Sentiment Analysis: An Approach for Opinion Extraction from Assamese Text

Submitted by

Chandana Dev

Ph.D Enrollment No-ENGE-02-2017

Ph.D Registration No-17096137 of 2017 - 2018

Date of Admission: 17-08-2017

Department of Electrical Engineering

Assam Engineering College

Jalukbari, Guwahati-781013, Assam, India

This thesis is submitted to

Gauhati University as requirement for the degree of

Doctor of Philosophy

Faculty of Engineering **August' 2024**

Declaration

I hereby declare that this thesis is the result of my own research work which has been carried out under the guidance of Prof (Dr.) Amrita Ganguly of Assam Engineering College. I further declare that this thesis as a whole or any part thereof has not been submitted to any university (or institute) for the award of any degree or diploma.

This thesis contains less than 90,000 (ninety thousand) words excluding bibliography and captions.

Place: (Chandana Dev)

Date

Acknowledgements

While carrying out the research work of my PhD programme, some persons have provided significant help and support which were the source of my motivation. At first, I would like to express my sincere gratitude from the deepest corner of my heart to my supervisor Dr. Amrita Ganguly, for keeping faith on me and for giving me the opportunity to carry out research work under her supervision. She has been encouraging me continuously during this period and provided all the necessary supports. Her expertise in the areas of interest helped me to maintain my moral high throughout the journey of this PhD work. I am thankful to the authorities of Assam Engineering College as well as Gauhati University for providing me all kinds of logistics and administrative support during the programme. I would like to express my sincere thankfulness to all the members of my Research Advisory Committee for continuously reviewing my progress during this period and for giving my research the correct direction. I would like to convey my heartfelt thanks to the faculty members and support staff of the department of Electrical Engineering of Assam Engineering College for their tremendous supports. I am also thankful to my best friends Dr. Rashi Borgohain and Dr. Bikramjit Goswami who provided me constant supports and suggestions at every moment whenever required. At this juncture of my PhD work, I want to express my gratitude towards my maa and sister for their continuous supports, motivation and encouraging me to achieve this goal of my life through their blessings. I would also like to express my honest and heartfelt thankfulness towards my husband Parag Kumar Das for being my constant support and strength throughout this journey. Finally, I would like to thank the almighty for showing me the correct path at every step of my research work and for providing me good health.

(Chandana Dev)

I would like to dedicate this thesis to my two daughters Idika and Harshika for being my constant source of happiness and motivation during my Ph.D journey.

Abstract

With the emergence of social media and the internet, people are able to communicate and express their views and thoughts in a variety of ways. Thus, there is an enormous amount of opinionated data available that needs to be analyzed and interpreted. Sentiment analysis fulfills such needs. It is also known as opinion mining. This is a methodical procedure that discovers, extracts, quantifies and categorizes the subjective content of textual data. It is based upon information extraction, analysis of text, computational linguistics and natural language processing. Sentiment analysis is a way to figure out if any reviews are positive, negative or neutral. It uses techniques like reading text, understanding language and computational techniques to categorize them according to the relevant sentiment. The goal is to understand how people feel about something by looking at reviews that are available in various platforms. This helps businesses, researchers and policymakers understand public opinion, customer satisfaction and make better decisions based on data. Overall, sentiment analysis helps us understand people's feelings and opinions mostly in the digital world. Sentiment analysis provides an understanding of how one feels, conveyed through writing. The majority of current research on sentiment analysis applications is focused on using machine learning and deep learning. Advanced deep neural networks and conventional machine learning techniques have both made a substantial contribution to the accuracy and scalability gains of sentiment analysis systems.

This research work presents a comprehensive framework for interpreting sentiments in Assamese text. It covers novel approaches to numerical expression of words, the use of dictionaries to interpret emotional contexts and various computer programs for example-

based learning. These advancements have a wide range of applications and aid in the understanding of Assamese sentiments. This research addresses preprocessing and modeling issues while offering a thorough method for sentiment analysis in Assamese text. Due to unavailability of any Assamese textual data, the first step in the research was to build an Assamese dataset customized for sentimental analysis. A novel approach for vectorizing textual Assamese data using TF-IDF is performed, with experimentation the mentioned approach shows meaningful outcomes. Lexicon named Assamese VADER, that is inspired by Bengali VADER and modified from English VADER is created to perform dictionary-based sentiment analysis of Assamese unstructured data. Its effectiveness in determining sentiment in Assamese texts is confirmed by experimental study.

Further, supervised machine learning approach is adopted for sentiment analysis of structured Assamese review data. N-gram and TF-IDF feature extraction algorithms are integrated with Decision Tree, K-nearest neighbor, Multinomial Naive Bayes, Logistic Regression and Support Vector Machine classifiers. This method works well for classifying sentiment in a variety of Assamese text domains. Three deep learning models, Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN) and hybrid LSTM-CNN model, are then presented. All of the proposed models, had shown accuracy rates higher than 98%, demonstrating their reliability at identifying complex sentiment patterns in Assamese text as compared to machine learning approaches.

The sentiment analysis techniques presented in this thesis have the potential to enhance a number of Assamese natural language processing (NLP) tasks. Even with some difficulties like lack of proper dataset, lack of standard reference of computational system, there is still much scope for research in this field for Assamese language. It is desirable

to experiment with different deep learning models and increase the dataset in order to develop a flexible system that can be used in a variety of applications. Expanding the size of the Assamese dataset may improve the sentiment analysis methods that are offered and enable the use of more advanced computational strategies. The incorporation of deep learning models has enhanced system performance, indicating prospects for incorporating more advanced learning methodologies such as hybrid machine learning and transfer learning.

(Chandana Dev)

1

Introduction

"*Sentiment is an attitude, thought, or judgment prompted by feeling*".

Sentiments can be appropriately viewed as emotionally loaded opinions **[1]**. Sentiment Analysis or opinion extraction is the computational study of people's opinions, attitudes and emotions towards an entity. It indicates the use of natural language processing, text analysis etc. to systematically identify, extract, quantify and study affective states along with subjective information of an entity. Sentiment analysis has a purpose to determine the attitude of a speaker, writer or other subject regarding some topic or the overall associated polarity or emotional reaction to a document **[2].** The basic objective of sentiment analysis is to identify the sentiment expressed for an entity and then analyze it according to the polarity such as 'positive', 'negative', 'neutral' etc **[3].** Computational techniques on different entities like Audio, Video, Images, and Texts are utilized by researchers to perform sentiment analysis. With an enormous data generated every day on the web, social media etc. analysis of sentiments and its associated applications have become a highly interesting topic to Natural language processing (NLP) researchers. With the rapid growth of textual data in multiple languages, sentiment analysis in multilingual data has opened a new research window to NLP researchers.

This chapter provides a brief introduction on sentiment analysis. **Section 1.1** provides an overview of the thesis by first highlighting the importance of the problem being addressed and the **motivation** behind the research. **Problem statement** is listed in **section 1.2**. **Research objectives** and **research contribution** are presented in **sections 1.3** and **1.4** respectively. Lastly, **thesis organization** is mentioned in **section 1.5**

1.1 Motivation

The method of examining user sentiment expressed in different forms such as text, audio, video or image is called Sentiment Analysis. People can now easily access the Internet and the digital revolution has led to a massive amount of user opinion data being shared on a variety of digital platforms. Nowadays, practically all businesses have an online presence and can be reached by customers with relative ease. Users are generating enormous amounts of opinion data in various forms on digital platforms. Online reviews have an impact on our daily decisions as well. Before making any decision, big or small, we usually seek feedback from others. Therefore, these online opinion data are crucial for any institution or organization that offers public services, directly or indirectly. These opinion data is viewed as raw materials that must be processed to be useful as they are typically available in large volumes and in an unstructured format. To understand the general sentiment of users, automatic sentiment analysis is becoming increasingly important for analyzing the polarity of written text reviews **[4].**

Analysis of textual data is no longer restricted to one or two languages. International languages like English, French, German etc and Indian languages like Bengali, Tamil, Oria, Kannada, Hindi etc, have been able to establish their existences in the field of NLP research for the last couple of years. Data scientists in this research area have moved into

multiple languages used across the globe. Very limited research work is made on the Assamese language, even though it is being one of the socially and culturally enriched languages of North-eastern India. Assamese is also a morphologically rich Indo-Aryan language mostly spoken mainly in the northeastern state of India, especially in Assam, where it is an official language, with around 14 million speakers **[5].** In terms of social and cultural aspects, many literally and cultural works have been manually documented in this language for many hundred years. But at the same time, in terms of computational aspects, it is still in the stage of almost unexplored for NLP researchers. Due to very limited or no available online resources in terms of datasets, compared to the available standard dataset for other languages, Assamese language is still in a very primitive level of research. Lack of available resources has not attracted researcher to explore sentiment analysis in this language. Data scientists are investigating the possibilities of sentiment analysis in the socially and culturally rich Assamese language of northeastern India with limited computational resources on this language. Research on sentiment analysis in Assamese data is important because various online portals which includes social media forums such as Meta (previously known as Facebook) and X (previously known as Twitter) are flooded with text data written in this language. The motivation here is to develop some computation techniques for sentiment analysis in Assamese language, as the rapid progress in the use of this language on online platforms has made this the need of the art. Also, from literature assessment on sentiment analysis in Assamese language, it is observed that no such work in this area has been done. Thus, the main research objective of this project is to employ computational methods to make it easier to utilize NLP applications in the Assamese language in future.

1.2 Problem Statement

In this work an attempt is made to develop computational techniques to perform SA in Assamese language. The work is divided in to three major tasks as mentioned below:

- Dataset collection/Preparation in Assamese language for NLP applications.
- Develop an appropriate approach to perform sentiment analysis with the collected/prepared dataset.
- Design a suitable machine learning and deep learning model to detect accurate sentiments from Assamese texts.

1.3 Research Objective

Sentiment analysis is the process of finding and extracting subjective information from the input data with the use of text analysis, natural language processing and computational linguistics. This is extensively used to reviews and social media for various applications, ranging from marketing to customer service. In other words, it attempts to identify the writer's or speaker's attitude toward a subject or the document's overall contextual polarity. The attitude could be the author's assessment or judgment, their emotional state at the time of writing, or the intended emotional message.

Every day, massive amount of data is generated from social networks, blogs and other media and they are all diffused in to the World Wide Web. This large volume of data contains very crucial opinion. The related information to these opinion can be used to benefit businesses and other aspects of commercial and scientific industries. Manual analysis and extraction of such useful information is time consuming and tedious. Hence,

fast computation technique in this regard is necessary. Sentiment Analysis is one such phenomenon which involves extracting sentiments or opinions from reviews provided by users about a particular subject, area or product. It is an application of natural language processing, computational linguistics, and text analytics to identify subjective information from source data. It clubs the sentiments in two categories like positive, negative or neutral. Thus, it determines the general attitude of the user associated to any topic or any context.

Researchers have conceived different works in sentiment analysis in relation to multiple data sources. Sentiment analysis in Bengali language has reached almost equal milestones in comparison to English language. NLP tasks in English language is already explored in many dimensions. Assamese as a language being similar to Bengali, has been almost unexplored one in the same field of research. In terms of dataset, researchers have a very easy excess of Bengali language as Google translator is incorporated with the same. Whereas, this kind of accessibility has been made available in Assamese language very recently (**12th may 2022**). Very few datasets which can be considered as negligible, are actually available to do any primitive to advanced levels of NLP tasks in Assamese language. As literature of previous work suggests, there is no available sentiment lexicon like SentiWordNet which is already available in different Indian languages such as Hindi, Bengali and Nepali etc. This brings the requirement to build some computational techniques to do sentiment analysis in Assamese textual data. With this objective, the presented research work has aimed to explore the possibilities of this low resourced and much less explored language in NLP field.

1.4 Research contribution

The major contribution of the proposed work reported in this thesis include the following:

1. **Development of a LEXICON based approach**: Assamese VADER has been proposed by modifying the existing English VADER taking Bengali VADER as backbone. From the experimental analysis it has been observed that the proposed model works efficiently on Assamese texts.

2. **Vectorization of the input textual data**: A novel approach to create TF-IDF vectors for Assamese Text has been proposed. A considerable number of experiments is performed throughout the process and significant result was found.

3. **Design Supervised Machine Learning model for Sentiment analysis:** Multiple supervised Machine Learning based classifier, including Decision Tree, K-nearest neighbour, Multinomial Naive Bayes, Logistic Regression, and Support Vector Machine, combined with n-gram and Term Frequency- Inverse Document Frequency (TF-IDF) feature extraction technique is employed for sentiment analysis. The suggested model is observed to be an effective one for sentiment classification in domain-independent Assamese text data.

4. **Design Deep learning Model for sentiment analysis:** Long short-term memory (LSTM), convolution neural network (CNN) and a hybrid convolutional neural network-long short-term memory (LSTM-CNN) model has been developed to do sentiment analysis. With all the proposed model, an accuracy of more than 98% have been achieved.

1.5 Thesis organization

The chapters of the thesis are organized as follows:

The **second chapter** describes about the theoretical background related with the Sentiment analysis. It includes about Sentiment analysis, theory of different sentiment analysis methods used in the work and different classifiers applied for the same.

In the **third chapter**, Lexicon based approach for categorization of sentiment is described by modifying very popular VADER lexicon primarily meant for social media datasets. By modifying existing VADER lexicon efficient performance has been observed and described in this chapter.

The **fourth chapter** reports the first methodology used for feature extraction method for Assamese text. Here, TFIDF techniques are discussed in relevance to Assamese texts.

In the **fifth chapter**, Supervised Machine learning method for detection of sentiment is reported for Assamese Review texts.

The **sixth chapter** discusses about the performance analysis of deep learning method for categorization of sentiment which is based on LSTM, CNN and LSTM-CNN classifier.

In the **seventh chapter**, the future scope about the reported work is discussed in details. The conclusion of the thesis is also illustrated in this chapter.

2

Theoretical Background

Natural language processing (NLP) is computational linguistic which combines various statistical and machine learning models to digitize, recognize, decipher and also generate textual content. This chapter covers theoretical background in sentiment analysis using machine learning and NLP. Fundamental introduction to **sentiment analysis** is discussed **in section 2.1**. While **section 2.2** discusses **different levels** of sentiment analysis. In **section 2.3 various approaches** to sentiment analysis are discussed. Different **steps involved** in performing sentiment analysis is listed and discussed in **section 2.4**. In **section 2.5** various challenges with sentiment analysis tasks were examined.

2.1 Sentiment Analysis

The process of analyzing people's views, feelings and opinions towards various entities like goods, services, products etc. is defined as sentiment analysis **[6]**. This subject has been a recent interest of study for many researchers in the area of computation linguistics. Researchers are greatly associated in building different types of computational techniques to perform sentiment analysis in different domains. With the use of these techniques, it has become possible to monitor public opinion towards a certain entity and obtain useful insights. Additionally, as pointed out in **[7],** the knowledge gained via

sentiment analysis contributes to the larger goal of understanding, explaining and forecasting social phenomena. Sentiment analysis explores the feelings that people express towards an entity, which helps to determine the general mood of the user and also produce insightful information about the same. This multidisciplinary field includes not only natural language processing and machine learning but also psychology, sociology. As data and computing power increased exponentially, machine learning became the dominant tool for sentiment analysis. There is huge volume of scientific literature on sentiment analysis using different machine learning techniques. Researchers from various fields have shown significant attention in the potential of sentiment analysis. Many studies have analyzed online user reviews for products, hotels, movies and similar areas. For instance, people often look for feedback from users and before making purchases. Sentiment analysis plays a crucial role on social media websites like Facebook and Twitter, where it helps understand the sentiment behind user tweets.

Basic steps followed in performing sentiment analysis is depicted in figure 1 below:

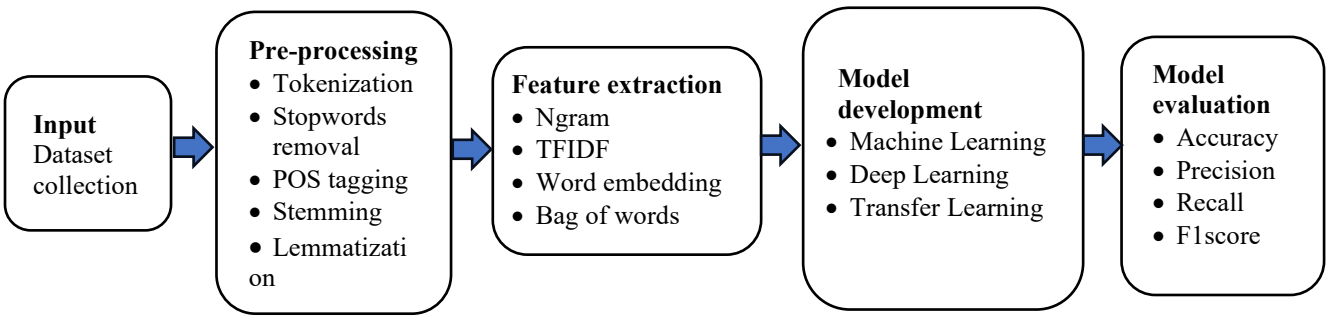


Figure 2.1: Basic steps for SA

- ***Input dataset collection***: Datasets are available in many languages for SA. However, there are some less explored languages around the world, where for performing SA, the very first step is collection and preparation of dataset. To do so researchers use various

web forums, translation tools etc. Experts in language and psychology have classified data from social media platforms like Twitter, YouTube, and Facebook for various studies. Data from social media, blogs and e-commerce sites is often unstructured and require processing for further computational tasks.

- ***Dataset Pre-processing***: Social media allows individuals to easily express their emotions. The unstructured nature of social media data makes sentiment and emotion analysis challenging for machines. Pre-processing is an essential step in cleaning data since it greatly affects succeeding approaches. This step involves removal of unwanted words which hold no sentiment in the input data. Tokenization, stop words removal, negation handling, part of speech tagging, stemming, lemmatization etc are the different method of data pre-processing.
- ***Model development:*** Various approaches are used to develop models to effectively analyse sentiment in input data. These include traditional machine learning approaches, as well as deep learning and transfer learning models.
- ***Model evaluation:*** This step involves assessing the performance of the developed models to determine their effectiveness in predicting sentiment from the input data. Common evaluation metrics include accuracy, precision, recall, F1 score, and AUC: Area Under Curve, ROC: Receiver Operating Characteristics. Additionally, confusion matrices can provide insights into the distribution of predicted sentiment labels and help identify misclassifications. Cross-validation techniques, such as k-fold cross-validation, are often employed to estimate the overall performance of models.

2.2 Levels of Sentiment Analysis

Sentiment analysis have different levels and different algorithms are formulated to address these levels. In this section different levels of sentiment analysis are discussed in details.

In general, there are 3 level of sentiment analysis which are:

a) Document Level b) Sentence Level c) Aspect Level

a) **Document Level:** At this level, the objective is to categorize whether an entire document conveys either a positive or negative sentiment [**8**]. For example, when dealing with a product review, the system evaluates whether the review portrays an overall positive or negative viewpoint regarding the product. This level of SA is performed mainly for review datasets. In regard to text reviews, individual review is assumed as a single document which are classified as negative, positive or sometimes neutral reviews. This level of analysis implies each document represents recommendations about a particular entity (e.g., products, services etc).

b) **Sentence Level:** At this level, the aim is to analyse individual sentence and evaluate whether it represents a favorable, negative, or neutral opinion. Neutral usually denotes "no opinion." This level of analysis is closely related to subjectivity categorization [**9**], which distinguishes between sentences that represent factual information and sentences that express subjective thoughts and opinions. However, subjectivity is not synonymous to sentiment, as many objective words can indicate opinions.

c) **Aspect level:** Document and sentence-level analyses do not reveal specific preferences or dislikes. The aspect level provides finer-grained analysis. The term "aspect level" was generally used to refer to feature-based opinion

mining and summarization. Aspect level analysis focuses on opinions rather than language constructs like papers, paragraphs, sentences, clauses, or phrases. Opinions are formed by combining a sentiment (positive or negative) and a target. An opinion is only useful if it has a clear aim. Recognizing the significance of opinion targets also helps us better comprehend the sentiment analysis problem. For example, the line "although the service is not that great, I still love this restaurant" clearly contains a positive.

2.3 Approaches of Sentiment Analysis

There are various approaches for performing sentiment analysis and it has been a significant part of many studies. Additional studies are still being conducted to identify better solutions in this regard. Sentiment classification is divided into mainly three approaches:

a) Machine learning approach b) Lexicon based approach c) Deep Learning approach

a) ***The Machine Learning (ML) approach***: This approach used for predicting the polarity of sentiments based on trained and test data sets. It applies the ML algorithms and uses linguistic features. The main advantage of this approach is its' ability to adapt and create trained models for specific purposes and contexts. However, the disadvantage is the low applicability of the method on new data as availability of labelled data is necessary that could be costly or even prohibitive **[10]**. Sentiment analysis is the practice of detecting and quantifying sentiment in text or audio using natural language processing, text analysis, computational linguistics, and other techniques **[11]**. To determine the polarity of text documents, an optimized sentiment analysis model is trained instead of being programmed. This model uses a variety of machine learning

algorithms. The type of problem determines selection of proper machine learning algorithm. Machine learning is incorporate in sentiment analysis to recognize the sentiments included in text. To identify if a text is good, negative, or neutral, one needs to consider the words that are used and their context. To make the computer learn to generate predictions on its own, one has to provide it with multiple number of examples that have labels (such as "positive" or "negative"). However, it's not always clear what a writer is trying to convey in term of sentiments. This is known as a "soft classification issue" where, each feeling is assigned a probability by the computer. Without explicit instructions, computers can learn new things which is being possible only by machine learning. Thus, sentiment analysis algorithms are able to comprehend context as well as words. Many machine learning based techniques are frequently employed in sentiment analysis to categorize text into positive, negative or neutral groups according to the sentiment represented. There are two primary Machine Learning approaches to sentiment analysis. They are discussed in below:

- Supervised learning

Most well-known machine learning method is supervised learning. This type of learning method originated as a type of "classical" machine learning, which depends on data scientists to create a task-specific algorithm for each function they want the machine to perform. This strategy uses labelled source data to train a model. The trained model can then make predictions for an output using new unlabeled input data. Because of its accuracy, supervised learning techniques are popular. These algorithms must be trained on a training set before being used on real data. popular supervised classification approaches for sentiment analysis, include Support Vector Machine (SVM), Naïve

Bayes (NB), Maximum Entropy (ME), Artificial Neural Network (NN), and Decision Tree (DT). Random Forest (RF), K-Nearest Neighbor (KNN), Bayesian Network (BN), and Logistic Regression (LR) are a few less popular algorithms **[12].** They are very briefly discussed below:

1. **Naive Bayes**: Simple probabilistic classifiers with strong feature independence requirements, naive Bayes classifiers are based on the Bayes theorem. Since they are effective, sentiment analysis jobs frequently utilize them as a baseline model.
2. **Support Vector Machines** (SVM): This supervised learning technique finds the optimum hyperplane to divide data points into classes. SVMs work well for problems involving sentiment analysis, particularly those involving high-dimensional data.
3. **Logistic Regression**: This is a statistical type model which is useful for binary classification tasks. This algorithm estimates the probability that a given instance belongs to a particular class based on its features. This is usually used in sentiment analysis due to its simplicity and interpretability.
4. **Decision Trees**: The feature space is recursively divided into regions linked to various class labels by decision trees, which are hierarchical structures. They are appropriate for sentiment analysis since they are simple to understand and intuitive, especially when working with categorical data.
5. **Random Forest**: This ensemble learning method builds several decision trees and aggregates the predictions from each to arrive at a final classification. When it comes to huge and noisy datasets, random forests excel in sentiment analysis tasks because it is mostly reliable and efficient.

6. **K-nearest neighbours (KNN)**: This is a basic machine learning technique that may be used for regression and classification problems. It works by examining the 'k' nearest labeled information points to a given data point to determine its label. It is non-parametric and intuitive, but it can be computationally significant when working with enormous amounts of data and is dependent on the value of 'k'.

Supervised learning has the benefit of giving the algorithm access to well-labeled training data, which enables it to learn and generate predictions which are based on known results. This enables the model to estimate new, unknown information with accuracy. Supervised learning algorithms are also capable of managing intricate connections between input characteristics and output labels, which makes them appropriate for a variety of tasks, including regression and classification.

- Unsupervised learning

Unsupervised learning train computers to learn from unlabeled data. This type of machine learning algorithms are used where there are no labelled data to train the classifier. These techniques rely on self-learning and have been demonstrated to be effective in the field of NLP, particularly in sentiment classification [**13], [14**]. The majority of the current unsupervised sentiment categorization approaches can be divided into two stages [**15], [16**]. In the first stage, the sentiment intensity of the text is calculated by estimating the sentiment strength of the terms and expressions used to express emotions. In the second stage, the sentiment categorization of data is achieved by referring to the sentiment strength of the data against the baseline value of '0'. Unsupervised learning involves instructing machines to utilize data that lacks labels or classifications. This suggests that the machine is engineered to autonomously learn

without the need for labelled training data. Operating without prior knowledge of the data, the machine must possess the capability to categorize it.

b) ***The Lexicon based approach***: Lexicon is the vocabulary of a particular language, field, social class, person, etc **[17].** In this context lexicons are words with predefined scores denoting the neutral, positive or negative nature of text, categorized by polarity and scores between +1 and -1. This lexicon-based method is viable for sentence and feature-level sentiment analysis, requiring no training data. However, its main drawback lies in its domain dependence, as words may have varied meanings across domains, resulting in differing positive or negative interpretations. To address this, a domain-specific sentiment lexicon or an existing vocabulary can be developed. While considered an unsupervised technique, its limitation lies in its domain-specific nature, restricting the applicability of words to other domains. This approach uses a sentiment lexicon to determine the valence (positive, negative and neutral) of given textual data. This approach is more understandable and can be easily implemented in comparison to machine learning based algorithms.

Few available standard lexicons are LIWC **[18]**, GI **[19]**, Hu-Liu04 **[20]**, ANEW **[21]**, Sentiwordnet **[22]**, Senticnet **[23]** VADER (Valence Aware Dictionary and sEntiment Reasoner)**[24]** etc

c) ***Deep learning approach***: This approach is a kind of machine learning inspired by the human brain. In recent years, it has become a highly effective strategy for natural language processing. It is becoming increasingly popular because of its capacity to achieve high accuracy. Over the few years, this type of learning has advanced in the sentiment classification hierarchy. To address the difficulty of managing numerous

hidden layers within a neural network, deep learning employs a multilayer strategy **[25]**. Deep learning refers to the higher number of layers of a neural network (NN). Early neural networks consisted of three layers: input, hidden, and output. Adding more hidden layers deepens the NN, allowing it to understand complicated relationships and solve complex problems. The number of 'hidden layers' determines the depth of the network. Neural networks handle word embeddings, not individual words. Deep learning allows for efficient feature learning by replacing handcrafted features with algorithms **[26]**. These can capture high-level features from input data. Neural networks can adapt to any domain by learning features from their current task. Deep neural networks can outperform classic machine learning methods with sufficient training data **[27]**. Deep learning has numerous uses, including sentiment analysis, computer vision, and automatic speech recognition. Many methods, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, Recurrent Neural Networks (RNNs), and Gated Recurrent Unit (GRU) networks, are commonly used in deep learning approaches. As these algorithms can extract significant characteristics from text, grasp context, and record sequential dependencies, they are well-suited to evaluate textual data and have been used widely in sentiment analysis. The other well-known deep learning algorithms, like Autoencoder, Deep Q-Network, DNN, Recursive Neural Network (ReNN), Capsule Network (CapN), and Generative Adversarial Network (GAN). These algorithms have only been employed in a small number of research and are not as popular. Additionally, some hybrid techniques are also used combining CNN and LSTM algorithms.

