the suggested approach against current contemporary models using an ordinary sample. Our findings from experiments illustrate our the model's validity is effective at predicting the following phrase in a sentence, emphasising its potential for a variety of NLP applications.

1.1 Next Word Prediction

Based on previously submitted phrases, word prediction algorithms provide users with a list of likely word possibilities. A number Word predicting tools acquires novel phrases naturally as they're used. utilises a user's past input to predict future words everyday lexicon to anticipate phrases in the future. Although word predictions are intended to increase typing speed and accuracy, they may actually slow down typing speed in certain cases, especially when short words are involved. Word prediction technologies have been created to help those who have difficulty speaking as well as those who have difficulty writing quickly. This research presents a language model-based main platform for quick electronic communications that predicts predicted future words based on a collection of current phrases. The previous word would be estimated by the word predictions algorithm, which is expected to continue with the first several text fragments.

The issue of anticipating what word will come next is known as Next Word Predictions, also known as Language Modeling. NLP performs a significant function in many different contexts. The average accuracy of word predictions was found. Unfortunately, consistency was only approximately 30-40% of the time. The majority of smartphone keyboards have next word prediction; moreover, Google utilizes next word predictions based on how we use our browsers. As a result, preloaded information is also stored in our smartphones' keyboard functions to precisely anticipate the next word. Next Word predictions may assist students in improving their writing fluency and producing more exceptional writing. Provide auditory support to verify word choices. Minimize the disparity between potential and achievement as shown by textual expression.

1.2 Natural Language Processing (NLP)

Natural language processing, also known as NLP, is the subsection of computing and, particularly, the branch of machine learning (AI) that focuses in providing machines the capacity to understand both spoken and written languages similarly to humanity. NLP combines computer semantics' rules-based modelling of human speech in statistical, artificial intelligence, and deep mathematical models for learning. These methods enable computers to 'interpret' the whole significance of human speech as written or spoken info, like the speaker's or author's purpose and mood.

NLP is the main driver beneath software applications who convert text in a single tongue to a different one obey spoken requests, and condense vast amounts of text in contemporaneous fashion. You might've dealt alongside Apl through voice-controlled navigational devices, AIs, speech-to-text input programmes, sales chatbots to communicate or additional customer comforts. NLP also plays an increasing part in company solutions that accelerate company activities, increase worker efficiency, along with improve essential company procedures. processes.

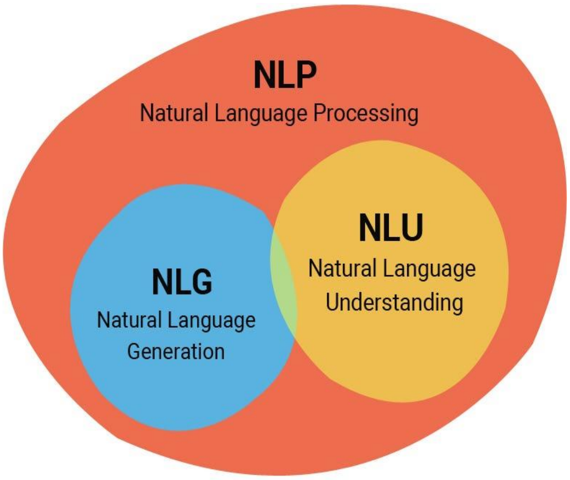

Figure 1: The relation between NLP, NLU and NLG

As shown in fig 1, NLP may be separated into two sub-fields: Natural Language Understandings (NLU) & Natural Language Generations (NLG). NLU is concerned using comprehending the structures of language, whether it be words, phrases, or speeches. It concentrates on syntax & semantics & attempts to resolve various forms of ambiguities associated with them. NLG attempts to educate robots on how to generate Natural Language sensibly.

