Sentiment Analysis Through CNN

MD.Arafat Muktadir[1], Md.Mahir Ahnaf Ahmed[2]
[1,2]Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh

***Abstract*—This paper explores the utilization of Convolutional Neural Networks (CNNs) for sentiment analysis, a facet of natural language processing. CNNs, originally designed for computer vision, exhibit significant efficacy in discerning sentiment from text through hierarchical feature extraction. The model leverages convolutional operations and pooling layers to capture local patterns and contextual relationships in sequential data. Preprocessing techniques, embedding layers, and transfer learning enhance its performance, addressing challenges like overfitting and imbalanced datasets. The abstract highlights promising results and underscores CNNs' potential in achieving state-of-the-art sentiment analysis across diverse domains, while acknowledging the need to address domain-specific nuances and challenges.**

***Index Terms*—Machine Learning, Deep Learning, CNN.**

I. Introduction

Text categorization and sentiment analysis are vital cornerstones in the quickly developing field of natural language processing, allowing computers to understand and interpret the subtleties that are buried in textual data. This study explores the revolutionary potential of Convolutional Neural Networks (CNNs) in text classification by analyzing how well they can decipher sentiment patterns. CNNs are a tempting option for identifying complex features and relationships within sequential textual material because of their extraordinary adaptability. CNNs were initially created for computer vision jobs. The increasing need for sophisticated text categorization models highlights CNNs' effectiveness in hierarchical feature extraction. The goal of this research is to better understand how CNNs are able to identify contextual links and local patterns, especially when it comes to sentiment analysis. The investigation goes beyond conventional techniques and presents a viable path to improve the precision and effectiveness of sentiment analysis jobs. The study is placed in the larger context of natural language processing research by the literature review. It is noteworthy that it incorporates ideas from foundational publications like "Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," which presents creative solutions to data constraints in NLP jobs. Furthermore, the analysis of "Natural Language Processing (almost) from Scratch" emphasizes a move away from task-specific designed characteristics and toward unified neural network structures. The dataset chosen for this investigation, sourced from the Internet Movie Database (IMDb), presents a real-world scenario for text classification. Comprising 50,000 labeled reviews categorized as favorable or negative, the dataset mirrors the challenges posed by imbalanced datasets and limited training data. Its inclusion of unlabeled data further enriches the learning environment, providing an opportunity for comprehensive model training. The proposed methodology unfolds systematically, encompassing dataset collection, preprocessing, strategic partitioning into training and testing sets, deep learning model training, testing, and result analysis. This structured approach aims to provide a robust framework for leveraging CNNs in text classification tasks, ensuring reproducibility and clarity in the experimental process. In presenting this exploration of text classification through CNNs, we aspire to contribute to the growing body of knowledge in natural language processing and offer insights into the potential of CNNs as powerful tools for sentiment analysis. The paper acknowledges the evolving landscape of NLP and sets the stage for a deeper understanding of CNNs' adaptability in unraveling the complexities embedded in textual data.

II. Literature Review

[1] has the title of Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks presents a set of simple and universal data augmentation techniques for natural language processing (NLP) tasks, specifically text classification. The authors introduce four text editing operations, including synonym replacement, random insertion, random deletion, and random swap, which can be applied to existing text data to generate new samples. The goal of EDA is to overcome the problem of limited training data, which is a common issue in NLP tasks. The authors evaluate EDA on five benchmark classification tasks and show that it provides substantial improvements on all five tasks, particularly for smaller datasets.

[2] titled "Natural Language Processing (almost) from Scratch" proposes a unified neural network architecture and learning algorithm that can be applied to various natural language processing tasks, including part-of-speech tagging, chunking, named entity recognition, and semantic role labeling. The authors emphasize the reduced reliance on a priori NLP knowledge and the avoidance of task-specific engineered features. Instead, the system learns internal representations on the basis of vast amounts of mostly unlabeled training data, resulting in a freely available tagging system with good performance and minimal computational requirements. The proposed system is faster and requires less memory compared to other tagging systems, and is more generalizable, meaning that it can be applied to multiple NLP tasks without the need for task-specific feature engineering. However, the authors acknowledge that their system is not perfect and that there is still room for improvement. They also note that their system may not be suitable for all NLP tasks and that there may be cases where task-specific feature engineering is necessary.

