Sentiment Analysis Using Feature Fusion and Convolutional Neural Networks

Oğulcan Altunörgü

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Engineering

Eastern Mediterranean University
August 2023
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science in Computer Engineering.

Prof. Dr. Zeki Bayram
Chair, Department of Computer Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Computer Engineering.

Assoc. Prof. Dr. Adnan Acan
Supervisor

Examining Committee

1. Assoc. Prof. Dr. Adnan Acan
2. Assoc. Prof. Dr. Mehtap Köse Ulukök
3. Asst. Prof. Dr. Ahmet Ünveren

ABSTRACT

The aim of this study is to use deep learning methods for sentiment analysis on a database of well-known documents. It aims to show the comparative success of the prepared algorithm against some published methods and to examine the data dependency using textual datasets.

Sentiment analysis (or opinion mining) is a method used in natural language processing to analyze text documents and determine the content of a text item. Natural language processing (NLP) is a branch of artificial intelligence (AI) that enables computers to understand and learn human language. When each word is expressed as a vector, the meanings of the words will be stored in the vectors, and the vectors of the words that have close meanings will be close to each other. Words are converted into mathematical expressions and their meanings are reduced to data that computers can process by expressing their proximity to another word mathematically. In this study, we aimed to use Word2Vec, LSA and CNN three popular text-based feature extraction methods used in natural language processing and machine learning. Word2Vec is trained on a large text dataset to obtain word vectors, ensuring that similar-meaning words are close together in the vector space. In this way, the relationships and meaning similarities between words are reflected. LSA is a text mining method used to extract semantic content from large text datasets. LSA uses matrix factorization technique to discover the structure between documents and words. We aim to perform sentiment analysis using deep learning models such as TextCNN and Conv1D.

TextCNN and Conv1D are convolutional neural networks effective for feature extraction and classification in text data. These models are used to extract features from text data and conduct sentiment analysis. The datasets used include Sentiment140, YELP, Amazon, and IMDB.

Keywords: word, machine learning, lsa, word2vec, textcnn, cnn.

ÖZ

Bu çalışmanın amacı bir dizi iyi bilinen belge veri tabanı üzerinde duygu analizi için derin öğrenme yöntemlerinin kullanılmasıdır. Hazırlanan algoritmanın bazı yayınlanmış yöntemlere karşı karşılaştırmalı başarısını göstermeyi, metinsel veri kümelerini kullanarak veri bağımlılığını incelemeyi amaçlamaktadır.

Duygu analizi, metin belgelerini analiz etmek ve metin öğesinin içeriğini belirlemek için doğal dil işlemede (NLP) bir yöntemdir. Doğal dil işleme yapay zekanın bir koludur ve bilgisayarların, insan dilini kavramasını, öğrenmesini ve idare etmesini sağlar. Her bir sözcük vektörel olarak ifade edildiğinde sözcüklerin anlamları vektörlerde saklanarak, birbirine yakın anlam taşıyan sözcüklerinde vektörleri birbirine yakın olacaktır. Sözcükler matematiksel ifadelere dönüştürülerek anlamları başka bir sözcüğe yakınlığı yine matematiksel olarak ifade edilerek bilgisayarların işleyebileceği verilere indirgenirler. Bu çalışmada doğal dil işleme ve makine öğrenmesi alanlarında kullanılan iki popüler metin tabanlı özellik çıkarma yöntemi olan Word2Vec, LSA ve CNN kullanmayı hedefledik. Word2Vec, kelime vektörlerini elde etmek için büyük metin veri kümesi üzerinde eğitilir ve benzer anlamlı kelimelerin vektör uzayında birbirine yakın olmasını sağlar. Bu şekilde, kelimeler arasındaki ilişkiler ve anlam benzerlikleri yansıtılır. LSA ise büyük metin veri kümelerinden anlamsal içerik çıkarmak için kullanılan bir metin madenciliği yöntemidir. LSA, belgeler ve kelimeler arasındaki yapıyı keşfetmek için matris çözümleme (matrix factorization) tekniğini kullanır. TextCNN ve Conv1D gibi derin öğrenme modellerini kullanarak duygu analizini gerçekleştirmeyi hedefliyoruz.

TextCNN ve Conv1D, metin verilerinde özellik çıkarma ve sınıflandırma için etkili olan evrişimli sinir ağlarıdır. Bu modeller, metin verilerinin özelliklerini çıkarmak ve duygu analizi yapmak için kullanılır. Kullanılan veri setleri Sentiment140, YELP, Amazon ve IMDB verilerini içermektedir.

Anahtar Kelimeler: word, machine learning, lsa, word2vec, textcnn, cnn

DEDICATION

To my family

ACKNOWLEDGMENT

I would like to express my gratitude to Assoc. Prof. Dr. Adnan ACAN for his supervision, advice, and guidance from the initial stages of this thesis, which provided me with exceptional experiences. He has been not only an academician but also a mentor and a source of inspiration for me. Throughout the working process, he continuously provided encouragement and support in various ways. His ideas and experiences truly contributed to my development. I am deeply indebted to him for all that he has done for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
CNN	Convolutional Neural Network
Conv1D	1D Convolutional Neural Network
IMDB	Internet Movie Database
LSA	Latent Semantic Analysis
NLP	Natural Language Processing
TextCNN	Text Convolutional Neural Network

Chapter 1

INTRODUCTION

1.1 Sentiment Analysis

The purpose of Sentiment Analysis is to determine whether comments about a topic are positive, negative, or neutral. These comments can be written or spoken, and there is no limitation on the subject of the documents. The main objective here is to enable machines to predict whether these comments are positive or negative, similar to a human, and do so automatically. Sentiment Analysis, also known as opinion mining, is an important field in Natural Language Processing (NLP) [1]. It focuses on determining the emotional content of textual data. This article provides a general overview of the methods, techniques, and various application areas of Sentiment Analysis. It explores the significance of understanding the emotional states expressed in texts. In the digital age, large amounts of textual data are generated daily, and understanding the emotions and opinions expressed in this data is crucial for businesses and organizations. Sentiment Analysis aims to automatically classify these texts into positive, negative, or neutral emotions. The techniques used in sentiment analysis offer insights into the efforts made to achieve accurate and efficient results. We aimed to use machine learning-based methods for Sentiment Analysis. Machine learning algorithms require labeled training data for sentiment classification. In recent times, especially deep learning techniques like Convolutional Neural Networks (CNNs) have shown promising results in capturing the context and meaning of textual data.

Sentiment Analysis has various applications in different fields. In the business world, companies use sentiment analysis to measure customer satisfaction, analyze product reviews, and understand brand perception [2]. Social media platforms employ sentiment analysis to understand users' opinions and emotional reactions. In political analysis, sentiment analysis is used to grasp the emotional state of the public.

Sentiment Analysis studies begin by taking a collection of documents in any file format (pdf, html, xml, word, etc.). This document input is simplified using preprocessing methods such as tokenization, part-of-speech tagging, information extraction, and establishing relationships between words [3]. To make the document more comprehensible, dictionaries specific to the document's language and other language-related resources can be used.

Once the document is simplified and analyzed through dictionaries, the decision is made on which Sentiment Analysis Approach to apply. These approaches may vary depending on the nature of the research. After applying the chosen method, the resulting data is prepared and presented in a way that the end-user can understand.

