UNIVERSITY OF MISKOLC
FACULTY OF MECHANICAL ENGINEERING AND INFORMATICS

INVESTIGATING THE DYNAMIC CONNECTION BETWEEN LANGUAGE AND SOCIAL MEDIA IN SENTIMENT ANALYSIS

PHD THESES

Prepared by

Aadil Gani Ganie

ENGINEERING OF INFORMATION TECHNOLOGY (BSC),
ENGINEERING OF COMPUTER APPLICATIONS (MSC)

JÓZSEF HATVANY DOCTORAL SCHOOL FOR COMPUTER SCIENCE AND ENGINEERING
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY
DECEMBER AND 2023

Head of Doctoral School

Dr. Jenő Szigeti
Full Professor

Scientific Supervisor

Dr. Samad Dadvandipour
PhD, Associate Professor

SUPERVISOR'S RECOMMENDATION

Aadil Gani Ganie, an accomplished Indian PhD student specializing in Computer Applications, focused on network security, particularly spam emails, during his M.Sc. studies. Commencing his doctoral journey at the University of Miskolc in 2020 as a Stipendium Hungaricum PhD student, Aadil has excelled in exams and conducted vibrant research in sentiment analysis (NLP).

In his PhD research, Aadil delves into sentiment analysis in social media, exploring the effectiveness of fine-tuning pre-trained transformer models, the influence of informal language on sentiment analysis, optimal configurations for Convolutional Neural Network-based sentiment analysis, and utilizing social media for mental health insights. With 16 publications showcasing expertise in neural networks, deep learning, and data science, Aadil has made notable contributions, including novel methods like clustering attention in deep learning models. His research extends to diverse domains, encompassing clinical trials and mental health prediction, offering a comprehensive analysis of sentiment analysis and valuable insights for future studies.

Aadil Gani Ganie, characterized by his enthusiasm, has authored 16 scientific publications, including 7 high-quality articles in Scopus indexed journals. Notably, one article is featured in an IF(10.865)/Q1, D1 international journal, and three in Q3 international journals, along with three in Q4 international journals. He has been an active presenter at national and international scientific conferences.

Based on these accomplishments, I, as Aadil Gani Ganie's supervisor (Dr. Samad Dadvandipour, associate professor), unequivocally consider his PhD studies to be a resounding success.

Dr. Samad Dadvandipour
Associate Professor
Institute of Information Sciences
Dept. of Applied Information Sciences
Faculty of Mechanical Engineering & Informatics
University of Miskolc- Hungary

Acknowledgments

I commence by expressing my profound gratitude to the divine presence for bestowing upon me the fortitude and resilience necessary to traverse this profound academic odyssey and accomplish this PhD thesis.

My deepest appreciation extends to my parents, whose boundless love, unwavering support, and relentless encouragement have been the cornerstone of my academic achievements. Their sacrifices and unwavering dedication to my education have been pivotal in my scholarly success.

I am indebted to my supervisor (Dr. Samad Dadvandipour), whose mentorship and counsel have been invaluable throughout this research expedition. Dr. Samad's wealth of expertise, profound knowledge, and unwavering support have profoundly shaped this thesis, and for that, I am profoundly grateful.

I want to express my gratitude to my colleagues, whose camaraderie and support have been a wellspring of inspiration. Their enduring friendship has imbued this academic journey with profound meaning and camaraderie.

I acknowledge with deep respect the steadfast dedication of the faculty members and reviewers. Their constructive critiques and sagacious guidance have played a pivotal role in refining my work, and for that, I am sincerely appreciative of their commitment to scholarly excellence.

Lastly, my heartfelt gratitude goes out to all those who, in myriad ways, supported me during the course of this PhD thesis. Their encouragement and support have been immeasurable, and I recognize that my journey has been enriched by their presence.

I am profoundly thankful for the unwavering support and guidance I have received throughout this odyssey. I earnestly hope that this PhD thesis stands as a testament to the collective diligence and dedication of all those who have contributed to its fruition.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

AI	Artificial Intelligence

NLP	Natural Language Processing
POS	Parts of Speech
HMM's	Hidden Markov Models
MEMM's	Maximum Entropy Markov Models
NP	Noun Phrase
VP	Verb Phrase
PP	Preposition Phrase
CRF's	Conditional Random Fields
NER	Named Entity Recognition
RNN	Recurrent Neural Network
ANN	Artificial Neural Network
CNN	Convolution Neural Network
BERT	Bidirectional Encoder
NMT	Neural Machine Translation
RoBERTa :	Robustly Optimized BERT Pre-training Approach
SVM	Support Vector Machines
AIIMS	All India Institute for Medical Sciences
CBI	Central Beureau of Investigation
US	United States
LLM's	Large Language Models
GRU	Gated Recurrent Unit
LSTM	Long Short-term Memory
BiLSTM	Bidirectional Long Short-term Memory
TRAC-1	Trolling Aggression Cyberbullying
GloVE	Global Vectors for Word Representation
OAG	Overly Aggressive
COA	Covertly Aggressive
NAG	Non-Aggressive
GPU	GraphicsProcessing Unit
ReLU	Rectified Linear Unit
ME	Maximum Entropy
CM	Confusion Matrix

FP	False Positives
FN	False Negatives
TP	True Positives
TN	True Negatives
ROC	Receiver Operating Characteristic
GPT	Generative Pr-trained Transformers
BPD	Bipolar Disorder
WHO	World Health Organization
EDA	Exploratory Data Analysis
API	Application Programming Interface
TRP	True Positive Rate
FPR	False Positive Rate
LR	Logistic Regression
NB	Naive Bayes
PTSD	Post Traumatic Stress Disorders
TfIDf	Term Frequency - Inverse Document Frequency
SGD	Stochastic Gradient Decent
PCA	Principle Component Analysis
GD	Gradient Decent
BGD	Batch Gradient Decent
MBGD	Mini Batch Gradient Decent
NAG	Nesterov Accelerated Gradient
CG	Conjugated Gradient
PC	Gradient Decent
MAE	Mean Absolute Error

Preface

This dissertation aims to explore the intricate interaction between language and social media within the domain of sentiment analysis. The exponential growth of social media data has necessitated the development of sophisticated techniques for automated sentiment analysis of textual content. This research endeavors to address five critical research questions that delve into various aspects of sentiment analysis of social media text.

The first inquiry investigates the efficacy of fine-tuning pre-trained transformer models on social media text for sentiment analysis, as compared to training models from scratch on social media data. Fine-tuning of pre-trained models has gained significant traction in the natural language processing arena, and optimizing such models for specific tasks can enhance their performance. However, the compatibility of these models with sentiment analysis of social media text remains unclear. Hence, this research seeks to provide a nuanced understanding of the strengths and limitations of fine-tuning pre-trained models on social media data. The study also explores the integration of clustering attention mechanisms with autoencoders in Natural Language Processing (NLP). The study achieves promising adaptability and performance across domains, demonstrating enhanced accuracy in financial sentiment analysis, fake news detection, and hate speech detection, with a focus on efficiency in resource-constrained environments.

The second research question focuses on the impact of informal language like emoticons, hashtags (Twitter), and slang- on the accuracy of sentiment analysis models applied to social media text. Informal language is ubiquitous in social media text, presenting a significant challenge for traditional sentiment analysis models designed for more formal text. This research aims to identify the role of informal language in sentiment analysis of social media text and identify methods to enhance the performance of sentiment analysis models in the presence of such language.

The third inquiry investigates how filter size and the number of filters affect the performance of Convolutional Neural Network (CNN)-based sentiment analysis models on three datasets: Amazon review, Amazon food review, and hate speech detection. The findings reveal that the optimal combination of filter size and number of filters varies with the dataset employed. Fine-tuning these parameters can improve the accuracy of the CNN model for sentiment analysis. The study underscores the criticality of considering these factors to attain optimal performance in sentiment analysis.

The fourth research question explores the use of social media posts to get valuable insights into the mental health of individuals and communities, using machine learning models capable of accurately predicting mental health status from social media text. With mental health being a growing concern, social media posts hold immense potential in providing a window into the mental health status of individuals and communities.

The fifth research question delves into the relationship between hate speech and mental health using logistic regression models trained on two preprocessed datasets - mental health and hate speech. The visualization of hate speech distribution in the mental health dataset suggests a potential relationship between hate speech and mental health, particularly in individuals with anxiety and bipolar disorders.

The final research question proposes a novel approach to gradient descent called "Sample Gradient Descent," which entails choosing an adequate data sample and subjecting it to batch gradient descent. The study uses Principle Component Analysis (PCA) to select the sample, resulting in faster convergence rates and reduced computation times compared to conventional batch gradient descent. The proposed method offers potential utility in various domains, including machine learning and optimization problems. What impact does fine-tuning have on the performance and interpretability of large language models across diverse textual domains, and how does it contribute to enhancing their adaptability to specific tasks. Overall, this dissertation offers a comprehensive analysis of sentiment analysis in social media and provides valuable insights into future research directions.

Chapter 1: Introduction

1.1 Natural Language Processing

Natural Language Processing (NLP) is a branch of artificial intelligence and computer science that studies natural language interactions between computers and people. It entails creating models and algorithms to comprehend, evaluate, produce, and work with human language. Natural language processing (NLP) finds wide-ranging uses in domains like question-answering systems, machine translation, sentiment analysis, text classification, and text summarization. A branch of linguistics, computer science, and artificial intelligence called natural language processing (NLP) studies how computers and human (natural) languages interact. Ultimately, it aims to make computers capable of reading, understanding, and producing text that is human-like.

1.2 From Linguistics to Natural Language Processing

The scientific investigation of language and its structural components, including grammar, semantics, and phonetics, is referred to as linguistics. Classical linguistics involves the creation and examination of rules governing language use. While significant progress has been made in the development of formal methods for syntax and semantics, the complexities of natural language comprehension often defy precise mathematical formalization. A linguist is generally understood to be a person who studies language, although the term may also be applied to individuals who engage in fieldwork and empirical research in the field. The use of mathematical techniques and principles to analyze and understand natural language is sometimes referred to as mathematical linguistics and may involve the application of discrete mathematical formalisms and theories, such as formal languages and automata theory, to the study of language. Computational linguistics, as a subfield of linguistics and computer science, involves the application of computational and statistical methods to the study of natural language and its structural and functional properties. With the advent of big data and advanced computing technologies, computational linguists have been able to leverage large datasets of natural language text and speech to discover novel patterns and insights that were previously beyond the reach of traditional rule-based approaches. In recent years, the statistical approach to natural language processing, which relies on probabilistic models and machine learning algorithms, has emerged as the dominant paradigm in the field, supplanting earlier rule-based approaches that were limited by their computational complexity and inflexibility. This shift towards a more empirical and data-driven approach has led to the use of the term "natural language processing" (NLP) to describe the field of computational linguistics, which seeks to develop computational models and algorithms that can facilitate the automatic analysis, interpretation, and generation of natural language data.

NLP tasks can be broadly divided into two categories:

Language understanding tasks: These tasks involve understanding and extracting meaning from natural language inputs. Examples include text classification, sentiment analysis, named entity recognition, and machine translation.

Language generation tasks: These tasks involve generating natural language outputs based on some input or intention. Examples include text summarization, machine translation, and dialog systems. In this work both language understanding tasks and language generation tasks have been employed. At a high level, NLP involves using computers to process, analyze, and generate human language. This involves several tasks, such as:

1.2.1 Tokenization

Tokenization is a process in NLP that involves breaking down text into smaller units, called tokens, to facilitate easier analysis. Tokens can be words, phrases, symbols, or even individual characters. The purpose of tokenization is to convert text into a more manageable format, so that NLP algorithms can operate on the text and extract meaningful information. For example, consider the sentence "The cat in the hat." If we tokenize this sentence, we would get the following tokens: "The", "cat", "in", "the", "hat". In this example, the tokens are individual words. Tokens, on the other hand, may consist of sentences, like "The cat in the hat," or even single characters, like "T," "h," "e," "," "c," "a," "t," "," "i," "n," ",", "t," "h," "e," "," "h," "a," and "t."

Tokenization is a crucial stage in natural language processing (NLP) as it enables algorithms to operate with more manageable, smaller text units instead of processing full documents or text blocks all at once. This can help to lower the number of errors that can happen while processing big volumes of text and considerably increase the efficiency of NLP algorithms. The terminology that is being utilized can also have an impact on tokenization. In NLP, there are often complex vocabularies, such as scientific or technical terms, that require special handling during tokenization. For example, in a scientific document, there may be many technical terms that should not be broken down into smaller tokens. In these cases, special algorithms or dictionaries may be used to preserve the integrity of the complex vocabulary during tokenization.

In summary, tokenization is a crucial step in NLP that involves breaking down text into smaller, more manageable units, called tokens, to facilitate easier analysis. The choice of tokens and the process of tokenization can be influenced by the vocabulary being used, which can impact the accuracy and efficiency of NLP algorithms.

1.2.2 Part-of-speech tagging

Assigning a part-of-speech (POS) tag to every word in a sentence is a crucial task in natural language processing. A phrase that designates a word's grammatical function in a sentence—noun, verb, adjective, adverb, etc.—is called a POS tag. Because it offers a structured representation of the words in a sentence that can be used for additional analysis, like syntactic parsing and semantic analysis, POS tagging is significant. For instance, it's simple to determine which nouns and proper nouns are marked as the subject in a sentence in order to extract the subject. In a similar vein, you can determine which verbs in a sentence are designated as the predicate in order to extract the action from it. POS tagging is the process of determining the POS tags for every word in a sentence using lexical resources like grammars and dictionaries as well as algorithms. For POS tagging, a variety of algorithms are available, such as rule-based, machine learning, and hybrid approaches that combine the best features of both machine learning and rule-based approaches. Among the phrases that are most frequently used in the POS tagging industry are:

Tokenization: The process of breaking down a sentence into its individual words, which are also known as tokens.

Ambiguity: The situation where a word can have multiple POS tags, depending on its context. For example, the word "bank" can be a noun (e.g., "I am going to the bank") or a verb (e.g., "I am going to bank my paycheck").

N-gram models: Statistical models that predict the POS tag of a word based on the POS tags of the previous n-1 words in a sentence.

Hidden Markov Models (HMMs): Probabilistic models that are used to solve sequence labeling problems, such as POS tagging, by modeling the probability of each word in a sentence given the previous words in the sequence.

Maximum Entropy Markov Models (MEMMs): Another type of probabilistic model that is used for sequence labeling problems, such as POS tagging. Unlike HMMs, which model the probability of each word given the previous words in the sequence, MEMMs model the probability of each word given the previous words and the current word.

Conditional Random Fields (CRFs): Another type of probabilistic model that is used for sequence labeling problems, such as POS tagging. CRFs model the probability of the entire sequence of words, rather than the probability of individual words, given the previous words in the sequence.

These are just a few of the terms that are commonly used in the field of POS tagging, but there are many more. The important thing to remember is that POS tagging is a critical step in the overall process of natural language processing that enables us to structure and analyze the words in a sentence.

1.2.3 Parsing

Parsing, also known as syntactic analysis, is a crucial step in the field of Natural Language Processing (NLP) that involves analyzing the structure of a sentence to understand its grammatical relationship among words. Parsing helps to identify the syntactic structure of a sentence and determine how words relate to each other semantically. In NLP, parsing is typically performed using grammatical rules or statistical models. Grammatical rules are specified in the form of context-free grammars, which are a set of production rules that define how words in a sentence can be combined to form well-formed phrases and sentences. However, in order to find patterns in the data and produce a statistical representation of a sentence's grammatical structure, statistical models are trained on sizable corpora of text.

One of the main outputs of parsing is a parse tree, which is a tree-like representation of the grammatical structure of a sentence. A parse tree encodes the constituent structure of a sentence, including its noun phrases (NP), verb phrases (VP), and prepositional phrases (PP). The tree also encodes relationships between words, such as subject-verb agreement and modifier-head relationships. Parsing is used in a variety of NLP applications, including information extraction, machine translation, and question answering. In these applications, parsing is used to extract meaningful information from a sentence and to generate an internal representation that can be easily processed by a computer program.

In NLP, parsing is often considered a challenging task due to the complexity of natural language and the variability of sentence structure. Parsers must be able to handle a wide range of grammatical constructions, including long-distance dependencies, idiomatic expressions, and ambiguity. Additionally, parsers must be able to process large volumes of text in real-time and deal with noisy or incorrect data, such as typos and grammatical errors.

