www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2014-2018

e-ISSN : 2583-1062
Impact Factor : 7.001

BIG DATA-DRIVEN DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING

Rakshit Dabral[1], Dr. Archana Kumar[2]

[1]Scholar, B. Tech. (AI&DS) 3nd Year Department of Artificial Intelligence and Data Science, Dr. Akhilesh Das Gupta Institute of Professional Studies, New Delhi, India.

[2]Professor, H.O.D (AI & DS) Department of Artificial Intelligence and Data Science, Dr. Akhilesh Das Gupta Institute of Professional Studies, New Delhi, India.

rakshitdabral1@gmail.com

DOI: https://www.doi.org/10.58257/IJPREMS36947

ABSTRACT

Sentiment analysis, a crucial task in natural language processing (NLP), aims to extract and classify sentiments expressed in textual data. This research delves into the application of deep learning techniques, powered by Big Data, to enhance sentiment analysis accuracy.

By leveraging a substantial Amazon review dataset, we train a simple feedforward neural network to classify sentiments as positive or negative. The model employs embedding layers to represent words as dense vectors, followed by a global average pooling layer to capture semantic information. A final dense layer with a sigmoid activation function predicts the sentiment probability.

The results demonstrate the effectiveness of deep learning in capturing complex linguistic nuances and achieving high accuracy. With an accuracy of 88.47%, the model outperforms traditional methods, showcasing the potential of Big Data and deep learning in sentiment analysis.

Future research directions include exploring more sophisticated architectures, addressing class imbalance issues, improving model interpretability, and incorporating domain-specific knowledge to further enhance sentiment analysis performance.

Keywords –Sentiment Analysis, Keras, Tensorflow, Hadoop , Deep Learning , Lemmatization , Vectorization , Stemming

Abbreviations - SST- Stanford Sentiment Treebank- **SA-** Sentiment Analysis- **NB-** Naive Bayes- **SVM-** Support Vector Machine- **NLP-** Natural Language Processing

1. INTRODUCTION

In the bustling digital marketplace of Amazon, a torrent of reviews flowed like a mighty river. Each review, a tiny ripple in the ocean of consumer sentiment, held the power to shape a product's fate. But amidst this vast sea of words, a challenge lurked: how to decipher the intricate nuances of human emotion? Enter the Amazon Whisperer, a sophisticated deep learning model trained on a colossal dataset of Amazon reviews. This digital oracle, armed with the power of Big Data, was capable of unravelling the hidden meanings behind the words. Imagine a product manager staring at a mountain of reviews. "This product is great!" one reviewer wrote. "It's terrible!" countered another. Without the Amazon Whisperer, deciphering these conflicting opinions would be like trying to find a needle in a haystack. But with Whisperer's help, the product manager could quickly identify the underlying sentiment behind each review. Was the reviewer expressing satisfaction, frustration, or perhaps a mix of both? The Whisperer could break down complex emotions into their constituent parts, revealing valuable insights that would otherwise remain hidden. Armed with this newfound understanding, the product manager could make data-driven decisions to improve the product, tailor marketing campaigns, and ultimately enhance the customer experience. The Amazon Whisperer, a testament to the power of deep learning and Big Data, had become an indispensable tool in the arsenal of modern businesses.

1.1 Challenges

Without sentiment analysis, harmful content like hate speech, cyberbullying, and misinformation would be harder to detect and mitigate. Brands and individuals would struggle to gauge audience reactions to their posts, limiting their ability to foster meaningful online interactions. Companies would struggle to identify and address customer concerns, leading to lower satisfaction levels.. Understanding consumer sentiment is crucial for market research. Its absence would hinder businesses' ability to make informed decisions about product launches, pricing, and marketing strategies.