1.2 Literature Review

Sentiment Analysis is an important task in Natural Language Processing (NLP) that aims to determine emotions, feelings, or opinions in textual data. Studies in this field have shown that deep learning techniques play a crucial role in achieving successful results for Sentiment Analysis.

In this text, we will focus on Turkish Sentiment Analysis studies conducted using deep learning approaches and examine some quotations to support these studies.

Introduction to Sentiment Analysis:

Deep learning techniques have achieved significant success in text-based tasks such as Sentiment Analysis [4]. These techniques automatically learn linguistic features and extract hierarchical representations to determine the sentiment of texts.

Tokenization in Sentiment Analysis:

Tokenization is an important step in processing text data for Sentiment Analysis [5]. It involves breaking down texts into smaller units, such as words or subwords, to facilitate the conversion of texts into numerical data.

Global Vectors for Word Representations (GloVe):

GloVe is a commonly used method to create numerical representations of words in texts [6]. By utilizing the relationships between word vectors, GloVe effectively represents the meaning of words.

Sentiment Analysis with Convolutional Neural Networks (CNNs):

Convolutional Neural Networks, which have been successful in image processing, are also employed in Sentiment Analysis [7]. CNNs identify important patterns and features in text content to classify sentiment effectively.

Sentiment Analysis with Long Short-Term Memory (LSTM) Networks:

LSTM networks are used to capture long-term dependencies in text data and yield successful results in Sentiment Analysis tasks [8]. LSTM networks understand the context in text content and accurately classify sentiment.

Sentic LSTM: Hybrid Networks for Targeted Aspect-Based Sentiment Analysis: Sentic LSTM is a hybrid network designed for targeted aspect-based sentiment analysis in text data [9]. This method achieves successful results in determining sentiment related to specific targets.

Deep learning approaches for Sentiment Analysis [10] serve as essential tools for effectively determining emotional content in textual data. Convolutional Neural Networks, Long Short-Term Memory networks, and GloVe representations are influential in improving sentiment analysis performance and are widely used in various text-based tasks. Combining and enhancing these methods present an exciting research area to further advance Sentiment Analysis.

1.3 Word2Vec

Word2Vec is a machine learning technique used for word embedding [11]. However, Word2Vec is often considered part of deep learning as one of its foundational components, based on neural network models, and is commonly trained using an approach called deep learning.

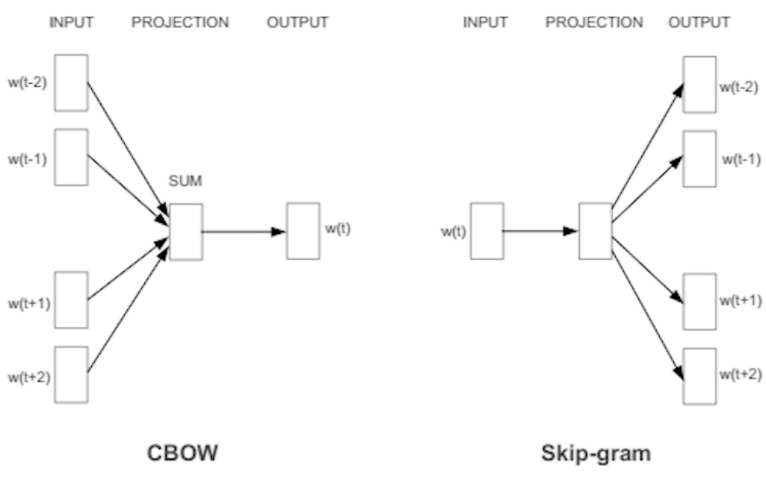

Figure 1: Architecture of Word2Vec models: CBOW and Skip-Gram

Word2Vec works by learning word vectors from large text corpora. These vectors are numerical representations that reflect the meaning, usage context, and relationships of a specific word in the language.

The Word2Vec algorithm aims for similar semantically meaningful words to have similar vectors, allowing the word vectors to capture semantic relationships in the language [12].

It falls under the category of machine learning as it works by taking examples from large datasets and building a model based on them. However, deep learning refers specifically to the fundamental neural network architectures used in the training of Word2Vec. In particular, Word2Vec models are often trained using "deep" neural networks, which are a type of artificial neural network.

Using Word2Vec for sentiment analysis can enhance the analysis of text data by considering the semantic similarities of words. For example, it can observe that certain words expressing emotions have similar vectors to other semantically related words. As a result, the model can better determine the emotional tone of a text by taking the context surrounding the word into account.

1.4 LSA (Latent Semantic Analysis)

Latent Semantic Analysis (LSA) is a statistical and natural language processing (NLP) method used to discover the hidden semantic structure in large text collections. This method is employed to uncover the semantic relationships between words in texts and to group similar content documents together [13].

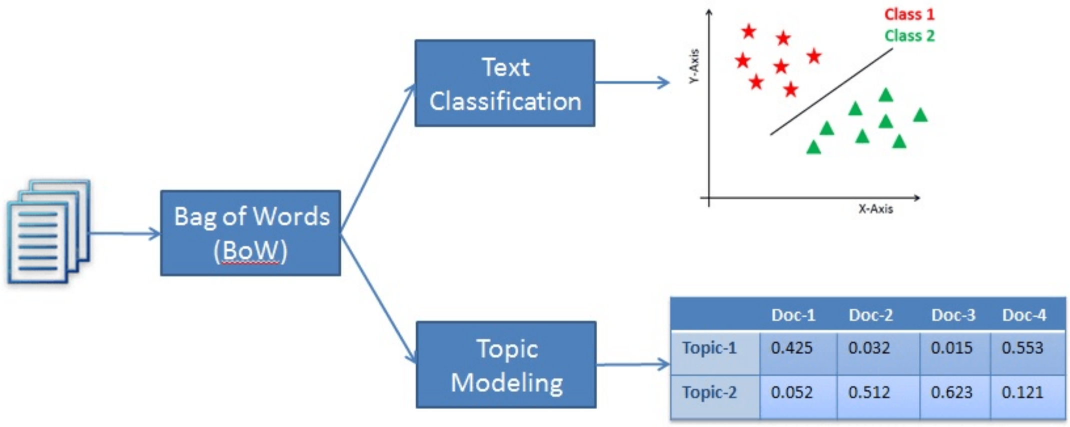

Figure 2: Word2Vec is a Machine Learning Technique

LSA uses mathematical techniques to represent texts in a high-dimensional vector space. These mathematical techniques are based on a matrix factorization method, typically referred to as Singular Value Decomposition (SVD). SVD helps reveal the hidden structure within the data by decomposing a matrix into three separate matrices.

The fundamental steps of LSA are as follows:

Text Data Preparation:

Firstly, the text documents are taken, and preprocessing steps are applied to remove irrelevant information, convert the text to lowercase, and eliminate stopwords (frequently used but typically uninformative words).

Term-Document Matrix Creation: A term-document matrix is constructed, where each row represents a document, and each column represents a word. The cells of the matrix contain the frequency of the word in the document or some other word representation measure.

Applying SVD: SVD is applied to the created term-document matrix. By decomposing the matrix into three matrices, SVD reveals the latent structure within the matrix: the left singular value matrix, the right singular value matrix, and the singular values matrix.

Dimension Reduction: Often, the dimension of the obtained singular values matrix is reduced. This helps reduce noise and achieve a more effective representation.