1.2.4 Named entity recognition

Within the field of natural language processing (NLP), named entity recognition (NER) is the automatic identification of textual entities, including people, places, dates, and organizations. NER systems usually locate named entities in text and group them into predefined classes (person, place, or organization, for example). Named entity recognition typically requires several stages, such as shallow parsing, tokenization, and part-of-speech (POS) tagging. In tokenization, text is divided into words, phrases, or other meaningful units called tokens. In POS tagging, each token is assigned a tag indicating its grammatical role in the sentence, such as noun, verb, adjective, or adverb. Shallow parsing, also known as chunking, involves the identification of phrases in the sentence that play a specific role, such as a named entity.

Following the first preprocessing stages, named entity identification in the text is usually accomplished by NER systems using machine learning algorithms. These algorithms may be based on statistical models like hidden Markov models (HMMs) or conditional random fields (CRFs), or they may be based on rule-based systems like regular expressions. In order to enhance their performance, certain NER systems also employ contextual data, such as the words that surround the named entities or their part-of-speech tags. In recent years, deep learning techniques, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have been used for named entity recognition with promising results. These techniques are capable of learning patterns in the data and can effectively capture the context of named entities in the text.

One of the main challenges in named entity recognition is dealing with the ambiguity of named entities. For example, a word like "Apple" can refer to a fruit, a company, or a school, and the meaning of the word depends on the context in which it appears. NER systems typically use various techniques, such as gazetteers, which are lists of named entities, or disambiguation algorithms, which resolve the ambiguity by considering the context of the named entities, to overcome this challenge. Overall, named entity recognition is an important step in NLP pipelines and plays a crucial role in a wide range of applications, including information retrieval, text classification, and question answering.

1.2.5 Sentiment analysis

This involves analyzing a piece of text and determining the overall sentiment (positive, negative, or neutral) expressed in it. Sentiment analysis is useful for tasks such as customer service and social media analysis.

1.2.6 Machine translation

The goal of machine translation, a branch of computational linguistics, is to translate text automatically between natural languages. The aim of machine translation is to generate text that is both semantically and fluently similar to the input. Because natural languages are frequently ambiguous and context-dependent, and because words and phrases can have different meanings depending on the context in which they are used, this can be a difficult task. Machine translation systems can be classified into two main categories: rule-based systems and statistical systems. Rule-based systems rely on a set of linguistic rules and dictionaries to translate text. These systems are typically more accurate for grammar and vocabulary, but can be limited by their ability to handle ambiguity and context-dependence.

In contrast, statistical machine translation systems acquire their translation skills through extensive parallel corpora of texts in both the source and target languages. Although these systems can handle a larger variety of input and are usually more data-driven, they occasionally have issues with fluency and grammar. Neural machine translation (NMT) has emerged as the cutting edge of machine translation in recent years. Deep neural networks are used by NMT systems to learn text translation. These systems are trained on large parallel corpora and can handle a wider range of input than rule-based or statistical systems. NMT systems have also been shown to produce translations that are more fluent and semantically equivalent to the input text.

In NMT, the input text is first tokenized and converted into a sequence of word vectors. The word vectors are then fed into a neural network, which uses an encoder to capture the meaning of the input text and a decoder to generate the translation. The decoder uses a context vector to keep track of what has been translated so far, and produces a probability distribution over the possible translations at each step. The final output is selected by sampling from the probability distribution. There are several key concepts and terminology used in the field of machine translation, including:

Source language: The language of the input text.
Target language: The language into which the text input is being translated.
Corpus: A collection of texts in a given language or set of languages.
Parallel corpus: A corpus that contains corresponding texts in the source and target languages.
Tokenization: The process of breaking up a text into individual words or phrases.
Word vectors: Representations of words or phrases as numerical vectors.
Encoder: An encoder is a part of a neural network that converts an input text's meaning into a context vector.
Context vector: A vector that summarizes the meaning of the input text and is used to generate the translation.
Decoder: A component of a neural network that generates the translation from the context vector.
Probability distribution: A function that assigns probabilities to different outcomes.
Sampling: The process of selecting an outcome from a probability distribution. Overall, machine translation is a complex field that requires a deep understanding of natural language processing, computational linguistics, and machine learning. However, the development of neural machine translation has greatly advanced the field and made it possible to produce high-quality translations for a wide range of applications.

NLP algorithms use a variety of methods and resources, including word embeddings—numerical representations of words that capture the connections between them—to carry out these tasks. To enhance their functionality, NLP algorithms frequently make use of machine learning methods like supervised and unsupervised learning. Numerous domains, including information retrieval, machine translation, customer service, and social media analysis, have found use for natural language processing (NLP). It has also seen significant advancements in recent years due to the availability of large amounts of annotated data and the development of deep learning models.

We investigate how language, social media, and context interact with sentiment analysis in this study. Sentiment analysis, sometimes referred to as opinion mining, is a branch of natural language processing (NLP) that uses computer methods to locate and extract subjective information from textual data in order to quantify and describe the opinions, feelings, and attitudes that are expressed in the text. This may involve the use of supervised or unsupervised machine learning algorithms, lexical resources, and computational linguistic methods to analyze the syntactic, semantic, and pragmatic features of the text, as well as contextual and cultural factors that may influence the interpretation of the text.

1.3 Sentiment Analysis

Sentiment analysis is a branch of natural language processing (NLP) that focuses on automatically recognizing and extracting subjective data from text. Determining whether a sentiment is positive, negative, or neutral in a textual document—like a news article, social media post, or product review—is the aim of sentiment analysis. Numerous applications, including market, political, and customer sentiment analysis, can benefit from this information.

Sentiment analysis, commonly referred to as opinion mining, is a branch of natural language processing (NLP) that uses computational methods to locate and extract subjective data from textual data in order to quantify and characterize the opinions, feelings, and attitudes that are expressed in the text. (Pang & Lee, 2008). This may involve the use of supervised or unsupervised machine learning algorithms, lexical resources, and computational linguistic methods to analyze the syntactic, semantic, and pragmatic features of the text, as well as contextual and cultural factors that may influence the interpretation of the text (B. Liu, 2012). At the most basic level, sentiment analysis involves the classification of text as positive, negative, or neutral, based on the presence or absence of certain keywords or phrases that are indicative of positive or negative sentiment (Hu & Liu, 2004) More advanced approaches may involve the identification and classification of specific emotions or affective states, such as joy, sadness, anger, fear, or surprise (Camras & Plutchik, 1980), or the quantification of the overall valence or intensity of the sentiment expressed in the text (Baccianella et al., 2008)

Sentiment analysis is widely used to obtain insight into the attitudes and opinions of large groups of people in a variety of domains, including social media analysis, market research, customer service, politics, journalism, and psychology. (Tumasjan et al., 2011) It has also been used to detect and classify online harassment and hate speech, as well as to analyze the impact of social media on mental health and well-being (De Choudhury et al., 2013) (Caliskan et al., 2017).The proliferation of technology, particularly the widespread access to the internet, has both positive and negative impacts on society. One the one hand, it facilitates the spreading and amplification of voices and movements (like the #metoo and LGBTQ+ movements) that aim to advance social justice and inclusivity. However, as was evident in the aftermath of actor Sushant Singh Rajput's death in India, where social media trends on Twitter exacerbated societal divisions, it also makes it easier for false or biased information to spread and for divisive narratives to be created.

At the most basic level, sentiment analysis involves the classification of text as positive, negative, or neutral, based on the presence or absence of certain keywords or phrases that are indicative of positive or negative sentiment. More advanced approaches may involve the identification and classification of specific emotions or affective states, such as joy, sadness, anger, fear, or surprise, or the quantification of the overall valence or intensity of the sentiment expressed in the text. In addition to its applications in social media analysis, market research, and customer service, sentiment analysis has also been used in various domains, including politics, journalism, and psychology, to gain insight into the attitudes and opinions of large groups of people.

1.4 Social Media and Sentiment Analysis

The exponential growth in digital text data and machine learning advancements have driven the field of Natural Language Processing (NLP) to rapid growth in recent years. One of the key applications of NLP is sentiment analysis, which aims to determine the polarity of a given piece of text, whether it is positive, negative, or neutral. Sentiment analysis has wide-ranging applications, including opinion mining, marketing research, customer relationship management, and monitoring public opinion, among others.

Social media has emerged as one of the most important sources of user-generated content and is an especially rich source of information for sentiment analysis. Users can freely and publicly express their opinions and feelings on social media platforms like Facebook, Instagram, and Twitter, which makes them a rich and varied source of data for sentiment analysis. The growth of social media and its impact on society have made sentiment analysis of social media data a critical research area. The challenge of sentiment analysis of social media data lies in the informal nature of the language used in such texts. Social media users often use informal language, including emoticons, hashtags, slang, and misspelled words, which can affect the accuracy of sentiment analysis models. Additionally, the multilingual and multicultural nature of social media can present challenges for sentiment analysis models, as sentiment expressions can vary greatly across different languages and cultures.

To address these challenges, researchers have explored different approaches to sentiment analysis of social media data. One approach is to fine-tune pre-trained transformer models, such as BERT and RoBERTa, on social media data. Another approach is to train models from scratch on social media data, leveraging the unique characteristics of the language used in social media texts. The performance of these models is then compared to determine which approach is more effective.

1.5 Research Questions

The aim of this thesis is to explore the intersection of language and social media in sentiment analysis. The following research questions are addressed:

- Can pre-trained transformer models be fine-tuned effectively on social media text for sentiment analysis, and if so, how does this compare to training models from scratch on social media data?
- How can clustering attention models, enriched by autoencoder-based dimensionality reduction, be optimized for dynamic clustering techniques and scaled to larger datasets while maintaining interpretability in NLP applications?
- How does the presence of informal language, such as emoticons, hashtags, and slang, impact the performance of sentiment analysis models on social media text?
- What is the significance of filter size and number of filters in Convolutional Neural Networks for sentiment analysis.
- Investigating the Use of Machine Learning for Assessing Mental Health through Analysis of Social Media Posts.
- Is there a relationship between hate speech and mental health, and if so, what are the potential mechanisms underlying this relationship?
- How does "Sample Gradient Descent" compare to conventional batch gradient descent in terms of convergence rates and computation times, and what are the potential applications of this method in machine learning and optimization problems?
- What impact does fine-tuning have on the performance and interpretability of large language models across diverse textual domains, and how does it contribute to enhancing their adaptability to specific tasks?
- How do different hyperparameters, such as learning rate, batch size, and layer-specific tuning, influence the fine-tuning process of large language models, and what optimal configurations exist for various natural language processing tasks

The formation of the research question in this study begins with a specific use case, which is the ‘sentiment analysis of the impact of news trending from untrusted sources, with particular reference to the recent death of an Indian actor’. The study starts by examining this use case and analyzing the results obtained from the analysis. We initiated our research with the objective of exploring the sentiment analysis on the impact of trending sourceless news, with a specific focus on the recent death of Indian actor Sushant Singh Rajput. To delve into the topic, we conducted a sentiment analysis on the tweets related to the event. We examined the outcomes to ascertain the efficacy of every approach using a variety of sentiment analysis classifiers, such as Support Vector Machines, Random Forest, Neural Networks, and Naïve Bayes. The Naive Bayes classifier emerged as the most accurate with over 82% accuracy, and a low false positive and negative ratio. The neural network, despite improving its accuracy through increased epochs, lagged behind with a 70.58% accuracy. The Random Forest and SVM classifiers exhibited similar results with an accuracy of 73.52%, and a similar false positive and negative ratio.

With regards to the passing of Sushant Singh Rajput, India was inundated with discussions surrounding his suicide, with some supporters claiming that he was murdered despite the declaration of his death as a suicide by the CBI team and medical professionals, including the head of AIIMS Forensics (*Sushant Singh Rajput Case: CBI Gets No Proof of Murder, Now Focusing on the Suicide Angle | Hindi Movie News - Times of India*, n.d.) (Wawre & Deshmukh, 2016). To understand the sentiments of the public towards this incident, a dataset was created by extracting trending tweets from Twitter and conducting a sentiment analysis on the cleaned data. This analysis was accomplished through the utilization of three classifiers: Naive Bayes, Support Vector Machine, Random Forest, and Neural Network, which categorized the tweets into positive, negative, and neutral sentiment.

To establish the validity of our outcomes, we contrasted them with multiple established methodologies. In a study published in (Samad & Gani, 2020), the sentiment analysis was performed on spam and ham emails to determine whether they were spam using SVM and Random Forest classifiers. The sentiment analysis was performed on both the content of the emails and their attachments. The SVM classifier exhibited a 96% accuracy rate, while the Random Forest classifier achieved a 97.66% accuracy rate. The SVM classifier recorded zero false positive results and 4% false negatives, while the Random Forest classifier recorded zero false positive results and 2.33% false negatives.

Table 1: *Results of (Wawre & Deshmukh, 2016)*

ALGORITHM	ACCURACY	FEATURE EXTRACTION	DATASET	NUMBER OF SENTIMENTS
SVM	45.71	Document level	IMDB	2
NAÏVE BAYES	65.75	Document level	IMDB	2

Table 2: *Results of (Gautam & Yadav, 2014)*

ALGORITHM	ACCURACY	FEATURE EXTRACTION	DATASET	NUMBER OF SENTIMENTS
SVM	85.4	Uni-gram	Tweeter	2
NAÏVE BAYES	88.2	Uni-gram	Tweeter	2
MAXIMUM ENTROPY	83.9	Uni-gram	Tweeter	2
SEMATIC ANALYSIS	89.9	Uni-gram	Tweeter	2

Table 3: *Results of (Le & Nguyen, 2015)*

ALGORITHM	ACCURACY	FEATURE EXTRACTION	DATASET	NUMBER OF SENTIMENTS
SVM	79.54	Uni-gram, bi-gram, object-oriented	Tweet	2
NAÏVE BAYES	79.58	Document-level	Tweet	2

Table 4: *Results of (Neethu & Rajasree, 2013)*

ALGORITHM	ACCURACY	FEATURE EXTRACTION	DATASET	NUMBER OF SENTIMENTS
SVM	90	Uni-gram	Tweet	2
NAÏVE BAYES	89.8	Uni-gram	Tweet	2
MAXIMUM ENTROPY	90	Uni-gram	Tweet	2
ENSEMBLE	90	Uni-gram	Tweet	2

Table 5: *Results of (Hasan et al., 2018)*

ALGORITHM	ACCURACY	FEATURE EXTRACTION	DATASET	NUMBER OF SENTIMENTS
SVM	70	Uni-gram,Sentence level	Tweet	3
NAÏVE BAYES	79	Uni-gram, Sentence level	Tweet	3

Table 6: *Results of (J. Singh et al., 2017)*

ALGORITHM	ACCURACY	FEATURE EXTRACTION	DATASET	NUMBER OF SENTIMENTS
NAÏVE BAYES	85.24	Uni-gram	Tweet	2
J-48	89.73	Uni-gram	Tweet	2
BFTREE	90.07	Uni-gram	Tweet	2
ONER	92.34	Uni-gram	Tweet	2

From the above assessments, it can be inferred that our model exhibits a competitive performance compared to the previously established models. By means of model optimization and increased data utilization, it is possible to attain superior outcomes. Given that our model accommodates for both conventional and ironic text, the resultant accuracy is commendable.

The way that technology is becoming more and more ingrained in our daily lives has created a situation where it can be both a boon and a bane. Because so many people now have access to the internet, stories are being created on a variety of online platforms and public opinions are being shaped by this. A significant fraction of the global population, estimated to be around 100 million individuals, actively utilizes platforms such as Twitter, which has become a crucial source of information for various purposes, including investigations and sentiment analysis. A study by Alexandre Bovet on US presidential elections found that a quarter of the 171 million tweets shared false or biased news, thereby underscoring the potential impact of online platforms in shaping public opinions.

The digital era has facilitated the spread of false information and scaremongering, with the World Economic Forum identifying online misinformation as a significant technological and geopolitical threat. In this era of big data, where access to information is just a click or swipe away, users are susceptible to being misled by malicious or misleading social media content. Studies on various social media platforms, including Facebook, YouTube, Twitter, and Wikipedia, as well as theoretical modeling, suggest the existence of echo chambers where users are likely to encounter and spread unsourced claims or facts that are consistent with their beliefs, leading to the amplification of disinformation and malicious posts.