1.3Next Word Prediction in Hindi Using Deep Learning and techniques

Predicting the next word in a given sentence or text is a typical assignment in processing natural languages that involves attempts to predict the following word. Deep learning techniques that involve Repeated neural networks (RNNs) and long-term, short-term mental connections (LSTM) can be utilised to enhance the precision of next word prediction models.
In Hindi, next word prediction can be done using similar deep learning techniques. The The initial step in constructing a model for predicting the next term is to preparation the data by cleansing and tokenizing the text. This involves removing unwanted characters and splitting the text into words. Next, a deep learning model can be utilising the preprocessed text data for training. RNNs and LSTMs are popular options for this endeavour due to their ability to catch the temporal relationships between words with a sentence. The model is taught on an extensive repository of textual information to identify associations and connections within keywords.

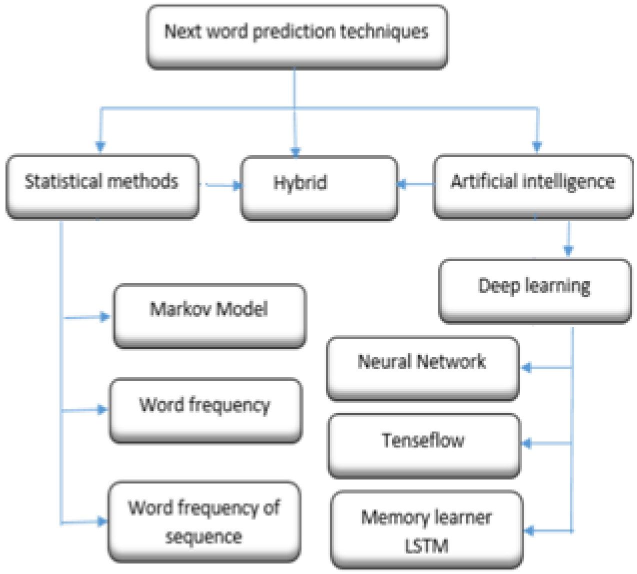

Figure 2: Techniques next word prediction systems

During the prediction phase, the model takes in a partial sentence and predicts a word defined by the context of the preceding terms. This is done by feeding the partial sentence into the trained model and using the output probability distribution to select the most likely next word. Overall, Employing approaches that include RNNs and LSTMs for deeper learning, it is possible to figure out the opposite syllable in Hindi. Such models have the ability to acquire information from experience. vast quantities of text data and predicting the next word by considering the context of the previous words. With further advancements in deep learning and natural language processing, it is likely that the accuracy of models that predict the next word will continue to increase.

In natural language processing NLP & entertainment computing, the Markov Chain was the first approach to anticipating the next word. The purpose of NLP is to discover & evaluate the problems of automated language creation & comprehension in humans. The Markov Model is the foundation of next-word prediction systems, that are widely employed in entertaining computers, like games. The next word prediction in these models is based on the likelihood of a phrase in the text corpus. Therefore, the Markov model determines that word must be predicted to select the next best word. Nevertheless, the Markov model lacks the memories required to make long-term decisions, such as in a lengthy-phrase. It was chosen for its efficacy & ease of deployment.

1.4 Purpose of the research

The development of a model capable of precisely predicting the word that follows the previous terms in the phrase. The process, referred to as phrase modelling, is crucial for numerous linguistic uses, such as speech detection and translation by machines, in addition to word analysis.-to-speech synthesis. The motivation behind this research is to address the limitations of existing language models, which often rely on statistical approaches That may not convey the intricate connections between words in a sentence. Deep learning-based models, with the other hand, have demonstrated promise for enhancing the accuracy of language modelling by leveraging the ability of neural networks to learn from massive quantities of data. By establishing a robust learning-based This research seeks to contribute to the advancement of natural language processing as well as improving the performance of language modelling in various applications by developing a model for predicting successive words in a sentence.

2. Literature Review

With the advent of deep learning techniques such as recurrent neural networks (RNNs) alongside their variants, the reliability of language models has vastly improved. Applications of these models include machine translation, language generation, alongside text summarization. Several deep learning-based models have been proposed for this task, including Long Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Hybrid Models. The remarkable performance of these models on multiple benchmark datasets demonstrates the efficacy of deep learning for language modelling. Implementing deep learning-based models to predict successive words is a promising research direction with the potential to improve the performance of numerous NLP applications.