Semantic Representation: In the final step, the reduced-dimension matrix is transformed into a vector space that contains lower-dimensional semantic representations of documents. This results in a more compact and efficient representation that captures the semantic similarities and relationships in the texts. LSA is used in various NLP tasks such as text-based information retrieval, text classification, summarization, word recommendation, and sentiment analysis.

When applied to large text collections, it can help in partitioning documents into groups with similar content or meaning and discovering hidden semantic relationships between documents.

1.5 LTSM (Long Short-Term Memory)

LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) architecture [14]. RNNs are used for operations based on sequential data (sequences).

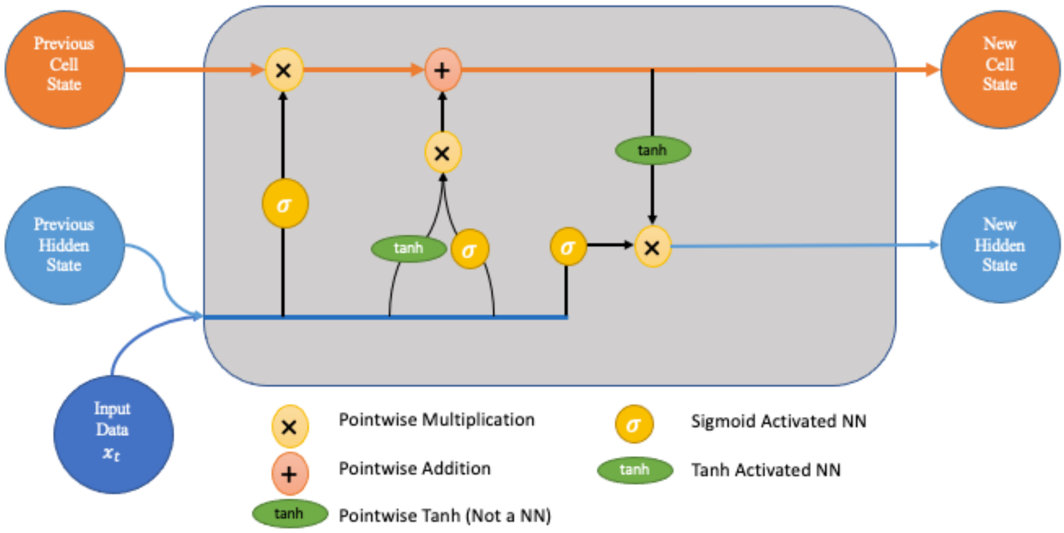

Figure 3: LSTM Model Diagram

However, traditional RNNs are known to have difficulty capturing long-term dependencies due to the vanishing gradient problem. LSTM is designed to address this issue and is particularly effective for tasks that require capturing long-term dependencies, such as natural language processing and other time series data analysis tasks.

LSTM was introduced in 1997 by Sepp Hochreiter and Jürgen Schmidhuber. The basic idea is to use a more complex cell structure compared to traditional RNNs to better retain long-term dependencies [15]. LSTM consists of three main components in this cell structure:

Cell State: The cell state represents the memory of the LSTM and can be thought of as a flow that stores information seen in the past. The cell state is updated at each step, allowing the model to retain important information over long periods.

Forget Gate: The forget gate decides which information in the cell state to forget. Clearing out old and irrelevant information prevents the formation of incorrect dependencies. The forget gate is calculated by multiplying a vector generated based on the past state and the current input.

Input Gate and Output Gate: The input gate controls how new information is added to the cell state. The output gate determines how the updated cell state will control the output. These gates are calculated using sigmoid activation functions (which produce values between 0 and 1).

LSTM learns through the input, forget, and output gates how much information to retain and how much to forget. This enables the model to capture long-term dependencies and effectively analyze relationships in sequential data. LSTM is a powerful deep learning architecture that has shown successful results in various tasks, including natural language processing, text generation, translation, language modeling, speech recognition, and time series analysis.

1.6 CNN (Convolutional Neural Network)

CNN (Convolutional Neural Network) is a deep learning architecture designed for computer vision and other visual data analysis tasks. It has been particularly successful in processing and analyzing image and video data, revolutionizing the field of computer vision.

CNNs differ from traditional fully connected structures as they incorporate special layers such as convolutional and pooling layers. These layers are designed to preserve local connections and the visual data structure while effectively extracting features [16].

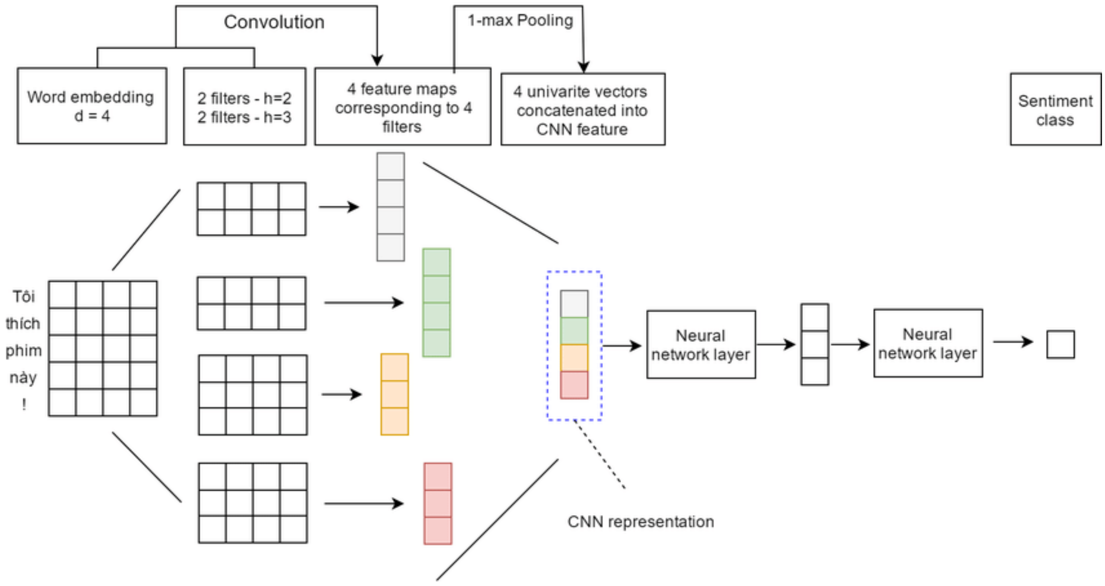

Figure 4: Illustration of our CNN model for sentiment analysis.

The key features of CNN are as follows:

Convolutional Layer: This layer performs a convolution operation on the input data using one or more filters (kernels).

The convolution process involves sliding the filter over the input data, creating feature maps. Filters are learned to capture various patterns in images and extract features effectively.

Activation Functions: Activation functions (e.g., ReLU) are often used after the convolutional layers. These functions allow the network to learn more complex and nonlinear relationships.

Pooling Layer: Pooling layers are used to downsample the outputs of the convolutional layers and reduce computational load.

They can also increase translation (displacement) invariance of the features. Typically, techniques like max pooling or average pooling are used [17].

Fully Connected Layers: After the convolutional and pooling layers, CNNs contain one or more fully connected layers. These layers are used for specific tasks such as classification, recognition, or prediction of the results.