Overall, the paper provides a solid foundation for future research in this area and is recommended for anyone interested in natural language processing and machine learning.

III. Research Objectives

Our research objective is to evaluate CNN in Sentiment Analysis. Also, to assess the effectiveness of Convolutional Neural Networks (CNNs) in sentiment analysis tasks, specifically focusing on their ability to extract hierarchical features from sequential textual data.

IV. Research Methodology

A. Proposed Workflow

Here, our proposed methodology consists of six parts. First, we will collect the dataset. Then, we will do preprocessing on the dataset. To elaborate, we will set the word limit to a constant for every sentence. Also, we will split the dataset into 2 parts. They are Training and Testing. The ratio would be 5:5. Now, we will be training the deep learning model. After that, we will be testing the deep learning model. Finally, we will acquire the results and analyze them. We have given two figures below for a better understanding. The first figure is the proposed methodology and the second figure is the proposed Deep Learning model.

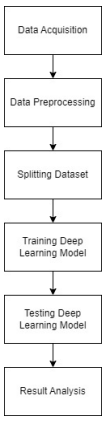

Fig. 1. Work Procedure

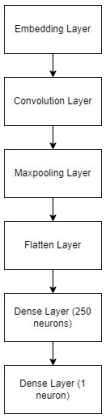

Fig. 2. Work Procedure of Feature Extraction From Training & Validation Dataset

B. Dataset

1) Dataset Details: The 50,000 reviews from the Internet Movie Database (IMDb) that have been classified as favorable or negative make up the IMDb Movie Reviews dataset, which is a binary sentiment analysis dataset. There are an equal amount of favorable and negative evaluations in the sample. Only really divisive reviews are taken into account. A review is rated ¡ 4 out of 10, if it is unfavorable, and ¿ 7 out of 10, if it is good. Reviews for each movie are limited to thirty. There is more unlabeled data in the dataset.

2) Dataset Preprocessing: Here, we have used an embedding layer which will take input as 450 words maximum and provide a 32 dimensional output of those words.

C. Convolutional Neural Network Architecture

We have used a custom CNN architecture that has 6 layers. The first layer is the embedding layer. The second layer is the convolution layer that has the dimension of 32×3. The activation function that has been used here is ReLU. The third layer is the max pooling layer.The fourth layer is the flatten layer. The fifth layer is dense layer, which has the activation function ReLU and has 250 neural units. The final layer is the dense layer, which has 1 neural unit.

V. Result And Analysis

Here, we have trained 1 convolutional neural network model. Then, we have tested it and acquired the accuracy.

A. Testing With the CNN Model

We have acquired the accuracy of 87.86% on the first epoch. Then on the second epoch we have acquired the accuracy of 88.64%. This is the final accuracy we have achieved.

B. Hyperparameter Values of the CNN Model

We have used the same hyperparameter values for all 4 CNN models that we have used. Here, we have shown them in table[I].

TABLE I
Hyperparameter Values

Hyperparameters	Values
Batch Size	128
Epoch	2
Loss Function	Binary Cross Entropy
Optimizer	Adam

VI. Conclusion

In summary, this paper explored the use of Convolutional Neural Networks (CNNs) in text classification, focusing on sentiment analysis. While highlighting CNNs' potential, we addressed challenges like limited sequential modeling and fixed input size. The study proposed a structured methodology, leveraging insights from related literature and real-world analysis using the IMDb dataset. Despite promising results, it's essential to acknowledge CNNs' limitations, emphasizing the

need for nuanced approaches in language-related tasks. Future work may involve hybrid models and increased interpretability efforts to enhance practical utility. This research contributes to advancing the understanding of CNNs in natural language processing, paving the way for more robust text classification solutions.

VII. Bibliography

1. https://aclanthology.org/D19-1670.pdf
2. https://dl.acm.org/doi/pdf/10.5555/1953048.2078186