The overall pictorial classifications of SA is shown in **Figure 2.2 .**

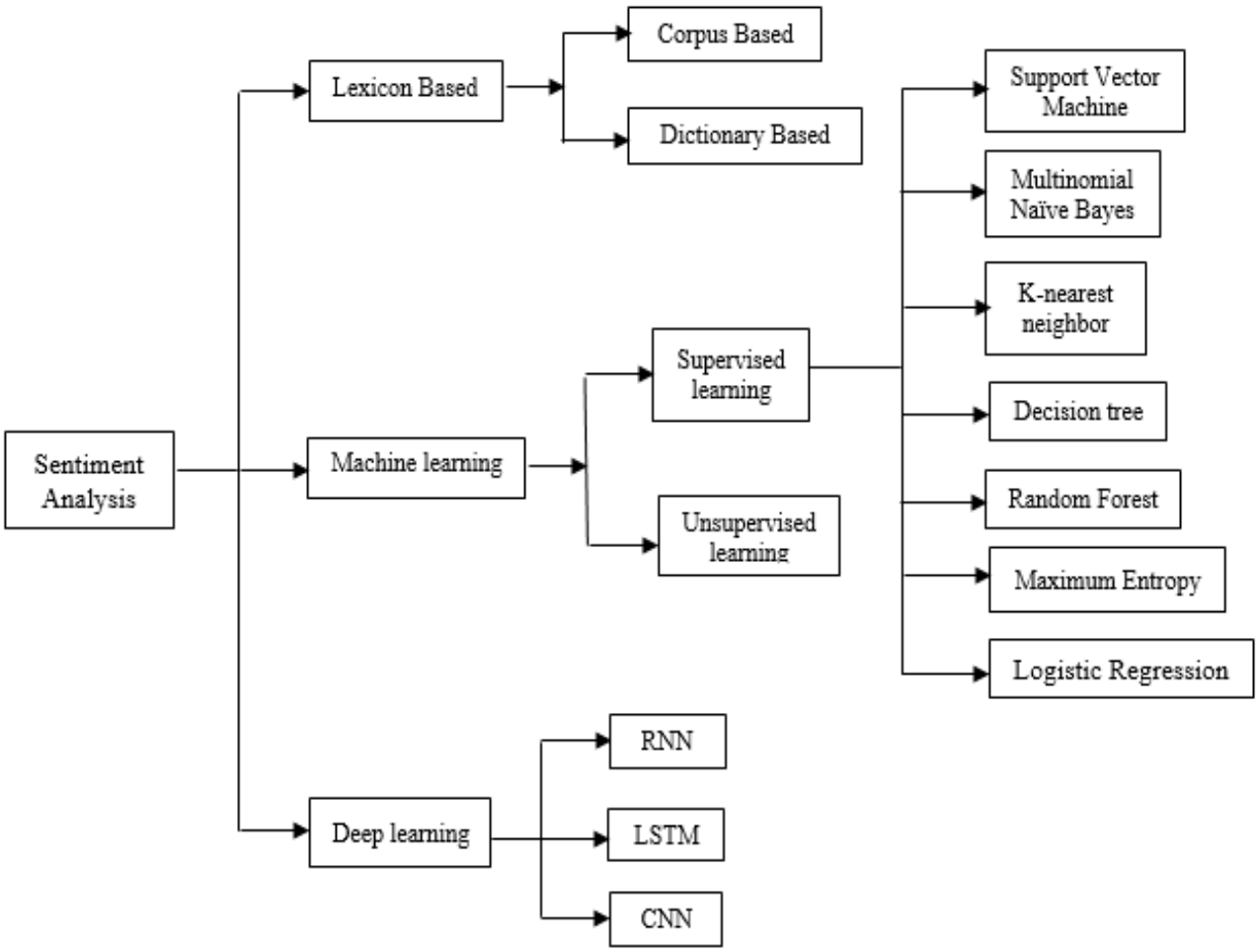


Figure 2.2: General classification of SA

2.4 Steps of sentiment analysis

Figure 1 shows the basic step involved in SA. Steps are briefly outlined below:

Step-1: Data Collection/Creation

Unlike many other Indian languages, there is no benchmark textual dataset available for NLP researchers in Assamese. This makes it difficult for researchers to do basic to advanced NLP research. To overcome this, a dataset was created for use in this work.

Details about the data collection procedures for Assamese language is elaborated in **chapter-4**

Step-2: Data Preprocessing

Initial data preprocessing steps in any NLP task includes removing stop words, negation handling, proper use of boosting words, tokenization etc. Tokenization is the method of breaking down a complete text sentence into smaller units. Each such units are called tokens. Tokens are nothing but a piece of single word that are present in a text. For e.g in a sentence like 'i am a human'; tokenization gives out four number of tokens which are- 'i', 'am', 'a', 'human'. Process of tokenization is very important as well as mandatory for any NLP task at the stage of data pre-processing.

Raw data typically has a lot of irrelevant text or string (such as hyperlinks, html tags, etc.) in it that is insignificant for machine learning algorithms. In order to obtain the true data for the sentiment analysis model, these noisy data must be cleansed. Stop words are those words which appear commonly in the review data but do not contribute much in terms of sentiment or polarity in the text data for the language model. Stop words in text data include, but are not limited to, commas, special characters, is, are, am, they, I, he, etc. These words can typically be excluded from sentiment analysis models since their polarity is neutral.

Commonly used pre-processing techniques for sentiment analysis are:

1. **Data cleaning** is the method of identifying and rectifying errors, inconsistencies, or inaccuracies in datasets to enhance their quality and reliability for analysis **[28].**
2. **Tokenization** is essential in both text analysis and natural language processing. It divides text into unique words or tokens **[29]**.

3. **Feature selection** seeks to improve model performance and decrease dimensionality by choosing the most relevant and instructive features from a dataset **[30].**

4. **Normalization** in machine learning refers to the process of bringing numerical data into a standard range, usually ranging from 0 to 1, in order to standardize measurements and enable dependable comparisons **[31].**

Step-3: Feature extraction

Features in natural language processing are primarily collections of words with sentiment or meaning that are used in documents **[32]**. One of the most important processes in handling the text data for sentiment analysis is feature extraction. Words used in the text that express people's opinions are considered features in text data for machine learning algorithms. Each such piece of text data is represented as a feature vector of real values before being fed into ML algorithms. Feature refers to the unit that is used by a classification algorithm for sentiment analysis. Feature selection is the task of selecting relevant terms from the text documents for the classification task.

Major function of feature extraction approaches is to convert unstructured text into meaningful representations that can be processed by machine learning algorithms. In the process of sentiment analysis, text documents are typically represented as a feature-document matrix. Features in a text document might be individual words. These words give us more information and help improve how we represent the relationship between features and documents.

However, different feature types carry a large number of features and relations between them, having diverse feature types leads to a high dimensionality problem. When we select features, we choose the ones that matter most to help classifiers do a better job.

This makes the process easier by narrowing down the feature set. As a result, we can create sentiment analysis applications that are effective and efficient. The several feature extraction techniques frequently used in sentiment analysis are discussed here briefly:

- **Bag-of-Words (BoW) Representation**: BoW ignores word order and grammar in favor of representing text as a collection of words. Based on word frequency, a vector is used to represent each document. BoW is efficient in capturing lexical information despite its simplicity.
- **TF-IDF Representation:** TF-IDF weights words according to how frequently they appear in a document and throughout the corpus. By eliminating repetitive words and emphasizing valuable content, it improves sentiment analysis precision.
- **Word Embeddings:** By taking into account a word's context in extensive text datasets, word embeddings such as Word2Vec and Glove produce dense vector representations of words. They excel at sentiment analysis jobs involving vast amounts of text data and grasp the semantic linkages between words.
- **N-gram Representations:** Word or character sequences in text are called N-grams. The most common types of them are trigrams, bigrams, and unigrams, which all represent local contextual information.

Step-4: Model development

An appropriate machine learning or deep learning model architecture is chosen for sentiment analysis, taking considerations for aspects like model performance and complexity. The selected model is trained with labeled dataset so that it can identify relationships between text features and sentiment labels.

Step:5 Model evaluation

Model evaluation is essential to verify the effectiveness of the build model. This is done by calculating few standard evaluation metrics such as accuracy, recall, f1 score etc**.**

2.5 Challenges in Sentiment analysis

Although sentiment analysis has come a long way, there are still many challenges that need to be resolved by researchers. Resource constraints, context awareness, domain adaptation, multimodal sentiment analysis, presence of emojis and emoticons roles in predicting sentiment analysis of any texts etc are few of them **[33]**. More research is necessary, in these mentioned areas. to improve sentiment analysis's adaptability and usefulness in a variety of dynamic scenarios. The most important challenges for performing sentiment analysis are outlined briefly in this section.

- **Resource (dataset) Constraints**: Since models used for sentiment analysis rely on data. Hence, creating a standard labeled dataset is a crucial step in the process. Low-resource languages are greatly affected by the scarcity of linguistic resources. When creating a dataset, information collected from different fields is mostly unstructured and misspelled. Additionally, data annotation takes a lot of effort and differs from person to person. A sentiment lexicon is also necessary alone with the dataset for doing sentiment analysis. However, it might introduce bias into more advanced analysis or decision-making. Therefore, by creating language resources from scratch via semi-supervised, unsupervised, and transfer learning techniques, the resource shortage can be addressed **[34]**. Resource limitations impact computing capacity, data accessibility etc. creating difficulties for sentiment analysis. Sentiment analysis faces many obstacles due to insufficient dataset resources, which have an additional impact on model performance

and development. Furthermore, models built with deep learning techniques require extensive resources and larger datasets while used for performing sentiment analysis. It becomes challenging for peoples and various organizations to fulfill the purpose of analyzing meaningful contexts with limited amount of data **[35]**.

In order to simplify sentiment analysis using limited resources various experiment can be performed. Models may be built to adapt well with reduced computer power usage. Semi-supervised learning can be employed, which require less training data. Furthermore, transfer learning can be applied to improve sentiment analysis's efficiency with smaller amounts of data. These techniques ensure that sentiment analysis functions successfully even in situations when resources are scarce.

- **Multimodal sentiment analysis**: Recently, a growing interest in multimodal sentiment analysis is seen trending among researchers. It is an alternative to typical text-based sentiment analysis, which examines sentiments expressed across several modalities like text, images, audio, and videos **[36]**. With the integration of many data sources, this method provides a more thorough knowledge of sentiment. Recent research progress has led to advanced models that can better predict sentiment by combining different types of information.

- **Domain adaptation**: It can be challenging to adapt sentiment analysis models to differentiate contexts such as social media and product evaluations etc. due to the wide variations in language and phrases. Sentiment analysis models frequently experience complications when they are applied to different domains, such as product reviews, social media, or news stories, where their accuracy is impacted by different vocabularies, writing styles, and sentiment expressions **[37].** The difficulty of this problem, known as domain adaptation, increases when working with multilingual text. It is difficult for

sentiment analysis algorithms to keep up with these variations, which makes it difficult to effectively predict sentiments in a variety of circumstances and languages.

- **Contexts understanding**: Sentiment analysis requires an understanding of context because it entails identifying the minute nuances that influence the sentiments expressed in text. There exists significant difficulty in understanding sentiments as it can be strongly influenced by context and factors like ambiguity, sarcasm, irony, negation, and word sense confusion. Advanced contextual analysis algorithms, enhanced word sense disambiguation methods and flexible machine learning models can overcome these challenges in sentiment analysis. For example, authors in **[38]** proposed a model that makes use of a dynamic configuration window function to distinguish between word senses by utilizing the context around problematic terms.

- **Emojis and emoticons**:

Emojis and emoticons present a significant barrier to sentiment analysis since they complicate the interpretation of text-based emotions. These visual representations might express feelings that contrast with the associated with words **[39]**.

- **Computational cost**: The computational cost of the built model grow exponentially with the training data size and with the complexity of the built model. Consequently, to train the deep model using a large corpus, a high-end GPU is needed. The computational cost of neural and attention architecture is higher than that of machine learning classifiers like naïve Bayes and support vector machines **[38][39].**

This chapter discusses about the theoretical background of SA. The subsequent chapters present different models for SA in Assamese language. Furthermore, three different approaches used to implement SA in the mentioned language namely the lexicon-based, machine learning and deep learning-based approaches are also elaborated in the next

chapters. A very first step of this research work i.e implementation of a very well-known lexicon named “VADER” based sentiment analysis is explained and discussed in the next chapter.

3

Lexicon Based Approach: Sentiment Analysis Using Modified VADER

The vocabulary meaning of a lexicon is the collection of meaningful words in a particular language. A polarity lexicon is a collection of words or terms classified according to the relevant attitude or emotional orientation, usually identifying whether they convey positive, negative, or neutral connotations. These lexicons have a list of words along with the initial level of polarities. Such lexicons are considered as one of the important resources for sentiment analysis in texts using computational techniques.

This chapter covers details about lexicon-based sentiment analysis approach using well known VADER. Related literature is reviewed in **section 3.1** followed by fundamental introduction to **lexicon** is mentioned **in section 3.2**. While **section 3.3** discusses **lexicon-based sentiment analysis**. In **section 3.4,** brief introduction about **VADER** is discussed. **Background involved** in performing sentiment analysis using VADER is mentioned in **section 3.5**. In **section 3.6 Proposed Assamese VADER** is elaborated. **Results and discussion** about the proposed VADER are finally outlined in **section 3.7.**

3.1 Literature review on related work

There are three main ways to construct polarity lexicons which are understanding existing lexicons from different languages, finding polarity lexicons from corpora **[24]** and annotating sentiments Lexical Knowledge Base **[40] [41] [42] [43].** In four major languages there are well-known manually constructed lexicons such as General Inquirer **[44],** Opinion Finder **[45],** SO-CAL **[46]** etc. In **[47]** and **[48]** authors investigated the techniques of translating Romanian and Spanish languages from English resources. As compared to existing English polarity lexicons such as SentiWordNet and VADER there exists hardly any polarity lexicon in Assamese language.

Enormous research works on lexicon based sentiment analysis in the English language have been performed such as VADER. VADER-tool is a combination of important lexical features obtained from five generalized rules that represent grammatical and syntactical conventions of human speech. It also holds the advantages of conventional sentiment lexicons such as LIWC **[49][50].**

3.2 Lexicon Based Sentiment Analysis

Lexicon-based sentiment analysis analyzes text data using established dictionaries or lexicons that include words labeled with their appropriate sentiment polarity, viz. positive, negative or neutral. In this method, each word in the text is compared to the sentiment lexicon entries to establish its sentiment orientation. By aggregating the sentiment scores of individual words, the overall sentiment of the text can be assessed. The leverage of Lexicon-based approaches are their simplicity and efficiency, while dealing with a massive number of text data. Nevertheless, there may be challenges like

difficulty in understanding context-dependent emotions, sarcasm, and subtle expressions. In spite of these limitations (constraints) lexicon-based sentiment analysis remains a pivotal method in gaining deep insights into the emotional tone of textual content across diverse domains, such as social media and consumer reviews etc. Lexicon-based sentiment analysis is an efficient method for determining how people feel about a subject based on the words they use. It is not ideal because words might have multiple meanings depending on the context. A large variety of sentiment analysis techniques rely heavily on an underlying sentiment (or opinion) lexicon. A sentiment lexicon is a collection of lexical features (for example, words) that are classified as positive or negative based on their semantic orientation **[48].** Manually developing and confirming such lists of opinion-bearing attributes takes a significant amount of time.

This chapter explores a technique for conducting Sentiment Analysis on Assamese texts, a morphologically intricate yet under-resourced Indo-Aryan language, by utilizing the principles of a widely-used sentiment analyzer known as "VADER". It contains a comprehensive list of words and phrases with sentiment scores ranging from -4 (extremely negative) to +4 (extremely positive), including an emphasis on emoticons and slang commonly found in social media content **[24].**

In this work, Assamese Vader has been formed employing "Bengali-Vader" as its foundation and following the standard procedure of a typical Vader tool. This entails the creation of a dictionary of negative booster words and the formulation of an Assamese Lexicon, pre-processing of data, enhancement of the valence of each word, calculation of valence, and text sentiment categorization based on the valence. The importance of a reliable dataset in this approach is crucial, and although there is a scarcity of good translation tools and resources for Assamese language to be used by the model, this model

was evaluated on a collection of Assamese texts from a famous Assamese novel titled "***Aximot jaar Heraal Heema*" written by Kanchan Baruah**". The comparison was conducted by manually translating the Assamese sentences to their Bengali and English equivalents, and the results were significant when contrasted with their Bengali and English counterparts.

Sentiment analysis of textual data is no longer limited to just one or two languages. Data scientists in this field have expanded their efforts to multiple languages worldwide. The growing amount of textual data in various languages has opened up new research possibilities for NLP researchers. In recent years, several Indian languages, such as Bengali, Tamil, Oriya, Kannada, and Hindi, have gained recognition in the field of NLP research.

Assamese language, which is a morphologically rich Indo-Aryan language mostly spoken mainly in the northeastern state of India, especially in Assam, where it is an official language, with around 14 million speakers **[5].** Due to very limited available resources in terms of datasets, sentiment analysis in Assamese language on this language is not yet being explored by data scientists.

3.3 VADER

VADER is a simple general sentiment analysis rule that is developed by comparing its performance against eleven benchmarks that reflect typical practice states. The General Inquirer (GI), SentiWordNet, Affective Norms for English Words (ANEW), Linguistic Inquiry and Word Count (LIWC), and machine learning-oriented techniques based on Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM-C) Classification and Support Vector Machine (SVM-R) Regression are the eleven common state-of-

practice standards **[24]**. The VADER tool is regarded as the gold standard for creating a sentiment lexicon since it considers both qualitative and quantitative techniques. This tool is an excellent choice for analyzing sentiments of data from social media. Strong sentiment lexicons improve sentiment analysis models' performance significantly. The built-up of the VADER English model follow the frameworks as mentioned here below:

1. **Data pre-processing**: For NLP tasks, such as sentiment analysis, data preprocessing is an essential step. The text data must be cleaned and prepared for analysis through a number of stages. The methods listed below outline how to prepare the data for the next steps of sentiment analysis. The different data preprocessing steps are discussed below:

- ***Tokenization***: Tokenization is the process of dividing the text into individual words, or tokens. The text data processing and analysis are made simpler by this step. For e.g. in a sentence like: “I love watching comedy movies” after tokenization of this sentence it breaks into five tokens as ‘I’, ‘love’, ‘watching’, ‘comedy’, ‘movies. As five words make the sentence hence there will be five tokens in the given sentence.
- ***Removal of Stopwords***: Words that appear frequently in any text but possess little to no sentiment value are known as stop words. Examples of these words are "is," "are," "of," "from," etc. Eliminating stop words makes the data less noisy and highlights the more important terms. This step helps to improve the efficiency and accuracy of applied algorithms by focusing on more meaningful words. However, based on the particular NLP task, language and domain, the stopwords selection may change. Different stopwords set may be found in different languages and areas. Furthermore, depending on the analysis's context, domain-specific stopwords may need to be included or excluded.

- ***Removal of Punctuation***: It is necessary to remove punctuation that frequently used but does not add to the text's sentiment, such as commas and exclamation points, semicolon etc. Eliminating punctuation makes the text simpler and increases effectiveness of sentiment analysis systems.

2. **Data Boosting**: After the data preprocessing step, each cleaned token undergoes a check for valence boosting. Tokens containing terms such as 'very', 'great', 'extremely', etc., are identified as boosting data. Due to presence of such boosting words, the valence of the adjacent word is enhanced in the given text. Phrases and idioms are then examined in the given texts for followed by the word 'but'. If the same is present, then valence of the word is additionally enhanced.

3. **Valence Calculation**: In this strep, valence of given statements are calculated. Valence signifies sentiment polarity which ranges from -4 to +4 **[24].** This range is further standardized to the interval from -1 to +1. Thus, a sentiment polarity is assigned to each sentence throughout the entire document or dataset.

The simplicity of VADER represents numerous advantages when compared to machine learning techniques. It is fast and computationally efficient without compromising accuracy. VADER may be used conveniently from a typical modern laptop with moderate specs (e.g., 3GHz processor and 6GB RAM), but it can take a long time to analyze the same corpus for more complex models like SVM (if training is needed) or pre-trained models.

Additionally, the lexicon and rules used by VADER are also directly accessible. VADER can therefore be easily inspected, understood, extended and modified as required. By utilizing both the lexicon and rule-based model, it enables the inner mechanism of the sentiment analysis tool more accessible and interpretable to wider ranges of users

community. Sociologists, psychologists or linguists who often use LIWC may also use VADER tool.

VADER also utilizes a human-validated sentiment lexicon. It has outstanding performance in diverse domains without the requirement of huge training data.

3.4 Proposed Assamese VADER

The main goal in doing this research work is to create a sentiment analysis method based on lexicons. The focus is on using a well-established vocabulary, a common reference lexicon, in order to achieve this purpose. The basis was established by VADER, a significant technology that is primarily designed for sentiment analysis in English text. VADER's capacity to support more languages, including Bengali, has grown over time. While VADER was originally developed for social media content, it has shown to be remarkably effective when used with textual data from many sources. In this work, the effectiveness of the modified VADER model is evaluated using text data, taken from a book published in Assamese. In the experiments, a small number of words from the English VADER vocabulary were translated into Assamese. The results were compared to Bengali, which is the closest relative of Assamese language. The accuracy and suitability of the proposed Assamese VADER model for sentiment analysis is determined through systematic testing.

3.4.1 Background

1. English Language: Substantial research work is done on sentiment analysis of English language, including VADER. To create a sentiment lexicon of gold-standard for this model, qualitative and quantitative methodologies are extensively used in recent times. According to literature review there exist numerous polarity lexicons for English e.g SentiWordNet and VADER which are available in open source platforms, however, in Assamese, such polarity lexicon has not been developed so far.

2. Assamese Language: Sentiment analysis in Bengali language has reached almost equal milestones in comparison to English language. Sentiment analysis in Bengali language have already been explored in many dimensions. Although the Assamese language being very much similar language to Bengali, same field of research is almost unexplored. In terms of the dataset, researchers have a very easy excess to Bengali language as Google Translator is incorporated with the same. Whereas Google Translator for Assamese language was unavailable at the time of this work. Very few datasets, that can be considered negligible, are available to do NLP tasks in the Assamese language. There is no sentiment lexicon like SentiwordNet available in Assamese, although the same is available in various Indian languages like Hindi, Bengali, Nepali etc. Authors of **[47]** have translated SentiWordNet into Bengali language and used its associated polarity to estimate the valence of Bengali textual data.

3.4.2 Creation of proposed Assamese Vader

In this work a model has been developed that is especially intended to identify the emotions conveyed in texts written in Assamese. The goal of the analysis is to

comprehend the feelings that these texts express at the phrase level. However, the manual creation of an Assamese lexicon was an essential part of the process. At first, the existing Bengali vocabulary was translated into Assamese in order to create this lexicon. Prior to the creation of the Assamese VADER sentiment analysis tool, the English-to-Bengali lexicon was translated using Google Translate. However, because language translation involves ambiguities and complexity, depending only on automated translation proved to be insufficient. To guarantee accuracy in expressing the sentiments in Assamese language and unavailability of any translation tool, word-by-word manual translation was performed. Word translation done manually required a great deal of time and effort. In order to address this issue, limited the number of terms (500 terms) were translated from the large list included in the English VADER model. In doing so each word were considered, selecting only those which would be most significant and effective for sentiment analysis in Assamese. The built Assamese VADER tool emphasizes the significance of paying close attention to detail and employing precise methodology when adapting sentiment analysis techniques to various linguistic settings. By offering a reliable tool for analyzing sentiment in Assamese texts namely Assamese VADER, researchers and practitioners will be provided with a valuable resource for accurately and effectively assessing the sentiments expressed in their work.

- **Dictionary creation:** Two dictionaries were created. One for negative words and another for booster words.

Negation Words Dictionary: The Negation Words Dictionary is a collection of terms commonly used in Assamese language to express negation in a sentence. In the context of sentiment analysis, these words can reverse the polarity or sentiment of the text.

Examples of negation words in the English language include "not," "no," "never," and "don't". Similarly, the Negation Words Dictionary for Assamese includes words commonly used in Assamese text that serve the same purpose. Some examples are ‘নাই’, ‘নকৰিব’, ‘নকৰো’, ‘নকৰিলে’, ‘নলয়’, ‘নহলে’, ‘নহব’, ‘নাই হোৱা’, ‘নাইকৰা’, ‘নাই লোৱা’, ‘কৰা নাই’, ‘হোৱা নাই’, ‘ভাল নহয়’, ‘বেয়া নহয়’, ‘নহয়’, ‘নহলে’, ‘নহবলগীয়া’ etc. When a sentence contains one or more negation words from this dictionary, then sentence’s polarity may be reversed : positive or negative, based on the context.

For instance, in the sentence "I am not happy" presence of the negation word "not" would interpret meaning of the sentence as negative.

Booster Words Dictionary: The Booster Words Dictionary is a collection of terms that amplify or boost the intensity of a sentiment in a sentence. These words can strengthen the polarity of the sentiment, making it more categorical. Booster words in English include "very," "extremely," "highly," and "absolutely" etc. Similarly, the Booster Words Dictionary for Assamese includes terms that serve the same purpose of intensifying sentiments. Such words are ‘বহুত’, ‘বেছি’, ‘খুব’, ‘অতি’, ‘অত্যাধিক’, ‘অধিক’ etc. When a sentence contains booster words from this dictionary, sentiment polarity may be amplified accordingly e.g, in the sentence "I am extremely happy" the presence of booster word "extremely" would make a stronger positive sentiment.

- **Assamese lexicon**: The original English Vader text comprises approximately 7000 words and around 500 emojis, which could not be translated manually. In this research project, while developing the Assamese Vader, there was no access to any machine translators at the time. Therefore, manual translation was the only viable option to achieve this objective. Polarity of all words has been maintained as specified in the original English Vader lexicon. Valence of all Assamese words used in this research are within the

range -4 to +4, -4 being the highest level of negativity and +4 being the highest level of positivity and '0' being neutral sentiments.

- **Pre-processing of data:** Basic steps for data pre-processing as discussed earlier were adopted for building the lexicon tool

Tokenization: Here, punctuation along with stop words are eliminated from the texts whose sentiments need to be evaluated. For instance in the following text: – 'He is not that good' in Assamese would appear as – 'সি', 'খুব', 'ভাল', 'নহয়' after tokenization. After punctuation removal, it would appear as same as above.

Assamese Stop Words Removal: Stop words are common words that do not carry significant meaning or sentiment in a sentence, such as articles, prepositions, and conjunctions. In the Assamese language, a list of stop words has been created, which includes words like 'সি' (he/she), 'মোৰ' (I), 'মই' (me), 'তেঁও' (he/she), etc. During preprocessing, these stop words are identified and removed from the sentence token list. This helps to eliminate noise and irrelevant information from the text data, allowing the sentiment analysis model to focus on words that convey meaningful sentiments. For example, the sentence 'সি খুব ভাল নহয়' (He/she is not very good) would be processed to 'খুব', 'ভাল', 'নহয়' after stop words removal.

Stemming: Stemming is the method of reducing derived words to their associated root or base form, known as the stem. In regard to the Assamese Vader model, stemming is performed to normalize words and make them comparable to the lexicon. Words with common suffixes like 'ৰ' (ra), 'টো' (to), 'ডাল' (dal), 'খন' (khan), etc., are identified for stemming. After stemming, a negation checklist is applied to the text to identify words with negation endings like 'নহয়' (not), 'নহব' (will not), 'নাই' (no), etc. For example,

the sentence 'বিপদৰ সময়ত নহয়' (Not in the time of danger) would be processed to 'বিপদ', 'সময়', 'নহয়' after stemming and negation check.

- **Boosting word valence**: Here, boosting words are searched in the texts. When such words appear in the text, their valence is intensified based on their position in the phrase. The 'bigram', trigram' method is used to check booster words position in the given text **[24]**. The n-gram technique searches two neighboring tokens. The trigram technique searches for three nearby tokens. Any word related to booster dictionary has its valence doubled by 0.9 for bigram tokens and 0.75 for trigram tokens. The existence of negation words has an impact on the sentence's overall mood. The multipliers are chosen by empirical observation, experimentation and also fine-tuning of the algorithm to achieve optimal performance.

Let's illustrate this with an example:

Sentence: " This book is very good "

In this sentence, "very" is a booster word which is included in the booster dictionary.

1. **Bigram Example:**

- The bigram is formed as "very good" by the booster word "very" and the word "good"
- As per the rule, the valence of "good" is multiplied by 0.9.
- If the original valence of "good" is 0.8 (positive), the adjusted valence would be 0.8 ×0.9 = 0.72 because of the presence of word "very"

2. **Trigram Example:**

- The trigram is formed by the words "is very good″
- As per the rule, the valence of "good" is multiplied by 0.75.
- If the original valence of "good" is 0.8 (positive), the adjusted valence would be

0.8× 0.75 = 0.6. because of the presence of word “very”

In both cases, the presence of the booster word "very" intensifies the sentiment valence of the word "good", but to different degrees depending on whether it forms a bigram or a trigram with the booster word.