CNN has achieved significant success in various visual data analysis tasks in computer vision, such as object recognition, face recognition, object detection, image classification, image segmentation, and style transfer. Adaptations have also been made for text-based tasks like text classification. CNN has made significant progress in the field of deep learning and is now effectively utilized in numerous applications and industries.

1.7 TextCNN (Convolutional Neural Network)

TextCNN is a deep learning model used for text classification and feature extraction tasks by employing deep learning techniques on text data. It stands for "Text Convolutional Neural Network."

TextCNN processes text data using a convolutional neural network (CNN) architecture. While CNNs are traditionally used for image recognition tasks, specialized variations like TextCNN have proven to be highly effective for analyzing text data.

In TextCNN, different-sized convolutional filters (kernels) are used to process text documents. These filters scan the document with varying window sizes to capture different features of the text.

TextCNN can classify text documents or extract features from them using the feature maps obtained after these convolution operations. It is particularly successful in text classification tasks, including sentiment analysis, text classification, and language detection, among others.

TextCNN is a valuable tool for extracting features from text data and automating text classification tasks, and it is widely used in such applications.

1.8 Conv1D (1D Convolutional Neural Network)

Conv1D stands for "1D Convolutional Neural Network," and it is a deep learning model used particularly for the analysis of one-dimensional data sequences, such as time series data.

Conv1D is a type of convolutional neural network (CNN) designed to process one-dimensional data, including tasks related to text processing or audio processing.

Conv1D is employed to perform feature extraction and classification tasks on one-dimensional data. It is especially effective in fields like pattern recognition in time series data, natural language processing, and speech recognition. The primary purpose of Conv1D is to identify significant features within a data sequence and utilize these features for data classification or analysis through convolutional operations.

Conv1D typically employs convolutional filters of various sizes to scan the data sequence. These filters are used to capture specific patterns or features within the data sequence. Subsequently, these learned features can be utilized for classification or analysis.

Conv1D is a deep learning model designed to perform feature extraction and classification on one-dimensional data sequences, and it is particularly useful for tasks involving deep learning approaches on such data types.

Chapter 2

SENTIMENT140

Sentiment140 is a large English text dataset collected from Twitter and labeled for sentiment analysis purposes. This dataset serves as a valuable source of training and testing data in machine learning and natural language processing (NLP) fields. It contains tweets from Twitter users, and each tweet is labeled as either positive, negative, or neutral. Each tweet is marked with a specific emotional category. If a tweet contains a positive sentiment, it may have the "positive" label, while if it contains a negative sentiment, it may have the "negative" label. This labeled dataset holds significant importance during the training phase of supervised learning algorithms used to train sentiment analysis models or classifiers. It enables the model to learn to detect emotional content, and later analyze new data [18].

The Sentiment140 dataset has been used in various studies focusing on sentiment analysis in fields such as social media analysis, product reviews, and understanding user reactions. Especially in sentiment analysis, large datasets are required for the development and validation of machine learning-based methods.

The dataset consists of 1,600,000 tweets extracted using the Twitter API. The tweets are labeled with numerical values (0 = negative, 4 = positive) and can be used for sentiment detection [19].

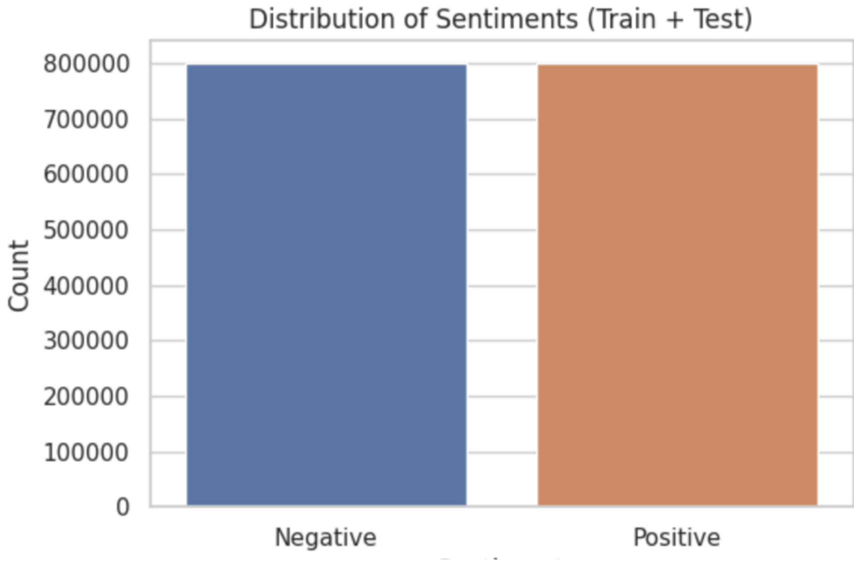

Figure 5: Negative and Positive Sentiment140 Data Count

Sentiment140 dataset includes the following 6 fields:

Target: The target field indicates the emotional state of the tweet. It is represented as a numerical value where 0 stands for negative sentiment, 2 for neutral sentiment, and 4 for positive sentiment.

Ids: The ids field contains a unique identifier for each tweet, such as "2087".

Date: The date field represents the date and time when the tweet was posted. For example, "Sat May 16 23:58:44 UTC 2009".

Flag: The flag field indicates the relevant query for the tweet. For instance, if there is a query associated with the tweet, it will be mentioned here (e.g., "lyx"). If there is no query associated, this field will be marked as "NO_QUERY".

User: The user field contains the name of the user who posted the tweet. For example, "robotickilldozr".

Text: The text field contains the actual content of the tweet. For instance, "Lyx is cool".

Each entry in the dataset includes these fields to describe the emotional context of the tweets, making it suitable for training and testing sentiment analysis models or classifiers.

Figure 6: Sentiment140 Dataset Includes The Following 6 Fields

	target	ids	date	flag	user	text
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....

Figure 7: Sentiment140 Top 5 Data

```
# The file path of the dataset
filepath = "/content/drive/MyDrive/sentiment140-Dataset/training.1600000.processed.noemoticon.csv"

# Load dataset
df = pd.read_csv(filepath, header=None, names=["target", "id", "date", "flag", "user", "text"], encoding='latin-1')
```

Figure 8: Loading The Dataset With Pandas

We split the data into training and test sets, using 80% for training and 20% for testing. Then, we tokenized the sentences. The "sentence tokenization" process involves breaking down the text into smaller components, such as words and punctuation marks. This step is essential for processing text-based data and performing linguistic analysis. Tokenization is a common first step in text processing and natural language processing projects.

The "word_tokenize" function parses a given sentence, taking into account words and punctuation marks, and returns a list. For example, the sentence "Hello, world!" can be tokenized as follows: ["Hello", ",", "world", "!"].

After tokenization, we used the Conv1D model to process and analyze text data, transforming textual information into a numerical format. Conv1D is a deep learning model used for processing text data and extracting features. This model applies convolution operations to convert text data into vectors, making it suitable for machine learning algorithms. At the end of the training process, the Conv1D model achieved an accuracy of 0.9328 on the training dataset and 0.7864 on the validation dataset. Conv1D is a crucial tool for extracting features and classifying text data by applying convolution operations to textual information.