1.2 Need of Sign Language Recognition

Sentiment analysis, also known as opinion mining, is a field of natural language processing that aims to identify and extract the sentiment expressed in text. It is crucial for various reasons Analysing public sentiment towards political

candidates, policies, and issues can help campaigns tailor their messages and strategies. Monitoring public sentiment towards brands can help companies identify and address negative feedback, protect their reputation, and improve customer satisfaction. Analysing customer reviews and social media posts can provide valuable insights into product preferences, satisfaction levels, and areas for improvement

1.3 Need of Big Data in Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a field of natural language processing that aims to identify and extract the sentiment expressed in text. It is crucial for various reasons Sentiment analysis often requires analysing data from various sources, including social media, customer reviews, news articles, and forums. Big data handles the variety and volume of these diverse datasets effectively. Accurate sentiment analysis models require large amounts of training data to learn complex patterns and nuances in language. Big data provides the necessary scale. Many applications of sentiment analysis require real-time or near-real-time insights. Big data technologies enable processing and analysing vast amounts of data quickly to meet these demands. Sentiment can be influenced by context, sarcasm, and other linguistic nuances. Big data helps models learn these complexities by exposing them to a wide range of examples.

2. LITERATURE REVIEW

[1]Venkat N Gudivana, research highlights the importance of large datasets in modern NLP research and applications. It states that the availability of massive training data can simplify models, leading to better performance than more complex models. The key to leveraging Web-scale data is to effectively utilise the available large-scale datasets.

[2] The study provides an overview of various techniques used in sentiment analysis (SA), focusing on recent research published until 2017. It discusses the importance of online opinions and comments and explores different approaches to extract them. The study highlights the popularity of Naive Bayes and SVM algorithms for sentiment classification and reviews the contributions of recent research in expanding the scope of SA. It also discusses the emerging use of big data (Hadoop) in SA and outlines future research directions in this area. The paper aims to be a valuable resource for new researchers entering the field of sentiment analysis, covering a wide range of techniques and illustrating different SA approaches for extracting and analysing sentiments.

[3] The study provides an overview of sentiment analysis, including different classification methods, their advantages and disadvantages, and the necessary steps involved in the process. It highlights the popularity of supervised machine learning methods, particularly NB and SVM, in sentiment analysis. The study discusses common application areas and explores the challenges associated with sentiment evaluation, emphasising domain dependence. It concludes by outlining future research directions to further expand the comparison and address the remaining challenges in sentiment analysis.

[4] The study provides a comprehensive review of sentiment analysis, examining its importance and challenges. It discusses key components like keyword extraction and classification methods, highlighting the potential of sentiment analysis in various domains. The study offers guidelines for developing effective sentiment analysis models, emphasizing the need for continuous improvement and exploration of new algorithms.

[5] The research paper highlights the importance of sentiment analysis in today's business landscape. It emphasizes the need for natural language processing (NLP) to bridge the gap between human language and machine understanding, enabling businesses to extract valuable insights from textual data. The paper also discusses the role of machine learning (ML) in sentiment analysis, exploring various techniques and algorithms that can be used to analyze sentiment effectively.

3 OBJECTIVES AND SCOPE OF WORK

3.1 Objectives

The primary objectives of research in Big Data and Deep Learning in NLP are: To develop more accurate and robust NLP models: By leveraging the vast amount of data available in the digital age, researchers aim to create NLP models that can better understand and process natural language, leading to more accurate and reliable results in various applications. To address the challenges of traditional NLP methods: Traditional NLP methods often struggle with large-scale datasets and complex linguistic phenomena. By incorporating deep learning techniques, researchers aim to overcome these

limitations and develop more effective NLP models. To enable new applications of NLP: The combination of Big Data and Deep Learning has the potential to unlock new applications of NLP, such as real-time language translation, personalized content generation, and advanced chatbots. To improve understanding of human language: By studying the patterns and relationships in large-scale language data, researchers can gain a deeper understanding of human language and communication.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2014-2018
e-ISSN : 2583-1062
Impact Factor : 7.001

3.2 Scope of Work

Phase 1

Researching	Researching about techniques and methods
Gathering Data	Looking for Big Data available online
Cuda	Installing Cuda and CUDN

Phase 2

Softwares	Installing required softwares
Libraries	Installing required libraries for the task

Phase 3

Importing Libraries	Importing Libraries in the Notebook
Importing Dataset	Importing Dataset
Cleaning Data	Cleaning the dataset and making it suitable and error free
Transforming Data	Converting the sentences to Tokens (Numerical Values)
Visualization	Visualising Data
Model Creation	Creating NLP model
Model Fitting	Training the model with the data
Model Testing	Testing the model performance on the test data
Model Saving	Saving the model architecture and weights

4 METHODOLOGY

4.1 Data Collection

Architecture:

The [6] dataset used in this study is a subset of the Stanford Sentiment Treebank (SST), a benchmark dataset for sentiment analysis tasks. SST consists of sentences extracted from movie reviews, along with their corresponding sentiment labels (positive, negative, or neutral). The size of this data set is 1,80,000 training set and 2,00,000 testing set with each polarity of positive and negative sentiment

4.2 Data Preprocessing:

SST consists of sentences extracted from movie reviews, product reviews etc along with their corresponding sentiment labels (positive, negative, or neutral). Now the sentences vary in length and vocabulary . To process it and pass to model we have to perform cleaning and preprocessing on it to convert string data into numerical values that can be fed into the model. This is done in few steps: Stop word removal is the process of eliminating common words, known as stop words, from the text. Stop words are words that do not carry significant semantic meaning, such as "the," "and," "a," and "in." By removing stop words, we reduce the dimensionality of the data and focus on the most informative words. Stemming is a simpler approach that involves removing suffixes and prefixes, while lemmatization is more sophisticated and attempts to find the base form of a word based on its part of speech and morphological rules. This allow us to normalise the words as we can have many different type of words Tokenization is the process of breaking down text into individual units, called tokens. These tokens can be words, subwords, or even characters, depending on the specific application.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2014-2018
e-ISSN : 2583-1062
Impact Factor : 7.001

4.3 Model Selection:

For this research we had gone with 'Simple Feed Forward Neural Network' designed for sentiment analysis. This architecture here consists of Embedding Layers, GlobalAveragePooling1D layer and at last Dense Layer. This Embedding layer converts input words into dense vectors, capturing the semantic relationships between words GlobalAveragePooling1D layer averages the embedding vectors for each word in the sequence, resulting in a single vector representing the overall sentiment. Dense layer is a fully connected layer with one neuron and a sigmoid activation function. It outputs a probability score between 0 and 1, representing the likelihood of the input text belonging to the positive class (e.g., positive sentiment).

4.4 Model Training:

The model is compiled with binary cross-entropy loss, Adam optimizer, and accuracy metric. This configuration is suitable for binary classification tasks like sentiment analysis, where the goal is to predict whether the input text belongs to one of two classes (positive or negative).

4.5 Model Evaluation:

Our model was evaluated with accuracy as its base parameter It measures the proportion of correct predictions made by the model out of the total number of predictions. In the context of sentiment analysis, a model is typically trained on a dataset of labelled reviews (positive or negative). During evaluation, the model is presented with new, unseen reviews, and it predicts their sentiment. The accuracy is calculated by comparing the model's predictions to the actual labels.