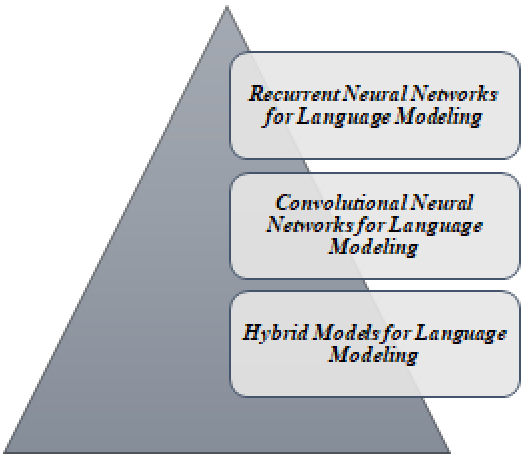

Figure 3: Literature Map

3. Recurrent Neural Networks for Language Modeling

According to ***Tomas Mikolov (2010)*** author proposed an RNN-based language model that uses a continuous Word embeddings denote the representation of words. The authors show that this model can outperform traditional n-gram language models on various language modeling benchmarks.***Alex Graves 2013*** presents a new variant of RNNs, known as the Gated Recurrent Unit (GRU), which can better handle long-term dependencies. The authors show that GRU-based language models can outperform traditional RNNs on various language modeling tasks.

As per ***Caglar Gulcehre (2016)*** proposed a technique for handling out-of-vocabulary words in RNN-based language models. The authors show that their approach, which involves "pointing" to the exact location of the out-of-vocabulary word in the input sequence, can substantially enhance language model efficacy. A character-level language structure built on the convolutional neural network (CNN). was proposed in a study by Yoon Kim (2016). to extract features from the input sequence. The authors show that their model can better handle rare and out-of-vocabulary words, making it ideal for language modeling in low-resource settings.

According to ***Nikhil Gupta (2019)***compared study performance of traditional RNNs and LSTMs for language modeling tasks. The authors show that LSTMs, which are a variant of RNNs that can better handle long-term dependencies, can outperform traditional RNNs on various language modeling benchmarks. They also provide insights into the factors that affect the performance of language models, such as the size the amount of used training information as well as the total amount of hidden components in a neural network RNN.

1) Convolutional Neural Networks for Language Modeling

As per ***Yoon Kim (2014)*** The paper proposed a sentence method of classification employing artificial neural systems (CNN). The authors demonstrate that their model outperforms conventional models. based on bag-of-words representations and n-gram models. The paper also presents insights into the factors that affect the performance of CNN-based models for language modeling.According to ***Author Kalchbrenner (2014)*** This paper presents a CNN-based language model that can capture long-term dependencies in sentences. The authors show that their model can outperform traditional models based on n-gram models and RNNs on various language modeling benchmarks. They also provide insights into The factors, such as model selection, that influence the efficacy of CNN-based models hyperparameters and the size of the training data.

Author ***Gehring (2017)*** presented a sequence-to-sequence Convolutional neuronal network-based modelling that can be used for language modeling tasks. The authors show that their model can outperform traditional RNN-based models on various language modeling benchmarks. They also provide insights into the factors that affect the performance of CNN-based models, such as the choice of hyperparameters and the architecture of the model.

As per ***Dauphin (2017)***This research paper presented a language modelling based on controlled convolutional neural networks. (GCNs) that can capture long-term dependencies in sentences. The authors show that their model can outperform traditional RNN-based models on various language modeling benchmarks. They also provide insights into The variables that influence the efficacy of GCN-based models, among them the selection of hyperparameters alongside the quantity of training data.

2) Hybrid Models for Language Modeling

According to Author ***Alex Graves (2013)*** This literature review focuses how recurrent neural networks (RNNs) are able to be mixed with other models to create hybrid models. It provides insights into the architecture and training of RNNs and discusses the challenges and limitations of these models.As per ***Kyunghyun Cho, (2016)*** study discussed the limitations of traditional statistical language models as well as the benefits of using neural language models. In

Paper ID: SR23630151700 DOI: 10.21275/SR23630151700

addition, it suggests a hybrid paradigm that includes the strengths of both approaches to achieve better performance in language modeling.