KFOLD

KFold is a cross-validation technique used to evaluate the model's performance in a more reliable way. It involves dividing the data into training and test sets to assess how well the model fits the data. We used K=10, which means the data is divided into 10 equal-sized folds. In each iteration, one fold is used as the test set while the other folds are used as the training set. This process is repeated 10 times, and at the end of each iteration, performance metrics such as accuracy are calculated. The final model performance is obtained by averaging the accuracy values from all iterations [20].

As a result, using KFold, we tested the model's performance 10 times, and the average accuracy obtained was 96.08%. This approach allows for a more reliable evaluation of the model's performance.

Chapter 3

YELP

Yelp is a publicly traded American company headquartered in San Francisco, California. The Yelp dataset is a large text dataset obtained from Yelp, an online business and service review platform, frequently used in the fields of natural language processing (NLP) and data mining. This dataset consists of user reviews and ratings of various businesses. It captures users' experiences with different establishments, including their opinions and comments [21].

The Yelp dataset contains evaluations of businesses, where each review includes a numerical rating and a text comment written by the user. Researchers commonly use this dataset for tasks such as text analytics, sentiment analysis, location-based services, social media analysis, and business evaluations. It aims to provide valuable insights into the quality of services offered by businesses, user experiences, product popularity, and more. The dataset is widely used for training and testing machine learning algorithms and developing various NLP techniques, such as sentiment analysis. It serves as a labeled dataset, facilitating the development of sentiment analysis and other NLP methods.

Glossary of terms

Yelp: It is a website where reviews are posted to support businesses, business owners and users.

Business: It lists many organizations such as Stores, Cafes, Home Supplies, Automotive sector.

Existing business owner: People who use Yelp to list their businesses and get feedback from users.

Future business owner: It is a person who wants to have a new business or start a business in the future.

User: They are people who write reviews after visiting different places or businesses, sign up for Yelp, or look for a job.

Analytics: They are the people who analyze and use the data in the system.

Review: These are the comments made by users after visiting the business. Comparison can be made out of 5 points.

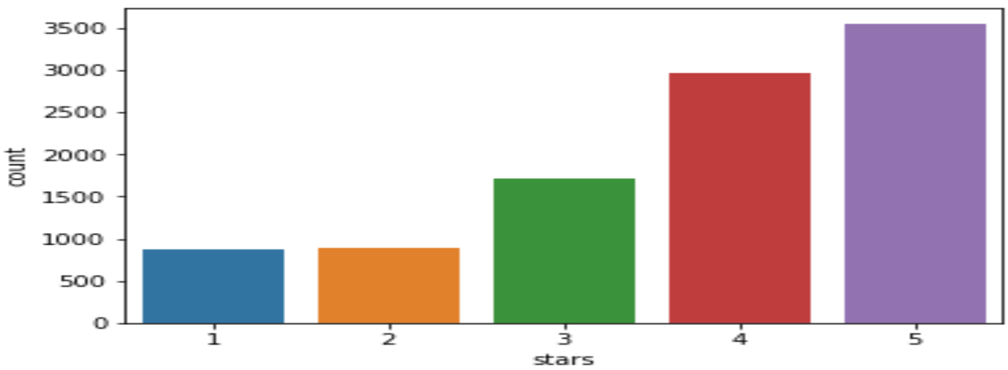

Figure 9: YELP Sample 10k Data Stars List

There are approximately 6,990,282 rows in the dataset.

We divided the data into training and test sets, using 80% for training and 20% for testing. Afterward, I tokenized the sentences. Following tokenization, I converted the sentences in the training and test data sets into vectors and saved them. The goal was to transform textual data into a numerical format, making it usable for machine learning algorithms.

Subsequently, I trained the model and calculated the accuracy. The Conv1D model achieved an accuracy of 99.03% on the training dataset. However, the accuracy value on the test dataset was 69.28%.

Conv1D is a deep learning model that applies convolution operations on text data, aiming to identify and classify text features. Therefore, the numerical representation and accurate processing of text data are crucial steps that impact the success of machine learning models.

Chapter 4

AMAZON

The Amazon dataset contains rich and diverse information about products available on the e-commerce platform. It includes details, features, and user reviews of various products across different categories. The dataset covers a wide range of product categories and individual products. Each product's detailed information includes its name, description, price, category, and other technical specifications. Additionally, customer ratings, scores, and user comments on the products are also part of the dataset [22]. The Amazon dataset is designed to be used in various fields such as natural language processing, sentiment analysis, product recommendation systems, price analysis, and market research. Using this dataset, one can identify popular products on the Amazon platform, understand customer trends, and analyze product features and prices. Due to its large size and labeled data, it serves as an ideal source for training and validating machine learning algorithms. Model developers can use this dataset to understand product evaluations and reviews, train sentiment analysis models, and predict product popularity. Overall, the Amazon dataset can provide a comprehensive understanding of Amazon's e-commerce platform.

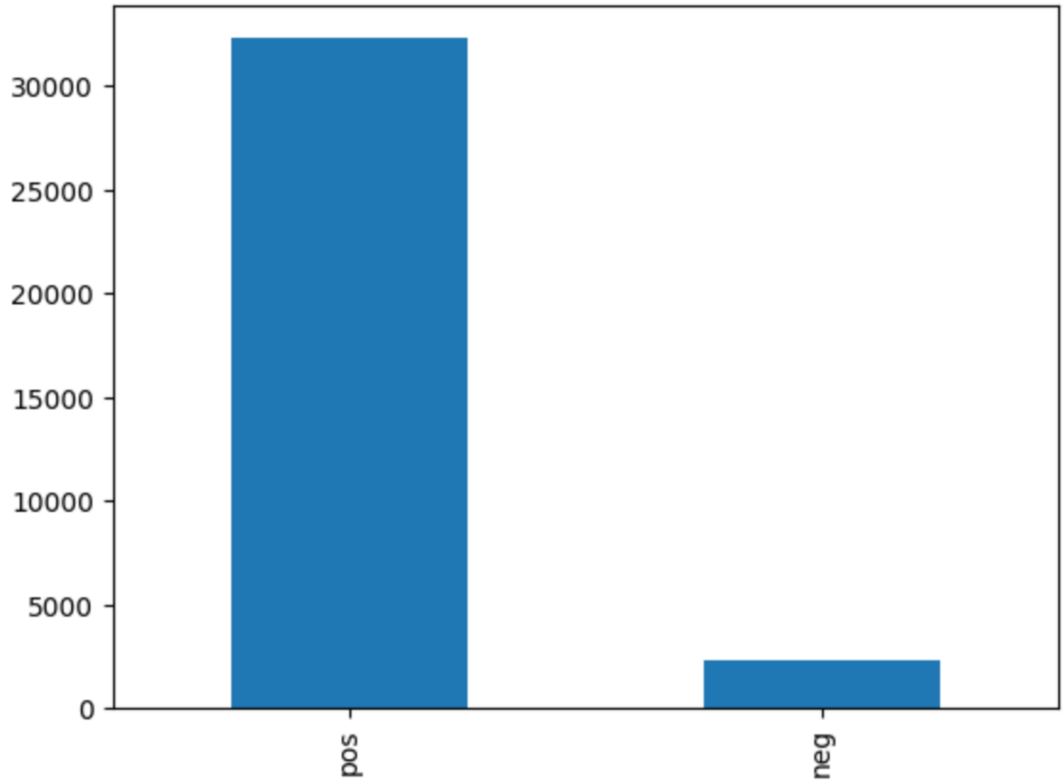

Figure 10: Amazon Negative and Positive Count

After splitting the data into training and test sets with an 80-20 ratio, I tokenized the sentences.

	reviews.rating	reviews.text	reviews.title	reviews.username
0	5.0	This product so far has not disappointed. My c...	Kindle	Adapter
1	5.0	great for beginner or experienced person. Boug...	very fast	truman
2	5.0	Inexpensive tablet for him to use and learn on...	Beginner tablet for our 9 year old son.	DaveZ
3	4.0	I've had my Fire HD 8 two weeks now and I love...	Good!!!	Shacks
4	5.0	I bought this for my grand daughter when she c...	Fantastic Tablet for kids	explore42

Figure 11: Sentiment140 Top 5 Data

I converted the sentences into training and test datasets and transformed them into vectors, saving the results. The goal was to convert text data into a numerical format suitable for machine learning algorithms. Afterward, I trained the TextCNN model and calculated the Accuracy value. The TextCNN model achieved an accuracy of 99.77% on the training dataset and 92.87% on the validation dataset.