In the context of negation words like "not," "no," or "never", when appear in a sentence, they can change how one feels about other words nearby. For example, if one says "not happy," the word "not" changes "happy" to mean the opposite of happy. However, in the Assamese language, negation words are employed differently than in English. In Assamese literature, these terms are typically used toward the end of the phrase, however in English, they are used in the midst of the text. Negation words are searched using the list of terms included in the negation words list, which is created at the beginning. The sentence's valence is multiplied by -1, resulting in the reversal of the sentence's present valence.

- Calculation of valence: First step to calculate the valence is to accurately calculate the score. The following example illustrates the method used for calculating the score. However, these scores require normalization which is calculated by using equation-1. The model then outputs sentiment polarity as positive, negative or neutral. However, it further requires normalization.

$$\boldsymbol{Normalized\ Score} = \frac{\boldsymbol{Score}}{\sqrt{\boldsymbol{score^2} + \propto}} \quad \textbf{----------(3.1)}$$

Where α =15 is the approximated maximum expected value and the score is calculated score to be normalized. Using an example written in Assamese, let's discuss how sentiment score is calculated and then normalized as mentioned above.

Sentence: “মই খুব সুখী” (I am very happy)

Here, individual words score is first calculated as per their sentiment polarity. Say, "মই" is a neutral sentiment word, whereas, "খুব" is a booster word and "সুখী" is a positive sentiment word.

Now, individual score is assigned to each word. Say, মই = 0 (neutral), খুব = +0.5 (booster and positive), সুখী=0.45 (positive). These values are choosen arbitrarily. Total score is calculated as

Score= 0 (মই) + 0.5 (খুব) + 0.45 (খুশি) = 0.95. Now normalization is done using equation no -1, and found the normalized score as,

$$\text{Normalized score } = \frac{0.95}{\sqrt{0.95^2+15}} = 0.24$$

- Sentence separation: A statement is classified as neutral, negative, or positive based on polarity determined from the valence of texts present in the sentence. If the he valence of the texts range from -1 to 0, indicating that it is negatively expressed. Here, '0' symbolizes the statement's neutral attitude and when the valence range is between 0 and +1, the sentence is regarded positively expressed.

The flow diagram below illustrates the proposed Assamese VADER for sentiment analysis:

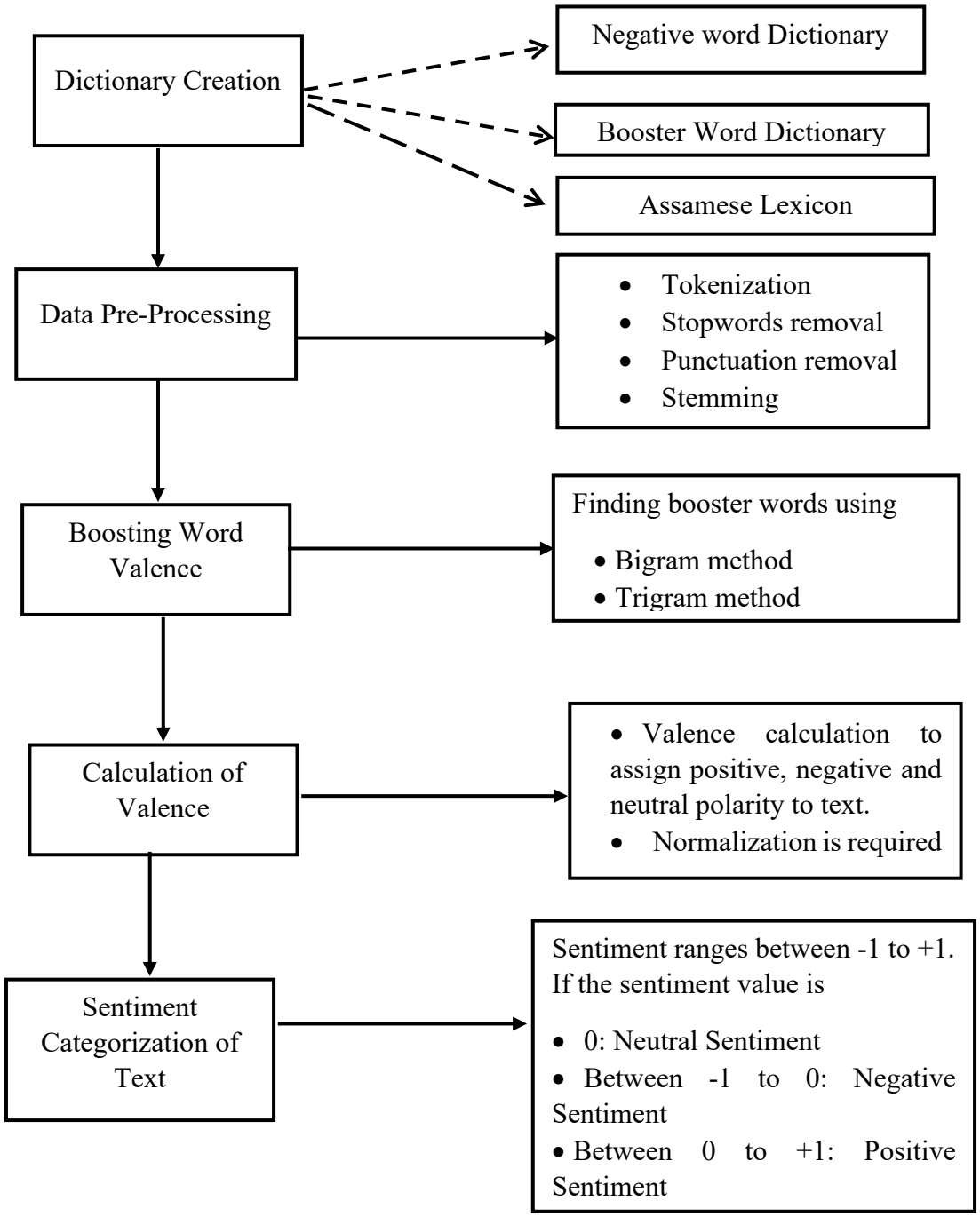


Figure 3.1: Flow diagram of Methodology

3.5 Results and Discussion

Initially, a set of randomly selected sample sentences were analyzed using the Vader tool, which yielded satisfactory results that matched human sentiment. Subsequently, extracts

from an Assamese famous novel mentioned earlier, were analyzed and the results were compared with those obtained using Bangla VADER **[51]** and English VADER **[24].** Manual translations were conducted for the sentences in both cases. Table 3.1 displays the comparison of the results obtained from the different models. The human sentiment was also extracted manually, and the sentiments expressed in the novel were predominantly neutral, which was reflected in the analysis results. The table indicates that the majority of the sentences in the novel expressed neutral sentiments. The analysis results were consistent with the human sentiment extracted from the novel.

Table 3.1: Sentiment Comparison of Different Vader Tools (Including Human Sentiment)

Serial No.	Sentences	Assamese Score	Bengali Score	English Score	Human Score
1	সি খুব বেয়া নহয় (He is not so bad)	0	0.659	0.4708	0
2	মই মোৰ মতে ভালে আছোঁ (I am fine in my opinion)	0.4404	0.6597	0.2732	1
3	ব্যর্থতা সাফল্যৰ চাবি (Failure is the key to success)	0	0.1027	0.1027	0
4	নিজৰ ওপৰত আত্মবিশ্বাস থকা ভাল (It is good to have self-confidence)	0	0.7096	0.6908	0
5	তেঁও এজন আদর্শ শিক্ষক (He is an ideal teacher)	0	0.5267	0.5267	1
6	সি ভুত ভয় কৰে (He is afraid of ghosts)	-0.4019	0.7003	0	-1
7	মোৰ সৌভাগ্য হোৱা নাই (I have not been fortunate)	-0.3818	-0.3818	-0.3412	-1

8	সি পৰিশ্রমী নহয় (He is not hardworking)	0	-0.4767	0	-1
9	কি ধুনীয়া খৱৰ (What a good news!)	0	0.4767	0.5994	1
10	ফান্দত ভৰি নিদিব (Don’t put your foot in the trap)	0	0.3182	-0.3182	0
11	সি পঢ়াত মনোযোগী নহয় (He is not attentive towards studies)	0	-0.3818	0	-1
12	বুদ্ধিমানৰ লগত লাগি থাকিলে আপুনিও সফল হব (If you stay with intelligent people, you will be successful.)	0.4215	0.7003	0.6486	1
13	মই মোৰ মাক বহুত ভাল পাওঁ (I love my mother a lot)	0.4927	0.6697	0.6369	1
14	সি খুব বেছি ভয় কৰা নাই (He is not much afraid)	0.4576	0.5034	0.6369	1
15	উজ্জ্বল বস্তু মাত্রেই সোন নহয় (Not everything that shines is gold)	-0.4404	0	0.4404	-1
16	এখন কপি থকা ৰামধেনু (A tremating rainbow)	0	0	0	0
17	উদ্দেশ্য নাই, চিন্তা নাই, ভাবনা নাই (No purpose, no worries, no thought.)	0.4404	0.4404	0	0

In this study, various methodologies were employed to determine the sentiment polarity of Assamese text data. The results obtained indicate whether a given sentence is positive, negative, or neutral in sentiment. To evaluate the accuracy of the approach, results were compared with those obtained using the VADER tool, which is available in English and Bengali languages. For evaluation purposes, Assamese texts were manually translated into English, and the sentiment polarity was compared between the translated texts and the result of the proposed method. While the English VADER tool

encompasses over 7000 words, the proposed Assamese VADER model utilizes only around 500 words. This disparity in resources contributes to differences in sentiment scores for the same sentence. For instance, consider the sentence "সি খুব বেছি ভয় কৰা নাই" (He/she is not very afraid). The proposed model assigns a sentiment score of 0.4576, Bengali VADER assigns 0.5034, and English VADER assigns 0.3412. Despite the variation in scores due to limited lexicon resources, the sentiment polarity classification remains consistent across all VADER tools and human sentiment detection, categorizing the sentence as positive. This example highlights the need for expanding the Assamese lexicon to improve the performance of sentiment analysis. Translating the entire English VADER lexicon into Assamese, can significantly enhance the accuracy and effectiveness of sentiment analysis for Assamese language text data.

A graphical representation of the scores is given in **Figure 3.2**. The figure depicts a bar chart of the sentiment scores of the statements in three languages, Assamese, Bengali and English.

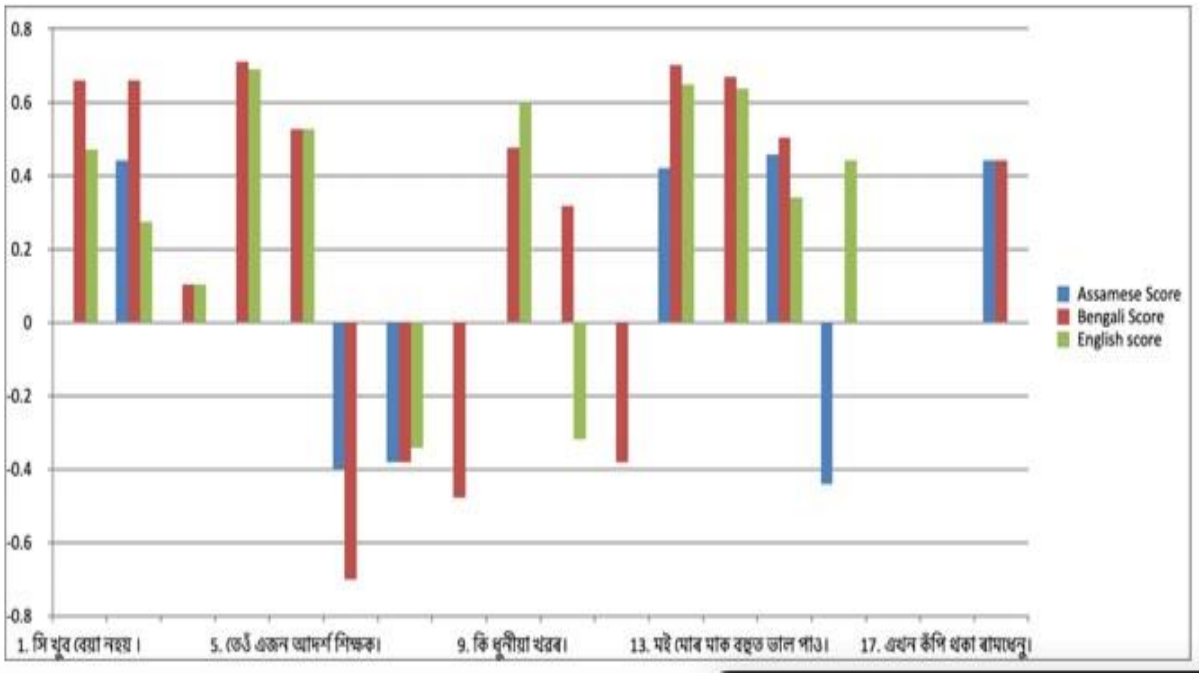


Figure 3.2: Bar chart of Comparison of Sentiment Scores in Assamese, English and Bengali Vader

Upon comparing values across all three languages, variations are evident due to the differing sizes of lexicons. Despite the limited lexicon resources available for Assamese, the proposed Assamese VADER model consistently yields satisfactory results, showcasing its effectiveness in sentiment analysis.

To further validate the results human sentiment assessment was made on the sentences used in the model. Figure 3, shows the comparison between the sentiment made by humans and those generated by proposed Assamese VADER tool. This comparison is presented through a bar graph, highlighting the sentiment polarity for a specific set of sentences. Since human sentiment assessments give a more realistic view of the polarity, valuable insights into the performance and accuracy in the Assamese Vader tool is obtained from this analysis. The comparative analysis offers a comprehensive

understanding of how well the proposed model aligns with human judgment, thereby validating its reliability and efficacy in sentiment analysis for Assamese language text data.

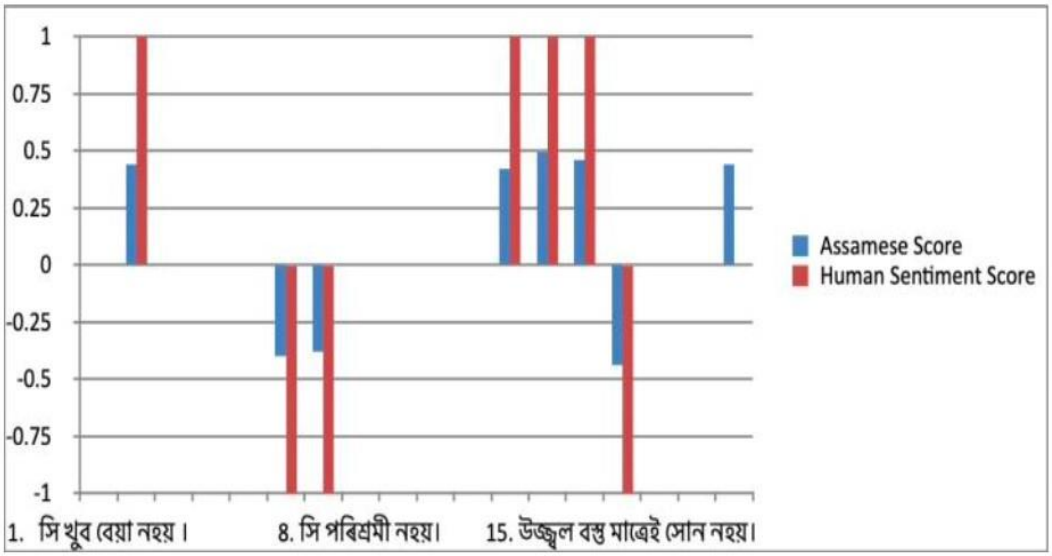


Figure 3.3: Bar chart of Comparison of Sentiment Scores in Assamese with Human Sentiment

As illustrated in the Figure, both the sentiment scores obtained from Assamese VADER and those manually determined for Human Sentiment show similar trends. However, it's important to note that this scenario is based on a limited lexicon for the tool, meaning that some variance may occur as the input size increases. This challenge can be addressed by expanding the lexicon of the VADER tool, which has the potential to significantly enhance performance.

3.6 Conclusion

This chapter presents various methodologies for the development of Assamese VADER, a sentiment analysis tool modified specifically for Assamese text, by applying

modifications to the existing English VADER framework with Bengali VADER as a reference. The primary objective was to adapt the English VADER model to accurately detect sentiment polarity in Assamese texts using an Assamese lexicon. To achieve this, boosting words were used and bigram and trigram techniques implemented to enhance the model's performance. The experimental analysis indicates that the proposed Assamese VADER model effectively analyzes sentiment in Assamese texts. However, due to the limited availability of resources in the Assamese language, initial testing of the model with minimal data posed challenges. Nevertheless, significant improvements were observed in performance as the size of the manually translated Assamese lexicon increased. This underscores the importance of expanding the lexicon with additional translated words to further enhance the model's accuracy and effectiveness. Machine learning approach for analyzing sentiments is then performed on different set of data and explained in next chapter.

4

Feature Extraction using TF-IDF from Assamese textual data

The most essential processes in handling text data for performing sentiment analysis is feature selection. In machine learning algorithms, the words in the text data that represent opinions of people are known as features in text data. This chapter discusses the basic feature extraction approaches in Natural Language Processing (NLP) for analysis of text written in Assamese. The requirement of feature extraction arises because machine learning algorithms cannot operate directly on unprocessed text data. Hence, there is a demand to translate text into a matrix or vector of features. Among the different methods of feature extraction, the most frequently employed ones are Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) .

Literature review related to TF-IDF technique is discussed briefly in **section 4.1**. Whereas, an summary of the **feature extraction** is illustrated in **section 4.2**. The **method of feature extraction** using **TF-IDF** is explained in **section 4.3**. **Methodology** of calculating TFIDF for Assamese language are presented in **sections 4.4. Results obtained** are outlined in **section 4.5**.

4.1 Literature Review on related work

The relationship between the specificity of terms in a document and the exhaustivity of its description was initially identified by **[52]**, which gave rise to the concept of TF-IDF. Authors of **[53]** provided a mathematical representation of the intrinsic relation by imprinting the IDF as the log of the total number of documents divided by the number of documents containing the word. Ever since its inception, TF-IDF has gained popularity as a word vectorization method for various natural language processing tasks, including keyword extraction, summarization, clustering, and classification. A classification approach for Chinese news segments was provided by **[54].** Term-Frequencies were utilized to construct the news terms database, which was then employed in the classifier. Similarly, authors of **[55]** proposed a classification method to classify Arabic text documents where the categorization was done by using TFIDF frequencies to find out the index terms, which were then used in the classifier.TF-IDF is also used in the classification work for Indian languages. Researchers of **[56]** work on Bengali text classification is noteworthy. Using TF-IDF with dimensionality reduction and 40% of the word's frequency, the features were calculated, and the study produced the maximum accuracy of 97.78% while classifying sports publications. With the combination of a modified TF-IDF and a Support Vector Machine **[57],** researchers in **[58]** were able to categorize a set of Bengali text texts with 92.57% accuracy. In addition to document classification, computation of TF-IDF is also used for processing of different Indian languages. In this research area, **[59]** have done significant work where, they have used TF-IDF along with GSS (Galavotti, Sebastiani, Simi) **[60]** co-efficient. They have used the same for calculating the sentence score for extraction of summarization of the input data. A summarizer has been developed by **[61]** where they ranked sentences in

accordance to the cumulative TF-IDF values. A comparison is then done between these values and the threshold (pre-defined) for finding the subjective terms of a given document. TF-IDF technique is used by **[62]** for vector space modelling the words with the cleaned(pre-processed) corpus. Three Dravidian languages viz. Tamil, Telugu and Kannada are taken by **[62]** for evaluating the importance of a word in a given document and also to model the vector space model of the vocabulary. Here, they have developed a keyword extraction method using TF-IDF.

With the studied literatures as mentioned above, it is observed that computation of TF-IDF plays a vital role in different levels of natural language processing tasks.

4.2 Feature extraction

In the study of Natural Language Processing (NLP), feature extraction is essential because it makes it possible for computers and algorithms to interpret and handle textual input, which is essentially made up of words, phrases, and letters. The conversion of text data into a numerical representation is necessary as it can be easily understood and used for a variety of NLP tasks. Moreover, algorithms and machines are not capable of directly interpreting raw text data [**63**].

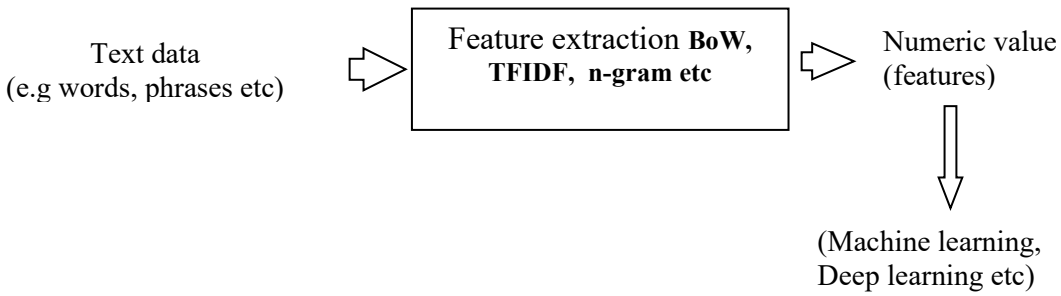


Figure- 4.1: Feature extraction for text analysis

Feature extraction as shown in the figure- 4.1, is essentially the method of converting

textual data into feature vectors which is numeric. A feature vector is a structured and understandable numerical representation that captures significant textual information. Machine learning models and algorithms use these feature vectors as input, which enables them to recognize patterns, anticipate outcomes, and carry out tasks based on the properties of the text data.

4.3 Term Frequency-Inverse document Frequency TF-IDF

The TF-IDF algorithm assesses the importance of a word within a document or corpus. When a term or phrase appears frequently in one document but rarely in others, it is considered valuable for classification as it can effectively differentiate between classes. Initially proposed by **[52,64]** TF-IDF is a statistical measure widely used in information retrieval and natural language processing. It evaluates the significance of a specific word in a document relative to the entire corpus. The underlying principle of TF-IDF is that a term that is common in a single document but uncommon across the corpus carries greater meaning than a word that is common in both the document and the corpus. TF-IDF relies solely on word frequency and does not consider the contextual meaning of the words.

Term frequency: It indicates how frequently a term appears throughout the document. It might be interpreted as the likelihood of discovering a term within the material. It computes the number of times a term *"t"* appears in a document "*d*", relative to the total number of terms in the document *"d"*,. The formulation is given in equation 4.1, as follows **[64] :**

$$f_{t,d} = \frac{No.of\ times\ t\ \ occurs\ in\ d}{Total\ no.of\ terms\ in\ \ d} \quad \text{.....} \quad (4.1)$$

Where, $f_{t,d}$ is the frequency of the term $\boldsymbol{t}$ in document $\boldsymbol{d}$.

For e.g. if a term "**t**" occurs **30 times** in a document having **300 total words**, then Term Frequency of "**t**" can be calculated as

$$TF = \frac{30}{300} = 0.1$$

Inverse Document Frequency (IDF): The inverse document frequency is a measure of how unusual or common a term is across all documents in the corpus. It emphasizes words that appear in very few documents across the corpus, or in plain terms, uncommon words have a high IDF score. IDF is a log-normalized value obtained by dividing the total number of documents 'D' in the corpus by the number of documents containing the term 't' and taking the logarithm of the overall term **[64]**.

$$idf(d, D) = log \frac{|D|}{\{d \in D : t \in d\}} \quad \text{.....(4.2)}$$

$|D|$ is the total number of documents in the corpus, $d \in D : t \in d$ is the count of documents in the corpus, which contains the term ***'t'***. Since the ratio inside the IDF's log function must always be larger than or equal to one, the value of IDF (and consequently TF-IDF) is greater than or equal to zero. When a term appears in a large number of documents, the logarithmic ratio approaches one and the IDF approaches zero.

For e.g, if there are total 3000 documents where only 300 documents contain the term "t", then Inverse Document Frequency (IDF) of the term 't' can be calculated as

$$IDF = log \frac{3000}{300} = 1$$

Term Frequency - Inverse Document Frequency (TF-IDF) is the combination of TF and IDF.

$$tfidf(t, d, D) = tf(t, d) \times idf(d, D) \quad \text{.....(4.3)}$$

For the example mentioned above, TFIDF will be calculated as,

$$tfidf = 0.1 \times 1 = 0.1$$

A phrase with a high frequency in a document but a low frequency in the corpus has a high TF-IDF score. For a word that appears in practically all documents, the IDF value approaches zero, bringing the TF-IDF closer to zero. The TF-IDF value is high when both IDF and TF values are high, which indicates that the term is rare throughout the document yet frequent within it. Let us use one example to better understand this:

1. Data science is interesting
2. Data science is excellent
3. Excellent topic

Each sentence in this example represents a separate document. Following table shows the calculations of TF-IDF.

Table-4.1 Calculations of TF-IDF

SL No.	Data	Science	Interesting	Excellent	Topic
1	(1/4)×log(3/2) =0.044	(1/4)×log(3/2) =0.044	(1/4)×log(3/1) =0.119	0	0
2	(1/4)×log(3/2) =0.044	(1/4)×log(3/2) =0.044	0	(1/4) ×log (3/2) =0.044	0
3	0	0	0	(1/2) ×log(3/1) =0.238	(1/2)×log(3/1) =0.238

It is seen that the term 'topic' and ‘interesting’ as rare (i.e. appears in only one document) as other terms, and so has a higher TF-IDF score. If a word is used repeatedly and appears in every document, it will eventually have a higher frequency of occurrence and a greater

value, which will make that word appear more frequently in sentences and be adverse to our analysis. Through the normalization of words that are frequently used in the collection of documents, TF-IDF aims to represent a word's significance to its document or phrase.

4.4 Methodology

This research approach to TF-IDF (Term Frequency-Inverse Document Frequency) extraction from documents is based on the standard procedures that computers utilize to model language. The following steps are used to assess a document's word importance.

4.4.1 Pre-processing

Data pre-processing involves various steps as this prepares the input raw textual data for further analysis. In other words, processing of raw data involves the conversion of the texts into single words or phrases. Additionally, removal of stop words, special characters etc which hold no meaningful context to the analysis, also needs to be done in the pre-processing steps.

Tokenization of the text data is done using the tokenization function available in the Natural Language Tool Kit (NLTK) library of Python package **[65]**. This includes converting the text data into single words known as tokens. The tokenized data is then further used for the next phase of pre-processing i.e. stop word removal. Stop words, or words that appear frequently in a language but usually have little value in the context of text analysis, are usually removed during the preprocessing step of natural language processing tasks. As there is no pre-defined function in the python programming environment for removing stop words, special characters etc. from the Assamese language, attempts were made to create a new function that could perform the aforementioned pre-processing task. The purpose of building the function is to remove

stop words from an Assamese text document. The function helps to simplify the text data and gets it ready for additional processing or analysis by eliminating stop words. The function reads the ‘stop word’ file to obtain the list of stop words. The stop words that have been used in this work has been collected on request from the Indian Language Technology Proliferation & Deployment Centre (TDIL-DC), for the Assamese language. The input text is tokenized during preprocessing and then linked with the user collected stop words list. Words which are included in the list of collected stop words are eliminated from the text and the words which are then left are tokenized for additional handling. This procedure aids in making sure that, during further analysis, only those terms are considered which holds meaningful contexts.

4.4.2 Computation of Term Frequency (TF)

For computing Term frequency, as discussed in the previous section, user-built function is used. The objective of building this function is to calculate the term frequency (TF) of words from pre-processed data. Term frequency counts how often a word is seen to appears in a given document from all the terms in the document. A word's TF is determined by using the equation no 4.1. Each word in the document undergoes this computation, producing a dictionary with every word linked to its matching term frequency value. Finally, the function returns this dictionary as its output, which contains the TF values for every word.

4.4.3 Computation of Inverse Document Frequency (IDF)

As explained earlier, IDF calculates a word's significance over a corpus of given documents. The inverse document frequency (IDF) of a word in a list of documents is determined by using the function named IDF. A list of documents, each represented as a dictionary of word frequencies, is fed into the function. It starts by setting all values to zero and initializing an IDF dictionary with keys derived from the words in the first document. The IDF dictionary is then updated by iterating through each document in the list and increasing the number of documents that include each term. The function first counts the document frequency for each word and then divides the total number of documents (N) by the word's document frequency to get the IDF value for each word. The equation no 4.2 is used for calculating the IDF.