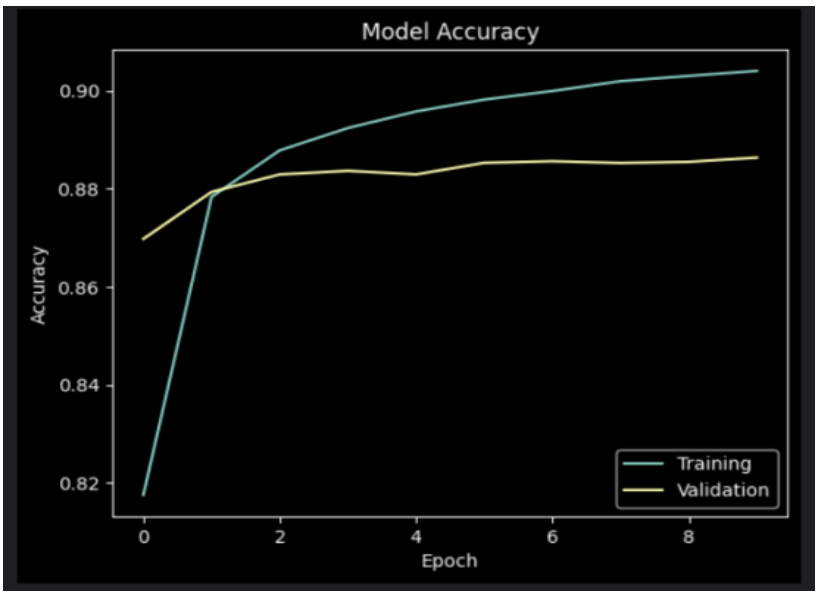

Fig. 1 Accuracy Graph

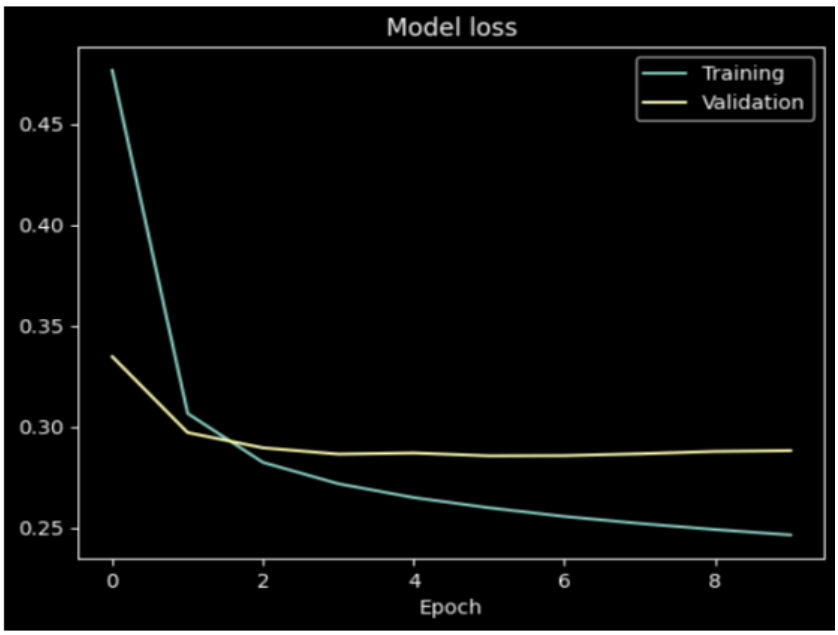

Fig. 2 LossGraph

5 CONCLUSION AND FUTURE WORK

5.1 Conclusion:

Here we investigated the application of a simple feedforward neural network for sentiment analysis on a large-scale Amazon review dataset. The model, trained on a dataset of 1,800,000 training samples and evaluated on 200,000 testing samples, achieved an accuracy of 88.4732%. The results demonstrate the effectiveness of deep learning models in capturing complex patterns and nuances in textual data, leading to accurate sentiment classification. The availability of a large-scale dataset was crucial for training a robust and generalizable model. The simple feedforward neural network architecture used in this study proved to be effective for sentiment analysis, highlighting the potential of even relatively basic models when combined with sufficient data. Future research could investigate more complex deep learning models, address class imbalance, improve interpretability, and incorporate domain-specific features. Overall, this research demonstrates the potential of deep learning for sentiment analysis on large-scale datasets

5.2 Future Work:

While the current research provides a solid foundation for sentiment analysis using deep learning on Amazon reviews, there is significant potential for further exploration. Incorporating domain-specific knowledge, addressing class imbalance, improving model interpretability, exploring more complex architectures, handling contextual nuances, and developing real-time sentiment analysis capabilities are all promising avenues for future research. By addressing these areas, we can continue to advance the state-of-the-art in sentiment analysis and its applications in various domains, such as customer satisfaction, market research, and brand management.

6 REFERENCES

[1] Venkat N Gudivana , “Big Data Driven Natural Language Processing Research and Application “ ,Published in HandBook of Statistics Aug 5 , 2015
[2] Ameen Abdullah Qaid Aqlan, B. Manjula and R. Lakshman Naik , “A Study of Sentiment Analysis: Concepts, Techniques, and Challenge “ , Published in Plant Long Non-Coding RNAs Jan , 2019
[3] Mayur Wankhade, Annavarapu Chandra Sekhara Rao & Chaitanya Kulkarni , “A survey on sentiment analysis methods, applications, and challenges “ Published in Artificial Intelligence Review Feb 7 , 2022
[4] Monali Bordoloi and Saroj Kumar Biswas , “Sentiment analysis: A survey on design framework, applications and future scopes“, Published in Springer Nature March 20 , 2023
[5] Sindhura Kannappan , “Sentiment analysis using product review data“, Published in Journal of Data Acquisition and Processing April , 2023
[6] AmazonReviewDataset(https://www.kaggle.com/datasets/kritanjalijain/am azon-reviews)