Chapter 5

IMDB

IMDb (Internet Movie Database) dataset contains 50,000 movie reviews collected from a well-known film database. IMDb is a widely used platform that provides a comprehensive movie database worldwide, allowing users to add reviews and ratings for films. Each example in the dataset includes reviews written by IMDb users about different movies. These reviews reflect users' opinions and thoughts about the films, and each review is accompanied by a rating given by the user. These ratings indicate how much users liked a particular movie according to IMDb's rating system [23]. The dataset is designed for text-based analyses such as natural language processing (NLP) and sentiment analysis.

By using this dataset, we can perform sentiment analysis on movie reviews, classify positive and negative sentiments, and conduct various analyses with test and train datasets. The IMDb dataset can also be utilized for training and validating machine learning models. The labeled reviews and ratings can be used for developing and testing text classification algorithms. This allows models analyzing movie reviews to better understand emotional and semantic content in natural language, leading to more accurate classification of films based on user sentiments.

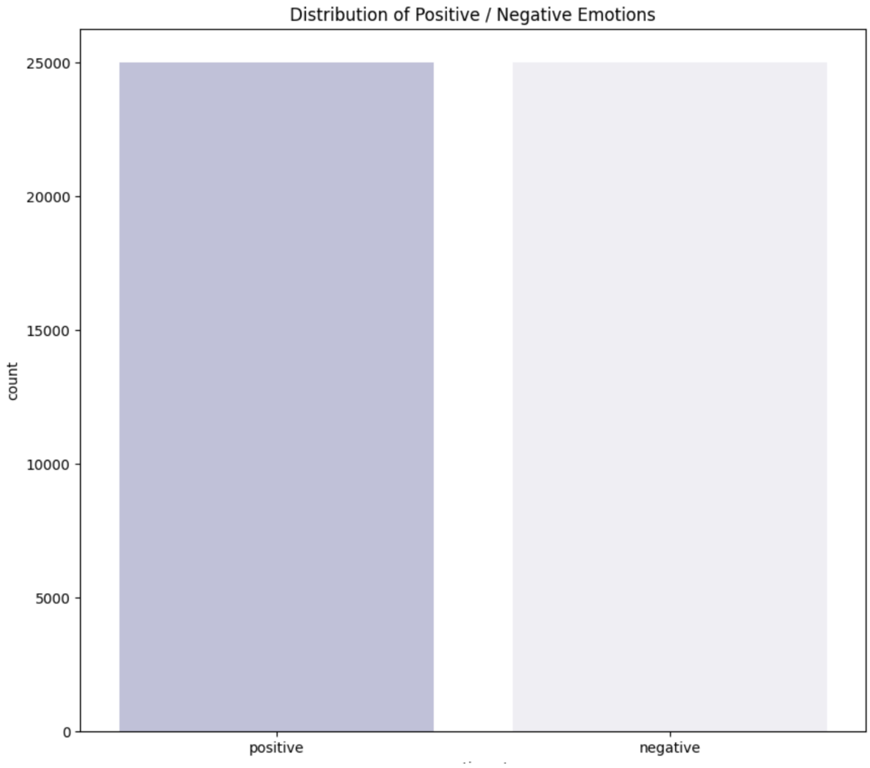

Figure 12: IMDB Negative and Positive Count

After splitting the data into training and test sets with an 80-20 ratio, I tokenized the sentences.

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive
5	Probably my all-time favorite movie, a story o...	positive
6	I sure would like to see a resurrection of a u...	positive
7	This show was an amazing, fresh & innovative i...	negative
8	Encouraged by the positive comments about this...	negative
9	If you like original gut wrenching laughter yo...	positive

Figure 13: IMDB Top 10 Data

I converted the sentences into text vectors in the Training and Test datasets and saved them. My aim was to transform text data into a usable numerical format for machine learning algorithms.

Subsequently, I trained the TextCNN model and calculated the accuracy. The TextCNN model achieved an accuracy of 99.81% on the training dataset. On the validation dataset, the accuracy value was 88.38%.

Chapter 6

METHODOLOGY

Sentiment analysis is a technique used to determine the emotional content expressed in text data. This method involves a multi-step process that combines natural language processing (NLP) and machine learning techniques. To perform sentiment analysis successfully, it is crucial to use appropriate data preprocessing, feature extraction, and classification methods.

The first step in sentiment analysis is to gather a suitable dataset, which may include text comments, tweets, or reviews from users. Once the data is collected, a data preprocessing step is performed to avoid data inconsistencies. In this stage, texts are normalized, special characters and numbers are removed, texts are converted to lowercase, and unnecessary stop words are eliminated.

To use machine learning algorithms for sentiment analysis, text data must be transformed into numerical vectors. This step is called "Feature Extraction." One of the most common methods used is "TF-IDF" (Term Frequency-Inverse Document Frequency) [24]. This method creates a vector by considering the frequency of a word in a text and its frequency in all documents in the dataset.

The vectors obtained from the feature extraction step are then classified into positive, negative, or neutral sentiments using classification algorithms. Among the most commonly used classification algorithms are Naive Bayes [25] and Deep Learning-based methods.

To evaluate the accuracy of the sentiment analysis model, the dataset is divided into training and testing data. After the model is trained on the training data, its performance is measured on the test data. The success rate of the model is evaluated using metrics such as accuracy, precision, and recall.

By correctly implementing these steps, the sentiment analysis model can successfully classify and analyze the emotional content in texts within the dataset.

6.1 Data Collection

In the Data Collection phase of Sentiment Analysis, you need to create a dataset containing text samples that will be used to determine the emotional direction (positive, negative, or neutral) of the texts. The following steps may be followed in the Data Collection phase:

Identifying Data Sources: First, you should determine the type of texts you want to use for Sentiment Analysis. For example, you may want to collect texts from social media platforms, product reviews, surveys, or specific news websites.

Creating the Dataset: Gather text samples from the identified data sources and create a labeled dataset. This dataset should include the corresponding labels for the texts (e.g., positive, negative, neutral).

Data Cleaning and Preprocessing: Clean and preprocess the collected text data. In this step, you can remove unnecessary characters, convert texts to lowercase, and remove special characters or links.

Data Labeling: Add sentiment labels to the texts in the dataset. These labels indicate whether the text is positive, negative, or neutral.