4.4.4 Computation of TF-IDF

The TF-IDF score is calculated after the term frequency (TF) and inverse document frequency (IDF) for each word in the document is obtained. The TF score, which indicates how frequently a word appears in the document, is multiplied by the IDF score, which reflects the word's relevance across all texts in the corpus. The method of calculating the TF-IDF score is as simple as multiplying a word's Term Frequency score by its matching IDF value. For every word in the text, this procedure is carried out again. The generated TF-IDF scores are then kept in a different dictionary, where each word has a corresponding TF-IDF score that has been calculated. In conclusion, the TF-IDF score indicates a word's relative importance within a given document compared to the entire corpus. This method weights often occurring words that may not have as much semantic

relevance to help locate terms that are unique or extremely significant to a particular article. The following simple formula is used to compute the same

$$\textbf{TF-IDF} = \textbf{TF} \times \textbf{IDF} \quad \text{.....(4.4)}$$

TF-IDF (Term Frequency-Inverse Document Frequency) is a method used to measure the importance of a term i.e a word in a document relative to its occurrence across a corpus. In a given example of a movie review containing 1000 words with the term "Awesome" appearing 10 times, the TF value is found as 0.01. Considering the term "Awesome" appears 1000 times in a given corpus of 1 million reviews, the IDF value is approximately 3. Multiplying the TF and IDF values yields the TF-IDF value of 0.03. This value signifies the significance of the term "Awesome" within the specific review compared to its occurrence across the entire corpus **[66]**.

4.5 Results and discussion

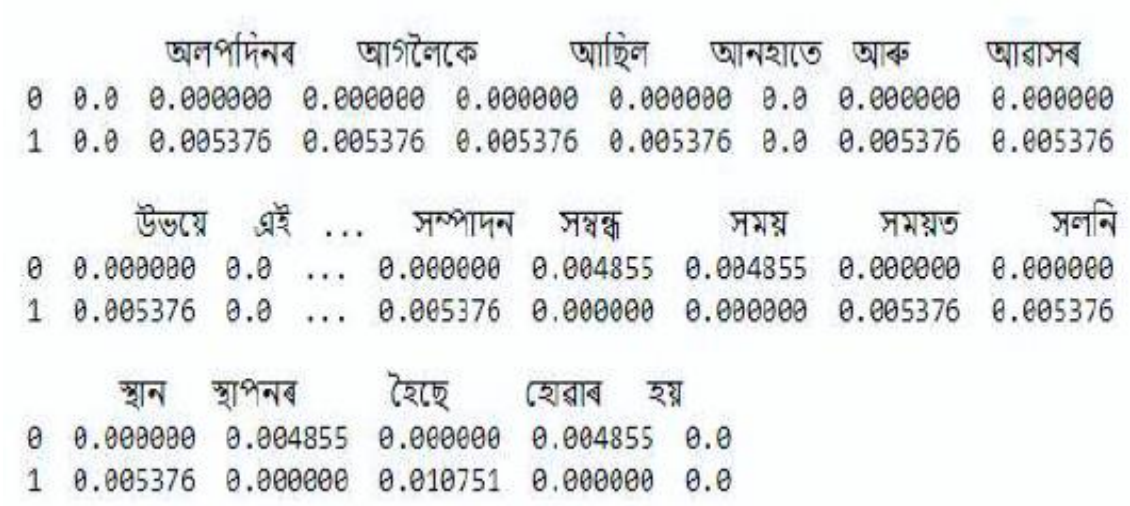
The functions implemented here in the experimental work are designed to handle a different type of tasks, including pre-processing and computing TF-IDF values, with inputs tailored to each specific task. These functions are engineered to produce calculated values as outputs, which are then utilized in subsequent stages of basic NLP tasks. The experimental methodology is carefully structured, beginning with the analysis of smaller text units such as sentences and gradually scaling up to larger units such as paragraphs and entire documents. The textual data used for experimentation is sourced from the Assamese General Text Corpus, which was obtained from the Indian Language Technology Proliferation & Deployment Centre (TDIL-DC). Specifically, the sentences under examination are extracted from an Assamese non-fiction text titled 'Axomiya Uponyasor Bhumika', which is part of the aforementioned corpus. The experimental findings are

presented through a series of figures 4.2 to 4.5. It is noted that as the size of the dataset increases, the execution time of the program also increases. However, the results obtained are satisfactory, confirming the program's ability to handle an arbitrary number of documents and compute the TF-IDF values of the words contained therein, demonstrating its scalability.

	অসম	কথা	কি	খায়	গছৰ	গোটা	ঠাই	দেশ
0	0.000000	0.0	0.000000	0.0	0.033448	0.0	0.000000	0.033448	0.033448
1	0.033448	0.0	0.033448	0.0	0.000000	0.0	0.033448	0.000000	0.000000

	দেশৰ	পাত	পাতে	বিষম	মানুহ
0	0.000000	0.000000	0.033448	0.0	0.033448
1	0.033448	0.033448	0.000000	0.0	0.000000

Figure-4.2 TF-IDF value for two sets of sentences



	অলপদিনৰ	আগলৈকে	আছিল	আনহাতে	আৰু	আৱাসৰ
0	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000
1	0.0	0.005376	0.005376	0.005376	0.005376	0.0	0.005376	0.005376

	উভয়ে	এই	...	সম্পাদন	সম্বন্ধ	সময়	সময়ত	সলনি
0	0.000000	0.0	...	0.000000	0.004855	0.004855	0.000000	0.000000
1	0.005376	0.0	...	0.005376	0.000000	0.000000	0.005376	0.005376

	স্থান	স্থাপনৰ	হৈছে	হোৱাৰ	হয়
0	0.000000	0.004855	0.000000	0.004855	0.0
1	0.005376	0.000000	0.010751	0.000000	0.0

Figure-4.3 TF-IDF value for words of two consecutive paragraphs (taken from one file)

	''পদুম	'অ'ল্ড	'অৰুণদই'ত	'অৰুণদই'ৰ	'আইভানহো'ৰ	'আলালেৰ
0	0.00000	0.00005	0.00005	0.00005	0.0001	0.00005	0.00005
1	0.30103	0.00000	0.00000	0.00000	0.0000	0.00000	0.00000

	'আসাম	'আসামবন্ধু'	'এলোকশী	...	ৰূপেই	ৰূপৰ	ৰূপৰেখাৰপৰা	\
0	0.00005	0.00005	0.00005	...	0.00005	0.00005	0.00005	
1	0.00000	0.00000	0.00000	...	0.00000	0.00000	0.00000	

	ৰেখাঙ্কন	ৰেখাপাত	ৰেবেকা	ৰেভাৰেণ্ড	ৰোগ	ৰোগগ্রস্ত,	ৰোগত
0	0.00005	0.00005	0.00005	0.00005	0.00005	0.00005	0.00005
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Figure-4.4 TF-IDF value for the words of one file

	''The	''এই	''খূর্ত	''গোৰক্ষ	''তহঁতৰ	''দৰিদ্র	\
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.000059	0.000029	0.000029	0.000029	0.000029	0.000029	0.000029

	''পদুম	''বুচু''	''মৃচ্ছকটিকম''	...	ৰেভাৰেণ্ড	ৰৈ	ৰোগ	\
0	0.00005	0.000000	0.000000	...	0.00005	0.000000	0.0	
1	0.00000	0.000029	0.000029	...	0.00000	0.000029	0.0	

	ৰোগগ্রস্ত,	ৰোগত	ৰোমাঞ্চকৰ	ৰোহসেনে	ৰোহসেনৰ	র্ভতৃ	ৱাৰিয়ে
0	0.00005	0.00005	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.00000	0.00000	0.000088	0.000029	0.000059	0.000029	0.000029

Figure-4.5: TF-IDF value for the words of two different files

The results that are obtained are encouraging and could be used to a number of Natural Language Processing (NLP) applications, such as document categorization and sentiment analysis. It is remarkable to notice, though, that TF-IDF values of some words are observed high in conjunction with an increase in documents, while other words show a different value. This finding clarifies the presence of particular terms that occurs frequently in documents, leading to greater TF-IDF scores.

This chapter presents a new approach to TF-IDF feature vector computation that is customized for texts written in Assamese. This method attempts to enable more complex Natural Language Processing (NLP) tasks, such as sentiment analysis, document categorization, and text summarization, through experiments carried out on an Assamese corpus file. The suggested methodology allows for additional NLP tasks by computing TF-IDF features for Assamese texts. TF-IDF vectors offer important insights into the relevance of particular phrases in various contexts by capturing the significance of words within documents about the overall corpus.

TF-IDF features is computed for texts written in Assamese allow for the pursuit of numerous NLP applications. Sentiment analysis, for example, can be used to identify the sentiment expressed in Assamese documents, which can help with interpreting social media sentiments, consumer feedback, and public opinion. Overall, the suggested method for calculating TF-IDF feature vectors for texts written in Assamese opens up possibilities for a range of complex NLP tasks, improving the data analysis and interpretation of Assamese language including the usage of text data for machine learning algorithms. The input textual data written in Assamese can be conveniently used as vector forms, by the various machine learning classifiers to perform any kind of advanced NLP tasks. Machine learning based sentiment analysis is discussed in the next chapter, where input data is vectorized using TFIDF feature extraction techniques

5

Sentiment Categorization of Assamese Textual Data Using Supervised Machine Learning Approach with Combined n-gram and TF-IDF Feature

A supervised machine learning based sentiment analysis of textual data is presented in this chapter. Sentiment categorization is considered as the major applications in the field of text analysis. It involves detecting the state of mind or viewpoint conveyed in textual material, including whether a piece of text contains neutral, positive, or negative sentiments. Text analysis or text classification, is a fundamental process in natural language processing (NLP) that requires labeling or classifying texts according to their content. Product reviews, social media monitoring, customer feedback analysis, political analysis, and brand monitoring are just a few of the fields in which sentiment analysis

finds use. It helps organizations in understanding feedback from customers, identifying common issues, and enhancing the consumer satisfaction. Sentiment analysis basically helps businesses to better understand consumer opinions from textual data and make data-driven decisions to improve goods and services. Sentiment analysis of review data using machine learning involves employing algorithms to examine and classify text reviews according to the sentiment they express. Sentiments are categorized as positive, negative and neutral using different machine learning techniques. Assamese, a language primarily spoken in northeastern India, has been used as the input textual data in this research study. Since, benchmark datasets in this language are limited and not usually available, (in contrast to other Indian languages like Hindi, Bengali, Kannada, Tamil, Telugu, etc.) the dataset used in this study was created by translating available Bengali resources into Assamese using Google Translator. For the sentiment analysis, a range of supervised machine learning methods are used, such as Support Vector Machine, Decision Tree, Multinomial Naive Bayes, K-nearest neighbor and Logistic Regression. Term Frequency-Inverse Document Frequency (TF-IDF) (as explained in chapter 4) and n-gram are the type of feature extraction techniques which are used to process the data. The experimental results obtained show that the Unigram model outperforms higher order n-gram models. Additionally, Multinomial Naive Bayes and Support Vector Machine produced results with above 90% accuracy in the sentiment analysis of Assamese textual data collected from different domains. These findings demonstrate that the proposed model is effective for sentiment categorization in Assamese textual data, irrespective of domain specialization.

A brief **Literature review on related work** on machine learning based sentiment analysis is provided in **section 5.1**. **Proposed SA using Supervised ML** is mentioned in **section**

5.2. **Methodology** used to implement the proposed work is explained in **section 5.3**. Using Assamese language dataset, **proposed model for SA in Assamese language** is elaborated in **section 5.4**. **Results and analysis** along with **conclusion** this chapter is illustrated in **section 5.5 and 5.6** respectively.

5.1 Literature review on related work

As early as 2000, sentiment analysis was mainly developed for the use of computational linguistic research **[4]**. When sentiment analysis was first developed, researchers used primarily written documents. However, the increase in online database that contain texts, tweets, blogs, news stories, reviews, and so on has prompted researcher working on sentiment analysis to use fast computational techniques to do the same. Various methods are employed to fulfill the purpose, including NLP, statistical analysis, and machine learning techniques **[67].** People use websites and social media to talk about different things, like movies, books, news, politics and business and share their thoughts and opinions.

It can be difficult to manually process enormous amounts of data for analysis. This is where sentiment analysis, becomes a very useful tool. SA research in natural language processing (NLP) has advanced to the point that it can now analyze text material methodically by automatically identifying and categorizing the sentences that are contained in the text. Researchers and analysts can process large amounts of textual data quickly and effectively with SA by using machine learning algorithms and statistical techniques. This gives the researchers an insights into the attitudes, sentiments and opinions of individuals or groups towards different topics, products, services, or events. Consequently, SA has developed into a vital tool for companies, scholars, and

organizations looking to collect insightful information from huge text collections. Sentiment analysis, also called opinion mining, investigates people's feelings and thoughts about specific entities **[68-78]**. It finds applications in various other areas, such as analyzing and predicting user opinions, devising marketing strategies, and forecasting stock market trends **[76]**.

As technology and digital platforms continue to advance rapidly, there has been a growing emphasis on analyzing large databases and text repositories through the expertise of natural language processing and sentiment analysis. Extracting sentiments from this huge amount of information requires the ability to effectively assess and interpret data in order to gain insight into consumer attitudes and opinions **[72].** Several languages, including Chinese **[77,78]**, Spanish **[79]** French **[80]** and numerous Indian languages, such as Bengali **[81]**, Hindi **[82]**, Tamil **[83]**, Telugu **[84],** Malayalam **[85,86]** etc., have seen a large number of sentiment analysis studies carried out in them.

The research on sentiments analysis (SA) in Assamese is very less, although it being a widely spoken language in northeastern India. Furthermore, there are very limited resources available in Assamese to be used for research in this domain. Sentiment analysis in low-resource languages like Assamese suffers greatly by this lack of resources, mostly because of the insufficient availability of datasets. As a result, there is comparatively very little research done specifically on the analysis of sentiment in Assamese text data, keeping the subject of sentiment analysis in Assamese underdeveloped. A domain-specific sentiment classifier needs a lot of labelled data to be trained in order to produce predictions that are correct. This extensively labeled data set requires a lot of resources and time to create **[87]**. The necessity to assess variety of opinions in social media, consumer reviews, and other text sources has called for the rise in popularity of sentiment

analysis in various regional and vernacular languages. The recent trend in these studies have been effectively conducted using conventional machine learning techniques.

In recent years, research on sentiment analysis of review data has gained popularity. This section has covered related work on sentiment categorization. Numerous contexts, such as product reviews, movie reviews, and customer feedback reviews, have all been researched in relation to sentiment categorization **[88,82]**. The majority of these studies have up till now concentrated on teaching machine learning algorithms to categorize reviews **[89]**. The authors of **[89]** conducted an experiment in which they divided attitudes into two categories: positive and negative, using a dataset of movie reviews. The n-gram technique has been utilized to perform SA utilizing three machine learning techniques: NB, SVM, and DT. It is found that NB performs worse than the other two algorithms when it comes to identifying the sentiments expressed in the texts. Authors in **[90]** have used SVM to classify sentiment in a dataset of movie reviews. Here, SA using a unigram-based feature model has been performed and the results are compared to previous works using k-fold cross-validation with k = 3,10. Researchers in **[91]** used SVM to perform SA on movie reviews while taking context valence shifters into account. It is found that including context valence shifters increases the system's accuracy. To increase the number of words in the term count method, authors have also used General Inquirer and Choose to Right Word. They implemented unigrams, bigrams, and adjectives as features for the machine learning algorithm and achieved an 86.2% accuracy using context valence shifters. In **[92]**, authors used three different machine learning classifiers NB, SVM and the n-gram model to perform SA on travel blog reviews. Using three-fold cross-validation in the experiments, it is observed that SVM and n-gram models outperform NB. In **[93]** authors used blogs, forums, and online reviews in English, French, and Dutch to conduct

sentiment analysis (SA). Using unigrams as features, they implemented machine learning techniques like Maximum Entropy (ME), SVM, and Multinomial Naive Bayes, with reported accuracy for English, French, and Dutch using unigrams as features of 83%, 68%, and 70%, respectively. Using Artificial Neural Network (ANN) and SVM machine learning methods, the authors of **[94]** carried out document-level sentiment classification on product and movie reviews. Here, sentiment is divided into two classes as positive and negative feelings using a bag-of-words approach. Sentiment categorization was carried out in **[95]** by authors using Twitter data sets, including the Obama-McChain Debate, SentiStrength Twitter data set, Sanders, and SemEval. To do sentiment analysis, they employed four well-known machine learning algorithms: Naïve Bayes (NB), Decision Tree, K-Nearest Neighbor (kNN), and Support Vector Machine (SVM). n-gram, Medical Research Council (MRC), Linguistic Inquiry and Word Count (LIWC), Apache OpenNLP toolkit, and Stanford Part-Of-Speech (POS) tagger have all used for text-to-feature vector conversion.

Tweet SA was carried out by **[96]**, where information acquisition, diagrams, and object-oriented extraction approaches make up the feature set for the application of NB and SVM. Researchers in **[97]** performed SA on tweets at the global and aspect levels in the Spanish language. A graph- based algorithm to extract the features and polarity lexicons to determine the sentiments has been used in this paper. The authors of **[98]** talk about using sentiment analysis to analyze Hindi movie reviews. Additionally, using a dataset of Bengali tweets, the authors in **[99]** developed a Bangla sentiment analysis model. Using Multinomial Naive Bayes and SVM classifiers using a feature set that included N-gram and SentiwordNet features, they classified the sentiment polarity conveyed in tweets. The created model showed that on the Bengali tweet dataset, an SVM classifier trained with

unigram and Sentiwordnet features outperformed other classifiers. In order to produce a favorable and negative review, aspect-based mining finds the element of a sentence and the user's opinions on these aspects **[100]**. The authors in **[100]** provide a novel approach where they created TF-IDF vectors for Assamese text. Within the news arena, authors in **[101]** created a sentiment polarity classification model for the low-resource Assamese language. Lexical elements including verbs, adverbs and adjectives are included in their model. An improved F1-score on the benchmark dataset indicates that the generated model performs better than the baseline. Due to the increasing need to understand feelings found in social media, consumer reviews, and other text sources, SA in Assamese is in high demand, just like it is in any other language. The focus of researcher **[102]** is doing SA on texts written in Assamese. Their study builds on the "Bengali-Vader" framework to modify the popular sentiment analyzer "Vader" into Assamese.

5.2 Proposed SA using Supervised ML

In order to begin this research project, there is a prerequisite of a corpus of labeled Assamese data. An empirical investigation of sentiment analysis on collected labeled Assamese text data from different web resources, is presented in this research work which can enable further sentiment analysis. In this study, a supervised learning model is developed to classify text data into positive and negative sentiments, specifically written in Assamese. For sentiment analysis, traditional machine learning (ML) techniques are used because of their adaptability and capacity to support extensive testing and analysis. This study uses a variety of traditional machine learning classifiers, including Random Forest (RF), k-nearest neighbors (kNN), Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT). Using the best possible combination of TF-IDF and n-grams for feature extraction from the labelled texts, these classifiers are trained and tested on

labeled datasets. Thus, the machine leaning algorithms are enabled to make accurate sentiment polarity predictions from the given texts collected from two different domains.

5.2.1 Classification techniques

For sentiment classification, this research has employed a variety of machine learning algorithms namely SVM, MNB, DT, KNN and RF. The various classification algorithms utilized in this research are briefly explained in the following sections.

A. Decision Tree (DT)

A supervised learning method for both regression and classification is decision tree analysis. It divides data iteratively until the leaf nodes have the fewest records by hierarchically dividing it according to conditions placed on its properties. Classification is then based on the class label that has the majority in the leaf node **[84]**. For both classification and regression tasks, this is a very effective supervised learning techniques. It constructs a tree structure that resembles a flowchart, with each internal node indicating a test on an attribute, each branch designating a test result, and each leaf node (terminal node) containing a class name. It is a powerful supervised machine-learning approach that may be applied to regression and classification issues equally. It is among the most efficient algorithms **[28,103]**.

B. Logistic Regression (LR)

Estimating a logistic model's parameters, like the coefficients in the linear combination, is done using logistic regression, commonly referred to as logit regression. Binary logistic regression makes up this kind of regression analysis, in which the independent variables might be continuous or binary values. As "0" and "1" in turn, the dependent variable is

also a binary indicator. Because the logistic function turns the log-odds into a probability, the associated probability of the "1" value will therefore fall between 0 and 1**[104,105].**

C. Multinomial Naïve Bayes (MNB)

One of the most fundamental and popular classification models in machine learning is the Naive Bayes classifier. Based on the fundamental concepts of the Bayes Theorem, it performs under the strong basis of feature independence, which streamlines the calculation process without compromising useful outcomes **[106]**. A method for updating probability in the context of new information is provided by the Bayes theorem. It is expressed as:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \ldots\ldots\ldots\ldots (5.1)$$

Where, $P(A|B)$ refers to the probability of event A given that event B has occurred.

$P(B|A)$ refers to the probability of event B given that event A has occurred.

The probabilities of events A and B are denoted by P(A) and P(B), respectively.

Using the assumption that terms occur independently, Multinomial Naïve Bayes (MNB) is a probabilistic classifier based on the Bayes theorem. When using Bayesian classification, a hypothesis is formed that a given set of data is part of a sentiment analysis system for a particular class, and the likelihood that the hypothesis is correct is evaluated **[106]**.

If a review data contains N number of sentences, where each sentence contains 'T' number of terms such that $S_i=\{t_1, t_2, t_3 \ldots\ldots\ldots\ldots t_T\}$, where S_i indicate the collection

of N sentences. Now the probability of S_i to occur in class C_k is calculated by the equation given below:

$$P(C_k|S_i) = P(C_k)\prod_{j=1}^{T} P(t_j|C_k) \ldots\ldots\ldots\ldots\ldots\ldots (5.2)$$

Here, $P(C_k|S_i)$ refers to the conditional probability of the term t_j to occur in a sentence of class C_k , and $P(t_j|C_k)$ refers to the prior probability of a sentence occurring in a class C_k . $P(t_j|C_k)$ and $P(C_k)$ are being calculated from training data.

D. Support Vector Machine (SVM)

The basic idea behind SVMs is to find the linear separators in the search space that may most effectively divide the various classes **[106]**. SVMs are the best option since text is sparse, meaning that only a small number of features are important but are frequently connected with one another. Additionally, text categories are usually structured in such a way that it facilitates linear separability. Thus, SVMs can process text data with ease, extracting relevant characteristics and correlations and classifying data into groups using the best possible hyperplanes **[3]**. Support Vector Machines are widely used in text classification due to their ability to function in high-dimensional spaces, relevance to all features, robustness in sparse sample sets, and linear separation for most problems. SVM outperforms other machine learning algorithms for determining people's opinions from text **[107]**. A hyperplane serves as the SVM's decision surface in linearly separable space. Text file reviews must be translated to numeric values before being classified using SVM, which requires input only in numerical vector forms. After converting the text file to a numeric vector, a scaling step may be performed to help manage the vectors and maintain them within the range of [1, 0].

E. K-Nearest Neighbor (kNN)

The sentiment analysis task may be assigned to the kNN (k-Nearest Neighbors) classifier because of its simplicity in binary classification tasks and its performance with enormous data sets. Sentiment analysis is a problem of binary classification in which text is classified as either positive or negative. By comparing fresh data points with existing ones, the kNN algorithm determines a label based on the majority of the closest neighbors. To train the classifier to differentiate between positive and negative feelings, a manually generated training set is employed. With all factors taken into account, the kNN classifier seemed an appropriate choice for sentiment analysis in this case due to its versatility, ease of use and efficacy in binary classification tasks. It could also manage massive volumes of data.

F. N -gram model

The continuous sequence of n items from a particular text sample is known as an N-gram. Based on the application, the elements may be base pairs, words, or letters **[108].** It is a technique for verifying "n" successive words or sounds from a given text sequence. This model facilitates in the prediction of the succeeding entity in a series. The n-gram model in sentiment analysis aids in determining the text's or document's sentiment. Unigram, bigram, trigram and so on are examples of higher n-grams where n=1,2,3 and so on respectively **[109]**

An example of a typical statement would be "I am an NLP researcher."

• Its unigram: ‘I’, ‘am’, ‘an’, ‘NLP’, ‘researcher’, where a single word is considered.

• Its bigram: ‘I am’, ‘am an’, ‘an NLP’, ‘NLP researcher’, where a pair of words are considered.

• Its trigram: ‘I am an’, ‘am an NLP’, ‘an NLP researcher’, where a set of words having a count equal to three is considered

5.3 Methodology

In this section, the sentiment classification technique utilized in this study is elaborated. The dataset comprises text reviews from two different domains which are movies reviews and restaurant reviews. However, to employ machine learning algorithms for sentiment classification, text reviews need to be converted into numerical matrices. This involves extracting Unigram, Bigram, and Trigram features from cleaned texts. The TF-IDF vectorizer is then used to transform each text document into a numerical vector. These vectors are then fed into supervised machine learning algorithms for sentiment classification, requiring a labeled input dataset.

This study employs various supervised machine learning classifiers which includes Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), Multinomial Naïve Bayes (MNB) and K Nearest Neighbor (kNN). The performance of these classifiers is evaluated using statistical measures such as Confusion Matrix, precision, recall, and F1 score. This comprehensive analysis enables the identification of the most effective classifier for sentiment classification in the dataset **[66]**.

Confusion Matrix:

The parameter of confusion matrix, used to evaluate the performance of the proposed model is shown in Table 5.1.

Table-5.1: Confusion Matrix Parameters and their meanings

TP	It represents the reviews which are positive and the classifier also classifies the reviews as positive
FP	It represents reviews which are positive, but classifier classifies the reviews as negative
TN	It represents the reviews which are negative and the classifier also classifies the reviews as negative
FN	It represents the reviews which are negative and the classifier also classifies the reviews as positive

As shown in Table 5.1, a comparative analysis of labels of different classes is done using four different terms which are, true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Based on the values of parameters obtained from the confusion matrix, other performance measuring attributes of the classifier such as Accuracy, F-measure, precision, and recall are calculated as given in Table 5.2.

Table-5.2: Performance evaluation parameters for Machine Learning classifiers

Performance measure	Description	Formula
Accuracy	Accuracy measures the fraction of appropriately classified samples.	$\frac{TP + TN}{TP + FP + TN + FN}$
Precision	It is described as the proportion of Truely predicted positive cases.	$\frac{TP}{TP + FP}$
Recall	It represents the proportion of correctly identified positive cases.	$\frac{TP}{TP + FN}$
F1 score	It is given by harmonic mean and computes the average of precision and recall.	$\frac{2 \times precision \times recall}{precision + recall}$

5.4 Proposed model for SA in Assamese language

Figure 5.1 below shows the operation of a proposed SA system for the Assamese language. It is divided into the following Sections: 'Data collection', 'Data preparation and Data Cleaning', 'Feature Extraction', 'Classification' and finally 'Model Testing'.