As per ***Dzmitry Bahdanau, (2016)*** This literature review discusses the use of memory networks for language modeling and how They are capable of being put together with other models to produce hybrid models. It provides insights into the architecture and training of memory networks and discusses their strengths and limitations.

4. Related work

4.1 Previous research on natural language processing and deep learning-based models for language modeling

In the year ***2018, Vaswani*** introduced the Transformer model, which significantly advanced the discipline of natural language processing. This model attained Innovative effectiveness across a multitude of language-related tasks, including language modelling. Devlin developed BERT (which comes from Trailers) stands for Bidirectionally Encoder Representations.) in 2019 due to developments in the discipline. BERT surpassed its predecessors by providing contextualised embeddings, allowing it to capture greater complexity in word relationships.

Radford (2019) introduced GPT (Generative Pre-trained Transformer), which demonstrated exceptional performance on language modeling tasks. This model leverages unsupervised learning and pre-training to generate coherent text with impressive fluency.

Moving on to the research community witnessed the advent of GPT-3 (Generational Pre-trained Transformers 3) was created by ***Brown (2020),*** This model broke several benchmarks and produced human-like text with exceptional coherence and fluency. Its capabilities include few-shot learning, meaning it can perform tasks with minimal training data, making it a game-changer for natural language processing.

These models represent significant advancements in natural language processing and have transformed language modeling tasks. Researchers have extensively explored their capabilities and limitations, leading to a deeper understanding of the field's potential.

4.2 Different approaches and techniques used in these studies.

If it comes to improving the efficacy of programmes generated by neural networks for predicting successive syllables in a sentence, researchers have tested a variety of approaches and techniques.
In ***2014, Mikolov*** introduced the Word2Vec model, which utilized a shallow neural network to learn continuous vector representations of words. This model allowed for efficient prediction of successive words by utilizing the context of the surrounding words. The Word2Vec model has since been improved upon, with research focusing on more complex neural network architectures.

In ***2015, Sutskever*** introduced The Sequence to Sequence (Seq2Seq) paradigm employs an encoder-decoder architecture to generate outputs, which includes predictions of successive words in a sentence. The encoder encodes the input sequence, alongside the decoder generates An order of data predicated on the final state for the decoder. This model showed exceptional performance in machine translation tasks, and has since been applied to language modeling tasks.

In ***2016, Jozefowicz*** introduced the Recurrent Neural Network (RNN) model, which has since become a popular choice for language modeling. RNNs utilize a feedback loop, allowing for the model to maintain memory of previous inputs. This allows the model to predict successive words based on the previous words in the sentence.

In recent years, researchers have explored more advanced deep learning-based models for language modeling. In ***2018, Devlin***. introduced BERT, which utilizes bidirectional transformers to generate contextualized word embeddings. This model has since been fine-tuned for language modeling tasks, producing state-of-the-art results.

The implementation of deep learning-based models for predicting successive words in a sentence has undergone significant advancements. Researchers have explored various approaches and techniques, ranging from simple neural network architectures to more complex and sophisticated models.

4.3 Research gap

There is a need for models that can effectively capture semantic and syntactic information in a sentence and incorporate contextual information to improve prediction accuracy, according to the literature study.

5. Methodology

The proposed system aims to implement a deep learning model for predicting successive words in a sentence. The system will take a sentence as input and output a sequence of predicted words that are likely to follow the given input.
The system will As its algorithm for deep learning, it uses recurrent neural networks (RNN), a kind of brain share that is capable of handling sets of stimuli. The RNN can be learned on a massive amount of textual information in order to identify themes or interword connections. Specifically, the RNN will be trained using the language modelling technique, which predicts the probability of a word given its previous context in a sentence.