The rise of online hate and disinformation has had negative impacts on various organizations, affecting sales, stock values, brand credibility, and customer loyalty. A recent example is the malicious social media campaign against Tanishq jewelers, resulting in a drop in the brand's stock value. Research has also highlighted the role of bots, or automated accounts, in spreading misinformation. In general, fragmentation in cultures and news consumption is often observed and aligned with political allegiances.

Chapter 2: Literature Review

The phenomenon of online harassment, primarily stemming from social media, is a cause for concern in contemporary society. A recent survey conducted by the Pew Research Center (*The State of Online Harassment | Pew Research Center*, n.d.) reveals that 31% of Americans, or 75% of online abuse targets, experienced their most recent instance of online hate on social media. This has led to the query of the efficacy of social media companies in mitigating online harassment, with 79% of the respondents suggesting that these companies are not adequately addressing the issue on their platforms. A similar survey conducted by the American Trends Panel (*The American Trends Panel | Pew Research Center*, n.d.) found that 41% of American adults have faced online hate and 25% have faced severe harassment. In light of these disturbing trends, researchers have been motivated to automate the detection and elimination of online harassment, which has led to the development of online hate detection methods using Natural Language Processing (NLP). The concept of abuse is challenging to institutionalize, as exemplified by Mishra (Mishra et al., 2018), who used NLP to analyze racism and sexism, and Nobata (Nobata et al., 2016), who studied hate speech, profanity, and derogatory language. The earliest reported method for abuse detection was that of Spertus (Spertus, 1997), who used hand-crafted rules over text to create feature vectors for machine learning. Dadvar (Dadvar et al., 2013) utilized a social feature engineering technique incorporating user identity traits to the likelihood of abusive behavior, referred to as user profiling.

This method also incorporated the user's age into the lexicon-based features to detect cyber-bullying. In (Dadvar et al., 2013), the authors used gender and character n-grams to detect sexism and racism in tweets, resulting in an improvement of the F1-score from 73.89% to 73.93%. (Waseem & Hovy, 2016) was the first to utilize deep learning models for online harassment detection, significantly outperforming existing traditional methods with an accuracy increase from 78.89% to 80.07%. (Badjatiya et al., 2017)utilized an LSTM model with GloVe for feature engineering to detect online abuse, achieving the best results with a weighted F1 of 93% through random embedding initialization. Park and Fung (Badjatiya et al., 2017) categorized comments by combining two datasets and found that combining the two granularities through two input channels improved accuracy, a conclusion also supported by researchers in (Djuric et al., 2015)(Park & Fung, 2017)(V. Singh et al., 2018). The authors in the GermEval shared task (C. Wang, 2018) made the winning submission with F1-scores of 76.95% and 53.59% for sub-tasks 1 and 2, respectively. Furthermore, (Z. Zhang et al., 2018) demonstrated that learning about the classification of emotions and detecting abuse leads to improved performance.

The domain of Natural Language Processing (NLP) has undergone a paradigm shift with the ascendancy of transformer-based models. Innovations like BERT (Devlin et al., 2018) and GPT (Brown et al., 2020) have ushered in a new era, eclipsing the traditional prowess of recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) across various NLP tasks. Central to their triumph is the attention mechanism, a pivotal component intricately woven into the fabric of the Transformer's self-attention mechanism (Vaswani et al., 2017).

The ability to conduct automated analysis of sentiment within textual content has gained immense importance across diverse domains including e-commerce, marketing, and social media. Often referred to as sentiment analysis or opinion mining, this process entails discerning the attitudes, opinions, and emotions expressed by individuals through textual data. The rapid expansion of social media platforms has propelled sentiment analysis into a prominent research area. Its applications span a wide array of domains, encompassing product reviews, customer feedback analysis, and even political sentiment analysis. (Derks et al., 2007).

However, sentiment analysis models face numerous challenges, one of which is the use of informal language within text data. Informal language, such as emoticons and slang, can significantly impact the performance of sentiment analysis models. Emoticons, commonly referred to as "emoji" or "smileys", are utilized to express emotions and sentiment within text messages, social media posts, and online communication. Emoticons have evolved into a prevalent method of expressing emotions online, with a discernible increase in their usage over time (Kwok & Wang, 2013) In contrast, slang refers to the informal language employed in casual communication, and it too can significantly impact the performance of sentiment analysis models (Poria et al., 2016) (Hutto & Gilbert, 2014). Prior studies have illustrated that integrating emoticon data can enhance the efficacy of sentiment analysis models (C. Liu et al., 2021) (Alsayat, 2022) (AlBadani et al., 2022). For instance, (Alsayat, 2022) proposed a framework that integrates emoticon information with word embeddings to augment the performance of sentiment analysis models. Similarly, (AlBadani et al., 2022) introduced a deep learning-based approach incorporating emoticon information to enhance the performance of sentiment analysis models, particularly when applied to Twitter data.

However, the presence of sarcasm within social media text can negatively impact the performance of sentiment analysis models. The detection of sarcasm is a challenging task for natural language processing (NLP) models (González-Ibánez et al., 2011) (Riloff et al., 2013) (L. Liu et al., 2019). For example, (L. Liu et al., 2019) proposed a neural network-based approach for sarcasm detection and demonstrated that it outperforms traditional machine learning-based approaches.

The field of sentiment analysis has garnered significant attention in recent years due to its various applications in areas such as opinion mining, customer feedback analysis, and natural language understanding. One of the most widely used techniques for sentiment analysis is the use of convolutional neural networks (CNNs), which have been shown to achieve state-of-the-art performance on a variety of datasets. However, there is still room for improvement in terms of understanding the impact of different architectural choices, such as filter size and number of filters, on the accuracy of CNN-based sentiment analysis models. Filter size refers to the length of the window that is used to scan the input data, while the number of filters is the number of different feature maps that are used to extract features from the input data. The selection of filter size and the quantity of filters employed profoundly influences the model's performance. Past research has demonstrated that the ideal filter size and number of filters are contingent upon the particular dataset and task being addressed.

In this research, we investigate the effect of filter size and number of filters on the accuracy of CNN-based sentiment analysis models using three datasets: Amazon review dataset, Amazon food review dataset, and hate speech detection dataset. The datasets were preprocessed using techniques such as stopword removal, stemming, noise removal, and spell check. We employed a Keras-based implementation of the CNN model, and evaluated its performance using a variety of metrics, including confusion matrix. Recent studies have shown that increasing the number of filters and filter size can improve the accuracy of CNN-based sentiment analysis models (Y. Zhang & Wallace, 2015) (X. Zhang et al., 2015) . However, there is limited research on the optimal filter size and number of filters for sentiment analysis, particularly on the impact of these choices on the performance of the model on different datasets. In this study, our objective is to bridge this gap by exploring the impact of filter size and the number of filters on the accuracy of sentiment analysis models based on Convolutional Neural Networks (CNNs). We conduct this investigation using the three datasets mentioned earlier.

Furthermore, our study makes a valuable contribution to the expanding body of literature pertaining to the utilization of deep learning techniques within the realm of natural language processing, with a particular focus on sentiment analysis. (Socher et al., 2013) (Aguzzoli et al., 2015) (Lin et al., 2015). In addition, the results of our research can inform the design of CNN-based sentiment analysis models, and provide insights into the trade-offs between filter size and number of filters for different datasets.

In recent years, there has been a growing interest in understanding the impact of different hyperparameters on the performance of CNNs in NLP tasks. For instance, in 2020,(Guo et al., 2020) proposed a multi-scale CNN model for sentiment analysis that utilizes different filter sizes to capture different levels of information. They showed that their model outperforms traditional CNNs and other state-of-the-art models on several benchmark datasets. Similarly, in 2021, (Tan et al., 2022) proposed a dynamic filter CNN model for sentiment analysis, which adapts the number of filters based on the input text. They showed that their model outperforms traditional CNNs and other state-of-the-art models on several benchmark datasets. Moreover, several studies have also examined the effect of different filter sizes and number of filters on the performance of CNNs in NLP tasks. For instance, in 2019, (Tran & Phan, 2019) proposed a CNN-LSTM model for sentiment analysis and showed that larger filter sizes and more filters result in better performance on benchmark datasets. Similarly, in 2020, (Y. Wang et al., 2020) proposed a CNN-GRU model for sentiment analysis and showed that larger filter sizes and more filters result in better performance on benchmark datasets.

However, most of these studies have focused on a single dataset (Ganie et al., 2021) or a single NLP task (Ganie & Dadvandipour, 2021). Therefore, in this research, we aim to investigate the generalizability of these findings by conducting a comprehensive study on the impact of filter size and number of filters on the performance of CNNs in sentiment analysis using three different datasets: Amazon review dataset, Amazon food review dataset, and hate speech detection dataset. There have been several studies that have investigated the use of CNNs for sentiment analysis, with some focusing on the architecture of the model (Severyn & Moschitti, 2015b) (L. Zhang et al., 2018) while others have focused on the preprocessing of the data (Xu et al., 2016) (Severyn & Moschitti, 2015a) In (Severyn & Moschitti, 2015b) the author proposed a convolutional neural network (CNN) for sentence-level sentiment classification. The proposed model has been tested on two datasets, and the results show that it outperforms traditional models such as support vector machines (SVMs) (Ganie & Dadvandipour, 2022) and maximum entropy (ME) classifiers.In the study conducted by (Zhang et al., 2015), a novel deep learning model for sentiment analysis was introduced, incorporating a fusion of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. Their proposed model underwent rigorous testing on two distinct datasets, demonstrating superior performance when compared to conventional models like Support Vector Machines (SVMs) and Maximum Entropy (ME) classifiers. Similarly, in the work by (Xu et al., 2016) , the authors also presented a sentiment classification model that integrated both Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for enhanced accuracy and effectiveness. The proposed model has been tested on two datasets, and the results show that it outperforms traditional models such as SVMs and ME classifiers. In (Severyn & Moschitti, 2015b) the authors proposed a new model for sentiment classification, based on a combination of convolutional neural networks (CNNs) and tree kernel-based features. The proposed model has been tested on two datasets, and the results show that it outperforms traditional models such as SVMs and ME classifiers.

In the study conducted by Zhang et al. (2015), a novel deep learning model for sentiment analysis was introduced, incorporating a fusion of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. Their proposed model underwent rigorous testing on two distinct datasets, demonstrating superior performance when compared to conventional models like Support Vector Machines (SVMs) and Maximum Entropy (ME) classifiers. Similarly, in the work by Xu et al. (2016), the authors also presented a sentiment classification model that integrated both Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for enhanced accuracy and effectiveness.

The use of social media has grown exponentially over the past decade, leading to a significant increase in the amount of user-generated content. This has given researchers an opportunity to analyze the vast amount of data and extract valuable insights. The digital realm provides individuals with the ability to communicate and express their opinions freely. Unfortunately, social media platforms are often exploited to disseminate harmful and aggressive messages, comments, and discriminatory language. This phenomenon is commonly known as online hate speech, encompassing any form of communication that disparages or diminishes an individual or a group based on attributes such as race, ethnicity, gender, sexual orientation, nationality, religion, or political affiliation. (Z. Zhang & Luo, 2019). Defamation (Coe et al., 2014), incitement of violence (Hanzelka & Schmidt, 2017), provocation through statements that promote discriminatory views on political or social issues (Bhavnani et al., 2009), and the spread of rumors and conspiracy theories (Sunstein & Vermeule, 2008) are among the behaviors that violate social norms.

According to (Chaudhry & Gruzd, 2020) study, an examination of 51,991 public comments posted on 119 news stories related to race, racism, or ethnicity on the Canadian Broadcasting Corporation News Facebook page revealed the widespread dissemination of hate speech directed towards indigenous and black individuals. This perpetuates dominant discourses on white identities, as highlighted by Ben-David and (Ben-David & Matamoros-Fernández, 2016), and reflects a racially segregated pattern of online traffic. Recent statistics indicate that a large number of people globally are afflicted with one or multiple mental disorders (Rehm & Shield, 2019). Early detection of mental illness has been shown to positively impact disease progression and treatment outcomes. In contemporary times, natural language processing (NLP), a specialized domain within artificial intelligence (AI) technologies, has played a pivotal role in the analysis and management of extensive textual data. NLP techniques have greatly facilitated diverse tasks, including information extraction and sentiment analysis. (Nadkarni et al., 2011), emotion detection, and mental health surveillance (Nadkarni et al., 2011) (Mukherjee et al., 2020). Identifying mental illness from text can be viewed as a text classification or sentiment analysis task, wherein the application of natural language processing (NLP) techniques enables the automated recognition of early warning signs of mental health issues. This capability supports early detection, prevention, and treatment efforts in the field of mental health.

Previous literature reviews have primarily focused on computational approaches to detecting mental illnesses, with a particular emphasis on specific disorders such as suicide (Rehm & Shield, 2019) (Franco-Martín et al., 2018) (Ji et al., 2021) and depression (Giuntini et al., 2020) (Mahdy et al., 2020) (Khan et al., 2018) or specific data sources like social media (Skaik & Inkpen, 2021) (Chancellor & De Choudhury, 2020) (Ríssola et al., 2021) or non-clinical texts (Calvo et al., 2017). Various machine learning models have been developed by combining multiple extracted features. The majority of papers using machine learning methods have employed supervised learning techniques such as Support Vector Machine (SVM) (Saleem et al., 2012) (Ziwei & Chua, 2019b) (Prakash et al., 2021), Adaptive Boosting (AdaBoost) (Prakash et al., 2021) (Almouzini & Alageel, 2019) (Mbarek et al., 2019), k-Nearest Neighbors (KNN) (Mbarek et al., 2019) Decision Tree (He & Luo, 2016), Random Forest (Prakash et al., 2021), Logistic Model Tree (LMT) (Briand et al., 2018), Naive Bayes (NB) (Deshpande & Rao, 2017), Logistic Regression (Deshpande & Rao, 2017), XGBoost (Deshpande & Rao, 2017), and ensemble models that combine several methods (Prakash et al., 2021), to detect mental illness.

The significance of mental health as a fundamental component of human well-being has become increasingly prominent in recent years. This heightened attention is driven by the alarming prevalence of mental health disorders and their adverse impact on both individuals and society at large. (*Mental Health and Substance Use*, n.d.). Social media platforms, such as Twitter and Facebook, have become increasingly popular for individuals to express their thoughts and emotions, making them a valuable source of data for predicting and classifying mental health status (De Choudhury et al., 2013) (Coppersmith et al., 2015) (Krishnamoorthy et al., 2021) ("Proc. - 2018 IEEE Int. Conf. Big Data Smart Comput. BigComp 2018," 2018) ("Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017," 2018) ("2015 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2015," 2015) ("ICNC-FSKD 2017 - 13th Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov.," 2018) (Chancellor & De Choudhury, 2020). Since 2013, research has been able to identify and evaluate the existence of clinically significant mental health conditions such as major depression (De Choudhury et al., 2013)(Tsugawa et al., 2015)(Reece & Danforth, 2017), suicidality (Coppersmith et al., 2016)(Burnap et al., 2015)(Shing et al., 2018), eating disorders (Chancellor et al., 2016)(T. Wang et al., 2017), and schizophrenia (Mitchell et al., 2015), through the utilization of various methodologies and techniques such as machine learning, artificial intelligence, natural language processing, and human-computer interaction.

The utilization of such techniques have been implemented by various social media platforms, such as Facebook, as a means of preventing suicide (*Facebook Is Using AI to Spot Users with Suicidal Thoughts - The Verge*, n.d.)(*Under the Hood: Suicide Prevention Tools Powered by AI - Engineering at Meta*, n.d.) and improving mental health outcomes. This has given rise to a burgeoning field known as "digital psychiatry" (Torous et al., 2014) which aims to harness these predictive signals for the betterment of mental health service provision. It has been demonstrated through various studies that individuals suffering from a plethora of mental disorders such as depression, psychosis, and other severe mental illnesses, tend to utilize social media platforms at rates comparable to the general population. This usage ranges from approximately 70% among middle-aged and older individuals to as high as 97% among younger individuals (K.A. et al., 2018) (Birnbaum et al., 2017)(Brunette et al., 2019) . Moreover, preliminary investigations have unveiled a noteworthy trend wherein a considerable portion of individuals grappling with mental illness opt to utilize social media platforms as a means to share their personal experiences. These platforms serve as avenues for them to acquire information related to their mental health and available treatment options. Additionally, they offer opportunities to both extend and receive support from others confronting similar mental health challenges. (J.A. et al., 2016) (Bucci et al., 2019).