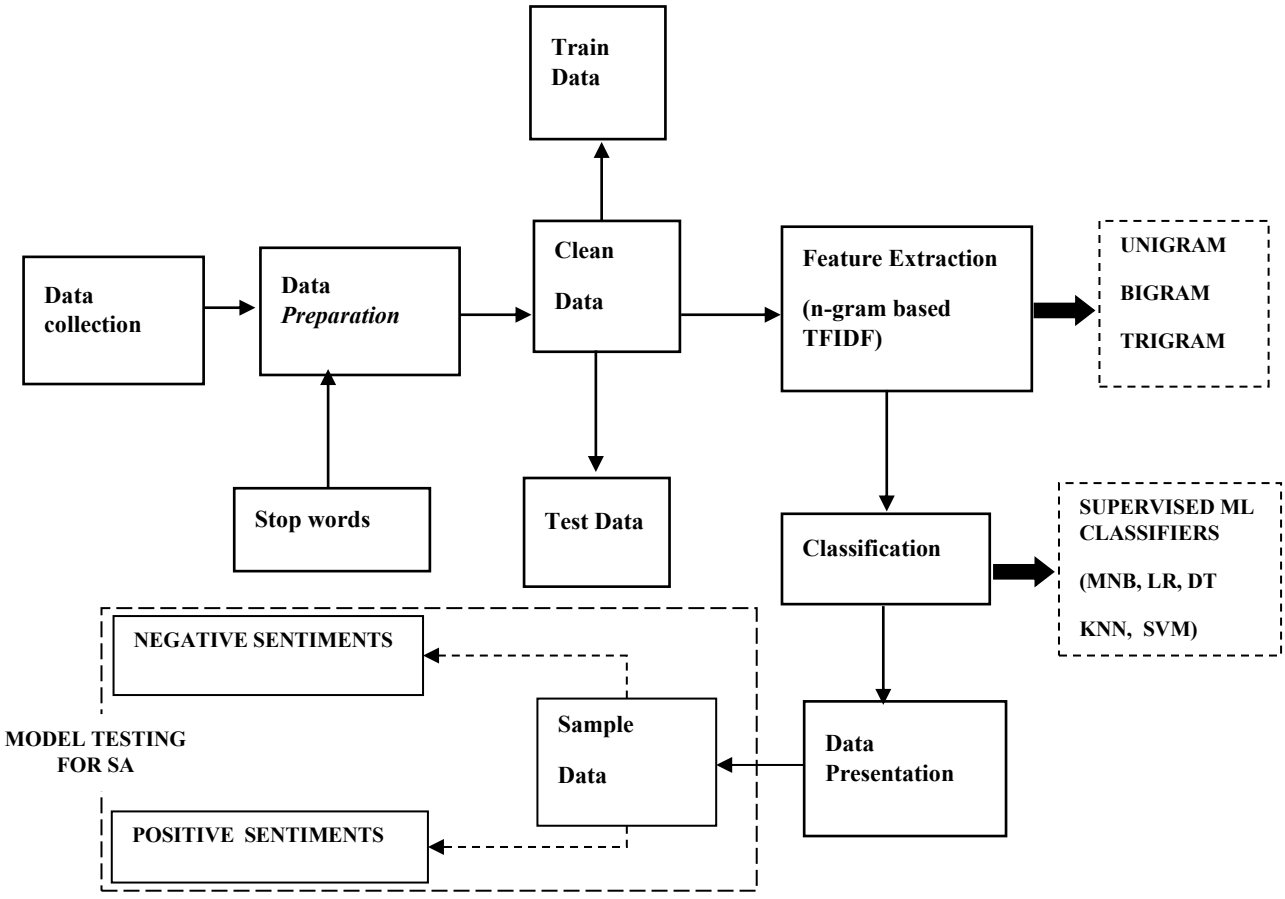


Figure-5.1: Proposed model of sentiment analysis

Step:1 Data Collection:

Because of the absence of a benchmark dataset, research in the field of Assamese language NLP is currently facing significant obstacles. Recently, it has been noted that to achieve their individual objectives, researchers are collecting data from different online sources. Additionally, researchers are utilizing translation tools to create Assamese

datasets from existing standardized benchmark datasets which are currently available in a range of other languages. In this research study, an Assamese corpus was prepared by employing the Google Translate to available Bengali datasets. In this work initially, a labelled dataset of Bengali restaurant reviews of around 1400 reviews are collected from diffrent social media groups where customers provide food and restaurant reviews **[110,111]**. Also, a labelled data set of around 2,000 movie reviews **[112,113,114]** is collected for this presented research work which is then translated through Google translator.

Step:2 Data Preparation & Data Cleaning

In this step data are pre-processed and cleaned for further processing. The pre-processing step involving tokenization, stop word removal, removal of punctuations etc. have been done. Thus, pre-processed data are cleaned by removing unnecessary symbols, tokens and numbers from the review texts.

Let (say) a training set $R = \{r_1, r_2, r_3, ..., r_n\}$ contains 'n' number of training reviews. Here, each review contains either positive or negative sentiment, denoted by the letters C_p and C_n respectively. A word vector $W[] = \{w_1, w_2, w_3, ..., w_l\}$ represents a review ' r_i ' with 'l' number of words. To remove inconsistencies from the dataset, all reviews are preprocessed. To remove the words 'w_i', which have no contribution in determining whether a review 'r_i' conveys positive (Cp) or negative (Cn) sentiment, a stop words list $S[] = \{s_1, s_2, s_3, ..., s_t\}$ with 't' stop words has been developed. Removal of stop words from a review is done by removing stop words $s_1, s_2, s_3, ..., s_t$ which are in the stop words set "S". Conjunctions, prepositions, interjections, pronouns, suffixes, and prefixes are all

examples of stop words. Some sample stop words in the Assamese language are given in Table 5.3.

Table- 5.3: Sample list of stop words

Stopwords	Type	examples
s_1	Pronoun	আপুনি
s_2	Preposition	ওপৰত
s_3	Interjection	বাহ্
s_4	Conjunctions	আৰু

A sample of cleaned data is shown below:

Movie Dataset:

Original: নাটকখন ভাল লাগিল, দৰ্শকক এনেকুৱা সুন্দৰ নাটক দিয়াৰ বাবে ধন্যবাদ।

Cleaned: নাটকখন ভাল লাগিল দৰ্শকক সুন্দৰ নাটক ধন্যবাদ

Here, stop words and punctuations like 'এনেকুৱা', 'দিয়াৰ' and 'বাবে' respectively which are removed from the original review.

Restaurant Dataset:

Original: কেৱল খাদ্যই বৰ ভালেই নহয় ঠাইখনো ভাল আৰু দৃশ্যটোও আচৰিত। দৃশ্যটো মোৰ বৰ ভাল লাগে।

Cleaned: কেৱল খাদ্যই বৰ ভালেই নহয় ঠাইখনো ভাল দৃশ্যটোও আচৰিত দৃশ্যটো বৰ ভাল

Here, stop words like ‘আৰু’, ‘মোৰ’, ‘লাগে’, respectively are removed from the original review. The final dataset prepared thus consists of two columns namely ‘Reviews’ and ‘Sentiment’. Where, reviews are listed under ‘Reviews’ column and sentiments are listed under ‘Sentiment’ column labelled as positive and negative. Sample dataset template is shown in Table 5.4 and Dataset statistics is summarized in Table 5.5.

Table- 5.4: Dataset Template

Index	Reviews	Sentiment
1	Review 1	Positive
2	Review 2	Positive
:	⋮	⋮
N	Review N	Negative

Table-5.5: Dataset Summery

Sl. No.	Reviews	Total Reviews	No. Reviews		No of words		Unique words		No. Documents	
			Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
1	Movie dataset	1990	977	1013	6432	6416	1470	2007	974	995
2	Restaurant dataset	1401	637	764	5669	14262	1449	3984	561	745

An 80-20 ratio was used to split the dataset into training and testing sets, with 80% of the data used for training the model and 20% for evaluating its performance. This method ensure that the model can generalize its performance on unknown data and has enough data to learn from.

Step:3. Feature selection methods:

Natural Language Processing (NLP) tasks such as texts analysis, retrieval of important information opinion mining etc. uses various feature selection methods. Among various feature selection methos Term Frequency and Inverse document frequency (TF-IDF) is a commonly used one.TF-IDF method is capable of estimating the relevance of a given word to that of a given document by calculating the probability of appearance of a word in the same document.

Basically, the higher value of TF-IDF for a word indicates the greater importance of the same word in the text **[100]**. In this work, reviews of the corpus are tokenized to prepare a vocabulary of Assamese terms.

"TF-IDF": The suggested model uses n-gram and frequency-inverse document frequency statistics as features. Here, used “TF-IDF” statistic is written as

$$tf\ idf\ (w,r) = tf(w,r) \log N|\{r€R : w€r\}| \ldots\ldots (5.3)$$

Here, $tf\ idf\ (w,r)$ = value of word ‘w’ in review ‘r’.

$tf(w,r)$ = frequency of word ‘w’ in review ‘r’.

N = total number of reviews.

$|r€R : w€r\}|$ = number of reviews containing ‘w’.

Step:4. SA using Machine Learning:

This research study has considered account various machine learning techniques, including MNB, SVM, and DT, LR, k-NN to classify the sentiments of reviews written in Assamese reviews of restaurants and movies are two distinct categories of data utilized

for the same purpose. Since the goal of the suggested model is to categorize the reviews, classification is the most significant component of the system. The retrieved features from the reviews are utilized to train the suggested model, which can classify reviews as positive or negative.

In this research work, number of experiments is done on textual reviews collected from the movie and restaurant genre. They are then evaluated using various traditional machine learning classifiers namely MNB, SVM, LR, DT, kNN. Using three different n-gram models (unigram, bigram, and trigram) on a review dataset, the experiments in this work aimed to examine the performance of machine learning classifiers in sentiment analysis tasks. In this work, we efficacy of machine learning classifiers in sentiment analysis tasks is investigated by conducting tests on a review dataset using three distinct n-gram models: unigram, bigram and trigram. Unigrams, which stand for single words, are the simplest type of textual data. By identifying word co-occurrences, bigrams, which take into account pairs of consecutive words and trigrams which take into account sequences of three consecutive words offer further context. Observations are made on how efficacy and accuracy of sentiment analysis are affected by different levels of textual context, by comparing the performance of classifiers trained with these various n-gram models. The purpose of the study is to determine if adding more intricate n-gram models would substantially improve the capacity of the classifiers to predict sentiment in the review data.

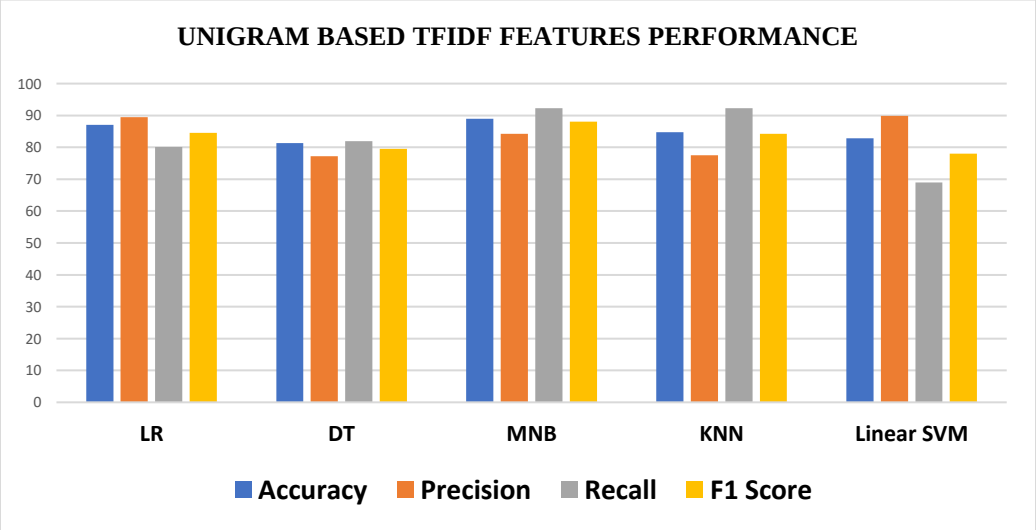
5.5 Results and analysis

The proposed model has also been tested with all the mentioned machine learning classifiers. Comparative analysis of these classifiers in terms of various performance evaluation parameters is discussed in the following sections

Classifier Report for N-gram based TFIDF Feature extracted model:

The following graphs depict the performance of different classifiers on (i) Restaurant Review and (ii) Movie Review datasets using unigram-based TFIDF features respectively.

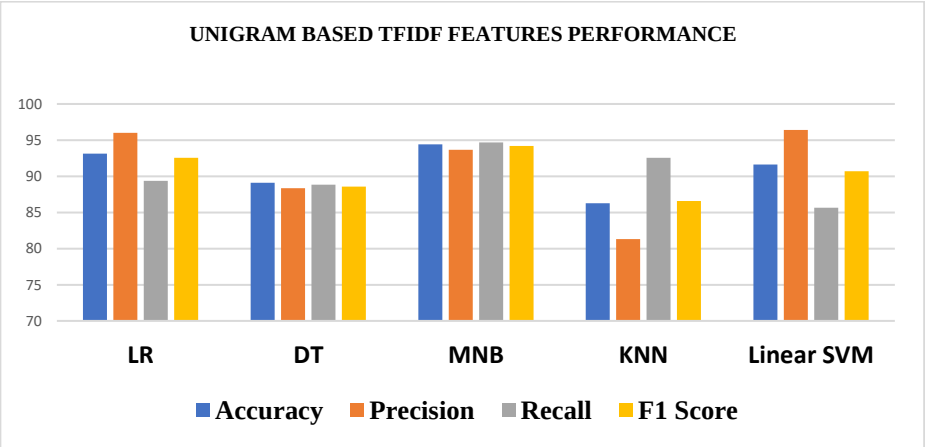
(i) Restaurant Review dataset:



Classifier	Accuracy	Precision	Recall	F1 Score
LR	87.02	89.42	80.17	84.55
DT	81.3	77.24	81.9	79.5
KNN	84.73	77.54	91.24	83.83
MNB	**88.93**	84.25	92.24	88.07
Linear SVM	82.82	89.89	68.97	78.05

Figure – 5.2 Unigram-based TF-IDF features performance for restaurant dataset

Movie Review Dataset:



Classifier	Accuracy	Precision	Recall	F1 Score
LR	93.15	96	89.36	92.56
DT	89.09	88.36	88.83	88.59
MNB	**94.42**	93.68	94.68	94.18
KNN	86.29	81.31	92.55	86.57
Linear SVM	91.62	96.41	85.64	90.7

Figure 5.3: Unigram based TF-IDF features performance for movie dataset

Observation: It is observed that out of the classifiers, MNB and SVM classifiers give the better performance. Classifier reports of MNB and SVM are illustrated in Table-5.6 for both datasets. It shows that as the dataset size increases the accuracy increases, showing the efficacy of both the classifiers. Also, higher value of F1 score indicates better classification, the values obtained show that MNB and SVM have performed better as compared to the rest of the classifiers for both datasets. Confusion matrix of MNB and SVM classifiers have been observed for the validation of correct predictions of the

sentiments for both datasets as shown in figures 5.4 to 5.7 below. Figure 8 shows the performance of the MNB and SVM classifiers for different N-gram features, where N=1, 2, 3.

Table-5.6 Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM) Classifier report

CLASSIFIER	DATASET	ACCURACY	PRECISION	RECALL	F1-SCORE
MNB Classifier	RESTAURANT (1400 reviews)	0.89	0.90	0.89	0.90
	MOVIE (2000 reviews)	**0.93**	0.97	0.88	0.92
SVM Classifier	RESTAURANT (1400 reviews)	0.82	0.89	0.69	0.78
	MOVIE (2000 reviews)	**0.91**	0.96	0.86	0.90

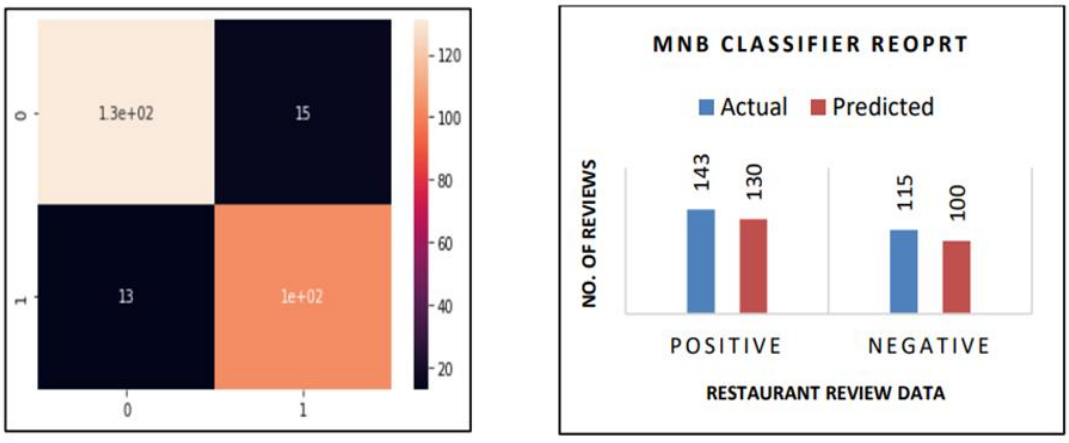


Figure 5.4: Confusion matrix evaluation of MNB classifier for Restaurent data

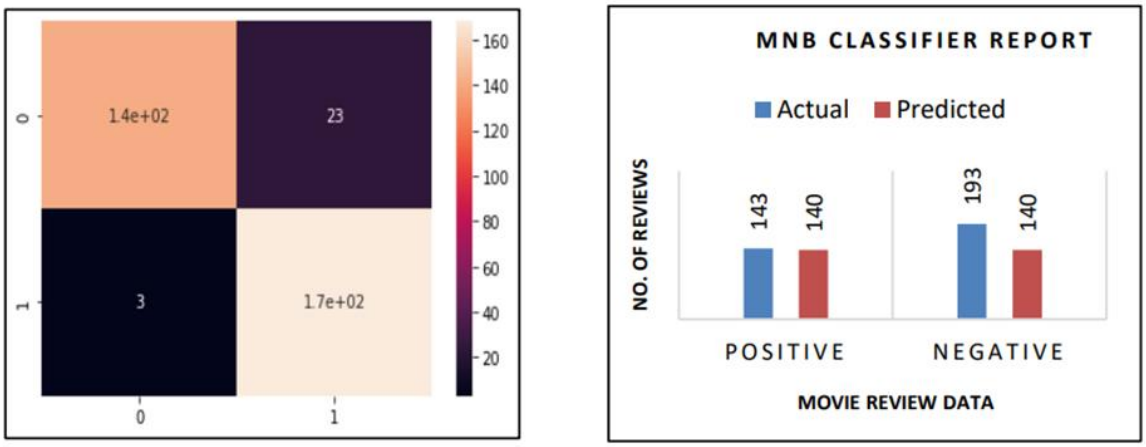


Figure 5.5: Confusion matrix evaluation of MNB classifier for Movie data

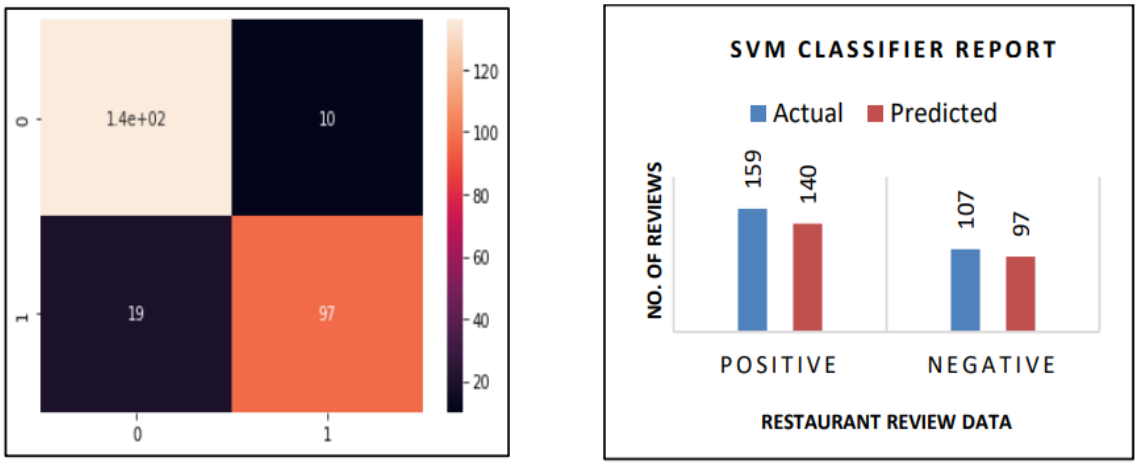


Figure 5.6: Confusion matrix evaluation of SVM classifier for Restaurent data

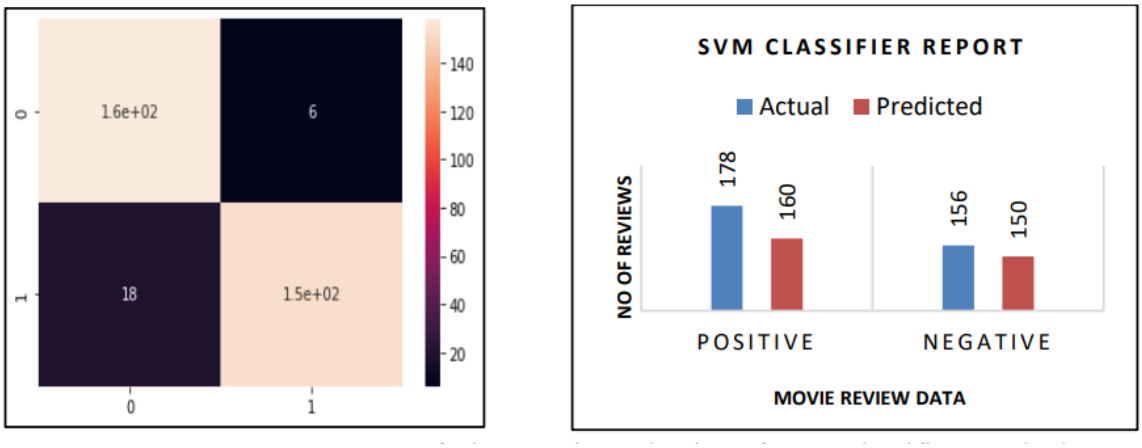


Figure 5.7: Confusion matrix evaluation of SVM classifier for Movie data

As illustrated in the results shown in figures 5.2-5.7 above, the following observations are obtained:

The **Multinomial Naïve Bayes** classifier achieves the **highest accuracy**, with an **average of over 85%** (88% for restaurant data and 94% for movie data). Support Vector Machine (SVM) also performs well, with an accuracy of over 80% (83% for restaurant data and 92% for movie data). In terms of the "**Precision**" assessment parameter, **Support Vector Machine (SVM) outperforms** all other classifiers, **scoring 89%** for restaurant data and **96%** for movie data. The **unigram-based TFIDF** feature extraction method had the **highest accuracy** for both classifiers. For both datasets utilized in this study, the model demonstrated equal performance in terms of machine learning classifier and feature extraction methodologies. The experimental design shows that **Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM)** classifiers employing **Unigram-based TFIDF** feature extraction can achieve **above 80% accuracy** for Assamese text data, regardless of domain.

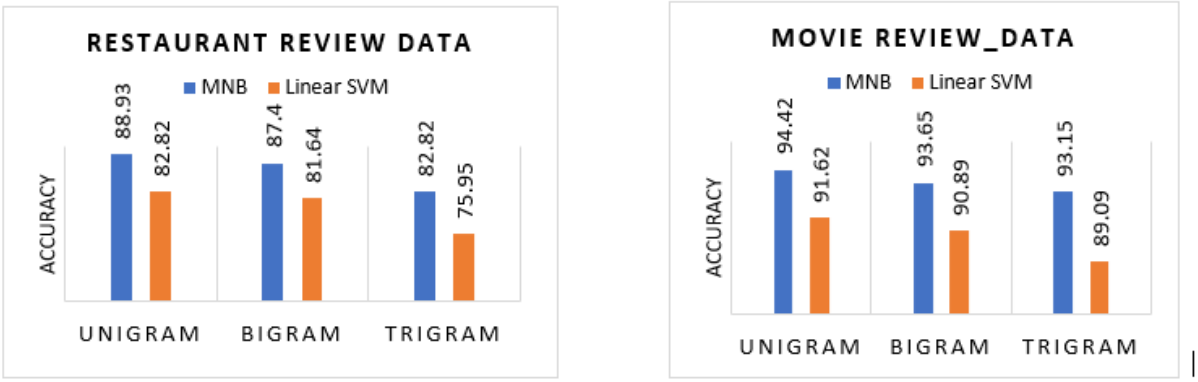


Figure 5.8: Comparative analysis of N-gram feature-based methods for MNB and SVM Classifier

Results Validation: An approach for evaluating the effectiveness and potential for development of machine learning models is cross-validation. Cross-validation techniques, namely K-fold cross-validation, were employed to enhance the assessment of machine learning classifiers in the experiments carried out for this research.

K-fold cross validation with K = 4, 6, 10 has been used here. Tables 5.7 and 5.8 examine the performance of these methods across different K values. According to the findings in the cross-validation report for the K values, MNB and SVM outperform the rest of the classifiers on both datasets. In the validation accuracy test, the classifiers achieve more than 85% accuracy across both datasets used in this experiment. Finally, all of the classifiers employed in this study's experimental design were tested on a small number of randomly selected sample reviews.

To assess the scope and constraints of the developed model, commonly annotated restaurant and movie reviews were used, along with all possible common comments posted on internet. The proposed algorithm correctly predicted sentiment categorization

using these review texts from both datasets. Furthermore, to validate the model's performance, the same sample texts were examined with an existing standard Natural Language Tool Kit (NLTK) sentiment analyzer. Table 5.9 provides a quick overview of our comparative investigation. Figure 5.9 shows the comparison results of our proposed model, NLTK and human sentiment interpretation on Assamese text data. This demonstrates the efficiency of the suggested model in the sentiment analysis of Assamese.

Table-5.7 Cross validation report for restaurant data

K-VALUE	CLASSIFIER	ACCURACY	PRECISION	RECALL	F1-SCORE
4	MNB	0.89	0.83	0.99	0.89
	SVM	0.90	0.90	0.95	0.90
6	MNB	0.91	0.85	0.99	0.91
	SVM	0.91	0.91	0.95	0.91
10	MNB	0.89	0.83	0.99	0.90
	SVM	0.90	0.90	0.92	0.91

Table-5.8 Cross validation report for movie data

K-VALUE	CLASSIFIER	ACCURACY	PRECISION	RECALL	F1-SCORE
4	MNB	0.86	0.85	0.88	0.87
	SVM	0.82	0.89	0.71	0.79
6	MNB	0.87	0.87	0.87	0.87
	SVM	0.84	0.91	0.76	0.82
10	MNB	0.88	0.87	0.89	0.87
	SVM	0.85	0.91	0.77	0.83

Table- 5.9 Comparative analysis of sample texts with NLTK and Human Score

Sl. No.	SAMPLE TEXTS	CLASSIFIER	SENTIMENT CLASS		
			OUR MODEL	NLTK	HUMAN OUTPUT
1	Assamese: খাদ্য বেয়া , পৰিৱেশ ভাল English: Food is bad, atmospheres is good	MNB	NEGATIVE	NEGATIVE	NEGATIVE
		LINEAR SVM			
2	Assamese: খাদ্য ভাল , পৰিৱেশ বেয়া English: Food is bad, atmospheres is good	MNB	POSITIVE	POSITIVE	POSITIVE
		LINEAR SVM			
3	Assamese: যদি আপুনি এটা সাধাৰণ সোৱাদ ভাল পায় ,এয়া এটা বিকল্প হ'ব পাৰে English: If u like a simple taste, it may be an option	MNB	NEGATIVE	NEGATIVE	NEGATIVE
		LINEAR SVM			
4	Assamese: ভাল সেৱা পাইছিলোঁ, খাদ্য সুন্দৰকৈ পৰিবেশন কৰা হৈছিল English: Got good service, food was nicely served	MNB	POSITIVE	POSITIVE	POSITIVE
		LINEAR SVM			
5	Assamese: চিনেমাৰ চিত্রনাট্য আছিল অতি উত্তম, এইটো এখন সুন্দৰকৈ চিত্রিত কৰা চিনেমা হৈছে English: The script of the movie was excellent; it is a beautifully portrayed movie	MNB	POSITIVE	POSITIVE	POSITIVE
		LINEAR SVM			
6	Assamese: চিনেমাখন বৰ সাধাৰণ আছিল,	MNB	NEGATIVE	NEGATIVE	NEGATIVE

	চিনেমাৰ অভিনেতা বেয়া আছিল English: movie was very average; movie cast was bad.	LINEAR SVM			

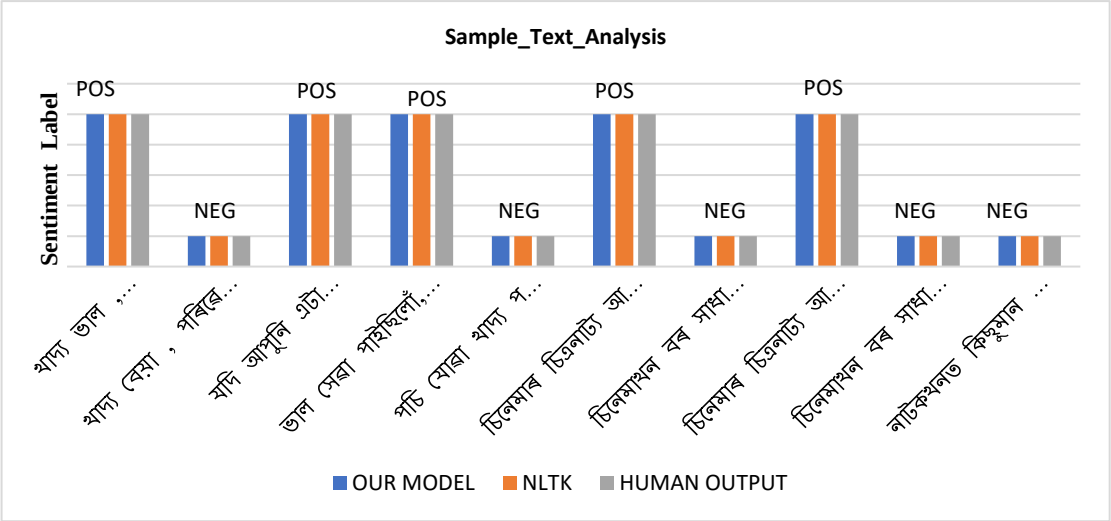


Figure-5.9: Comparison of sentiment labels for sample texts

In this presented study, five different machine learning classifiers are considered to test a proposed model for sentiment analysis of restaurant and movie reviews. Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) classifiers regularly outperform other classifiers, with accuracy rates above 85% across a range of dataset sizes. The model's effectiveness is proved through precise sentiment analysis of test texts from restaurant and movie reviews written in Assamese-language, which is compared to the Natural Language Toolkit sentiment analyzer and human interpretation. The model's efficacy is further demonstrated via cross-validation and comparison with other sentiment analysis approaches. It is observed that for sentiment classification, mostly Support

Vector Machine (SVM) and Multinomial Naïve Bayes (MNB) perform better than other classifiers.