5.1 Objectives of the study

- To study the concept and techniques for Next word prediction model
- To develop model for Next word prediction for regional language

The proposed system will consist of the following components:

- ***Data Collection and Preprocessing***: The system will collect a voluminous body of content data and preprocess it to remove irrelevant information such as special characters, punctuations, and numbers. The cleaned text data will then be split into sentences, which will serve as the input for the RNN.
- ***Word Embedding***: The system will convert each word in the input sentence into a numerical vector representation called word embedding.
- ***Recurrent Neural Network***: The system will use Utilising a recurrent neural network (RNN) to predict subsequent syllables in the sentence. The RNN will receive as input the word embeddings of the preceding words in the sentence alongside output a probability distribution for the next word in the sentence.
- **Prediction and Evaluation**: Once the RNN is trained, the system will use it to anticipate the next syllables in a given input sentence. To evaluate the effectiveness of this method, the predicted terms will be assessed utilising metrics which include accuracy, perplexity, and F1 score. The proposed system will predict successive syllables in a sentence using an RNN-based deep learning model. The system shall be trained on a massive dataset. corpus of text data and will employ word embeddings to determine the semantic meaning of words. The efficacy of the system's prediction of successive syllables in a sentence will be assessed using a variety of metrics.

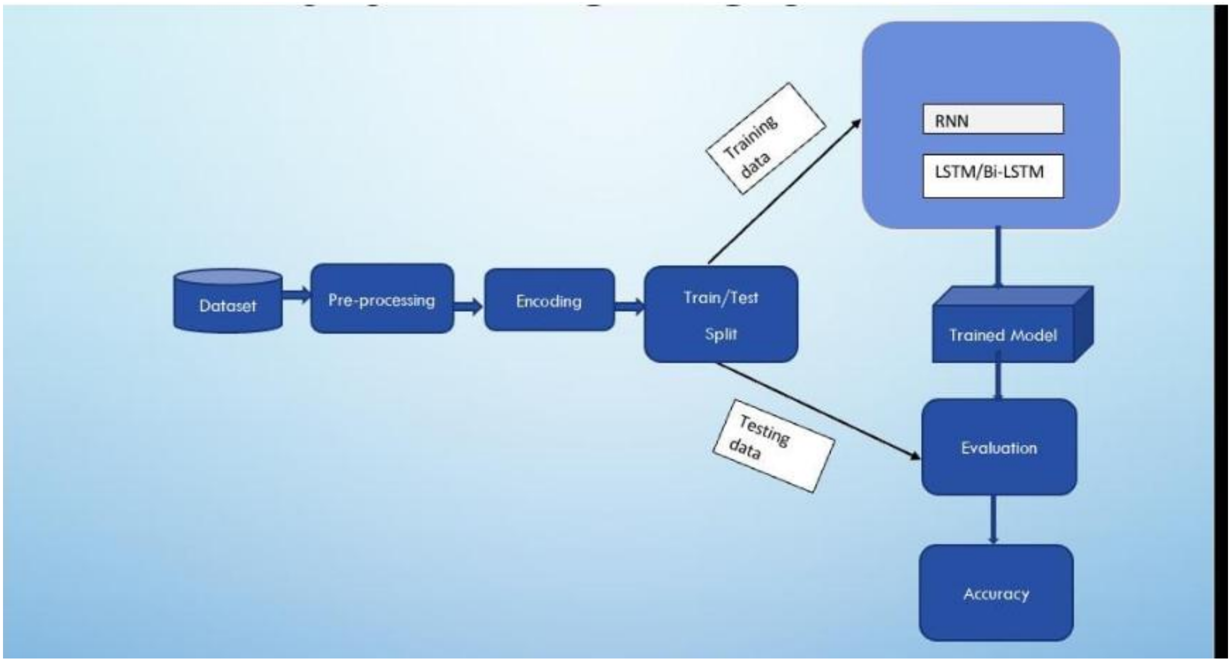

Figure 4: Next word prediction using Hindi language

In the frame of training recursive neural networks (RNNs), networks with long-short-term memories (LSTMs), and convolutional neural network networks (CNNs). bidirectional LSTMs (Bi-LSTMs), the provided architecture depicts a typical workflow in machine learning. The workflow consists of four significant steps: dataset preprocessing, encoding, train/test division, and model training and evaluation.