Numerous studies have utilized machine learning techniques to classify mental health status based on social media data. One study employed support vector machines and random forests to classify depression and anxiety in Twitter users, achieving an accuracy of 73% (De Choudhury et al., 2013). Another study employed a combination of natural language processing and machine learning techniques to classify depression, anxiety, and stress in social media users, achieving an accuracy of 72.5% (Coppersmith et al., 2015). Other studies have employed decision tree algorithms (“Proc. - 2016 IEEE Int. Conf. Big Data, Big Data 2016,” 2016) convolutional neural networks (T.Sajana, 2021)(“10th International Conference on Social Computing and Social Media, SCSM 2018 Held as Part of HCI International 2018,” 2018), and deep learning approaches (Nijhawan et al., 2022)(Burrows et al., 2022) to categorize mental health issues in data from social media. The focus of other research has been on social media platform suicidal ideation prediction. One study utilized a combination of machine learning and network analysis to identify individuals at risk of suicidal ideation on Twitter, achieving an accuracy of 83% (De Choudhury et al., 2013). Another study employed a deep learning approach to predict suicidal ideation in Reddit posts, achieving an accuracy of 82.8% (“Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018,” 2019). Machine learning approaches have been utilized in other studies to predict suicide behavior in social media data. (De Ávila Berni et al., 2018)(Weller et al., 2021) and to identify individuals at risk of suicide (Castillo-Sánchez et al., 2020)(Ryu et al., 2019)(Chancellor et al., 2019).In this research, we used cutting-edge machine learning methods to infer people's mental health from their social media posts. We classified the posts in our dataset as either suicidal or non-suicidal using another dataset containing posts from individuals with suicidal thoughts and the cosine similarity measure. We applied logistic regression, support vector machines.

The gradient descent (GD) is a widely used optimization algorithm in machine learning (Bottou, 2012). It is an iterative method that minimizes a cost function by finding the steepest descent direction. GD has two main variants, batch gradient descent (BGD) and stochastic gradient descent (SGD). BGD updates the parameters using the average of all the training examples while SGD updates the parameters using a single training example at a time. However, both BGD and SGD have their limitations. BGD requires all the training data to be loaded into memory, which can be computationally expensive for large datasets. SGD can be unstable and may converge to a suboptimal solution (Bottou, 2012). To overcome the limitations of BGD and SGD, several new optimization techniques have been proposed in the literature. One of these techniques is the mini-batch gradient descent (MBGD), which is a compromise between BGD and SGD. MBGD updates the parameters using a small random sample of training examples (a mini-batch) at a time. This approach reduces the memory requirements of BGD and the instability of SGD (Goodfellow I., 2016).

Another optimization technique that has gained popularity in recent years is the Adam optimizer (Kingma & Ba, 2015). The Adam optimizer stands as a variant of stochastic gradient descent, distinguished by its utilization of adaptive learning rates tailored for each parameter. Additionally, it integrates momentum to expedite the convergence process. Extensively proven in numerous deep learning applications, the

Adam optimizer has demonstrated its effectiveness and versatility in optimizing complex neural networks. (Kingma & Ba, 2015). Scholars have suggested alterations to the gradient descent algorithm, in addition to these optimization methods. For example, the Nesterov accelerated gradient (NAG) descent uses an accelerated gradient method to estimate the gradient at the next step (NESTEROV, 1983). This modification has been shown to improve convergence rates compared to standard GD (Sutskever et al., 2013). Another modification of GD is the conjugate gradient (CG) descent, which uses conjugate directions to estimate the gradient (Shewchuk, 2019). CG has been shown to converge faster than standard GD for certain types of problems (Shewchuk, 2019).

The use of second-order methods, such as the Hessian matrix, has also been proposed to optimize the cost function. However, these methods are computationally expensive and are not widely used in practice (Goodfellow I., 2016). Apart from optimization techniques, several studies have focused on the choice of activation functions. The sigmoid and ReLU (rectified linear unit) activation functions are the most often utilized ones. Sigmoid functions are smooth and differentiable but suffer from the vanishing gradient problem, while ReLU functions are non-smooth but do not suffer from the vanishing gradient problem (Goodfellow I., 2016). Recent studies have also focused on the use of convolutional neural networks (CNNs) for image recognition (Krizhevsky et al., 2012). CNNs use a hierarchical architecture that learns features at multiple levels of abstraction. They have shown to be effective in various image recognition tasks, such as object recognition and segmentation (Krizhevsky et al., 2012). In addition to optimization techniques and activation functions, several studies have investigated regularization techniques. Regularization is used to prevent overfitting by adding a penalty term to the cost function. Common regularization techniques include L1 and L2 regularization, dropout, and early stopping (Goodfellow I., 2016).

1.5 Thesis Outline

This thesis is organized as follows in the remaining chapters:

Thesis Outline:

Chapter 2: Literature Review

Chapter 3: Pre-trained Transformer Models for Sentiment Analysis on Social Media Text
3.1 Overview of Pre-trained Transformer Models
3.2 Fine-Tuning Pre-trained Transformer Models for Sentiment Analysis on Social Media Text
3.3 Training Models from Scratch on Social Media Data and Comparison of Fine-tuning and Training from Scratch
3.4 Enhancing Language Comprehension: An Innovative Approach Incorporating Clustered Attention In Transformer Models.

3.5 Findings and Conclusions

Chapter 4: The Impact of Informal Language on Sentiment Analysis Models on Social Media Text
4.1 Overview of Informal Language in Social Media Text
4.2 Emoticons, Hashtags, and Slang
4.3 The Effect of Informal Language on Sentiment Analysis Models
4.4 Findings and Conclusions

Chapter 5: and Optimizing Filter Size and Number of Filters for Sentiment Analysis using CNN: A Comparative Study across
5.1 Overview of Filter Size and Number of Filters
5.2 Explanation Of The Role Of Filter Size And The Number Of Filters In Cnn Based Sentiment Analysis Model
5.3 Conclusion

Chapter 6: Investigating the Relationship between Hate Speech and Mental Health Using Natural Language Processing Techniques
6.1 Prologue
6.2 Findings and Discussions
6.3 Conclusion

Chapter 7: Social Media Posts as a Window into Mental Health: A Machine Learning Approach and
7.1 Overview of Mental Health and Social Media
7.2 Methodology for using Social Media Posts to Predict Mental Health
7.3 Findings and Conclusions

Chapter 8: From Big Data to Smart Data: A Sample Gradient Descent Approach for Machine Learning and Fine Tunning LLM's
8.1 Foreword
8.2 Findings and Conclusions

Chapter 9: Conclusion
9.1 Summary of Findings
9.2 Contributions
9.3 Implications and Recommendations
9.4 Future Work

Chapter 10: References

Chapter 3: Pre-trained Transformer Models for Sentiment Analysis on Social Media Text

3.1 Overview of Pre-trained Transformer Models

One kind of deep learning architecture that has gained popularity for natural language processing tasks, such as sentiment analysis, is pre-trained transformer models. Compared to training from scratch, these models can be refined to accomplish particular tasks using a far smaller amount of data because they have already been pre-trained on enormous volumes of text data. The Transformer architecture, which was presented in the 2017 paper "Attention Is All You Need" by Vaswani et al., serves as the foundation for Transformer models. (Vaswani et al., 2017). The key innovation of the Transformer architecture is the attention mechanism, which allows the model to attend to different parts of the input sequence when making predictions. This allows the model to effectively capture long-range dependencies in the input data, making it well-suited for tasks such as language translation.

Before delving into the intricacies of Transformers, it is imperative to comprehend the evolutionary trajectory of the field, starting from the rudimentary Simple Recurrent Neural Networks (RNNs) to the present-day Transformer models. This expository odyssey will provide insights into the underlying principles and key advancements that have culminated in the emergence of Transformers as the state-of-the-art approach for various Natural Language Processing (NLP) tasks.

3.1.1 RNN

In this section, we will delve into the intricacies of Recurrent Neural Networks (RNNs) and compare them with Simple Artificial Neural Networks (ANNs). Specifically, we will explore the key distinctions between the two and highlight the most substantial difference between them. The capacity of Recurrent Neural Networks (RNNs) to process sequential data has led to a notable increase in the popularity of this class of neural networks in recent years. In this type of network, the hidden state from one time step is passed on to the next time step, allowing information to persist over a sequence of inputs. RNNs are commonly used for tasks such as language modeling, machine translation, speech recognition, and text classification, among others.

Before the advent of RNNs, feedforward neural networks (also known as simple Artificial Neural Networks (ANNs)) were commonly used for many applications. However, these networks were not designed to handle sequential data, as they lacked the ability to capture long-term dependencies between inputs. To address this issue, RNNs were introduced, which allow the network to use information from past time steps to inform its predictions at the current time step.

The key difference between RNNs and simple ANNs is that the former can handle sequential data and the latter cannot. In simple ANNs, the input data is processed through multiple layers of artificial neurons, with the output of one layer being the input to the next layer. On the other hand, in RNNs, the input data is processed through a hidden layer that takes as input the current input and the hidden state from the previous time step. This allows RNNs to capture dependencies over long sequences of inputs and to make predictions based on this information.

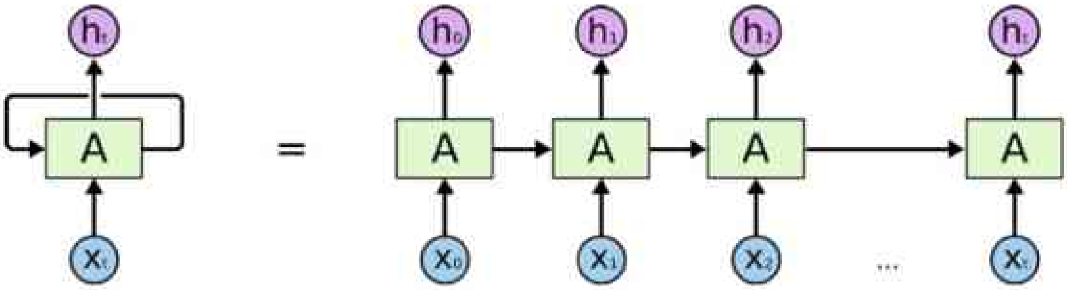


Figure 1. *RNN Architecture, source:* (*Understanding LSTM Networks -- Colah's Blog*, n.d.)

The fundamental concept behind feedforward networks is the ability to retain information learned during the training process. Nevertheless, Recurrent Neural Networks (RNNs) possess this characteristic, but with the added advantage of being able to remember prior inputs while generating outputs. This ability to incorporate information from past inputs is what sets RNNs apart from basic feedforward networks.

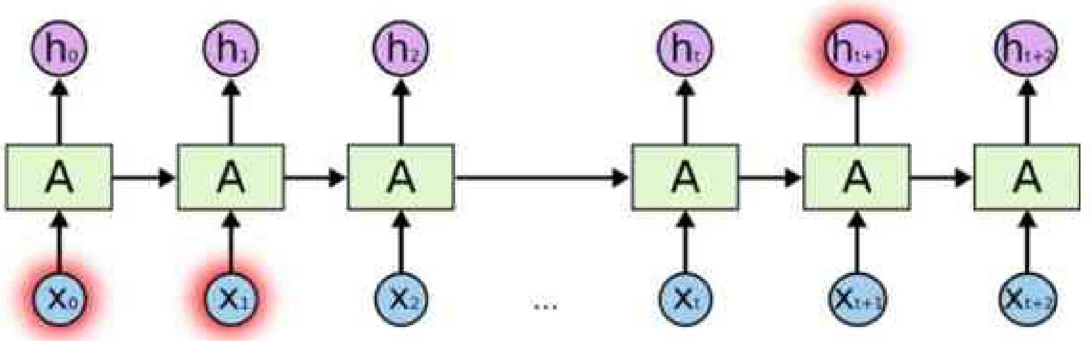


Figure 2. *Image illustrating long-term dependencies, source:* (*Understanding LSTM Networks -- Colah's Blog*, n.d.)

Many types of models, such as vector-sequence, sequence-vector, and sequence-to-sequence models, use the Recurrent Neural Network (RNN) architecture. Vector-sequence models are used to convert fixed-sized input vectors into output vectors of varying lengths, such as image captioning, where an image is transformed into a textual description. Sequence-vector models, on the other hand, process input vectors of variable size and produce fixed-sized output vectors, such as sentiment analysis of movie reviews. Finally, the widely utilized sequence-to-sequence model takes a sequence of inputs and produces another sequence of outputs, including language translation and time series data prediction in the stock market. Despite its widespread usage, RNNs are plagued by certain disadvantages, such as slow training times and the vanishing gradient problem, which limits its ability to effectively handle long sequences and maintain long-term dependencies. The mathematical formulation of RNNs can be described as follows:

Let W_xh, W_hh, and W_ho be the weight matrices that connect the input layer to the hidden layer, the hidden layer to itself, and the hidden layer to the output layer, respectively. Let X_t be the input at time step t, and H_t be the hidden state at time step t. Then, the equations that describe how the hidden state and the output are computed are:

$$H_t = \tanh(W_{xh} * X_t + W_{hh} * H_{t-1}) \tag{1}$$

$$Y_t = W_{ho} * H_t \tag{2}$$

where Y_t is the output at time step t and tanh is the hyperbolic tangent activation function.

The vanishing gradient problem, which arises when the gradients used to update the weights get extremely small, is the primary difficulty in training RNNs. The gradients must be propagated back through numerous time steps in order for the weights to be updated using backpropagation through time, which causes this issue. To address this problem, various techniques have been proposed, including the use of long short-term memory (LSTM) cells and gated recurrent units (GRUs). In conclusion, RNNs are a powerful tool for handling sequential data and have been widely used in various applications, including language modeling, speech recognition, and machine translation. The mathematical formulation of RNNs, based on a set of recursive equations, provides a way to model the dependencies between inputs and outputs over time.

3.1.2 LSTM

A kind of recurrent neural network (RNN) called the Long Short-Term Memory (LSTM) architecture was created to address the vanishing gradient issue in conventional RNNs. Hochreiter and Schmidhuber introduced the LSTM model in 1997. Unlike traditional RNNs, which only have a single layer, LSTMs consist of a series of memory cells that are interconnected through gates, allowing the model to maintain long-term memory in addition to processing information in real-time.

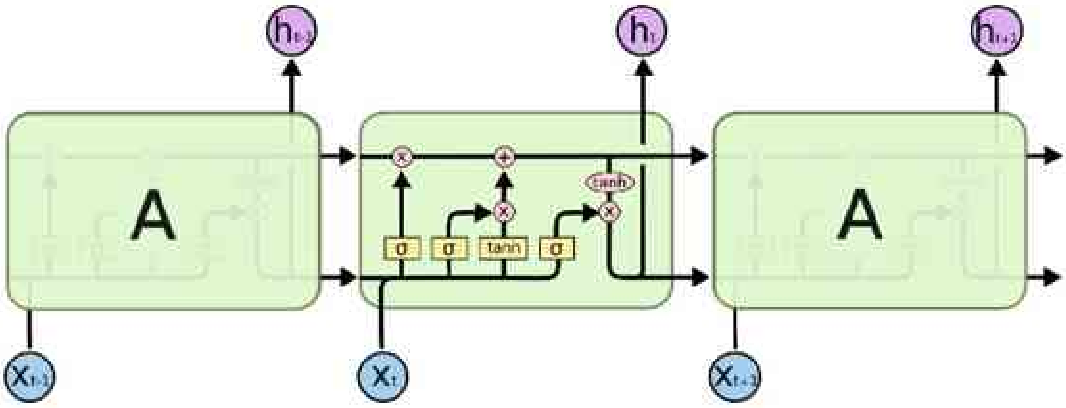


Figure 3. LSTM Architecture, source: (*Understanding LSTM Networks -- Colah's Blog*, n.d.)