Data Balancing (Optional): If there is class imbalance in your dataset, you can take additional steps to balance the classes. For example, you can generate synthetic data to increase the minority class.

Data Splitting: Divide the dataset into three parts: training, validation, and test data. This split is important to evaluate how well the model generalizes to real-world data compared to the training data.

To build a successful Sentiment Analysis model, you need a labeled and cleaned dataset. This collected dataset will be used to train and evaluate the machine learning model.

6.2 K-FOLD CROSS VALIDATION

K-Fold Cross Validation is a cross-validation method used to evaluate the performance of the machine learning model. It is used to detect how well the model generalizes to real-world data and to detect problems such as overfitting or underfitting.

The K-Fold Cross Validation process includes these steps:

The dataset is randomly divided into pieces: In the first step, the dataset is divided into k equal-sized pieces (layers). K usually takes a value such as 5 or 10, but different values are optionally available.

Model training and evaluation: Then, as the dataset is divided into k parts, a total of k operations are performed. In each iteration, one layer is used as the test set and the remaining k-1 layers are used as the training set. The model is trained on the k-1 layer and evaluated on the test data.

Average of performance measures: K-Fold Cross Validation is used to estimate the overall performance of the model by averaging the performance measures obtained at each iteration. K-Fold Cross Validation is often preferred when working with a limited amount of data and when it is necessary to reliably evaluate the model to determine how it fits to real-world data. It is also preferred in case of instability of the dataset because the chance of each class being represented in each layer increases.

This method helps to more reliably evaluate the overall performance of machine learning models and increases the probability that the model will succeed.

Chapter 7

RESULTS AND DISCUSSIONS

7.1 Preparation of the Working Environment

The data in the datasets has been used to create a CNN (Convolutional Neural Network) model and perform training and testing processes. The CNN model is a type of deep learning model commonly used in text processing. During the training process, the model's weights and learnable parameters are adjusted using the examples in the dataset. The testing process is used to evaluate the performance of the trained model on a separate dataset.

The goal is to train and test the model on several personal computers using these images. However, the image processing power of the CPU on personal computers was found to be slow and insufficient, making it impossible to train and test all the data. Further research revealed that GPUs are faster for text data analysis and processing. Therefore, GPU-Tensorflow and GPU-Keras were installed on personal computers to work with GPUs. However, the memory capacity of the GPUs on personal computers was not sufficient, and working with GPUs was not possible.

As a solution, a virtual machine was sought, and Google Colab's Pro version was rented. Google Colab Pro provides access to GPU, TPU, and high memory usage.

All the data was uploaded to Google Drive so that it could be accessed through Google Colab Pro. While attempting to perform training and testing using all the images on Google Colab, two main issues were encountered.

Firstly, due to the large amount of data, there were timeout issues with the read requests. Secondly, the GPU and memory capacity of Google Colab were insufficient for all the data. However, these issues were resolved with code optimizations. Subsequently, the models were trained and tested successfully on Google Colab using the provided data.

7.2 General Results

Sentiment Analysis, a significant subject in the field of Natural Language Processing, involves the development of numerous methods to determine the sentiment or emotion expressed in texts. In this section, I will share information about general results from Sentiment Analysis studies conducted using popular datasets such as Yelp, Amazon, Sentiment140, and IMDB.

Yelp Dataset:

The Yelp dataset consists of user reviews about restaurants, hotels, and various businesses. This dataset is known for its diversity, covering a wide range of topics. The results of Sentiment Analysis studies on Yelp reviews generally indicate a predominance of positive sentiments. User reviews often praise delicious meals, commendable service, and overall positive experiences at these establishments. However, due to variations in customer experiences, negative reviews are also present. It is important to mention that Conv1D was used in this dataset.

Conv1D is a deep learning model used to apply convolution operations to text data. This model can be employed in tasks such as extracting features and determining emotional tones in text data. Studies conducted on the Yelp dataset using Conv1D may prove to be useful in examining the emotional content of text data.

Amazon Dataset:

The Amazon dataset contains user reviews about various products from the e-commerce platform Amazon. Sentiment Analysis can be employed to determine how products are generally evaluated based on these reviews.

Results from Sentiment Analysis studies on Amazon data usually reveal a significant proportion of positive sentiments. However, depending on specific products or contexts, negative reviews can also be found. It is important to mention that TextCNN was used in this dataset. TextCNN is a deep learning model utilized for text classification and feature extraction tasks using deep learning methods on text data. Studies conducted on the Amazon dataset using TextCNN have been successful in determining the overall evaluation of user reviews.

Sentiment140 Dataset:

Sentiment140 is a dataset comprising tweets collected from Twitter. Given the nature of tweets, short and intense texts, Sentiment Analysis presents some challenges in this dataset. Studies conducted on the Sentiment140 dataset demonstrate that tweets mostly contain emotional content. Users often express their positive or negative experiences and share their emotional reactions through tweets.

It is important to mention that Conv1D was used in this dataset. Conv1D is a deep learning model used to apply convolution operations to text data.

This model can be employed in tasks such as extracting features and determining emotional tones in text data. Studies conducted on the Yelp dataset using Conv1D may prove to be useful in examining the emotional content of text data.

IMDB Dataset:

IMDB is a popular film database that includes user reviews about movies. Sentiment Analysis can be utilized to determine the overall evaluations of movies based on these reviews. Research on the IMDB dataset indicates that reviews are predominantly positive. Popular films with strong narratives and successful performances often receive positive reviews, but some films may receive more mixed reactions.

In conclusion, Sentiment Analysis studies conducted on Yelp, Amazon, Sentiment140, and IMDB datasets show that text-based sentiment analysis can yield successful results depending on the textual content and the algorithms employed. The prevalence of positive sentiments, the propensity of users to share positive experiences, and the prominence of emotional content in texts are noteworthy findings. However, since datasets and text contents are diverse, it is crucial to design Sentiment Analysis methods and models that accommodate these variations. It is important to mention that TextCNN was used in this dataset. TextCNN is a deep learning model utilized for text classification and feature extraction tasks using deep learning methods on text data.

Studies conducted on the IMDB dataset using TextCNN have been successful in determining the overall evaluation of user reviews.

By doing so, Sentiment Analysis becomes an effective tool for understanding customer feedback, social media reactions, and other text-based sentiment expressions.

7.3 Discussion

All obtained results were compared with similar studies conducted in recent years. The following comparison table, the neural network used in the studies and the results obtained are presented in the table below.

Table 1: IMDB Accuracy Results Table

Authors	Year	Techniques	Accuracy
Zhilin Yang [26]	2019	XLNet	96.21
Sinong Wang [27]	2021	EFL	96.1
Charaf Eddine Benarab [28]	2021	AlexNet	87
Oğulcan's Work	2023	CNN	88.38

Table 2:YELP Accuracy Results Table

Authors	Year	Techniques	Accuracy
Zhilin Yang [29]	2019	XLNet	72.95
Chi Sun [30]	2019	BERT ITPT FiT	70.58
Ameya Prabhu [31]	2019	ULMFiT	67
Oğulcan's Work	2023	CNN	69

Table 3: Amazon Accuracy Results Table

Authors	Year	Techniques	Accuracy
Jianqing Zhang [32]	2020	TLSAN	97.73
Zeping Yu [33]	2019	SLi-Rec	84.94
Oğulcan's Work	2023	CNN	92.87

Table 4: Sentiment140 Accuracy Results Table

Authors	Year	Techniques	Accuracy
Athindran [34]	2018	Naive Bayes	77
Hemakala, T. [35]	2018	AdaBoost	84.5
Oğulcan's Work	2023	CNN	78,64

Chapter 8

CONCLUSION

In the studies, an examination was made to determine the emotional aspects of the text data by using sentiment analysis methods. For analysis, collected text data was used and texts were classified as positive, negative or neutral.