5.6 Conclusion

This chapter elaborates about the main research objective of this presented study which attempted to predict sentiment in two distinctive categories i.e positive and negative, by utilizing randomly generated review data from the domains of movies and restaurants. This research represents the first endeavor to analyze Assamese text data in such a manner. Upon evaluating the model's performance with randomly selected reviews from both domains, encouraging results were observed.

Presented experimental methodology confirms that the utilization of sophisticated feature extraction methods and the incorporation of additional machine learning approaches have the scope to enhance the accuracy and precision of the classification process. Furthermore, the same research has focused on investigating deep learning models and expanding the considered dataset to larger samples sourced from various domains such as social media, product reviews, and news. Through these efforts, advancement in the understanding of sentiment analysis in Assamese text and predictive capabilities of proposed models is presented in next chapter.

6

Sentiment Analysis of Assamese Text: A Deep learning Approach

Deep learning (DL) methods, inspired by artificial neural networks **[115] [116]**, are now widely used in natural language processing. A deep neural network has two main layers: input and output, with additional hidden layers in between. A neural network is similar to the biological brain, consisting of neurons arranged in layers that work together to process information **[117]**. It may be trained to do tasks like classification or predictive modeling by altering the weights of its neurons. Deep learning (DL), a subset of machine learning simulates how the human brain interprets and uses data. It learns from large amounts of data to match inputs with specified labels instead of relying on human-made rules. Deep learning enables computational models with multiple processing layers to learn and represent data at various levels **[118]**.

Deep Learning has shown promising results for natural language processing (NLP) applications like sentiment analysis (SA) **[119]**. The primary idea behind Deep Learning approaches is to use deep neural networks to identify complicated properties retrieved from massive amounts of data with minimal external involvement. Deep learning algorithms for sentiment analysis are becoming increasingly popular. They offer automatic feature extraction, as well as greater representation capabilities and higher performance than classic feature-based techniques **[120]**.

The volume of online content distributed via social media platforms has increased significantly as information and communication technology (ICT) has advanced rapidly. This expansion has generated a significant interest among researchers from various sectors, including academia, government, and private enterprise, in the field of sentiment analysis based on online assessments. Notably, sentiment analysis has developed as a key area of research in the interdisciplinary domains of Machine Learning (ML) and Natural Language Processing (NLP). The capacity of deep learning algorithms to provide excellent results has led to their recent popularity in this field. These algorithms, which are at the most advanced stage of research and development, are being used to improve the effectiveness of sentiment analysis approaches.

Assamese, a language with limited resources yet rich in morphology, has not yet benefited much from the advancements of NLP research. Therefore, this study aims to provide a novel perspective on deep learning based sentiment analysis. The experimental outcome based on deep learning techniques has reflected that in spite of Assamese which is less explored as well as lacks significant resources for NLP researchers, it has given effective results.

This chapter presented an experimental investigation where Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN) along with a hybrid model combining LSTM-CNN, are used for performing sentiment analysis in Assamese text reviews. The effectiveness of these models is tested and compared to classic machine learning methods using Assamese text review datasets as examples. The results found in the experimental investigation shows that the suggested deep learning models perform better in sentiment analysis, emphasizing their potential in identification of sentiments polarity in the

considered review dataset. An accuracy of more than 95% is achieved with all the proposed deep learning models.

6. 1 Literature review on related work

Presented research work attempts to determine if a review for a given entity holds positive or negative sentiment. Generally, entity may be a product, services related to business, entertainment etc. (*like movie, restaurant, hotels etc.*). The proposed work has taken movie reviews as the entity to investigate users' sentiment for the same. In this context, a variety of approaches have been used to obtain the results. Various researchers have done various different level of computational investigations in this field. Deep learning in this regard has evolved into a sophisticated machine learning system capable of learning multiple layers of data representations or features and producing cutting-edge prediction outcomes **[121,122,123]**. Deep learning has seen many uses in sentiment analysis in the last several years, expanding on its successes in different application areas**[124]**. Deep learning is a subset of artificial neural networks that acquires data using numerous layers for learning. These neural networks act similarly to neurons in living creatures, processing data in a way that mimics biological brain networks **[125]**. It has shown enormous potential in NLP applications, such as SA **[126].**

This study presents a deep learning-based sentiment categorization approach for textual reviews written in Assamese language. User feedback can significantly impact customer satisfaction. However, evaluating each review individually is a time-consuming process. Furthermore, administering such surveys necessitates a significant commitment of both financial and human resources. To keep up with the growing number of visitors and user preferences, an automated system is needed to understand the contextual polarity of

reviewer comments across many platforms. The authors in **[127]** created a total of 8435 Bengali annotated corpuses for sentiment analysis using deep learning techniques, where they found BiLSTM performs well with an accuracy of 91.35%. Assamese language has distinct cultural and linguistic characteristics which makes it different from other languages, also making sentiment analysis task challenging for this language. Furthermore, because there is no standard collection of data for this type of research, it opens up opportunities for new and creative research work in this field. The absence of a standardized dataset and limited existing research in sentiment analysis for the Assamese language present a unique opportunity for pioneering advancements in this field. By addressing this gap, researchers can introduce innovative methodologies and groundbreaking approaches.

In this context, the primary objective is to develop a novel framework for sentiment analysis specifically tailored for Assamese language using deep learning methodologies. The development of a robust sentiment analysis framework customized for Assamese has the potential to enhance understanding and interpretation of sentiment dynamics within this linguistic community, opening up new paths for targeted and effective communication strategies and decision-making processes. Opinion mining for sentiment analysis in qualitative information is proving to be a emerging field of study for many researchers **[128,129]**.

Sentiment analysis is currently the most popular type of approach for measuring consumer sentiment and viewpoints in text and it serves as a foundation for developing innovative models. The research field of sentiment analysis encompasses a various different range of technical aspects, and is an emerging area of research that merges knowledge fusion through machine learning. It holds substantial potential as a research

topic in the realm of artificial intelligence **[130,131]**. The study by authors in **[132]** investigates the function of emotions in online discourse, specifically for Bangladeshi users who communicate their feelings in Bangla. It provides four models that combine different Word Embeddings with Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) approaches. The top-performing model improves text-based interactions on social media platforms by integrating Word2Vec embedding with a hybrid CNN-LSTM architecture. It achieves an amazing accuracy of 90.49% and an F1 score of 92.83%. Authors in **[133]** examines sentiment analysis of tweets written in Bengali, focusing on the classification of tweet sentiment polarity into positive, negative, or neutral. Various deep learning techniques like LSTM, BILSTM, and CNN are compared for their effectiveness in categorizing emotional tones. The study also compares traditional machine learning approaches with deep learning techniques to determine which method performs better in analyzing and classifying sentiments in Bengali tweets. Overall, aim of the research is to enhance performance of sentiment analysis in Bengali text data available on social media. The usefulness of different types of machine learning approaches, deep learning approaches and hybrid model based approaches, in text classification for the English and Bangla languages is examined in **[134]**. Main aim of the study is sentiment analysis and it does so by examining the feedback and comments on the well-known e-commerce site "DARAZ" made in Bengali. Seven machine learning models and several deep learning models namely, Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), Convolutional 1D (Conv1D), and a combination Conv1D-LSTM, are implemented as one part of the research technique. With an average accuracy of 82.56% for sentiment analysis in English text and

that of 86.43% for Bangla text when uses the Porter stemming algorithm, Support Vector Machine (SVM) models outperform other models, according to the major findings.

The efficiency of sentiment analysis models trained on a single language on multilingual tweets is investigated in **[135]** The work makes use of a multilingual twitter dataset as well as experimental results from CNN, RNN, and combination CNN-RNN models. CNN produces the best results, with an accuracy of 85.91% and an F1-score of 84.61%. The model also performed well on unseen tweets in European languages other than the original languages of the tweets used for training.

Several techniques and technologies were employed by researchers in **[136]** to analyse Bengali text, including Glove word embedding, the Adam based optimizer, a deep learning classifier based on Convolutional Neural Networks (CNN) and Glove-BiLSTM. The mentioned study demonstrated a high accuracy of 99.43% in evaluating Bengali texts.

In **[137]** authors explore the use of deep neural networks, specifically Long Short-Term Memory (LSTM), for sentiment analysis in text written in Hindi. It uses word embeddings and fine-tunes parameters to achieve accurate sentiment classification in Hindi. The study provides highlights on the importance of sentiment analysis in Hindi for organizations in India, providing useful insights for a product, service etc.

The authors of **[138]** propose a sentiment classification system for tweets using deep learning techniques. The system categorizes positive emotions into subcategories like enthusiasm and happiness, while negative emotions are categorized into anger, boredom, and sadness. The method was tested using Recurrent Neural Networks and Long Short-Term Memory on three datasets. The results shows higher accuracy, with the LSTM model achieving 88.47% for positive/negative classification and 89.13% and 91.3% for

positive and negative subcategories, respectively. This research contributes to emotion recognition using deep learning techniques.

The researcher of **[139]** explores the challenges faced by under-resourced languages like Sinhala in using deep learning methods for sentiment analysis in Natural Language Processing (NLP) applications. It uses old and new sequence models, a dataset of 15,059 Sinhala news comments, and a 9.48 million token corpus, making it the largest sentiment-annotated Sinhala dataset available.

The paper **[140]** presents a novel approach combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for sentiment analysis of short texts. The model, which captures coarse-grained local features and learns long-distance dependencies, outperforms existing methods on benchmark datasets, achieving accuracy rates of 82.28%, 51.50%, and 89.95%, demonstrating its effectiveness in handling texts with limited contextual information.

Latest research work in SA in different languages using various technology with different types of datasets is outlined in **Table 6.1.**

Table 6.1 Related Recent Literature on Sentiment Analysis Using Online Text Reviews

Sl. No.	Authors	Domain	Language used	ML Techniques
1	Sezgen et al. (2019) **[141]**	Airline	English	Latent semantic analysis
2	Banerjee and Chua (2016) **[142]**	Airline	English	Statistical analysis
3	Kumar and Zymbler (2019) **[143]**	Airline	English	SVM, ANN, CNN
4	Ye et al. (2014) **[144]**	Hotel	English	LR
5	Chen et al. (2019) **[145]**	Hotel	English	LSTM
6	Dey et al. (2016) **[146]**	Hotel	English	NB, KNN
7	Lee et al.(2018) **[147]**	Hotel	English	DT, NB, SVM
8	Guo et al. (2017) **[148]**	Hotel	English	Latent Dirichlet analysis
9	Siering et al. (2018) **[149]**	Airline	English	NB, SVM, NN
10	Rehman et al. (2020) **[150]**	Airline	English	LR, DT, NB, SVM, LSTM, CNN, CNN-LSTM
11	Shrivastava et al.(2020) **[151]**	Movie Review	Hindi	LSTM, GA-GRU
12	Bhowmik et al(2022)**[152]**	Cricket review	Bengali	LSTM
13	Chowdhury et al. **[153]**	Movie review	Bengali	SVM LSTM
14	Dev et al.**[154]**	Restaurant and movie reviews	Assamese	SVM, NB, LR, DT, KNN

Considering the limited research on sentiment analysis in low-resource languages like Assamese, this work contributes in the following ways:

- Creation of dataset of reviews for use in deep learning.
- Develops deep learning systems namely LSTM, CNN and hybrid model LSTM-CNN.
- The dataset is tested and trained using these models.
- Evaluates how well the built models perform.
- Compares the built deep learning models effectiveness to classical machine learning approaches.

The rest of the chapter is ordered into several sections. **Section 6.2** presents **theoretical background** of the deep learning models used for the proposed work followed by the discussion of **methodology** of the proposed work in **section 6.3. Section 6.4** discusses about the **results** obtained on proposed model and its comparison with existing counterparts. Finally, **section 6.5** concludes with **future directions** for research and acknowledges the limitations of this study.

6.2 Theoretical background

6.2.1 Long Short-Term Memory (LSTM)

Recurrent or very deep neural networks can be challenging to train because they frequently come across an issue where the gradients expand excessively large or small, which makes learning difficult. In order to address this issue with long-term patterns, scientists developed the LSTM model **[155]**. This model's remarkable ability to learn from data sequences, both in theory and in practical applications, has had a significant influence

on numerous fields **[156][157]**. LSTM is a deep neural network model which is an improved recurrent neural network (RNN) that can understand order dependence. The data can be stored in LSTM model for a very long time. LSTM is basically used for the processing, prediction and for classification of time-series data **[158]**. LSTM is intended to produce longer-term dependencies more effectively than a typical RNN **[159]**. It is a powerful class of neural networks. It is supervised to learn from labeled samples and is well-known for efficiently compressing text in natural language. Its feedback mechanism modifies weights, or internal settings, to help it learn from errors and gradually become more efficient. LSTM's memory units are made for long-term dependencies, which helps it handle sequential data well, such as values over time or words in a sentence**[160]**. Figure-1 depicts the typical LSTM architecture. There are three inputs out of which two originate from the prior state; one is $\mathbf{h_{t-1}}$ which is a hidden state from the earlier time step, which originated from an earlier LSTM, and the other is $\mathbf{C_{t-1}}$, a memory that originated from an earlier time step. The most important part of the structure is the gate, which consists of two gates: the memory gate and the forget gate.

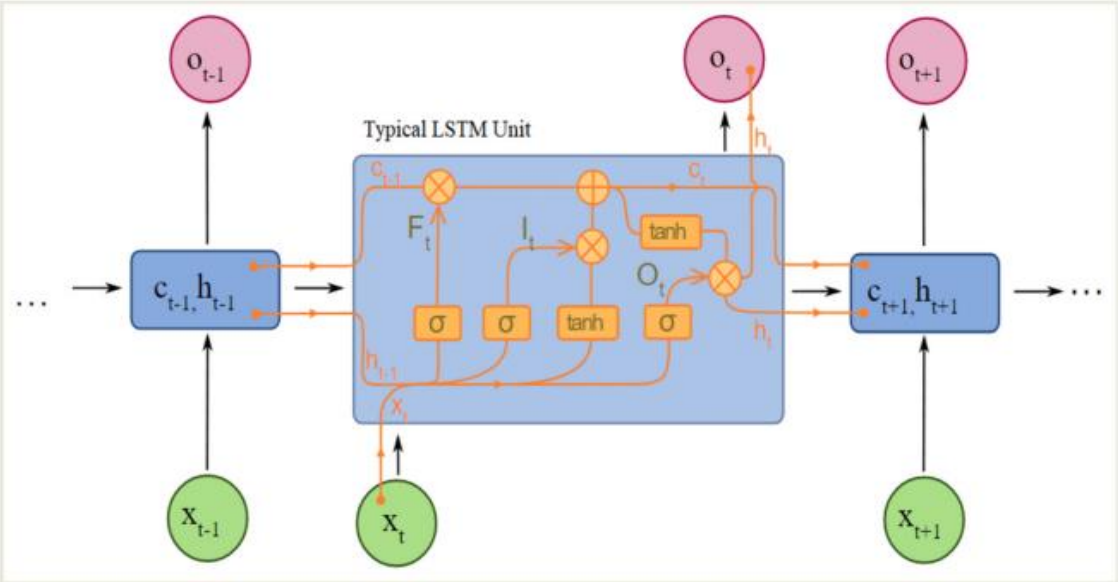


Figure- 6.1: Internal Structure of LSTM model

The equation 6.1 for LSTM can be described as follows:

$$
\begin{aligned}
I_t &= \delta\left(w^i x_t + u^i h_{t-1}\right) \\
F_t &= \delta(w^f x_t + u^f h_{t-1}) \\
O_t &= \delta(w^o x_t + u^o h_{t-1}) \\
C_t &= tanh(w^c x_t + u^c h_{t-1}) \\
C_t &= i_t \underline{\odot} \tilde{c_t} + f_t \underline{\odot} \tilde{c}_{t-1} \\
h_t &= O_t \underline{\odot} \tanh\left(\tilde{c}_t\right)
\end{aligned}
\qquad \textbf{(6.1)}
$$

Here, I_t represents the input gate, F_t represents forget gate, O_t represents the output gate, δ represents the sigmoid function, $w^i, w^f, w^o, w^c \; and \; u^i, u^f, u^o, u^c$ represent the weight matrices and x_t is the vector input to timestep t, h_t current is exposed hidden state and C_t represents memory cell state and '$\underline{\odot}$' represents the element-wise multiplication. In this research work proposed LSTM model uses 128 units of neurons.

6.2.2 Convolution Neural Network (CNN)

Neural networks that are particularly good at processing spatial information are called CNN. CNN gives better performance in classifying and segmenting image and video data. However in recent years, CNN has become more effective at NLP tasks. CNN performed well on a number of text classification tasks that used in **[161]**.

There are basically four layers in a standardised model of CNN **[162]**:

1. ***Convolutional layers***: They are created by moving a window across a text and then successively applying the same convolutional filter to each window. The

dimension of the convolutional window is '**s × d**', where '**s**' denotes the region's size in a filter matrix, while '**d**' denotes the dimensionality of the input data **[121]**.

2. ***Pooling layers***: They are used for down sampling. It reduces the input parameters by doing dimensionality reduction. In order to accomplish pooling process, initial selection of a window is done and a pooling function is used to apply to the input items contained within that window. An additional output vector is produced by the pooling function. There are a few pooling strategies available, such as max-pooling and average pooling, the most popular one is max pooling, which considerably reduces map size **[163]**.
3. ***Fully connected layers***: This layer is analogous to the fully connected network in the conventional models. The output of the first phase, which typically includes pooling and convolution layers, is fed into the fully connected layer and the final output is obtained by computing the dot product of the input vector and the weight vector **[164]**.
4. ***Activation Function****:* An appropriate activation function is capable of enhancing the performance of CNN optimized for any specific type of application **[165]**. Essentially, activation functions act as the "switch" in artificial neurons, determining whether or not to activate the neuron in response to the input's weighted sum. This is similar to how neurons function in the human brain, they may fire or may not fire. This biological comparison helps in better understanding of how activation mechanisms work inside a CNN **[166]**.

6.2.3 LSTM-CNN (The Hybrid model)

An LSTM and a CNN system are included in the proposed model. The input to the model receives is the word embedding vector as applied to text reviews. The word embedding

strategy was employed with CNN and LSTM models since it improves performance for NLP tasks by representing words as vectors **[167].** As deep learning algorithms require numerical input hence, we employ Keras (a word embedding technique) is employed to create word embedding, which transforms the text data into vector values. In word embeddings, if two words are found to have similar contexts, they tend to have similar meanings, leading to comparable vector representations. For instance, the terms "dog," "puppy," and "pug" are frequently used in comparable contexts and will have vector representations in common with words like "cute," "fluffy," or "bite." These vectors are fed into the suggested CNN and LSTM.

6. 3 Methodology

The sentiment analysis of Assamese text review data which are collected from various sources is the main emphasis of this presented research project. Here, a hybrid LSTM-CNN technique, CNN and LSTM are used in a classification model. The effectiveness of the recommended models in contrast to traditional machine learning models is determined through a thorough analysis of the outcomes produced by the aforementioned deep learning models. Figure 6.2 below shows the proposed work model.

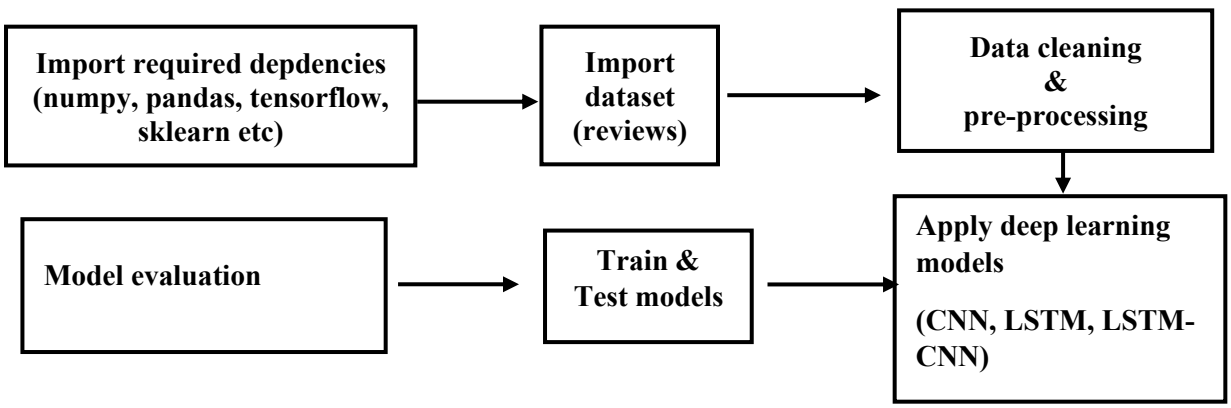


Figure -6.2: Proposed work plan

The research presented here utilizes Python as the mainstream programming language to compute sentiments from a dataset of Assamese reviews. To fulfill the mentioned objective, various dependencies available with latest python version such as Pandas, NumPy, TensorFlow, matplotlib and scikit-learn etc. has been incorporated. Overall proposed plan of presented work has been accomplished through different steps which are illustrated in next section.

- ***Data Collection/Creation*:** The most essential element of any deep learning computational task is a standardized benchmark dataset. As per various literatures that are currently available, it indicates that there exists no benchmark annotated dataset available for the Assamese language. The most important phase for a limited resourced language like Assamese is the collecting and building of datasets. Construction of dataset for analyzing sentiment from review data written in Assamese language is discussed in details in ***Chapter-5***. The quantity of the available dataset was found insufficient for using in deep learning techniques. As a result, the current amount of data (approximately 2000) has been enhanced with additionally collected data from various web forums. For this research project, finally a total of about 4000 reviews have been created.

- ***Data pre-processing/cleaning*:** Data cleaning is performed in this step. Various basic NLP tasks which includes tokenization, stop word removal, punctuation removal, etc., are implemented to clean/pre-process the dataset. The objective of data preprocessing is to enable the deep learning models to use the common format of the dataset. This step aids in removing words and phrases that are unnecessary or have no significance on sentiment analysis. For Example, words like "i," "she," "you," "for,"

"from," which in Assamese is termed as “মই”, “তেওঁ”,“তুমি”,“ বাবে”, “ পৰা” and so on have no sentimental meaning in any phrase.

- ***Deep learning model development*:** The proposed model use three different sets of deep learning algorithms. The deep learning models namely, LSTM, CNN, and hybrid LSTM-CNN are developed using TensorFlow and Keras libraries of python package. Detail discussion of mentioned models is described in results and discussion section of this chapter.

- ***Data training and testing*:** In order to perform model evaluation dataset training and testing is necessary. With this objective splitting ratio of 80:20 where, 80% data is used for training and 20% is used for testing by the proposed models as mentioned above.

- ***Model Evaluation:*** For optimal consistency and to improve comparability, a specific set of hyperparameters needs to be selected for the model evaluation. These hyperparameter choices are vital as they have a direct impact on how well the models function and perform during the training and testing stages. Analyzing the models' performance in these conditions is important in addition to defining the hyperparameters. This involves observation of measures like accuracy, testing and training loss. These performance metrics parameters show how well the models are predicting sentiment appropriately and how effectively they are learning from the training data. Model evaluation based on these metrics are all depicted in **Result and discussion** section of this chapter. In summary, the proposed models can be thoroughly assessed according to the combination of specified hyperparameter values and performance analysis via

visualizations. This method ensures stability and dependability while evaluating the models' suitability for sentiment analysis assignments.

6.4 Results and discussions

The overall performance of the proposed models is evaluated for movie reviews dataset written in Assamese language. To do this conventional parameter are used such as accuracy, testing and training loss of the proposed model. Moreover, to observe the proposed models performance with keras embedding method, all of the variables listed in Table 6.1 have been taken into consideration.

Table-6.2: Hyperparameter Settings for the proposed models

No of epochs	10
Batch size	128
Activation Function	Sigmoid
Train test splitting ratio	80:20
Loss function	Binary Cross entropy
Optimizer	Adam
Dropout	0.2

All three proposed models are summarized with designed values, the accuracy of the models for both testing and training data and the loss values of the models for both training and testing data which are illustrated in the **sections 6.4.1, 6.4.2, 6.4.3** below.