3.1.3 Transformers and Attention Networks

When a transformer model is pre-trained on copious amounts of textual data, it can acquire valuable representations of the text that can be optimized for particular applications, like sentiment analysis. BERT (Bidirectional Encoder Representations from Transformers) is one of the most well-liked pre-trained transformer models. It was first presented in a 2018 paper by (Devlin et al., 2019). BERT has been optimized for numerous NLP tasks, including sentiment analysis, and has produced state-of-the-art results in many of them. Adding a new layer on top of an already-trained transformer model and training it on a small amount of labeled sentiment analysis data is the usual method for fine-tuning the model for sentiment analysis. The model is able to use the pre-learned representations of the text because the pre-trained weights are maintained constant. This fine-tuning process can significantly reduce the amount of labeled data needed to train a high-performing sentiment analysis model. The Transformer model is a novel neural network architecture introduced in 2017 by Vaswani et al. in their paper "Attention is All You Need." It is a type of deep learning model that operates on sequences of data, such as text, audio, or video. The Transformer model is unique in its use of attention mechanisms, which allow the model to focus on specific parts of the input sequence while generating the output.

The attention mechanism in a Transformer model is based on the concept of scaled dot-product attention. The attention mechanism calculates a weighted sum of the input representations, where the weights are calculated based on the dot-product between the query and key vectors, which are learned during training. The result is a context-aware representation of the input, which the model uses to generate its output.

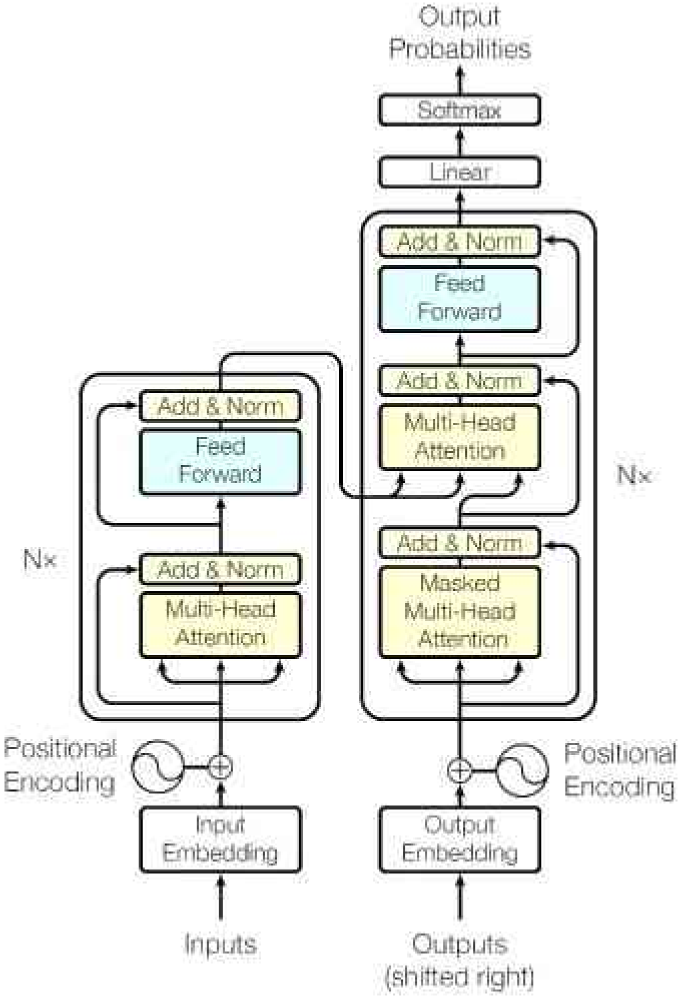


Figure 4. *Transformer Model Architecture, Source:* (Vaswani et al., 2017)

The mathematical formulation of the Transformer model's attention mechanism can be expressed as follows:

Given an input sequence x = (x1, x2, ..., xn), where xi is the representation of the i-th input element, the Transformer model calculates the attention score eij between each pair of elements xi and xj using the following equation:

$$e_{ij} = \text{softmax(Qi * Kj^T / √d\_k)} \quad (3)$$

where Qi and Kj are the query and key vectors for the i-th and j-th input elements, respectively, and d_k is the dimension of the key vector.

The attention scores are used to calculate the weighted sum of the input representations, which becomes the context-aware representation of the input:

$$\text{c\_i = ∑\_j eij * Vj} \quad (4)$$

where c_i is the context-aware representation of the i-th input element, and Vj is the value vector for the j-th input element.

The Transformer model uses this context-aware representation as input to a feed-forward neural network to generate the final output. In summary, pre-trained transformer models have become a popular choice for sentiment analysis tasks due to their ability to effectively capture long-range dependencies in the text data and their ability to be fine-tuned on small amounts of task-specific data. BERT is one of the most widely used pre-trained transformer models and has achieved state-of-the-art results on a variety of NLP tasks, including sentiment analysis.

3.2 Fine-Tuning Pre-trained Transformer Models for Sentiment Analysis on Social Media Text

Fine-tuning pre-trained transformer models for sentiment analysis on social media text is an effective approach that leverages the powerful language representation capabilities of these models to solve NLP tasks. The main idea behind fine-tuning is to use the pre-trained weights of a large transformer model as the starting point for training on a new task, rather than training the model from scratch. This approach has been widely adopted in NLP because it can significantly reduce the amount of data and computation required to train a model, while still allowing the model to perform well on new tasks. Sentiment analysis is a popular NLP task that involves classifying the sentiment of a text as positive, negative, or neutral. In the context of social media text, sentiment analysis can be used to identify patterns and trends in public opinion, understand the tone and mood of social media conversations, and inform decision-making in various industries. To fine-tune a pre-trained transformer model for sentiment analysis, the first step is to prepare a training dataset that consists of social media text and corresponding sentiment labels. This dataset can be annotated manually or obtained from existing annotated datasets.

Once the training dataset is prepared, the next step is to fine-tune the pre-trained transformer model on the sentiment analysis task. This involves updating the model's weights using backpropagation and gradient descent to minimize the cross-entropy loss between the predicted sentiment labels and the true sentiment labels. The fine-tuning process can be performed using transfer learning, where only the final layer of the model is trained, or using full-model fine-tuning, where all the layers of the model are updated. In addition to fine-tuning, attention networks can be used to enhance the performance of the transformer model for sentiment analysis. Attention networks allow the model to focus on different parts of the input text to make predictions, rather than using a fixed representation for the entire text. This can improve the model's ability to capture the relevant information in the text, leading to more accurate sentiment predictions.

In conclusion, fine-tuning pre-trained transformer models for sentiment analysis on social media text is an effective approach that leverages the power of these models to solve NLP tasks. By fine-tuning the model and using attention networks, it is possible to achieve high accuracy in sentiment analysis, even with limited training data.

3.3 Training Models from Scratch on Social Media Data and Comparison of Fine-tuning and Training from Scratch

Fine-tuning pre-trained transformer models and training models from scratch on social media data are two different approaches in sentiment analysis. The former approach involves utilizing a pre-trained transformer model and fine-tuning it for the specific task of sentiment analysis on social media text. On the other hand, the latter approach involves training a model from scratch on social media data. Both of these approaches have their own advantages and disadvantages. The benefit of fine-tuning pre-trained transformer models is that the model's performance on the target task can be greatly enhanced by utilizing the pre-trained model's knowledge and high-level representations. The researchers fine-tuned a pre-trained Bert model to identify online harassment in the study "Identification of online harassment using ensemble fine-tuned pre-trained Bert," and they got encouraging results.

3.3.1 Problem Definition: Identification of online harassment using ensemble fine-tuned pre-trained Bert

The identification of online hate speech is a critical issue for the field of natural language processing (NLP), particularly as social media has amplified this phenomenon by providing a virtual platform for online harassment. This study aims to address this issue by utilizing the trolling aggression and cyber-bullying dataset from the shared tasks workshop. To accomplish this, the study employs an extreme pre-processing methodology and implements an ensemble approach in the model building process. The study also evaluates the performance of existing algorithms such as random forest, logistic regression, and multinomial Naïve Bayes. The results demonstrate that logistic regression is the most efficient algorithm, achieving an accuracy of 57.91%. However, the ensemble bidirectional encoder representation from transformers demonstrates even more promising results, yielding a precision of 62%, outperforming most existing models.

3.3.2 Application of BERT model

The widespread use of social media platforms has made it possible for people to engage and communicate with one another on a worldwide level. But the harmful use of social media has increased the prevalence of online harassment, which can take the form of cyber-bullying, which is characterized by repeated actions intended to terrorize, terrorize, or shame others. (*Cyberbullying: What Is It and How to Stop It | UNICEF*, n.d.). In order to address this issue, researchers in the field of Natural Language Processing (NLP) have sought to develop automated solutions for detecting online harassment.

There has been considerable work performed in this area using traditional supervised machine learning methods such as Support Vector Machines (SVMs), Long Short-Term Memory (LSTM) networks, logistic regression, and decision trees (Dinakar et al., 2011) (Reynolds et al., 2011) (Badjatiya et al., 2017). However, the majority of this work has been carried out using data in the English language. The present study seeks to address this challenge by using fine-tuned Bidirectional Encoder Representations from Transformers (BERT) networks to detect online harassment in multilingual data (Malte & Ratadiya, 2019). To that end, the authors of this study have focused on performing extensive pre-processing on the data, including lemmatization, stop-word removal, and Part-of-Speech (PoS) tagging, in order to ensure the accuracy of the input fed into the model. The study uses the TRolling Aggression and Cyber-bullying (TRAC)-1 dataset, which has been shown to yield high accuracy results. The results of this study demonstrate that fine-tuning pre-trained BERT networks using an ensemble approach can achieve accuracy rates comparable to state-of-the-art results, with greater accuracy than baseline results obtained using traditional classifiers.

3.3.3 Dataset

For the purpose of this study, the data was obtained from the shared task on aggression identification dataset, which was organized at the "Trolling, Aggression, and Cyber-Bullying Workshop" (Ganie & Dadvandipour, 2021). The training data consisted of 10,799 Facebook comments that were randomly selected and annotated into three categories: Overly Aggressive (OAG), Covertly Aggressive (COA), and Non-Aggressive (NAG). The validation or test data comprised of 1200 samples.

3.3.4 Results and Discussions

This study aimed to identify instances of online harassment in the TRAC-1 dataset using an ensemble approach with pre-trained BERT. The dataset was collected from the "Shared Task on Aggression Identification" workshop and comprised 10,799 randomly selected Facebook comments that were annotated into three categories: Overly Aggressive (OAG), Covertly Aggressive (COA), and Non-Aggressive (NAG). The validation data consisted of 1,200 samples. The identification of abusive observations is crucial as it validates the experiences of victims and provides insight into the extent of the problem.

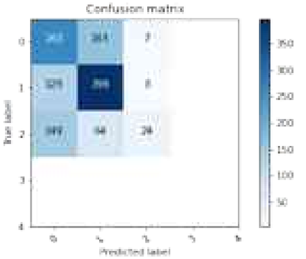
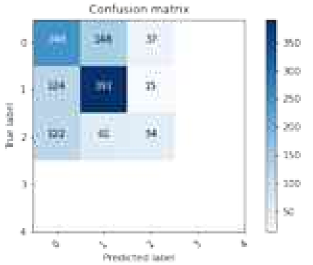
However, the recent study by (Malte & Ratadiya, 2019) failed to incorporate crucial preprocessing steps such as handling of NaN values, stopword removal, PoS tagging, contractions, stemming, and lemmatization, which could have led to over-fitting of their model (Paraschiv & Cercel, 2020) (Rajamanickam et al., 2020) (Ganie et al., 2021) (Modha et al., 2018). Additionally, they did not consider fine-tuning strategies (Al-Hafadhi & Krallics, 2021), which could have further improved the performance of the model. In contrast, this study performed all of the aforementioned preprocessing steps, as well as attempted to identify abuse in multilingual text. The experiment was conducted on a Tesla T4 GPU with a Core i5 processor and 12 GB of RAM, as detailed in Table 7.

***Table** 7: Data Preprocessing*

ID	TEXT	LABEL	CLEAN	NO CONTRACTIONS	STOPWORDS	POSTAG	LEMMATIZED
FACEBOOK_CORPUS_MSR_466073	Most of Private Banks ATM's Like HDFC, ICICI etc are out of cash. Only Public sector bank's ATM working	NAG	most of private banks atm like hdfc, icici etc are out of cash. only public sector bank atm working	['most', 'of', 'private', 'banks', 'atm', 'like', 'hdfc,', 'icici', 'etc', 'are', 'out', 'of', 'cash.', 'only', 'public', 'sector', 'bank', 'atm', 'working']	['private', 'banks', 'atm', 'like', 'hdfc', 'icici', 'etc', 'cash', 'public', 'sector', 'bank', 'atm', 'working']	[('private', 'JJ'), ('banks', 'NNS'), ('atm', 'VBP'), ('like', 'IN'), ('hdfc', 'NN'), ('icici', 'NN'), ('etc', 'FW'), ('cash', 'NN'), ('public', 'NN'), ('sector', 'NN'), ('bank', 'NN'), ('atm', 'IN'), ('working' , 'VBG')]	most of private banks atm like hdfc icici etc are out of cash only public sector bank atm working

The researchers attempted to classify online harassment using pre-existing algorithms such as SVM, Naive Bayes, logistic regression, and random forest, but the results were not as desired. Despite this, the accuracy was still higher than the baseline of 35.53% as demonstrated in Fig. 3. Due to the limited performance of these algorithms, the researchers turned to a deep learning approach and fed the preprocessed data into a pre-trained BERT model with a multi-head attention mechanism. BERT is a widely used language representation pre-training method that can be downloaded and utilized by NLP practitioners. There are two ways to approach the problem using BERT, either using existing models to extract high-quality language features from text data, or fine-tuning them to produce state-of-the-art predictions for specific NLP tasks such as classification, entity identification, and question answering.

MultinomialNB Results SVM Results

Accuracy	Confusion matrix	Accuracy	Confusion matrix
57.08		57.75	

Logistic Regression Random Forest

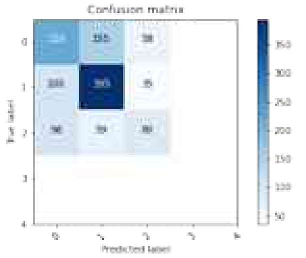
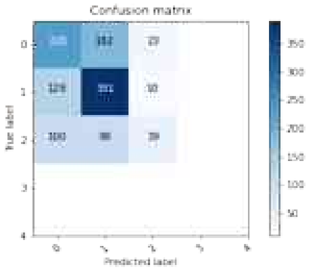
Accuracy	Confusion matrix	Accuracy	Confusion matrix
57.91		54.83	

Figure 5. *Machine Learning Results*

Table 8: *Learning rate= 2e-5, batch size =32*

EPOCH	TRAINING LOSS	VALIDATION LOSS	VALIDATION ACCURACY	TRAINING TIME	VALIDATION TIME
1	0.90	0.83	0.59	0:02:04	0:00:04
2	0.64	0.85	0.6	0:02:00	0:00:05
3	0.31	1.11	0.6	0:02:01	0:00:05
4	0.12	1.31	0.6	0:02:02	0:00:05

Table 9: *Learning rate = 5e-5, batch size = 64*

EPOCH	TRAINING LOSS	VALIDATION LOSS	VALIDATION ACCURACY	TRAINING TIME	VALIDATION TIME
1	0.91	0.85	0.60	0:01:31	0:00:08
2	0.61	0.95	0.60	0:01:36	0:00:09
3	0.23	1.31	0.60	0:01:38	0:00:09
4	0.07	1.60	0.60	0:01:39	0:00:09

1. Learning rate = 2e-5, adam epsilon = le-8, warm steps = 0, batch size = 64

Table 10: *Learning rate = 2e-5, batch size = 64*

EPOCH	TRAINING LOSS	VALIDATION LOSS	VALIDATION ACCURACY	TRAINING TIME	VALIDATION TIME
1	0.92	0.85	0.59	0:01:43	0:00:04
2	0.70	0.86	0.60	0:01:48	0:00:04
3	0.45	0.97	0.60	0:01:50	0:00:04
4	0.26	1.07	0.60	0:01:51	0:00:04