According to our results, sentiment analysis methods were able to classify texts into emotional categories with great accuracy. This shows that such algorithms are effective in understanding the emotional tone of text data.

However, some difficulties were also encountered in our analysis. In particular, grammatical features such as the complexity of texts, the polysemy of emotional expressions, and irony can affect the accuracy of the analysis. Therefore, the development of algorithms that are more advanced in emotional analysis and that better understand language features will be an important step.

As a result, sentiment analysis methods are a powerful tool for understanding the emotional content of text data. However, further research and development is required to improve the accuracy of the analyses. This study is an important step in showing the future potential of sentiment analysis.

The results in this study may form the basis for improving emotional content analysis and expanding its use in various applications. Extensive studies involving more data and different language features can accelerate developments in sentiment analysis and further improve our ability to understand the emotional tone of text data.

Chapter 9

IN THE FUTURE

In this study, a deep learning model was trained and tested with 4 Datasets. In the future, it is predicted that Word2Vec LSA techniques can be used to obtain higher results than these obtained results.

In addition, although the CNN model is successful in sentiment analysis, it has been observed that higher results can be obtained with TLSAN and XLNET when some of the literature is examined. It is predicted that high results can be obtained by using XLNET and TLSAN instead of CNN in the future.

REFERENCES

[1] Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture* (pp. 70-77).

[2] Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM,* 56(4), 82-89.

[3] Mhatre, M., Phondekar, D., Kadam, P., Chawathe, A., & Ghag, K. (2017, July). Dimensionality reduction for sentiment analysis using pre-processing techniques. In 2017 International Conference on Computing Methodologies and Communication (ICCMC) (pp. 16-21). IEEE.

[4] Aggarwal, C. C. (2018). *Machine learning for text: An introduction. Machine Learning for Text*, 22-40. https://doi.org/10.1007/978-3-319-73531-3_1

[5] Chakravarthy, S. (2020, July 10). *Tokenization for natural language processing*. https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4

[6] Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory. Neural Computation*, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[7] Ma, Y., Peng, H., Khan, T., Cambria, E., & Hussain, A. (2018). *Sentic LSTM: A hybrid network for targeted aspect-based sentiment analysis. Cognitive Computation,* 10(4), 639-650. https://doi.org/10.1007/s12559-018-9549-x

[8] Pennington, J., Socher, R., & Manning, C. (2014). *Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* https://doi.org/10.3115/v1/d14-1162

[9] Pennington, J., Socher, R., & Manning, C. (2014). *Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* https://doi.org/10.3115/v1/d14-1162

[10] Tang, D., Qin, B., & Liu, T. (2015). Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*(6), 292-303.

[11] Ma, L., & Zhang, Y. (2015, October). Using Word2Vec to process big text data. In 2015 *IEEE International Conference on Big Data (Big Data)* (pp. 2895-2897). IEEE.

[12] Church, K. W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155-162.

[13] Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (2004, November). Latent semantic analysis. *In Proceedings of the 16th international joint conference on Artificial intelligence* (pp. 1-14).

[14] Heryadi, Y., Wijanarko, B. D., Murad, D. F., Tho, C., & Hashimoto, K. (2023, February) *Comparing the Effect of Attention Mechanism and Aspect Embedding on the Performance of Long Short-term Memory Model for Aspect-Based Sentiment Analysis*. In 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE) (pp. 880-885). IEEE.

[15] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

[16] Muhammad, P. F., Kusumaningrum, R., & Wibowo, A. (2021). Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews. Procedia Computer Science, 179, 728-735.

[17] Aizawa, A. (2003). *An information-theoretic perspective of tf–idf measures. Information Processing & Management*, 39(1), 45-65.

[18] Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). *Naïve Bayes. Encyclopedia of machine learning*, 15(1), 713-714.

[19] Yue, W., & Li, L. (2020, December). *Sentiment analysis using Word2vec-CNN-BiLSTM classification.* In 2020 seventh international conference on social networks analysis, management and security (SNAMS) (pp. 1-5). IEEE.

[20] Saleena, N. (2018). *An ensemble classification system for twitter sentiment analysis.* Procedia computer science, 132, 937-946.

[21] Kamyab, M., Liu, G., & Adjeisah, M. (2021). *Attention-based CNN and Bi-LSTM model based on TF-IDF and glove word embedding for sentiment analysis.* Applied Sciences, 11(23), 11255.

[22] Berrar, D. (2019). *Cross-Validation*.

[23] Nakayama, M., & Wan, Y. (2019). *The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews.* Information & Management, 56(2), 271-279.

[24] Haque, T. U., Saber, N. N., & Shah, F. M. (2018, May). *Sentiment analysis on large scale Amazon product reviews.* In 2018 IEEE international conference on innovative research and development (ICIRD) (pp. 1-6). IEEE.

[25] Tripathi, S., Mehrotra, R., Bansal, V., & Upadhyay, S. (2020, September). *Analyzing sentiment using IMDb dataset*. In 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 30-33). IEEE.

[26] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). *Xlnet: Generalized autoregressive pretraining for language understanding*. Advances in neural information processing systems, 32.

[27] Wang, S., Fang, H., Khabsa, M., Mao, H., & Ma, H. (2021). *Entailment as few-shot learner.* arXiv preprint arXiv:2104.14690.

[28] Benarab, C. E. (2021). *Classifying Textual Data with Pre-trained Vision Models through Transfer Learning and Data Transformations*. arXiv preprint arXiv:2106.12479.

[29] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). *Xlnet: Generalized autoregressive pretraining for language understanding*. Advances in neural information processing systems, 32.

[30] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). *How to fine-tune bert for text classification?* In Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18 (pp. 194-206). Springer International Publishing.

[31] Prabhu, A., Dognin, C., & Singh, M. (2019). *Sampling bias in deep active classification: An empirical study*. arXiv preprint arXiv:1909.09389.

[32] Zhang, J., Wang, D., & Yu, D. (2021). *TLSAN: Time-aware long-and short-term attention network for next-item recommendation.* Neurocomputing, 441, 179-191.

[33] Yu, Z., Lian, J., Mahmoody, A., Liu, G., & Xie, X. (2019, August). *Adaptive User Modeling with Long and Short-Term Preferences for Personalized Recommendation.* In IJCAI (pp. 4213-4219).

[34] Athindran, N. S., Manikandaraj, S., & Kamaleshwar, R. (2018, November). *Comparative analysis of customer sentiments on competing brands using hybrid model approach*. In 2018 3rd International Conference on Inventive Computation Technologies (ICICT) (pp. 348-353). IEEE.

[35] Rathi, M., Malik, A., Varshney, D., Sharma, R., & Mendiratta, S. (2018, August). *Sentiment analysis of tweets using machine learning approach*. In 2018 Eleventh international conference on contemporary computing (IC3) (pp. 1-3). IEEE.