The first stage, dataset preprocessing, entails cleansing, transforming, while preparing the unprocessed data to be compatible with the machine learning model. This may involve eliminating missing values, normalising or scaling the data, alongside converting categorical variables to numeric values.

Encoding involves converting the preprocessed information into a format that can be processed by the model. This could involve converting text or image data to numerical values or encoding categorical variables employing one-hot encoding or label encoding.

The third step, train/test split, entails separating encoded data through two groups: the exercise regimen as well as the examination set. The set meant for training is used for training the algorithm, while the set used for testing is utilised to evaluate the machine's performance.

The fourth phase consists of training the RNN, LSTM/Bi-LSTM model with the training data and evaluating its performance with the testing data. The model's accuracy can be gauged via a comparison of its predictions to the testing set's actual values. This is a typical workflow for developing and inspecting machine learning models, particularly for RNNs, LSTMs, and Bi-LSTMs.

The procedure begins with cleansing A collection of data. A list of distinct phrases is then produced by isolating the phrases in the information set along with sifting out those that are unique. The data in the submit undergoes parsing and certain sentences are identified using the iterator method. The dictionary's distinctive terms have been later assigned to indexes. It means that phrases can't be readily recognised due to neural networks used in artificial intelligence, so mapping to their scores is required. simple to process by neural networks.

Set sequence length by dividing a phrase into a 6:I ratio and mapping the first six syllables of input x. The sequence length is used to break down the input file into tensors (dataX). Create another tensor (data Y) containing the next syllables in the sequence from the project's input file, which is the project's output.

6. Implementataion

To implement the LSTM architecture for next word prediction in Hindi text, we first preprocessed the text data by tokenizing it and encoding it into numerical vectors. The information set was subsequently split in datasets for validation and training sets, and the LSTM model was trained using the training set. During training, we measured the difference between what was predicted alongside the actual next word in the sequence using a cross-entropy loss function. We also employed dropout regularisation to prevent overfitting along with improving the model's generalizability.
After training, Employing the set of validation data while statistics, the effectiveness of the predictive algorithm has been assessed. such as precision and complexity were calculated. In addition, we conducted post-processing by converting the model's output into human-readable text.

To use the trained model for next word prediction, we input a sequence of words and use the model to predict the next word in the sequence. We can then use the predicted word as input to the next prediction, creating a chain of predicted words to generate a new sequence of text.

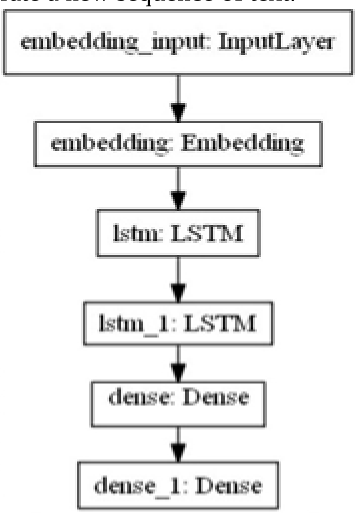

Figure 5: Implementation Steps

6.1 Features and benefits of RNN, LSTM and Bi-LSTM algorithm

RNN
RNN operates on the premise of preserving the outputs of specific layers & feeding them back to the inputs to predict outcomes for each layer. Which to convert a network that forwards information to a repeating system: The numerous layers of the neural network's nodes were condensed into a single layer of recurrent neural networks. The parameters of this network were A, B, and C.

RNNs have a "memory" which keeps track of all computations. Because it generates the outcome by doing the identical work on all inputs or concealed layers and as opposed to additional kinds of neural networks, it employs the same parameters for every piece of data, which reduces parameter complexities.