2. Learning rate = 5e-5, adam epsilon= le-8, warmup steps= 1, batch size= 16

Table 11: *Learning rate = 5e-5, batch size = 16*

EPOCH	TRAINING LOSS	VALIDATION LOSS	VALIDATION ACCURACY	TRAINING TIME	VALIDATION TIME
1	0.90	0.86	0.59	0:02:13	0:00:05
2	0.61	0.94	0.60	0:02:19	0:00:05
3	0.27	1.34	0.61	0:02:21	0:00:05
4	0.11	1.85	0.62	0:02:21	0:00:05

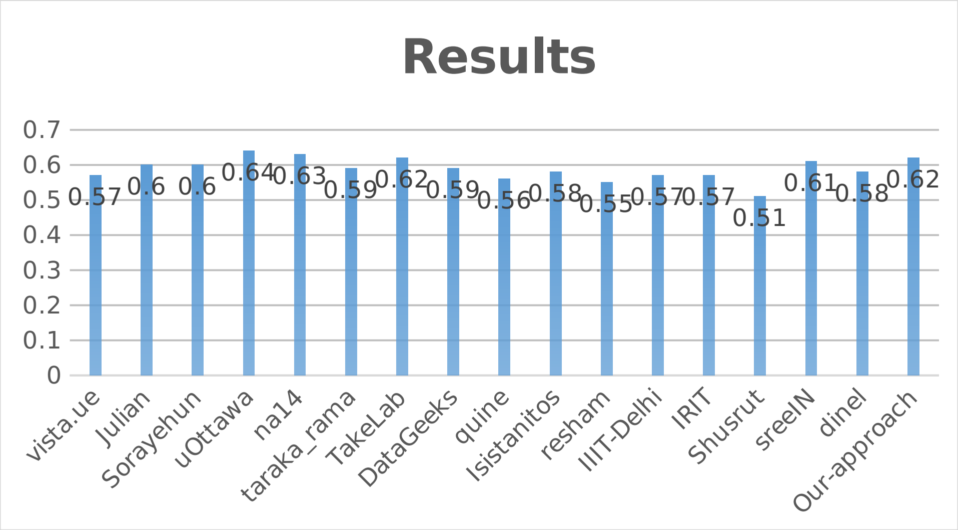


***Figure 6**. Result comparison on test data*

3.4 Enhancing Language Comprehension: An Innovative Approach Incorporating Clustered Attention in Transformer Models

This study pushes the boundaries of text classification through the introduction of an innovative clustering-based attention mechanism seamlessly integrated with embeddings, autoencoders, and inspired by transformer architectures. Evaluating its impact on financial sentiment analysis, fake news detection, and hate speech detection datasets, our research reveals consistent improvements in accuracy compared to conventional attention mechanisms. The incorporation of autoencoders contributes to performance refinement by reducing dimensions and optimizing both time and space efficiency. Notably, our model maintains comparable performance while exhibiting significantly reduced complexity, as evident from execution times and parameter counts. This research makes a significant contribution to the evolving landscape of attention mechanisms in natural language processing, showcasing the synergistic benefits of clustering strategies, embeddings, autoencoders, and transformer-inspired architectures in elevating the effectiveness of text classification models. The amalgamation of these components emerges as a promising paradigm, paving the way for advancements in precision, interpretability, and efficiency within the realm of text classification.

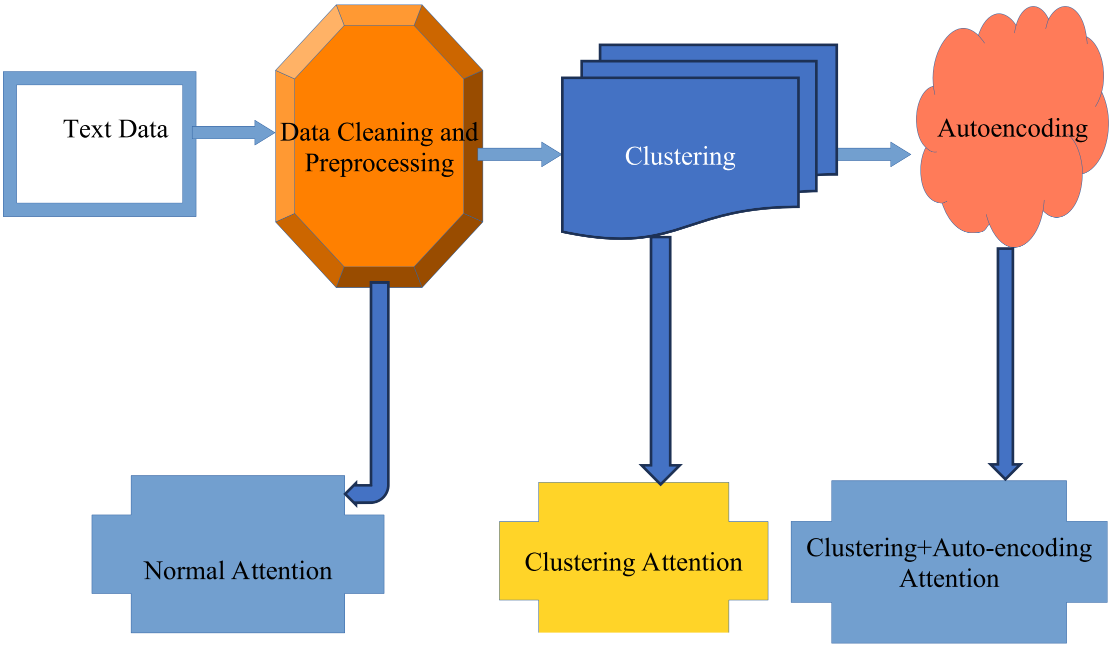


***Figure** 7: Workflow of clustering attention*

The attention mechanism involves the formulation of key (K), query (Q), and value (V) matrices through linear transformations. These matrices facilitate dynamic computation, enabling models to capture nuanced dependencies within sequential data. While these linear transformations have proven effective, their innate capacity to harness latent structural nuances might be optimized further. Our research delves into the augmentation of attention mechanisms by integrating clustering techniques into the generation of K and Q matrices. The integration of clustering adds a layer of semantic grouping, potentially enhancing the model's acuity in discerning intricate patterns within data. Clustering has exhibited promise across various domains (Xu & Tian, 2015; Chen et al., 2018), and our endeavor is to channel its potential within the intricate landscape of NLP. This fusion is poised to contribute to a more subtle understanding of the intricacies encoded within language data.

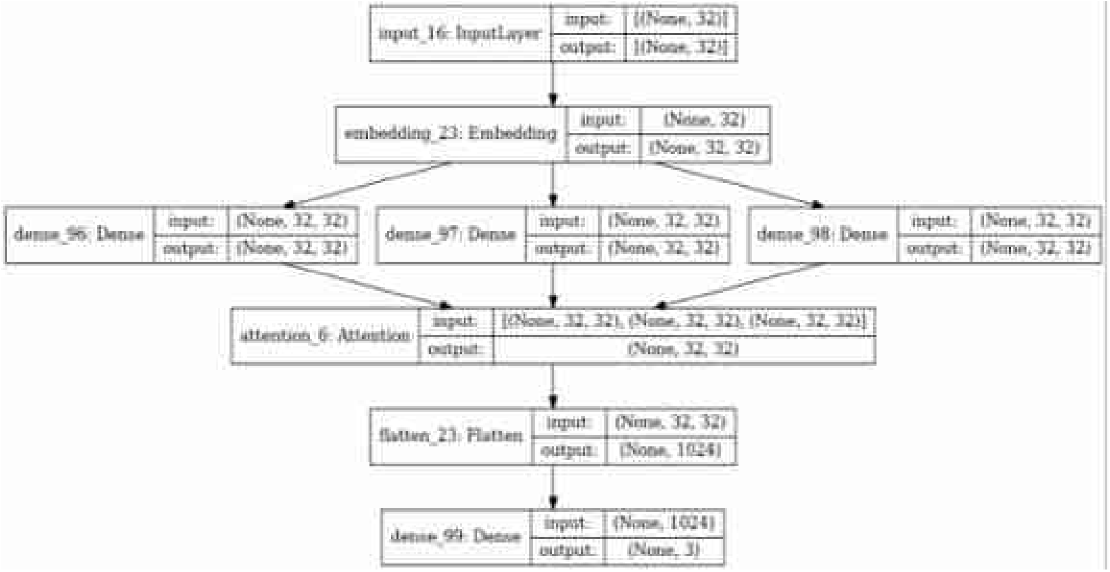


Figure 8: *Normal_attention_model_architecture*

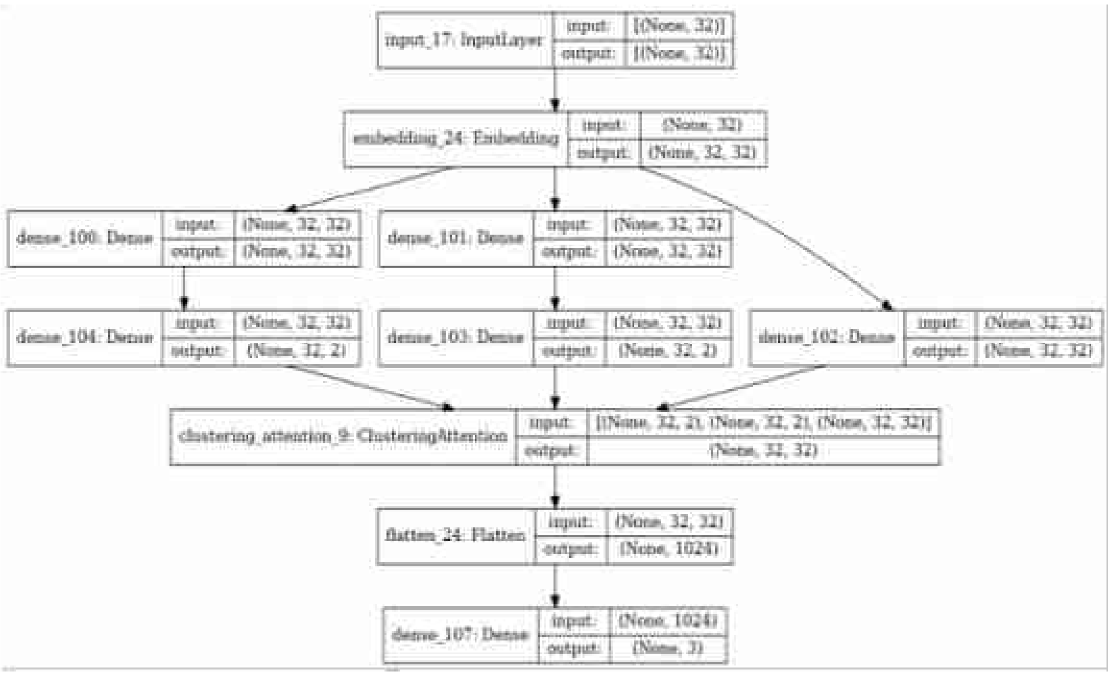


Figure 9: *AttentionwithCluster_model_architecture*

Furthermore, our exploration extends into the realm of cross-domain scenarios. Here, we scrutinize the knowledge transferability of clustered embeddings across diverse datasets, investigating how insights garnered from one domain influence the performance of models when transposed onto another. In essence, this research constitutes a comprehensive investigation into the interplay between attention mechanisms, embeddings, and clustering. The overarching objective is to push the boundaries of NLP applications, leveraging the synergy of these components to enhance model understanding and performance. Amidst this exploration, we introduce autoencoders into the fold, aiming to streamline the model's complexity and enhance efficiency. The subsequent evaluation on distinct datasets—financial sentiment analysis, fake news detection, and hate speech detection—reveals consistent accuracy improvements over traditional attention mechanisms. Noteworthy is the model's ability to maintain comparable performance while significantly reducing execution times and parameter counts, underscoring the potential of this amalgamation. This study not only advances the understanding of attention mechanisms but also highlights the transformative impact of synergizing clustering, embeddings, and autoencoders within the evolving landscape of NLP.

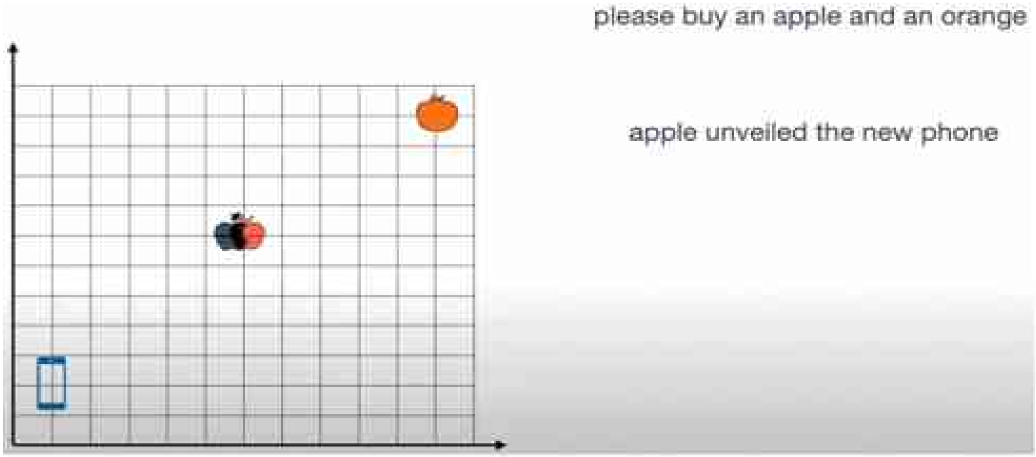

please buy an apple and an orange
apple unveiled the new phone

Figure 10: *Working Principal of Attention Mechanism*

In our quest to enhance attention mechanisms through clustering, we introduce a novel objective function that refines the conventional attention scoring mechanism. Our focus is on the key (K) and query (Q) matrices, integral components in attention computation. Let's start with the mathematical background. Consider a set of data points [illegible] with dimensions D. Each data pointx_i can be represented as a vector x_i in R^D . In the attention mechanism, the attention score α_{ij}between a query $q_i$and key $k_j$is traditionally computed as the scaled dot product

$$\alpha_{ij}=\frac{q_i \,.\, k_j}{\sqrt{D}} \tag{5}$$

This formulation ensures that the gradients remain stable during training. To incorporate clustering into the attention mechanism, we propose clustering the matrices $K$and Q. Let $K^c$and Q^cdenote the clustered version of $K$and . The clustering operation can be expressed as:

$$K^c = Cluster(K)$$

$$Q^c = Cluster(Q)$$

This operation groups similar vectors together, fostering a more structured attention computation. Now, we define our novel objective function as the mean squared error between the attention scores with and without clustering:

$$l=\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\alpha_{ij}\left(K^{c},Q^{c}\right)-\alpha_{ij}\left(K,C\right)\right)^{2} \tag{6}$$

where N is the number of data points.

Algorithm: Custom Attention Training Loop

1. Input:

3. **num_epochs:** Number of training epochs.
4. **data_batches:** Batches of input data for each iteration.

2. Initialize Model Parameters:

1. Initialize or load the model parameters.

3. Define Functions:

1. **scaled_dot_product_attention(q, k, v):** Define the scaled dot-product attention mechanism.
2. **cluster_matrix(matrix):** Define a function to perform clustering on the input matrix.
3. **custom_attention(q, k, v):** Define the custom attention mechanism using clustered matrices and compute the loss as the mean squared error (MSE) between the attention mechanisms.

4. Training Loop:

A. For each ***epoch*** in the range ***num_epochs***:
   . For each ***batch*** in ***data_batches:***
      a. Preprocess the batch to obtain query ***(q)***, key ***(k)***, and value ***(v)*** matrices.
      b. Compute the clustered attention and loss using the ***custom_attention*** function
      c. Update the model parameters using the loss gradient (not explicitly shown in the code snippet).

5. Output:

- The trained model parameters.

Note: The actual implementation of the clustering algorithm (cluster_matrix) is not provided in the code snippet. The choice of clustering algorithm would depend on the specific requirements of the problem, and the implementation would need to be completed based on the chosen clustering method.

Our proposed objective function encourages the attention mechanism to adapt its focus based on clustered representations of keys and queries, fostering more structured and context-aware attention scores. The subsequent experiments and results will delve into the efficacy of this approach in real-world NLP tasks, providing insights into the potential benefits of clustering-enhanced attention mechanisms. In our empirical exploration, we meticulously conducted a series of experiments to assess the efficacy of the clustering attention mechanism in comparison to the baseline attention model. The experiments were strategically designed to cover diverse domains, encompassing financial sentiment analysis, fake news detection, and hate speech detection. Leveraging these distinct datasets, we sought to unravel the inherent adaptability and performance nuances of the clustering attention mechanism across various contextual landscapes.