6.4.1 Proposed LSTM model

Table -6.3: Proposed LSTM model summary

```
Model: "LSTM"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 86, 128)           128000

 spatial_dropout1d (Spatial  (None, 86, 128)           0
 Dropout1D)

 lstm (LSTM)                 (None, 128)               131584

 dense (Dense)               (None, 2)                 258

=================================================================
Total params: 259842 (1015.01 KB)
Trainable params: 259842 (1015.01 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

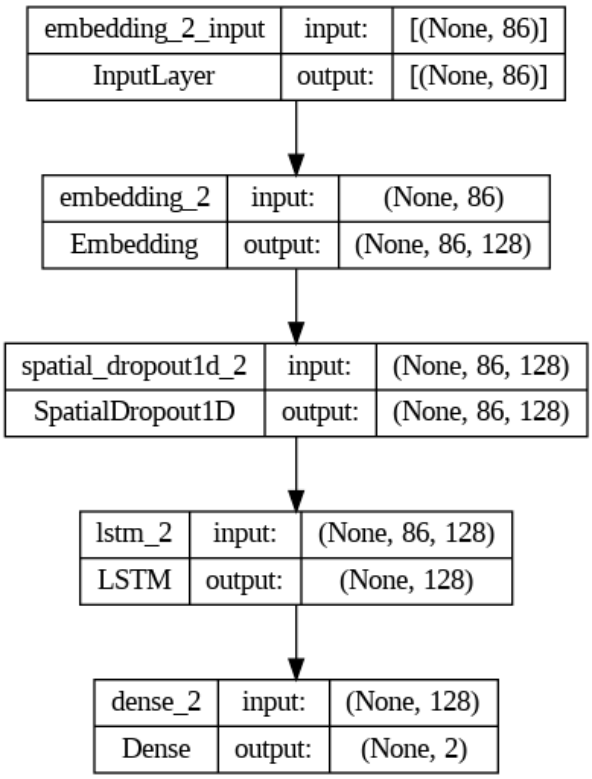


Figure: 6.3: Layer architecture of proposed LSTM model

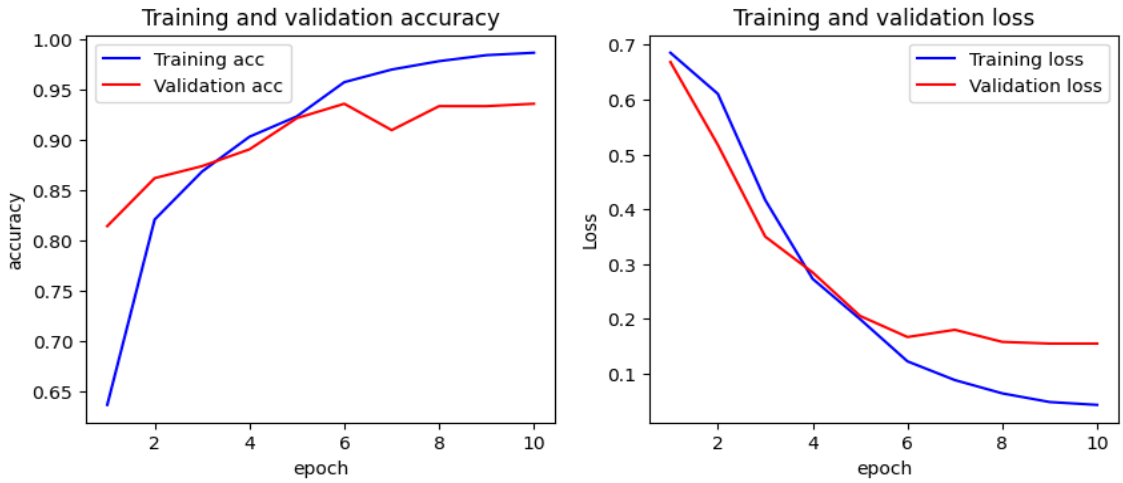


Figure- 6.4: Results of the accuracy and validating accuracy of proposed LSTM model

6.4.2 Proposed CNN

Table- 6.4: Proposed CNN model summary

```
Model: "CNN"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_5 (Embedding)     (None, 100, 128)          1280000

 conv1d_5 (Conv1D)           (None, 96, 128)           82048

 global_max_pooling1d_5 (Gl  (None, 128)               0
 obalMaxPooling1D)

 dense_5 (Dense)             (None, 1)                 129

=================================================================
Total params: 1362177 (5.20 MB)
Trainable params: 1362177 (5.20 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

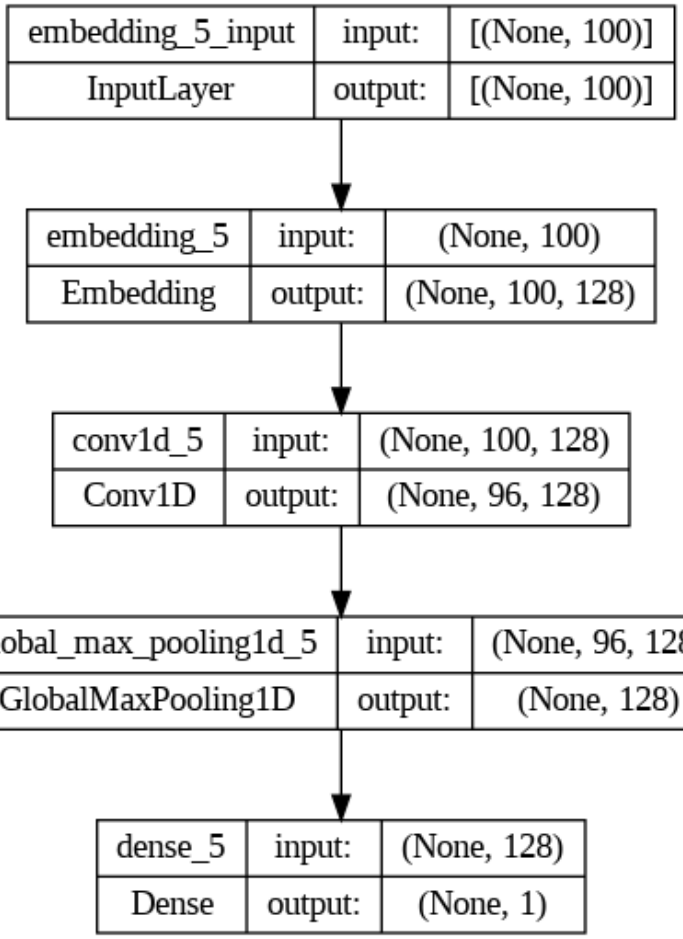


Figure-6.5: Layer architecture of proposed CNN model

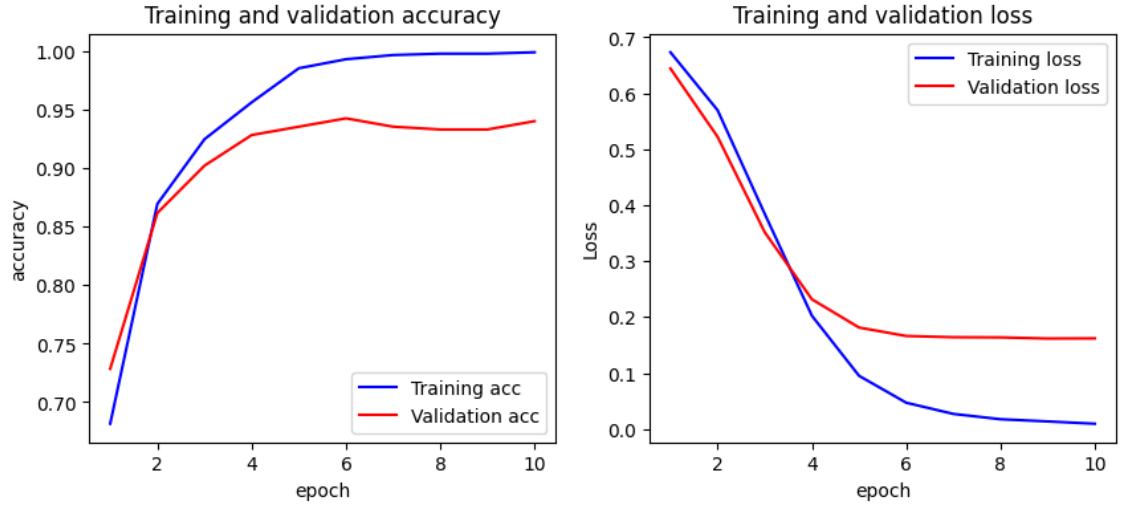


Figure-6.6: Results of the accuracy and validating accuracy of proposed CNN model