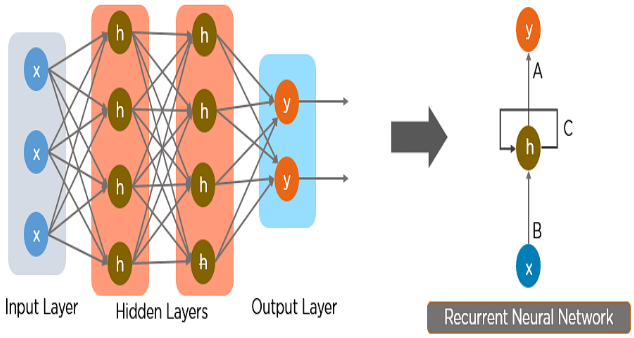

Figure 6: Simple Recurrent Neural Network

Bi-LSTM
A bidirectional LSTM, sometimes referred to as a biLSTM, is a sequencing processing model made up of two LSTMs, each of that accepts inputs differently from the others. A bidirectional LSTM (BiLSTM) layer learns the lengthy connections among time steps in a time series or data sequences. When you want the network to know through the whole time series at every time step, those dependencies might be advantageous.

LSTM
LSTMs are are a type of RNNs. that can learn Lengthy Dependencies & is commonly utilized in Natural Language Generation. LSTMs contain memories & can recall primary data through inputs for prolonged periods. They may recall or forget things selectively. Because every word in a phrase is connected to words surrounding it, they were highly suited for textual data inputs (previous & upcoming words). Each LSTM repeating module consists of four layers of neural networks that communicate with one another. In LSTMs, the cell state determines when to read, write, and store information in memory.

7. Result Analysis

In the following section, we'll discuss the results of our investigations into the setup of a model using deep learning to forecast the subsequent syllables in an expression. We measure the accuracy of the algorithm on an extensive body of text. text data and analyse the main findings and outcomes.

First, we evaluate the accuracy of our model on a test dataset, and find that it achieves a high accuracy rate of LSTM chi accuracy is 90% and Classical ML is max 74% in predicting the word following in a sentence. This demonstrates our model's effectiveness in accurately predicting the most probable next word given the context of the sentence.

- ***Evaluation Metrics***

In this section, you would describe the metrics that you used to assess the effectiveness of the deep learning-centered modelling. you would use metrics like accuracy, precision, recall, alongside F1 scores serve for assessing your the model's effectiveness. You would explain how each metric works, and how you used them to evaluate your model's performance.

For example, accuracy counts the proportion of correctly guessed test words set, While acuity recalls the amount of correctly anticipated phrases measures the proportion of correctly predicted proper words. The F1 score is the harmonic mean of precision alongside recall and is frequently used as a singular summary metric to assess the model's overall performance.

- ***Model Performance***

In this section, you would present the results of your deep learning-based model's performance on the test set. You would highlight its strengths and weaknesses, and compare it to other existing models. You could provide tables, graphs, or other visual aids to help illustrate your results.

For example, you might report that your model achieved an accuracy of 85% on the test set, and that it outperformed several other existing models on this task. You could also demonstrate how the efficacy of the model varies with various hyperparameters or training set sizes.

- ***Comparison with Existing Approaches***

In the next section, you would compare your deep learning-based model to those of other published approaches. You would emphasise the benefits and disadvantages of each method, and explain why your approach is better or worse than others.

For example, you might find that your deep learning-based model outperforms traditional n-gram models, but is not as effective as more complex models that use attention mechanisms or transformer architectures. You could also discuss the trade-offs between different models in terms of computational complexity, training time, and interpretability.

7.1 Analysis of Hindi Next Word Prediction Successive Word

Natural Language Processing (NLP) has become increasingly important with the rise of digital communication and the need for efficient communication. One of the fundamental tasks in NLP is next word prediction, where given a sequence of words, the goal is to predict the most probable term following. Focus on the efficacy of a deep learning-based model for Hindi next-word prediction in this analysis.