Table 12: *Experiment1_Financial_SentimentAnalysis*

Metrics	Normal Attention	Clustering Attention
Test Accuracy	0.778	0.79
Execution Time (s)	2.303	2.476
Space Used (params)	325,410	325,730

Table 13: *Experiment2_HateSpeech*

Metrics	Normal Attention	Clustering Attention
Test Accuracy	0.778	0.79
Execution Time (s)	2.303	2.476
Space Used (params)	325,410	325,730

Table 14: *Experiment3_FakeNews*

Metrics	Normal Attention	Clustering Attention
Test Accuracy	0.849	0.844
Execution Time (s)	2.648	2.469
Space Used (params)	326,243	326,563

In the domain of financial sentiment analysis, the clustering attention model exhibited a commendable convergence trajectory in both training accuracy and loss, as depicted in Figures 10 and 11. The nuanced confusion matrices (Figures 12 and 13) underscored the model's ability to discern complex sentiment patterns. Comparative analysis against the baseline attention model revealed comparable performance metrics, affirming the clustering attention mechanism's robustness in capturing intricate sentiment dynamics.

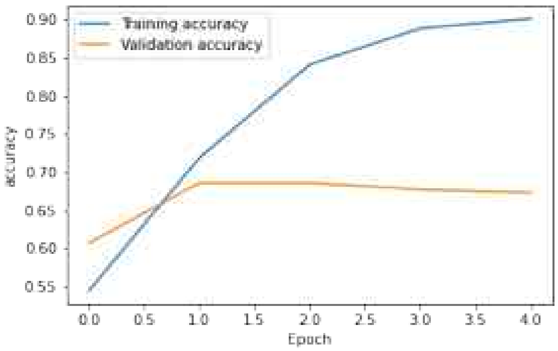


***Figure** 11: Training accuracy normal*

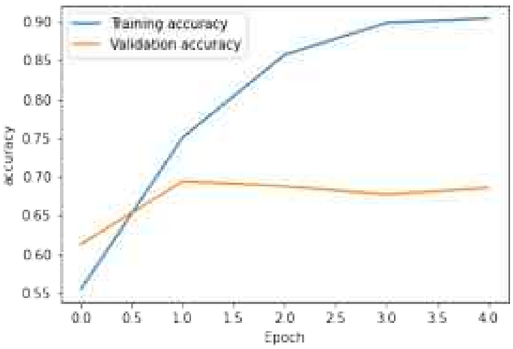


***Figure** 12: Training accuracy cluster*

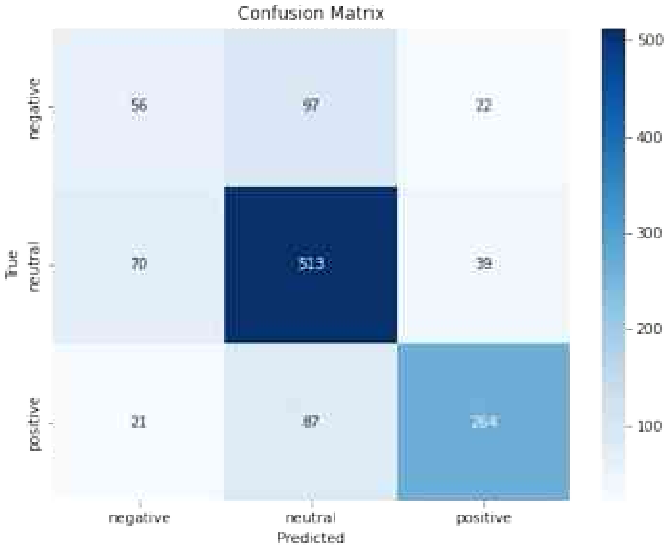


Figure 13: *Confusion_matrix_normal*

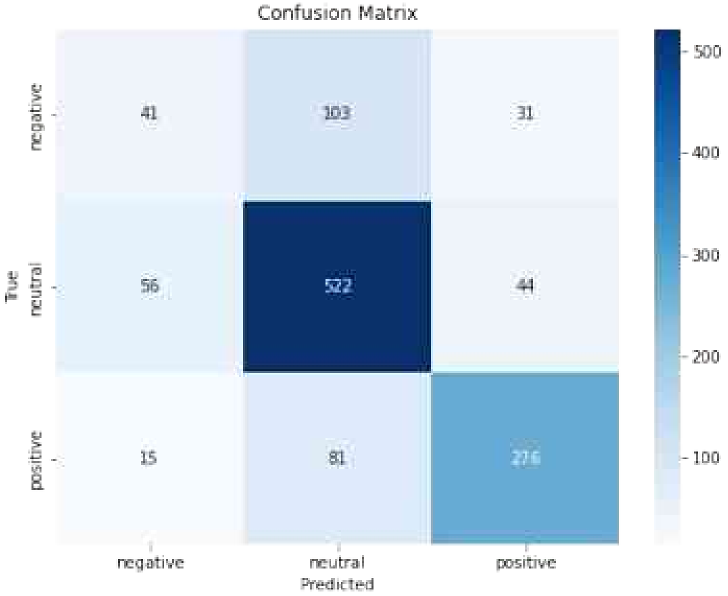


Figure 14: *Confusion_matrix_cluster*

Cross-Domain Adaptability:

Expanding our investigation beyond financial sentiment analysis, we delved into the domains of fake news detection and hate speech detection. Figures 14 and 15 showcase the clustering attention model's adeptness in navigating the challenges posed by fake news detection, demonstrating its resilience and comparable performance against the baseline attention mechanism.

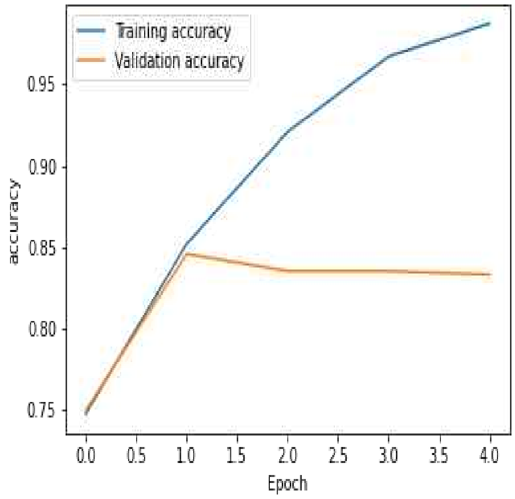


Figure 16: Training_accuracy_normal_hatespeech

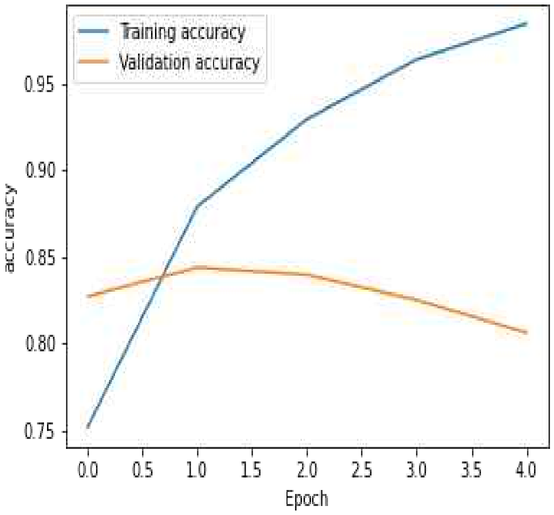


***Figure** 15: Training_accuracy_cluster_hatespeech*

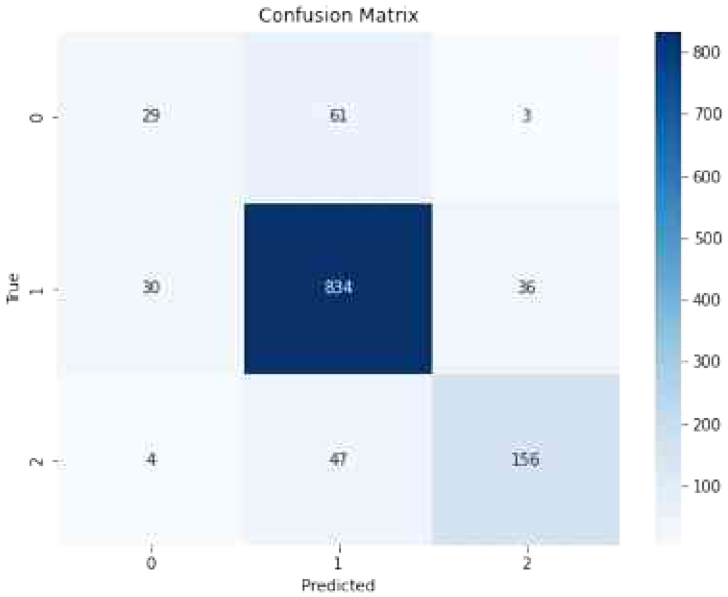


***Figure** 17: Confusion_matrix_normal_hatespeech*

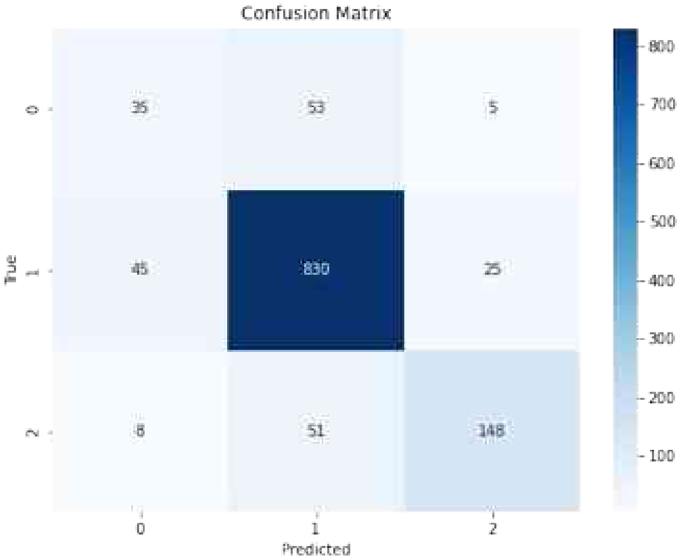


***Figure** 18: Confusion_matrix_cluster_hatespeech*

In the realm of hate speech detection (Figures 18 and 19), the clustering attention model once again proved its mettle, mirroring the baseline attention model's efficiency. The confusion matrices vividly capture the model's discriminative power, highlighting its potential in addressing the intricacies of hate speech classification.

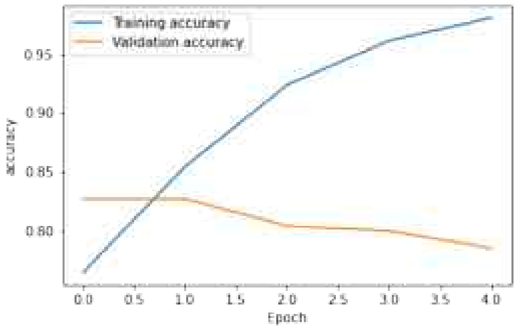


***Figure** 20: Training_accuracy_normal_fakenews*

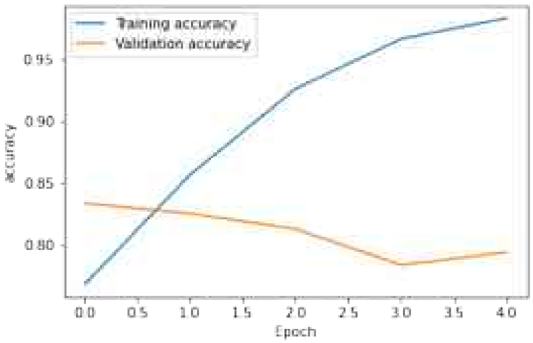


***Figure 19:** Training_accuracy_cluster_fakenews*

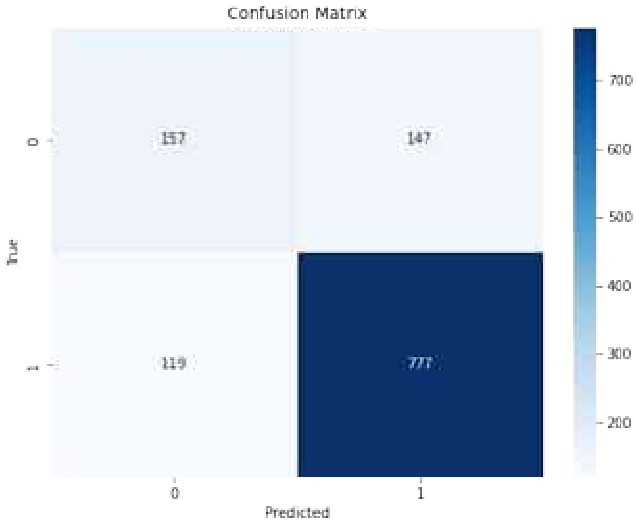


***Figure** 21: Confusion_matrix_normal_fakenews*

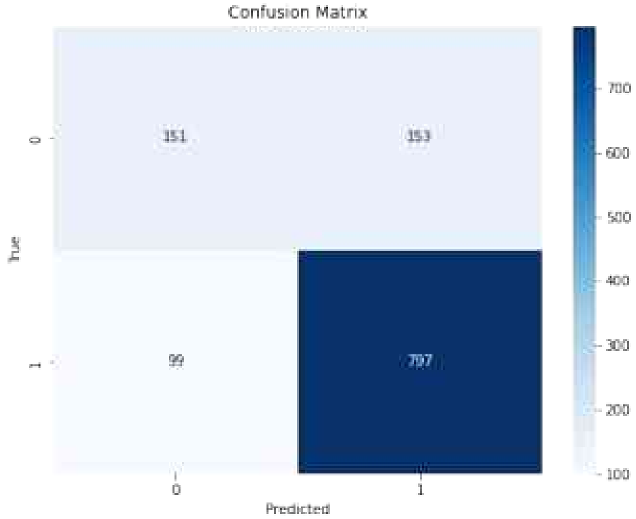


***Figure** 22: Confusion_matrix_cluster_fakenews*

Effect of Autoencoders on Model Performance:

The incorporation of autoencoders in the attention mechanism framework yields compelling enhancements in both efficiency and model complexity. The evaluation of two attention models—normal attention and clustered+autoencoder attention—underscores the impact of this augmentation. The results, as summarized in the table below, demonstrate noteworthy improvements in accuracy, execution time, and parameter count.

Table 15: *Clusterd plus Autoencoder Attention Comparison*

Model	Test Accuracy	Execution Time (seconds)	Space Used (parameters)
Normal Attention	0.755	0.1759	165,410
Clustered+Autoencoder Attention	0.7467	0.1185	138

The introduction of autoencoders within the attention mechanism not only maintains competitive accuracy levels but also substantially reduces execution time and model parameters. In the normal attention model, execution time is measured at 0.1759 seconds, whereas the clustered+autoencoder attention model achieves comparable accuracy with a significantly faster execution time of 0.1185 seconds. Furthermore, the approximate space used by the clustered+autoencoder attention model is drastically reduced to 138 parameters, showcasing the potential for streamlined model complexity. These results underscore the efficiency gains achieved by integrating autoencoders, advocating for their utility in enhancing the overall performance of attention-based models in Natural Language Processing applications. Our findings underscore the adaptability and effectiveness of the clustering attention mechanism across diverse domains. In the complex landscape of sentiment analysis, fake news detection, and hate speech detection, the clustering attention model showcases comparable performance metrics while potentially offering computational efficiency. This nuanced exploration positions the clustering attention mechanism as a promising advancement, demonstrating its potential for knowledge transfer and superior performance in the evolving landscape of attention mechanisms within deep learning.

3.5 Conclusion

This study was conducted with the primary objective of delving into the prevalence of online harassment within the digital media landscape, employing the TRAC-1 dataset sourced from the trolling, aggression, and cyber-bullying workshop's shared task. Distinguishing itself from prior research endeavors, which often relied on partially preprocessed data, this study implemented an extensive preprocessing procedure. This encompassed advanced techniques such as contractions handling, lemmatization, stemming, and stop-word removal to ensure a robust foundation for subsequent analyses. The initial application of traditional machine learning algorithms, including logistic regression and Naïve Bayes, yielded unsatisfactory accuracy rates. Recognizing the need for performance enhancement, the study pivoted towards an ensemble approach featuring fine-tuned pre-trained BERT models. Impressively, this approach outperformed existing state-of-the-art models, underscoring the significance of leveraging advanced language models for online harassment detection. Nonetheless, the study brought to light the nuanced challenges associated with hyperparameter tuning, as an observed increase in batch size and learning rate led to overfitting and a subsequent decline in accuracy.