6.4.3 Proposed LSTM-CNN

Table -6.5: Proposed LSTM-CNN Summery

```
Model: "LSTM-CNN"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_5 (Embedding)     (None, 165, 128)          2560000

 bidirectional_5 (Bidirecti  (None, 165, 256)          263168
 onal)

 conv1d_4 (Conv1D)           (None, 163, 100)          76900

 conv1d_5 (Conv1D)           (None, 161, 64)           19264

 global_max_pooling1d_1 (Gl  (None, 64)                0
 obalMaxPooling1D)

 dense_5 (Dense)             (None, 1)                 65

=================================================================
Total params: 2919397 (11.14 MB)
Trainable params: 2919397 (11.14 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

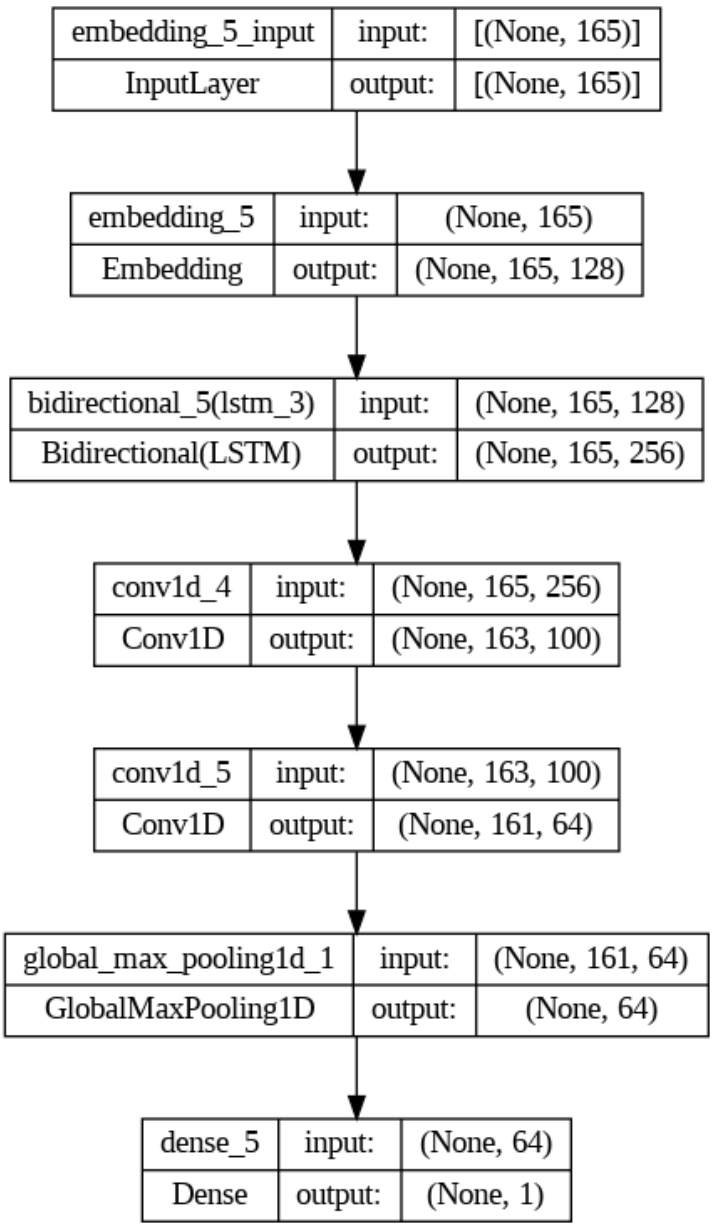


Figure-6.7: Layer architecture of proposed LSTM-CNN model

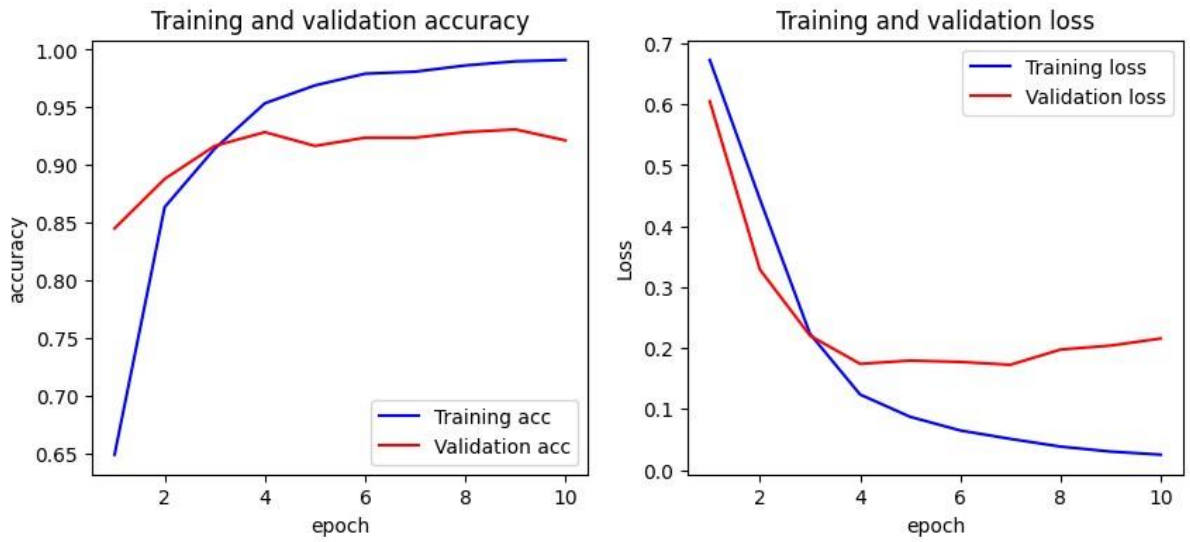


Figure-6.8: Results of the accuracy and validating accuracy of proposed LSTM-CNN model Movie data

Table- 6.6: Comparison on proposed model with existing models with different datasets and languages

Classifier	Accuracy	Dataset used	Language used
CNN **[142]**	85%	Airline sentiment	English
LSTM **[142]**	85.2%	Airline sentiment	English
CNN-LSTM **[142]**	90.2%	Airline sentiment	English
CNN **[142]**	87.1%	Twitter Airline sentiment	English
LSTM **[142]**	88.2%	Twitter Airline sentiment	English
CNN-LSTM **[142]**	91.3%	Twitter Airline sentiment	English
LSTM **[150]**	72.1%	Cricket review	Bengali
SVM **[151]**	88.9%	Movie review	Bengali
LSTM **[151]**	82.42%	Movie review	Bengali
SVM **[152]**	82%	Restaurant Reviews	Assamese
MNB **[152]**	89%	Restaurant Reviews	Assamese

SVM **[152]**	81%	Movie review	Assamese
MNB **[152]**	93%	Movie review	Assamese
Proposed LSTM	**98%**	Movie Reviews	Assamese
Proposed CNN	**99%**	Movie Reviews	Assamese
Proposed LSTM-CNN	**99%**	Movie Reviews	Assamese

Table 6.6 shows that the three proposed models work excellently when it comes to evaluating sentiment in Assamese movie reviews. They show resilience and effectiveness, when used with the hyperparameter settings as mentioned in Table 6.1.

The CNN and LSTM-CNN architectures exhibit higher accuracy, with both models reaching an outstanding 99% accuracy rate. The obtained results shows that the proposed models can recognize and categorize the emotions expressed in Assamese movie reviews with accuracy. This improved performance highlights the significance of using deep learning techniques for sentiment analysis tasks, particularly in the context of languages like Assamese. Additionally, with a 98% accuracy rate, the LSTM model performs quite effectively. The high accuracy rate of CNN and LSTM-CNN model highlight the benefits of applying deep learning architectures to sentiment analysis tasks, particularly in languages like Assamese where linguistic complexities could make conventional machine learning techniques difficult to use. Furthermore, the LSTM model's competitive performance validates the potential of recurrent neural networks in identifying contextual information and temporal connections seen in textual data.

6.5 Conclusion

Three types of deep learning models are proposed for sentiment analysis in this research study. Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and

an LSTM + CNN hybrid. These models, which concentrate on industries like movies, are made to examine the sentiment conveyed in textual reviews. In order to use these models, the textual reviews are initially transformed into numerical vectors using Keras embedding techniques. By using these vectors as input data, the models are able to learn and interpret the sentiment expressed in the text.

When compared to traditional machine learning models, the experimental analysis carried out in presented study indicates that the suggested deep learning based models attain excellent levels of accuracy. The report does, however, recognize the need for more research on sentiment analysis in a number of different areas related movie domains. Although the majority of the attention is directed toward sentiment analysis of text reviews from the restaurant and movie industries, there are many more contexts in which customer sentiment can be investigated and analyzed. To expand the application of sentiment analysis, future studies may include examining reviews written in Assamese languages in multimodal contexts.

7

Conclusion and Future Scope

7.1 Conclusions

This thesis attempts to provide a research contribution to the progress of sentiment analysis techniques and future improvements by expanding awareness of different methodology and their applicability primarily customized for the Assamese language. This thesis discusses the necessity for sentiment analysis specifically for the Assamese language, thereby identifying several research gaps that can encourage researchers to seek solutions and bridge the gaps in computational analysis.

This thesis investigates or reviews the previous research works related to sentiment analysis in a variety of languages using different types of data. Henceforth, it also presents a summary of studies on natural language processing (NLP) in Indian languages, including Tamil, Telugu, Malayalam, Odia, Nepali, Hindi, Bengali, and Assamese. While analyzing numerous Indian languages, various computational linguistics techniques have been identified along with their accessibility through internet sources and published works. This study primarily focuses on the machine learning (ML) methods applied in different research works which are relevant to sentiment analysis primarily in Indian languages.

This thesis presented a lexicon-based approach to perform sentiment analysis on Assamese textual data. Experiments for this include consideration of the existing standard benchmark lexicon called VADER developed for the Bengali language which is an

adaptation of the English VADER tool. This is followed by the use of a few 100 words from the Bengali VADER model to build Assamese VADER. The developed Assamese VADER lexicon performs very effectively with the dataset considered in the Assamese language. The addition of booster words and the use of bigram and trigram, has resulted in performance enhancement of the developed lexicon in terms of sentiment analysis. Initially, dataset collection for these experiments had been a difficult task which had overcome with the help of an available web source named The Indian Language Technology Proliferation & Deployment Centre (TDIL-DC), a national portal for dataset repository. A comparative analysis is also presented in the thesis work to verify the performance of the build lexicon in Assamese with the existing counterparts designed in other languages with large datasets.

The presented thesis explores feature extraction techniques for the input dataset written in Assamese. Amongst various existing feature extraction techniques, the very widely and commonly used TFIDF technique is used for this purpose. The proposed approach for calculating the TFIDF feature vector for the input dataset opens numerous scopes for different types of NLP tasks. The extracted features are used as input to the classical machine learning classifiers to perform sentiment analysis on considered Assamese text data.

This thesis presented the application of machine learning techniques including Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), Logistic Regression (LR), and K-Nearest Neighbors (KNN) were employed for sentiment classification. The dataset was divided into 80% for training and 20% for testing. To assess the performance of these techniques, various metrics such as Precision, Recall, F-measure, and Accuracy were utilized. For experiments randomly generated review data from two different

domains viz, restaurant and movie have been used. From these text reviews, sentiments have been predicted as positive and negative. It is observed from these experiments that out of all the considered classical machine learning classifiers MNB and SVM perform with the highest accuracy of 85% and above. A comparative analysis has been done to validate the effectiveness of these best-performing classifiers with natural language tool kit sentiment analyzer and human interpretation. This thesis presents three deep-learning models for sentiment analysis in textual reviews. These models efficiently analyze sentiment expressions by converting textual data into numerical vectors through the use of Keras embedding techniques. The outcomes of the experiments show that the accuracy is higher than that of conventional machine learning models. The study highlights the need for more extensive research on sentiment analysis in a variety of contexts outside of dining establishments and motion pictures. Subsequent studies could investigate sentiment analysis in multimodal Assamese languages and assess other hybrid models and optimization techniques to improve performance. All things considered, the results offer a fundamental structure for additional investigation into Assamese text sentiment analysis, especially in deep learning hybrid network models.

The presented thesis work is summarized as follows:

- The advancement of sentiment analysis methods for the Assamese language is the main objective of this thesis.
- It outlines research needs in computational analysis and emphasizes the significance of sentiment analysis for Assamese.
- To identify techniques and applications, existing relevant studies on sentiment analysis in a variety of languages including Indian languages are reviewed.
- A lexicon-based method for Assamese sentiment analysis is proposed, utilizing the VADER lexicon that is being modified from Bengali.
- To prepare data for sentiment analysis using machine learning classifiers such as SVM, DT, NB, LR, and KNN, feature extraction techniques, in particular, TFIDF are investigated.
- The accuracy of the experimental results is found promising. MNB and SVM exhibit the best performance amongst all other considered classifiers.
- Three different types of deep learning models built for sentiment analysis that outperform traditional machine learning techniques in terms of accuracy.

7.2 Future Scope

The presented sentiment analysis approaches have the potential to enhance various NLP tasks in the Assamese language. Despite various limitations, there exists a tremendous scope of research in this area specially customized for the Assamese language. To build a system that is versatile enough to be utilized in other applications, it is possible to

experiment with the different deep learning based models in addition to expanding the dataset to include reviews of products, social media, political analysis, stock prediction and other topics. With the expansion of datasets in the Assamese language, presented sentiment analysis approaches may be upgraded to more advanced computational techniques. With the implementation of deep learning models, the built system has shown performance enhancement. This predicts the scope for inclusion of more advanced learning processes for the same NLP tasks viz transfer learning, hybrid machine learning etc. Additionally, the same system has the prospect of experimenting with numerous other forms of datasets extracted from different domains of interest.

Future prospects of the presented research work can be outlined as follows:

- Advancement in data pre-processing, cleaning etc. can be incorporated for enhancement of the system performance.
- For building lexicon-based approach, more updated lexicons may be adapted as reference for experimenting the flexibility of the system.
- Additional, advance type of feature extraction method may be used to experiment with the considered dataset.
- Machine learning based approach of sentiment analysis system have scope of use of hybrid-based classifiers for exploring the applicability of the system.
- In addition to the deep learning, transfer learning-based approach may also be adopted for the analysis of sentiment with the increased size of datasets.
- Last but not the least dataset written in different existing native dialects of Assamese language can be built and used for the aforementioned methods.

  Sentiment analysis has explored almost all Indian and international languages to their individual possible extends. But the same application in different dialect of

a low resourced language like Assamese is completely unexplored one. Social media, web forums are now a days have opened up for all types of users giving them the freedom to post comments on everything and anything in their own ways. Hence sentiment analysis on such raw, unfiltered, multidomain comments/reviews will definitely bring a new horizon in this area of research.

References

1. Mika V. Mäntylä, Daniel Graziotin, Miikka Kuutila, “The evolution of sentiment analysis—A review of research topics, venues, and top cited papers”, Computer Science Review, Volume 27, February 2018, Pages 16-32, ISSN 1574-0137
2. https://en.wikipedia.org/wiki/Sentiment_analysis
3. Walaa Medhat, Ahmed Hassan Hoda Korashy, “Sentiment analysis algorithm & applications: A survey”, Ain Shams Engineering Journal,Volume 5, May 2014, Pages 1093–1113.
4. Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr., 2, 1-135.
5. Goswami, G. C., & Tamuli, J. (2003). Asamiya. The Indo-Aryan languages, 391-443.
6. Liu, Chao & Jarochowska, Emilia & Du, Yuansheng & Vachard, Daniel & Munnecke, Axel. (2015)
7. Pozzi, Federico & Fersini, Elisabetta & Messina, Vincenzina & Liu, B.. (2017). Challenges of Sentiment Analysis in Social Networks.
8. Pang, Lee, and Vaithyanathan, 2002; Turney, 2002
9. Wiebe, Bruce, & O'Hara, 1999
10. Alessia D’Andrea, Fernando Ferri, Patrizia Grifoni, Approaches, Tools and Applications for Sentiment Analysis Implementation International Journal of Computer Applications (0975 – 8887) Volume 125 – No.3, September 2015
11. Artificial Intelligence Review (2021) 54:4997–5053 https://doi.org/10.1007/s10462-021-09973-3. Akhtar et al., 2016.

12. Aydogan, Ebru and M. Ali Akcayol. “A comprehensive survey for sentiment analysis tasks using machine learning techniques.” 2016 International Symposium on Innovations in Intelligent Systems and Applications (INISTA) (2016): 1-7.
13. S. M. Al-Ghuribi, S. A. Mohd Noah, and S. Tiun, ‘‘Unsupervised semantic approach of aspect-based sentiment analysis for large-scale user reviews,’’ IEEE Access, vol. 8, pp. 218592–218613, 2020, doi: 10.1109/ACCESS.2020.3042312.
14. A. Yadav, C. K. Jha, A. Sharan, and V. Vaish, ‘‘Sentiment analysis of financial news using unsupervised approach,’’ Proc. Comput. Sci., vol. 167, pp. 589–598, Jan. 2020, doi: 10.1016/j.procs.2020.03.325.
15. J. Rothfels and J. Tibshirani, ‘‘Unsupervised sentiment classification of English movie reviews using automatic selection of positive and negative sentiment items,’’ CS224N-Final Project, vol. 43, no. 2, pp. 52–56, 2010.
16. N. Raghunathan and K. Saravanakumar, "Challenges and Issues in Sentiment Analysis: A Comprehensive Survey," in IEEE Access, vol. 11, pp. 69626-69642, 2023, doi: 10.1109/ACCESS.2023.3293041.
17. https://www.dictionary.com/browse/lexicon.
18. Agarwal, A, Xie, B., Vovsha, L, Rambow, O, & Passonneau, R. (2011). Sentiment analysis of Twitter data. In Proc. WLSM-11s
19. Akkaya. C, Wiebe, J, & Mihalcea, R. (2009). Subjectivity word sense disambiguation. In Proc. EMNLP-09.
20. Baccianella, S., Esuli. A., & Sebastiani, F. (2010). SentiWordNet 3.0. In Proc. of LREC-10.
21. Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings.

22. Cambria, E, Havasi, C, & Hussain, A. (2012). SenticNet 2. In Proc. AAAI IFAI RSC-12.

23. Cambria, E, Speer, R., Havasi, C, & Hussain, A. (2010). SenticNet. In Proc. of AAAI SCK-10.

24. C. H. E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in Eighth International Conference on Weblogs and Social Media (ICWSM14). Available at (20/04/16) http://comp. social. gatech. Edu/papers/icwsm14. vader. hutto. pdf, 2014.

25. Seshadri, S. & Kumar, M. & Kp, Soman. (2016). Analyzing sentiment in Indian languages micro text using recurrent neural network. 7. 313-318.

26. Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. Aspect based Sentiment Analysis in Hindi: Resource Creation and Evaluation. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 2703–2709, Portorož, Slovenia. European Language Resources Association (ELRA).

27. Ligthart, A., Catal, C. & Tekinerdogan, B. Systematic reviews in sentiment analysis: a tertiary study. Artif Intell Rev 54, 4997–5053 (2021). https://doi.org/10.1007/s10462-021-09973-3

28. C. Dev, A. Ganguly and H. Borkakoty, "Assamese VADER: A Sentiment Analysis Approach Using Modified VADER," 2021 International Conference on Intelligent Technologies (CONIT), Hubli, India, 2021, pp. 1-5, doi: 10.1109/CONIT51480.2021.9498455.

29. Al Amin, Imran Hossain, Aysha Akther and Kazi Masudul Alam, Bengali VADER: A Sentiment Analysis Approach Using Modified VADER, in International

Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

30. Rahm, E., Do, H.H., et al., 2000. Data cleaning: Problems and current approaches. IEEE Data Eng. Bull. 23 (4), 3–13.
31. Aliwy, A.H., 2012. Tokenization as preprocessing for Arabic tagging system. Int. J. Inf.Educ. Technol. 2 (4), 348.
32. Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3 (Mar), 1157–1182.
33. Patro, S., Sahu, K.K., 2015. Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462.
34. Jamin Rahman Jim, Md Apon Riaz Talukder, Partha Malakar, Md Mohsin Kabir, Kamruddin Nur, M.F. Mridha,Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review, Natural Language Processing Journal, Volume 6,2024,100059,ISSN 2949-7191,https://doi.org/10.1016/j.nlp.2024.100059.
35. Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems 226 (2021), 107134
36. Soleymani, Mohammad; Garcia, David; Jou, Brendan; Schuller, Björn; Chang, Shih-Fu; Pantic, Maja (September 2017). "A survey of multimodal sentiment analysis". Image and Vision Computing. 65: 3–14. doi: 10.1016/j.imavis.2017.08.003
37. Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., Hussain, A., 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Inf. Fusion 91, 424–444.

38. Fang, Z., Lu, J., Liu, F., Zhang, G., 2022. Semi-supervised heterogeneous domain adaptation: Theory and algorithms. IEEE Trans. Pattern Anal. Mach. Intell. 45 (1),1087–1105.

39. Bhatia, S., Kumar, A., Khan, M.M., 2022. Role of genetic algorithm in optimization of Hindi word sense disambiguation. IEEE Access 10, 75693–75707.

40. V. Hatzivassiloglou and K. R. McKeown, Predicting the semantic orientation of adjectives, in Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics. Association for Computational Linguistics, 1997, pp. 174–181.

41. S.M. Kim and E. Hovy, Determining the sentiment of opinions, in Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004, p. 1367.

42. J. Kamps, M. Marx, R. J. Mokken, M. De Rijke et al., Using wordnet to measure semantic orientations of adjectives. in LREC, vol. 4. Citeseer, 2004, pp. 1115–1118.

43. H. Liu and P. Singh, Concept net—a practical common sense reasoning tool-kit, BT technology journal, vol. 22, no. 4, pp. 211–226, 2004.

44. Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

45. Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie C, Riloff E & Patwardhan, S. (2005, October). Opinion Finder: A system for subjectivity analysis. In Proceedings of HLT/EMNLP 2005 Interactive Demonstrations (pp. 34-35).

46. M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon based methods for sentiment analysis,” Computational linguistics, vol. 37, no. 2, pp. 267–307, 2011.

47. R. Mihalcea, C. Banea, and J. Wiebe, “Learning multilingual subjective language via cross-lingual projections,” in Proceedings of the 45th annual meeting of the association of computational linguistics, 2007, pp. 976–983.

48. B. Liu, Sentiment analysis and subjectivity. Handbook of natural language processing, vol. 2, pp. 627–666, 2010.

49. J. W. Pennebaker, M. E. Francis, and R. J. Booth, “Linguistic inquiry and word count: Liwc 2001,” Mahway: Lawrence Erlbaum Associates, vol. 71, no. 2001, p. 2001, 2001.

50. J. W. Pennebaker, R. J. Booth, and M. E. Francis, “Linguistic inquiry and word count: Liwc [computer software],” Austin, TX: liwc. net, 2007.

51. S. Chowdhury and W. Chowdhury, Performing sentiment analysis in Bangla microblog posts, in Informatics, Electronics and Vision (ICIEV), 2014 International Conference on. IEEE, 2014, pp. 1–6.

52. Jaiswal, A. (2022, January 21). NLP Tutorials Part ii: Feature extraction. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2022/01/nlp-tutorials-part-ii-feature-extraction/

53. Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520. https://doi.org/10.1108/00220410410560582

54. Chan, T. Y., & Chang, Y. S. (2017, November). Enhancing classification effectiveness of Chinese news based on term frequency. In *2017 IEEE 7th*

International Symposium on Cloud and Service Computing (SC2) (pp. 124-131). IEEE.

55. Abu-Errub, A. (2014). Arabic text classification algorithm using tfidf and chi square measurements. *International Journal of Computer Applications*, *93*(6), 40–45. https://doi.org/10.5120/16223-5674
56. Dhar, A., Dash, N. S., & Roy, K. (2018). Application of TF-IDF feature for categorizing documents of online Bangla web text corpus. In Intelligent Engineering Informatics: Proceedings of the 6th International Conference on FICTA (pp. 51-59). Springer Singapore.
57. Kecman, Vojislav. (2005). Support Vector Machines – An Introduction. 10.1007/10984697_1
58. Islam, M. S., Jubayer, F. E. M., & Ahmed, S. I. (2017, February). A support vector machine mixed with TF-IDF algorithm to categorize Bengali document. In *2017 international conference on electrical, computer and communication engineering (ECCE)* (pp. 191-196). IEEE.
59. Jayashree, Ranga Reddy. (2011). Document Summarization in Kannada Using Keyword Extraction. Computer Science & Information Technology. 1. 121-127. 10.5121/csit.2011.1311.
60. Galavotti L, Sebastiani F, Simi M (2000). Experiments on the use of feature selection and negative evidence in automated text categorization. In: International conference on theory and practice of digital libraries. Springer, Berlin, Heidelberg, pp 59–68.
61. Sarkar K (2012) Bengali text summarization by sentence extraction. arXiv preprint arXiv:1201.2240

62. Hannan A, Boro SR, Sarma JPSSK (2015) An approach to Bodo document clustering. Int J Innovative Res Sci Eng Technol (ISSN: 2319-8753) 4(12), https://doi.org/10.15680/IJIRSET.2015.0412069

63. Hanumanthappa M, Narayana Swamy M, Jyothi NM (2014) Automatic keyword extraction from Dravidian language. Int J Innovative Sci Eng Technol 1(8):87–92

64. Salton, G., & Yu, C. T. (1973). On the construction of effective vocabularies for information retrieval. Proceedings of the 1973 Meeting on Programming Languages and Information Retrieval - SIGPLAN '73, 48–60. https://doi.org/10.1145/951762.951766

65. Feature extraction techniques-NLP. (2020, March 3). GeeksforGeeks.http:// www.geeksforgeeks.org/feature-extraction-techniques-nlp/

66. A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n- gram machine learning approach," Expert Systems with Applications, vol. 57, pp. 117– 126, Sep. 2016.

67. S. Rani, "Sentiment Analysis of Social Media for Hindi Language" Ph.D dissertation, Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology,India,2020.

68. Almatrafi, Omaima, et al. "Application of Location-Based Sentiment Analysis Using Twitter for Identifying Trends towards Indian General Elections 2014." *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, ACM, 2015, pp. 1–5. *DOI.org (Crossref)*, https://doi.org/10.1145/2701126.2701129. Available: https://dl.acm.org/doi/10.1145/2701126.2701129.

69. Kim, Soo-Min, and Eduard Hovy. “Determining the Sentiment of Opinions.” Proceedings of the 20th International Conference on Computational Linguistics - COLING ’04, Association for Computational Linguistics, 2004, pp. 1367-es. DOI.org(Crossref),https://doi.org/10.3115/1220355.1220555.Available: http://portal.acm.org/citation.cfm?doid=1220355.1220555.

70. Indurkhya, Nitin, and Frederick J. Damerau, editors. Handbook of Natural Language Processing. Chapman & Hall/CRC, 2010.

71. Liu, Bing, et al. “Opinion Observer: Analyzing and Comparing Opinions on the Web.” Proceedings of the 14th International Conference on World Wide Web - WWW ’05, ACM Press, 2005, p. 342. DOI.org(Crossref),https://doi.org/10.1145/1060745.1060797.Available: http://portal.acm.org/citation.cfm?doid=1060745.1060797.

72. Patodkar, Vaibhavi N., and Sheikh I.R. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining.” IJARCCE, vol. 5, no. 12, Dec. 2016, pp.32022.DOI.org(Crossref),https://doi.org/10.17148/IJARCCE.2016.51274.Available: http://ijarcce.com/upload/2016/december-16/IJARCCE%2074.pdf.

73. Pang, Bo, and Lillian Lee. “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.” *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL ’04*, Association for Computational Linguistics, 2004, pp. 271-es. *DOI.org (Crossref)*,https://doi.org/10.3115/1218955.1218990.Available: http://portal.acm.org/citation.cfm?doid=1218955.1218990

74. Pang, Bo, and Lillian Lee. “Opinion Mining and Sentiment Analysis.” *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, 2008, pp. 1–135. *DOI.org*

(Crossref),https://doi.org/10.1561/1500000011.Available: http://www.nowpublishers.com/article/Details/INR-011.

75. Turney, Peter D. "Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews." *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Association for Computational Linguistics, 2001,p.417. *DOI.org (Crossref)*, https://doi.org/10.3115/1073083.1073153. Available: http://portal.acm.org/citation.cfm?doid=1073083.1073153.

76. Whitelaw, Casey, et al. "Using Appraisal Groups for Sentiment Analysis." *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, ACM, 2005, pp. 625–31. *DOI.org (Crossref)*, https://doi.org/10.1145/1099554.1099714. Available: https://dl.acm.org/doi/10.1145/1099554.1099714.

77. Rani, Sujata, and Parteek Kumar. "A Sentiment Analysis System to Improve Teaching and Learning." Computer, vol. 50, no. 5, May 2017, pp.3643.DOI.org(Crossref),https://doi.org/10.1109/MC.2017.133. Available: https://academic.oup.com/dsh/article/34/3/569/5146723.

78. Wu, Shih-Jung, et al. "Development of a Chinese Opinion-Mining System for Application to Internet Online Forums." The Journal of Supercomputing, vol. 73, no. 7, July 2017, pp. 2987–3001. DOI.org (Crossref), https://doi.org/10.1007/s11227-016-1816-6. Available: http://link.springer.com/10.1007/s11227-016-1816-6.

79. Li, Zhouyang, et al. "Analysis of Customer Satisfaction from Chinese Reviews Using Opinion Mining." 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), IEEE, 2015, pp. 95–

99.DOI.org(Crossref),https://doi.org/10.1109/ICSESS.2015.7339013. Available: http://ieeexplore.ieee.org/document/7339013/.

80. Henriquez Miranda, Carlos, and Jaime Guzman. "A Review of Sentiment Analysis in Spanish." TECCIENCIA, vol. 12, no. 22, Dec. 2016, pp. 35–48. DOI.org (Crossref), https://doi.org/10.18180/tecciencia.2017.22.5.

81. Rhouati, Abdelkader, et al. "Sentiment Analysis of French Tweets Based on Subjective Lexicon Approach: Evaluation of the Use of OpenNLP and CoreNLP Tools." Journal of Computer Science, vol. 14, no. 6, June 2018, pp. 829–36. DOI.org(Crossref),https://doi.org/10.3844/jcssp.2018.829.836.

82. Banik, Nayan, and Md. Hasan Hafizur Rahman. "Evaluation of Naïve Bayes and Support Vector Machines on Bangla Textual Movie Reviews." 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), IEEE, 2018, pp. 1–6. https://doi.org/10.1109/ICBSLP.2018.8554497.

83. Rani, Sujata, and Parteek Kumar. "A Sentiment Analysis System for Social Media Using Machine Learning Techniques: Social Enablement." Digital Scholarship in the Humanities, vol. 34, no. 3, Sept. 2019, pp. 569–81. https://doi.org/10.1093/llc/fqy037.

84. Thavareesan, Sajeetha, and Sinnathamby Mahesan. "Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation." 2019 14th Conference on Industrial and Information Systems (ICIIS), IEEE, 2019, pp. 320–25. https://doi.org/10.1109/ICIIS47346.2019.9063341.

85. Naidu, Reddy, et al. "Sentiment Analysis Using Telugu SentiWordNet." 2017 International Conference on Wireless Communications, Signal Processing and

Networking (WiSPNET), IEEE, 2017, pp. 666–70. https://doi.org/10.1109/WiSPNET.2017.8299844.

86. Nair, Deepu S., et al. “SentiMa - Sentiment Extraction for Malayalam.” 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2014, pp. 1719–23. https://doi.org/10.1109/ICACCI.2014.6968548.

87. S., Soumya, and Pramod K.V. “Sentiment Analysis of Malayalam Tweets Using Machine Learning Techniques.” ICT Express, vol. 6, no. 4, Dec. 2020, pp. 300–05. https://doi.org/10.1016/j.icte.2020.04.003.

88. Das, Ringki, and Thoudam Doren Singh. “A Multi-Stage Multimodal Framework for Sentiment Analysis of Assamese in Low Resource Setting.” Expert Systems with Applications, vol. 204, Oct. 2022, p. 117575. https://doi.org/10.1016/j.eswa.2022.117575.

89. M. Gamon, “Linguistic correlates of style: authorship classification with deep linguistic analysis features,” in Proceedings of the 20th International Conference on Computational Linguistics - COLING’04, Geneva, Switzerland: Association for Computational Linguistics, 2004, pp. 611-es. doi: 10.3115/1220355.1220443.

90. B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP’02, Stroudsburg, United States: Association for Computational Linguistics, 2002, pp. 79–86. doi:10.3115/1118693.1118704.

91. V. S and T. S. N, “Breast Cancer Diagnosis and Classification Using Support vector machines With Diverse Datasets,” International Journal of Computer Sciences and

Engineering, vol. 7, no.4, pp.442–446, Apr. 2019, doi: 10.26438/ijcse/v7i4.442446. Available: http://www.ijcseonline.org/full_paper_view.php?paper_id=4054.

92. A. Kennedy and D. Inkpen, "sentiment classification of movie reviews using contextual valence shifters," Computational Intell, vol. 22, no. 2, pp. 110–125, May 2006, doi: 10.1111/j.1467-8640.2006.00277.x.

93. P. De Pelsmacker, S. Van Tilburg, and C. Holthof, "Digital marketing strategies, online reviews and hotel performance," International Journal of Hospitality Management, vol. 72,pp. 47–55, Jun. 2018, doi: 10.1016/j.ijhm.2018.01.003.

94. E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," Inf. Retrieval, vol. 12, no. 5, pp. 526–558, Oct. 2009, doi: 10.1007/s10791-008-9070-z.

95. S. Al-Natour and O. Turetken, "A comparative assessment of sentiment analysis and star ratings for consumer reviews," International Journal of Information Management, vol. 54, p. 102132, Oct. 2020, doi:10.1016/j.ijinfomgt.2020.102132.

96. A. C. E. S. Lima, L. N. De Castro, and J. M. Corchado, "A polarity analysis framework for Twitter messages," Applied Mathematics and Computation, vol. 270, pp. 756–767, Nov.2015, doi: 10.1016/j.amc.2015.08.059.

97. B. Le and H. Nguyen, "Twitter Sentiment Analysis Using Machine Learning Techniques," in Advanced Computational Methods for Knowledge Engineering, H. A. Le Thi, N. T. Nguyen, and T. V. Do, Eds., Cham: Springer International Publishing, 2015, pp. 279–289. doi: 10.1007/978-3-319-17996-4_25.

98. O. Araque, I. Corcuera-Platas, J. F. Sánchez- Rada, and C. A. Iglesias, “Enhancing deep learning sentiment analysis with ensemble techniques in social applications,” Expert Systems with Applications, vol. 77, pp. 236– 246, Jul. 2017, doi:10.1016/j.eswa.2017.02.002

99. C. Nanda, M. Dua, and G. Nanda, “Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning,” in 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai: IEEE, Apr. 2018, pp. 1069–1072. doi: 10.1109/ICCSP.2018.8524223.

100. K. Sarkar and M. Bhowmick, “Sentiment polarity detection in Bengali tweets using multinomial Naïve Bayes and support vector machines,” in 2017 IEEE Calcutta Conference (CALCON), Kolkata: IEEE, Dec. 2017, pp. 31–36. doi: 10.1109/CALCON.2017.8280690.

101. H. Borkakoty, C. Dev, and A. Ganguly, “A Novel Approach to Calculate TF-IDF for Assamese Language,” in Electronic Systems and Intelligent Computing, P. K. Mallick, P. Meher, A. Majumder, and S. K. Das, Eds., Singapore: Springer Singapore, 2020, pp. 387–393. doi: 10.1007/978-981-15-7031-5_37.

102. R. Das and T. D. Singh, “A Step Towards Sentiment Analysis of Assamese News Articles Using Lexical Features,” in Proceedings of the International Conference on Computing and Communication Systems, A. K. Maji, G. Saha, S. Das, S. Basu, and J. M. R. S. Tavares, Eds., Singapore: Springer Singapore, 2021, pp. 15–23. doi: 10.1007/978- 981-33-4084-8_2.

103. Kalaivani. K S, “Performance enhancement of machine Learning approaches for document level Sentiment classification” Ph.d dissertation, Faculty of Information and Communication Engineering, Anna University, India,2021

104. https://www.geeksforgeeks.org/decision-tree/

105. J. Tolles and W. J. Meurer, “Logistic Regression: Relating Patient Characteristics to Outcomes,” JAMA, vol. 316, no. 5, p. 533, Aug. 2016, doi: 10.1001/jama.2016.7653.

106. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, Third edition. in Wiley series in probability and statistics, no. 398. Hoboken, New Jersey: Wiley, 2013.

107. Joachims T. Probabilistic analysis of the rocchio algorithm with TFIDF for text categorization. In: Presented at the ICML conference; 1997.

108. M. Rushdi Saleh, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. A. Ureña-López, “Experiments with SVM to classify opinions in different domains,” Expert Systems with Applications, vol. 38, no. 12, pp. 14799–14804, Nov.2011, doi:10.1016/j.eswa.2011.05.070.

109. https://www.geeksforgeeks.org/n-gram-language-modelling-with-nltk/---110

110. A. Sharma and S. Dey, “A comparative study of feature selection and machine learning techniques for sentiment analysis,” in Proceedings of the 2012 ACM Research in Applied Computation Symposium, San Antonio Texas: ACM, Oct. 2012, pp. 1–7. doi: 10.1145/2401603.2401605.

111. P. Lak and O. Turetken, “Star ratings versus sentiment analysis–a comparison of explicit and implicit measures of opinions,” in System Sciences (HICSS), 2014 47th Hawaii International Conference on, IEEE, 2014, pp. 796–805.

112. O. Sharif, M. M. Hoque, and E. Hossain, “Sentiment Analysis of Bengali Texts on Online Restaurant Reviews Using Multinomial Naïve Bayes,” in 2019 1st International Conference on Advances in Science, Engineering and Robotics

Technology (ICASERT), Dhaka, Bangladesh: IEEE, May 2019, pp. 1–6. doi: 10.1109/ICASERT.2019.8934655.

113. A. Sharma and S. Dey, “A comparative study of feature selection and machine learning techniques for sentiment analysis,” in Proceedings of the 2012 ACM Research in Applied Computation Symposium, San Antonio Texas: ACM, Oct. 2012, pp. 1–7. doi: 10.1145/2401603.2401605.
114. Pang, B., and Lee, L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting of the ACL(Barcelona, Spain, July 21–26, 2004). 2004, 271–278.
115. Prabha MI, Srikanth GU (2019) Survey of sentiment analysis using deep learning techniques. In: 2019 1st International conference on innovations in information and communication technology (ICIICT). IEEE, pp 1–9
116. Chandra Y, Jana A (2020) Sentiment analysis using machine learning and deep learning. In: 2020 7th International conference on computing for sustainable global development (INDIACom).IEEE, pp 1–4
117. Zhang L, Wang S, Liu B (2018) Deep learning for sentiment analysis: a survey. Wiley Interdiscip Rev Data Min Knowl Discov8(4):1253
118. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015). https://doi.org/10.1038/nature14539
119. Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Systems with Applications, 77, 236–246. https://doi.org/10.1016/j.eswa.2017.02.002

120. Sahoo, C., Wankhade, M., & Singh, B. K. (2023). Sentiment analysis using deep learning techniques: A comprehensive review. International Journal of Multimedia Information Retrieval, 12(2), 41. https://doi.org/10.1007/s13735-023-00308-2

121. K. Lyu and H. Kim, "Sentiment Analysis Using Word Polarity of Social Media," Wireless Personal Communication, vol. 89, no. 3, pp. 941-958, Aug. 2016.

122. G. Wang, P. Pu, and Y. Liang, "Topic and Sentiment Words Extraction in Cross-Domain Product Reviews," Wireless Personal Communication, vol. 102, no. 2, pp. 1773-1783, Sep. 2018.

123. F. Xu, Z. Pan, and R. Xia, "E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework," Information Processing & Management, vol. 57, no. 5, pp. 102221, Sep. 2020.

124. L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," WIREs Data Min & Knowl, vol. 8, no. 4, pp. e1253, Jul. 2018.

125. H. N. Mhaskar, "Neural networks and approximation theory," Neural Networks, vol. 9, no. 4, pp. 721-722, Jun. 1996.

126. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of machine learning research, 12, 2493-2537.

127. E. Hossain, O. Sharif, M. M. Hoque, and I. H. Sarker, "SentiLSTM: A Deep Learning Approach for Sentiment Analysis of Restaurant Reviews," in Hybrid Intelligent Systems, vol. 1375, A. Abraham, T. Hanne, O. Castillo, N. Gandhi, T. Nogueira Rios, and T.-P. Hong, Eds., Cham: Springer International Publishing, pp. 193-203, 2021.

128. T. Mandhula, S. Pabboju, and N. Gugulotu, “Predicting the customer’s opinion on amazon products using selective memory architecture-based convolutional neural network,” J Supercomputing, vol. 76, no. 8, pp. 5923-5947, Aug. 2020.

129. F. Tavazoee, C. Conversano, and F. Mola, “Recurrent random forest for the assessment of popularity in social media: 2016 US election as a case study,” Knowl Inf Syst, vol. 62, no. 5, pp. 1847-1879, May 2020.

130. M. Tsytsarau and T. Palpanas, “Survey on mining subjective data on the web,” Data Min Knowl Disc, vol. 24, no. 3, pp. 478–514, May 2012

131. Po-Jen Chen, Jian-Jiun Ding, Hung-Wei Hsu, Chien-Yao Wang, and J.-C. Wang, “Improved convolutional neural network based scene classification using long short-term memory and label relations,” in 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, Hong Kong: IEEE, Jul. 2017, pp. 429–434.

132. M. V. Mäntylä, D. Graziotin, and M. Kuutila, “The evolution of sentiment analysis—A review of research topics, venues, and top cited papers,” Computer Science Review, vol. 27, pp. 16–32, Feb. 2018

133. M. Hoq, P. Haque, and M. N. Uddin, “Sentiment Analysis of Bangla Language Using Deep Learning Approaches,” in Computing Science, Communication and Security, N. Chaubey, S. Parikh, and K. Amin, Eds., in Communications in Computer and Information Science. Cham: Springer International Publishing, 2021, pp. 140–151.

134. K. Sarkar, “Sentiment Analysis of Bengali Tweets Using Deep Learning,” in Computational Intelligence in Data Science, vol. 578, A. Chandrabose, U. Furbach, A. Ghosh, and A. Kumar M., Eds., Cham: Springer International Publishing, 2020, pp. 71–84.

135. R. K. Das, M. Islam, M. M. Hasan, S. Razia, M. Hassan, and S. A. Khushbu, “Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models,” Heliyon, vol. 9, no. 9, p. e20281, Sep. 2023

136. S. Mboutayeb, A. Majda, and N. S. Nikolov, “Multilingual Sentiment Analysis: A Deep Learning Approach:” in Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning, Kenitra, Morocco: SCITEPRESS - Science and Technology Publications, 2021, pp. 27–32.

137. Md. S. Mahmud, Md. T. Islam, A. J. Bonny, R. K. Shorna, J. H. Omi, and Md. S. Rahman, “Deep Learning Based Sentiment Analysis from Bangla Text Using Glove Word Embedding along with Convolutional Neural Network,” in 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Oct. 2022, pp. 1–6.

138. B. K. Shrivash, D. K. Verma, and P. Pandey, “An Effective Framework for Sentiment Analysis Using RNN and LSTM-Based Deep Learning Approaches,” in Advances in Computing and Data Sciences, vol. 1848, M. Singh, V. Tyagi, P. K. Gupta, J. Flusser, and T. Ören, Eds., Cham: Springer Nature Switzerland, 2023, pp. 340–350.

139. P. C. Shilpa, R. Shereen, S. Jacob, and P. Vinod, “Sentiment Analysis Using Deep Learning,” in 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Feb. 2021, pp. 930–937.

140. L. Senevirathne, P. Demotte, B. Karunanayake, U. Munasinghe, and S. Ranathunga, Sentiment Analysis for Sinhala Language using Deep Learning Techniques,” 2020

141. X. Wang, W. Jiang, and Z. Luo, “Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts,” in Proceedings of COLING

2016, the 26th International Conference on Computational Linguistics: Technical Papers, Y. Matsumoto and R. Prasad, Eds., Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 2428–2437.

142. E. Sezgen, K. J. Mason, and R. Mayer, "Voice of airline passenger: A text mining approach to understand customer satisfaction," Journal of Air Transport Management, vol. 77, pp. 65–74, Jun. 2019,
143. S. Banerjee and A. Y. K. Chua, "In search of patterns among travellers' hotel ratings in TripAdvisor," Tourism Management, vol. 53, pp. 125–131, Apr. 2016
144. S. Kumar and M. Zymbler, "A machine learning approach to analyze customer satisfaction from airline tweets," J Big Data, vol. 6, no. 1, p. 62, Dec. 2019
145. Q. Ye, H. Li, Z. Wang, and R. Law, "The Influence of Hotel Price on Perceived Service Quality and Value in E-Tourism: An Empirical Investigation Based on Online Traveler Reviews," Journal of Hospitality & Tourism Research, vol. 38, no. 1, pp. 23–39, Feb. 2014
146. L.-C. Chen, C.-M. Lee, and M.-Y. Chen, "Exploration of social media for sentiment analysis using deep learning," Soft Computing, vol. 24, no. 11, pp. 8187–8197, Jun. 2020
147. L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier," IJIEEB, vol. 8, no. 4, pp. 54–62, Jul. 2016
148. P.-J. Lee, Y.-H. Hu, and K.-T. Lu, "Assessing the helpfulness of online hotel reviews: A classification-based approach," Telematics and Informatics, vol. 35, no. 2, pp. 436–445, May 2018

149. Y. Guo, S. J. Barnes, and Q. Jia, “Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent Dirichlet allocation,” Tourism Management, vol. 59, pp. 467–483, Apr. 2017

150. M. Siering, A. V. Deokar and C. Janze, “Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews,” Decision Support Systems, vol. 107, pp. 52–63, Mar. 2018

151. A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, “A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis,” Multimedia Tools Appl, vol. 78, no. 18, pp. 26597–26613, Sep. 2019

152. K. Shrivastava and S. Kumar, “A Sentiment Analysis System for the Hindi Language by Integrating Gated Recurrent Unit with Genetic Algorithm,” IAJIT, vol. 17, no. 6, pp. 954–964, Nov. 2020

153. N. R. Bhowmik, M. Arifuzzaman, and M. R. H. Mondal, “Sentiment analysis on Bangla text using extended lexicon dictionary and deep learning algorithms,” Array, vol. 13, p. 100123, Mar. 2022

154. C. Dev and A. Ganguly, “Sentiment Analysis of Assamese Text Reviews: Supervised Machine Learning Approach with Combined n-gram and TF-IDF Feature,” ADBU Journal of Electrical and Electronics Engineering (AJEEE), vol. 5, no. 2, pp. 18-30, Sep. 2023.Accessed:Feb.10,2024.[Online].Available:https://journals.dbuniversity.ac.in/ojs/index.php/AJEEE/article/view/4134

155. S. Hochreiter and J. Schmid Huber, "Long Short-Term Memory," in Neural Computation, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

156. John F. Kolen; Stefan C. Kremer, "Gradient Flow in Recurrent Nets: The Difficulty of Learning Long Term Dependencies," in A Field Guide to Dynamical Recurrent Networks, IEEE, 2001, pp.237-243, doi: 10.1109/9780470544037.ch14.

157. Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. Artificial Intelligence Review, 53(8), 5929-5955.

158. Shrivash, B. K., Verma, D. K., & Pandey, P. (2023). An effective framework for sentiment analysis of Hindi sentiments using deep learning technique. Wireless Personal Communications, 132(3), 2097–2110. https://doi.org/10.1007/s11277-023-10702-y

159. Tristan Stérin, Nicolas Farrugia, Vincent Gripon. An Intrinsic Difference Between Vanilla RNNs and GRU Models. COGNITIVE 2017 : Ninth International Conference on Advanced Cognitive Technologies and Applications, Feb 2017, Athènes, Greece. pp.76 - 81.

160. P. K. Jain, V. Saravanan, and R. Pamula, “A Hybrid CNN-LSTM: A Deep Learning Approach for Consumer Sentiment Analysis Using Qualitative User-Generated Contents,” ACM Trans. Asian Low Resource. Lang. Inf. Process., vol. 20, no. 5, pp. 1-15, Sep. 2021.

161. B. Liu, Sentiment analysis: mining opinions, sentiments, and emotions, Second edition. in Studies in natural language processing. Cambridge; New York: Cambridge University Press, 2020.

162. Indolia, S., Goswami, A. K., Mishra, S. P., & Asopa, P. (2018). Conceptual understanding of convolutional neural network- a deep learning approach. Procedia Computer Science, 132, 679–688. https://doi.org/10.1016/j.procs.2018.05.069

163. Lee, K. B., Cheon, S., & Kim, C. O. (2017) "A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes." IEEE Transactions on Semiconductor Manufacturing 30 (2): 135-142.Lima, E., Sun, X.

164. Zhou, Y., Wang, H., Xu, F., and Jin, Y. Q. (2016) "Polarimetric SAR image classification using deep convolutional neural networks." IEEE Geoscience and Remote Sensing Letters 13 (12): 1935-19.

165. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. Pattern Recognition, 77, 354–377. https://doi.org/10.1016/j.patcog.2017.10.013.

166. English, M. L. in P. (2023, June 21). Convolutional neural network — lesson 9: Activation functions in cnns. Medium. https://medium.com/@nerdjock/convolutional-neural-network-lesson-9-activation-functions-in-cnns-57def9c6e759

167. Antonio Gulli and Sujit Pal. 2017. Deep Learning with Keras. Packt Publishing Ltd.

A

PUBLICATIONS

1. Borkakoty, H., Dev, C., Ganguly, A. (2020). A Novel Approach to Calculate TF-IDF for Assamese Language. In: Mallick, P.K., Meher, P., Majumder, A., Das, S.K. (eds) Electronic Systems and Intelligent Computing. Lecture Notes in Electrical Engineering, vol 686. Springer. https://doi.org/10.1007/978-981-15-7031-5_37.

2. C. Dev, A. Ganguly and H. Borkakoty, "Assamese VADER: A Sentiment Analysis Approach Using Modified VADER," *2021 International Conference on Intelligent Technologies (CONIT)*, Hubli, India, 2021, pp. 1-5, https://doi: 10.1109/CONIT51480.2021.9498455.

3. Dev, C. & Ganguly, A. (2023). Sentiment Analysis of Assamese Text Reviews: Supervised Machine Learning Approach with Combined n-gram and TF-IDF Feature. *ADBU Journal of Electrical and Electronics Engineering (AJEEE), 5*(2), 18-30. Retrieved from https://journals.dbuniversity.ac.in/ojs/index.php/AJEEE/article/view/4134.

4. Dev, C. & Ganguly, A. (2023). Sentiment Analysis of Review Data: A Deep Learning Approach Using User-Generated Content. *Asian Journal of Electrical Sciences*, *12*(2), 28–36. https://doi.org/10.51983/ajes-2023.12.2.4119.