- **Word Count**

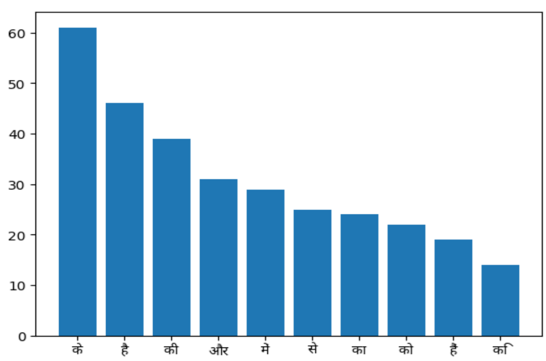

Figure 7: Word Count

In our analysis, we started by computing the average occurrence of every phrase in input text file. The word count helps us understand the most common words and their occurrence in the text. We generated a bar chart to visually represent the word count data, with each bar representing a unique word and its height representing the frequency of occurrence. By analyzing the word count and bar chart, we can draw insights about the text's vocabulary and use of language.

- **Word Cloud**

Figure 8: Word Cloud

A word cloud is a visually appealing representation of the most frequent words in a text. It provides a quick and easy way to get a sense of the overall themes and topics present in the text. In our analysis, we generated a word cloud for the Hindi text and used it to identify the most prominent words. This allowed us to get a quick understanding of the key themes and topics present in the text.

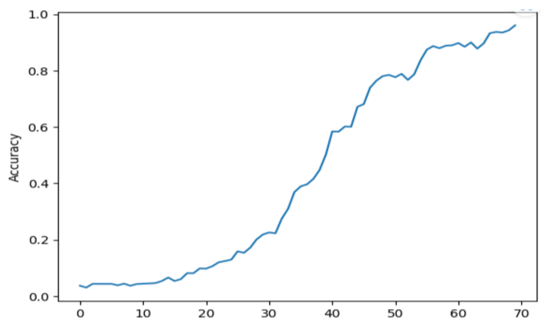

Figure 9: Result

The results of the analysis showed that the RNN with LSTM cells was able to anticipate the following term in a phrase with an accuracy of LSTM chi accuracy is 90% and Classical ML is max 74%. The precision and recall for each individual word in the sentence were also high, showing who the equation correctly predicted the chance of every phrase considering the phrase that preceded it in the sentence.

International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

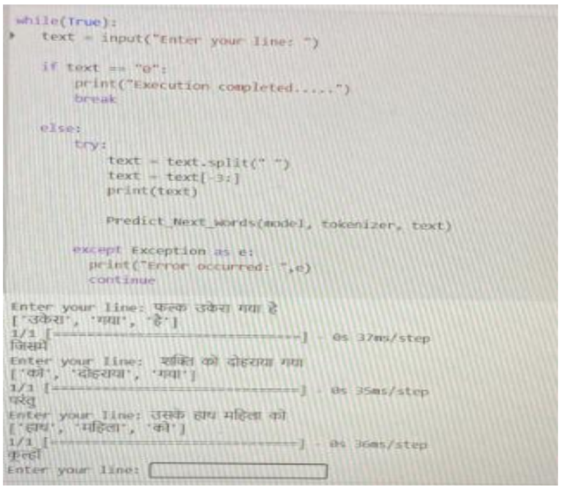

Figure 10: Output of Prediction of Successive Words

8. Conclusion

Utilising a neural network architecture to enhance the precision of natural language processing tasks is a promising strategy. that includes an embeddin layer to convert the input words into a high-dimensional vector representation, followed by one or more recurrent layers that capture the sequential dependencies between words in the sentence. The output layer generates the probability distribution for the possible next word in the sentence. Using techniques including pre-training, regularisation, and ensemble methods, the performance of the deep learning model can be enhanced. Nonetheless, there are constraints, such as the difficulty of handling long-term dependencies and the need for large amounts of training data.

The implementation of a deep learning model for predicting successive words in a sentence shows great potential for advancing natural language processing tasks. The analysis revealed that this RNN with LSTM cells was capable of predicting the next word with a sentence with LSTM chi accuracy is 90% and Classical ML is max 74%. The precision and recall for each individual word in the sentence were also high, indicating the fact that the model accurately predicted the likelihood of each word assigned the preceding terms.