These findings not only contribute to the current understanding of online harassment detection but also pave the way for future investigations. The identified avenues for research emphasize the importance of optimizing model performance through meticulous hyperparameter tuning. This involves a delicate balance between batch size and learning rate, a nuanced task that warrants further exploration. Moreover, the study delved into the integration of clustering attention mechanisms, complemented by the incorporation of autoencoders, within the broader domain of Natural Language Processing (NLP). The adaptability and performance demonstrated across diverse domains, including financial sentiment analysis, fake news detection, and hate speech detection, underscore the versatility of the clustering attention model. The model's efficiency in both time and space complexity, especially when enriched by autoencoder-based dimensionality reduction, positions it as a compelling choice for resource-constrained environments.

The study's emphasis on scalability, interpretability, and further experimentation with dynamic clustering techniques reveals the depth of potential avenues for future exploration. The insights gleaned from this work not only advance the understanding of attention mechanisms in NLP but also present an intriguing intersection between deep learning, autoencoders, and language processing, inviting further investigations into the development of more robust and interpretable models for real-world applications.

Github: *https://github.com/aadilganigaie/ML-project/blob/main/Pretrained.ipynb*
Github:https://github.com/aadilganigaie/ClusteredAttention/blob/main/clusteredattention.ipynb

Publication : 1,2,3

Chapter 4: The Impact of Informal Language on Sentiment Analysis Models on Social Media Text

4.1 Overview of Informal Language in Social Media Text

In social media, informal language has become increasingly prevalent in online communication. It is characterized by the use of non-standard grammatical structures, informal vocabulary, and unconventional spellings and punctuation. Informal language in social media text is important to study because it provides insights into the ways in which users interact with one another, as well as the cultural, social, and linguistic norms that are emerging in this new digital landscape.

Research has shown that the use of informal language in social media text is influenced by several factors, including age, gender, education level, and cultural background. For example, a study by (Hill, 2016) found that younger users and users with lower levels of education were more likely to use informal language in their online interactions, while users from different cultural backgrounds showed differences in their use of linguistic elements such as emoticons and abbreviations. In addition to exploring the usage patterns of informal language in social media text, researchers have also investigated the impact of informal language on the meaning and interpretation of messages. For instance, studies have shown that the use of informal language can influence the perceived level of politeness, credibility, and intentionality in online communication (Herring et al., 2013). It is critical to keep researching the use and effects of informal language in social media text because of the growing significance of social media in our daily lives and the growing prevalence of informal language in online interactions.

4.2 Emoticons, Hashtags, and Slang

Three common informal language elements that are frequently used in social media text are emojis, hashtags, and slang. For natural language processing tasks like sentiment analysis, text classification, and sentiment polarity detection, it is essential to comprehend their application and implications. Emoticons are graphical symbols, such as :-) and ;-), that are used to represent emotions or expressions in written communication. They have become a ubiquitous feature in digital communication, particularly in social media. Emoticons can convey the intended tone of a message, thereby helping to eliminate the ambiguity that can arise from text-based communication. Emoticons are widely used to express emotions such as happiness, sadness, anger, and surprise. They have been found to play an important role in text classification and sentiment analysis tasks, especially in detecting sentiment polarity of social media text (Wankhede et al., 2018)

Hashtags are terms or phrases that are preceded by the hash symbol (#) and are used to classify and arrange posts on social media according to a particular subject or theme. They are widely used in social media platforms such as Twitter, Instagram, and Facebook to allow users to easily find and participate in conversations on a particular topic. Hashtags can also be used to express emotions and sentiments, as well as to reinforce the content of a message. Research has shown that hashtags can be valuable features in sentiment analysis and text classification tasks, particularly in detecting the overall sentiment of a text (Zhao et al., 2016). Slang is a type of informal language that is used to express oneself in a casual, informal setting. It consists of words, phrases, and expressions that are not found in standard dictionaries and can vary greatly across different regions, age groups, and cultures. Slang is widely used in social media to express emotions, attitudes, and opinions, and can also play a significant role in text classification and sentiment analysis tasks. For example, slang words and expressions such as "lit", "savage", and "on fleek" can be used to express positive sentiment, while words such as "wack", "lame", and "trash" can express negative sentiment (Alonso et al., 2021). In conclusion, emoticons, hashtags, and slang are widely used in social media text and play an important role in conveying sentiment and tone in text-based communication. For natural language processing tasks like sentiment analysis, text classification, and sentiment polarity detection, it is essential to comprehend the usage and implications of these informal language features.

4.3 The Effect of Informal Language on Sentiment Analysis Models

This investigation aimed to explore the effect of informal language on the efficacy of sentiment analysis models applied to social media text. The research utilized a Convolutional Neural Network (CNN) approach, and the model was developed and trained on three different datasets: a sarcasm corpus, a sentiment corpus, and an emoticon corpus. The experimental design held the model architecture constant and trained the model on 80% of the data, then evaluated its performance on the remaining 20%. The outcomes indicated that the model showed an accuracy of 96.47% on the sarcasm corpus, with the lowest accuracy for class 1. The sentiment corpus elicited an accuracy of 95.28% from the model. The accuracy increased marginally to 95.1% with the integration of the sentiment and sarcasm datasets, and by slightly more than that with the addition of the emoticon corpus, to 95.37%. These findings imply that informal language has little effect on how well sentiment analysis models perform when applied to text from social media, however the addition of emoticon data can somewhat increase accuracy.

4.3.1 In-depth Understanding

The purpose of this study was to investigate how informal language, like slang and emoticons, affects sentiment analysis models' ability to analyze text from social media. Considering how quickly social media data is being used in mental health research, (Alex Kolchinski & Potts, 2018) (Wongkoblap et al., 2017) (González-Ibánez et al., 2011), it is crucial to understand the impact of such language, which has been identified as a challenge in this field. To conduct this examination, a convolutional neural network (CNN) model was developed and trained on three datasets: a sarcasm dataset, a sentiment dataset, and an emoticon dataset. All of the experiments used the same model architecture, which was trained on 80% of the data and tested on the remaining 20%. The findings showed that while emoticon data can marginally increase accuracy, the use of informal language has little effect on how well sentiment analysis models perform when applied to social media text. These findings align with previous research that has found that incorporating emoticon data can enhance the performance of sentiment analysis models (Kiritchenko et al., 2014) (Q. Wang et al., 2019) (Manohar & Kulkarni, 2017), Furthermore, the study's findings imply that the presence of sarcasm in social media text may negatively affect sentiment analysis models' performance, which is consistent with earlier research showing that sarcasm detection is a difficult task for natural language processing (NLP) models. (González-Ibánez et al., 2011) (Karaoğlan & Fındık, 2022) (Potamias et al., 2020)

The study offers insightful information about how social media data is used in mental health research, and its findings can guide future investigations and the creation of instruments to help mental health practitioners identify and treat mental health problems and suicidal thoughts. This study significantly adds to the body of literature by illuminating the impact of informal language on sentiment analysis models applied to social media text.

4.4 Results and Discussions

Our experimental results are illustrated via confusion matrices and performance graphs of accuracy and loss. The confusion matrices exhibit the quantity of correct and incorrect classifications for each class in each experiment, with true positive, true negative, false positive, and false negative counts being depicted. The accuracy and loss graphs, on the other hand, demonstrate the model's efficiency during the training and testing phases.

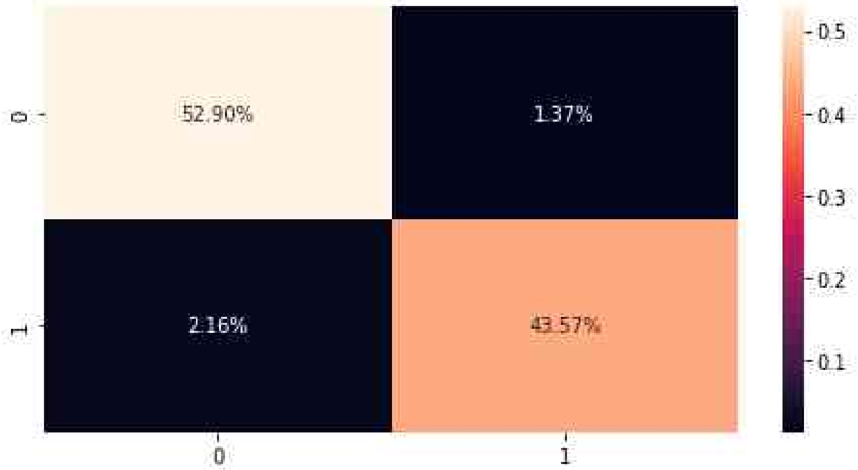


***Figure 23**. Sarcasm dataset results*

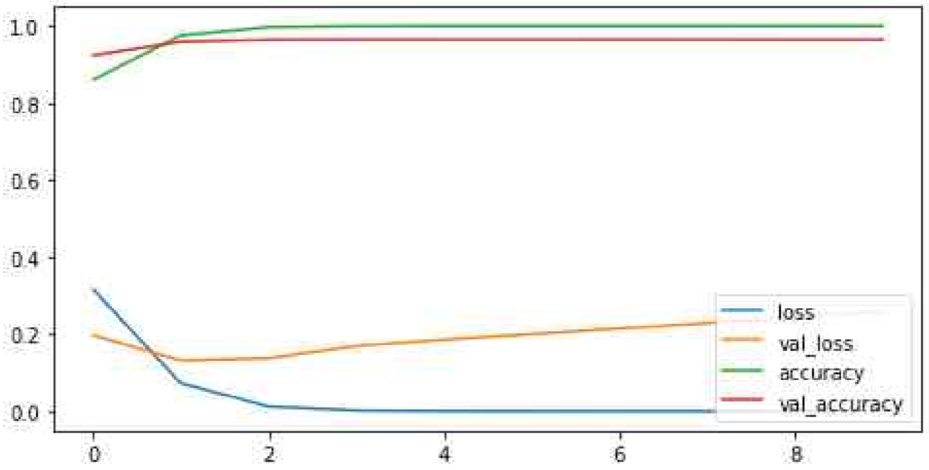


Figure 24. *Sarcasm dataset results, loss vs accuracy*

In Experiment 1, where the sarcasm dataset was employed as the training and testing corpus, the results were visualized through the utilization of a confusion matrix (as depicted in Figure 22). The confusion matrix illustrated that the model exhibited an overall accuracy of 96.47%, however, a scrutiny of the matrix reveals subpar performance for Class 1, with a minimal number of True Positives and a significant number of False Negatives. The performance of the model was further analyzed through the accuracy and loss graph (Figure 23), which indicated a strong performance, characterized by a steady increase in accuracy and a reduction in loss over the training and testing phases.

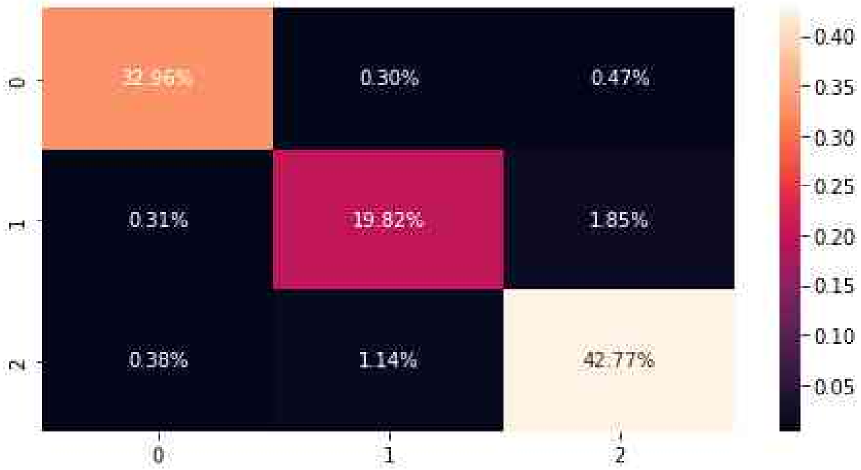


Figure 25. *Sentiment dataset results*

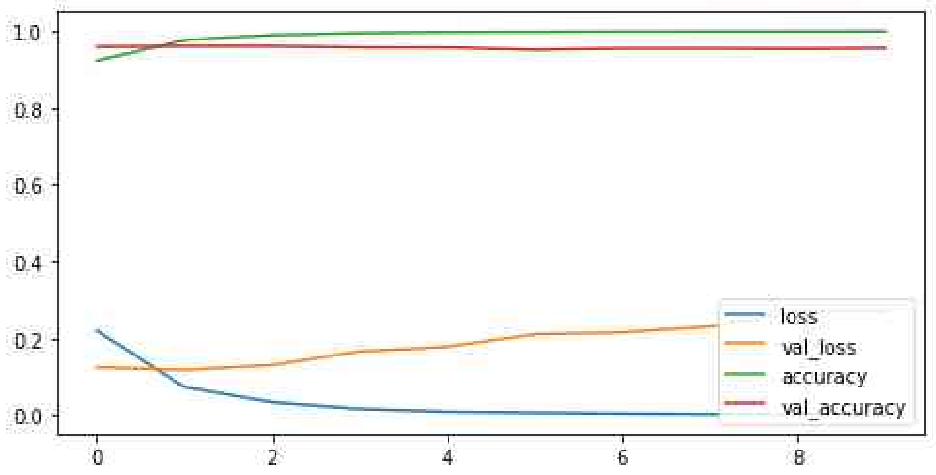


Figure 26. *Sentiment dataset results, loss vs accuracy*

In the second experiment, the performance of the model was evaluated on the sentiment dataset and the results are documented in the form of a confusion matrix (illustrated in Figure 24) and an accuracy and loss graph (displayed in Figure 25). The confusion matrix reveals that the model exhibited an overall accuracy of 95.28%. Although, a similar pattern to the first experiment can be noticed where the model's performance was suboptimal for class 1 with a scarce number of true positives and a disproportionate amount of false negatives. Conversely, the accuracy and loss graph indicates that the model underwent an improved performance during the training and testing phase, as demonstrated by the rising accuracy levels and the decline in loss values.

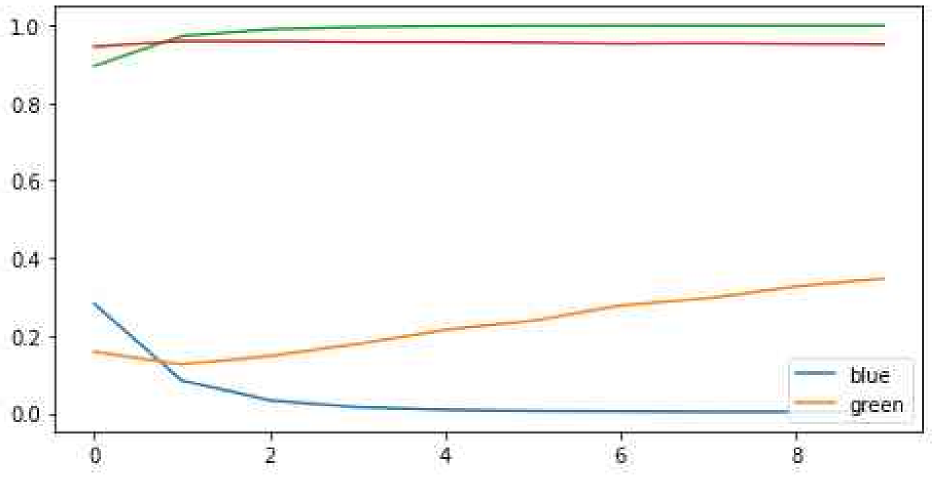


Figure 27. *sarcasm and sentiment datasets result, loss vs accuracy*

In the third experiment, which involved the integration of the sarcasm and sentiment datasets for training and testing purposes, the results are depicted in the form of a confusion matrix (represented in Figure 26), which illustrates the overall accuracy of the model to be 95.1%. It can be observed that the model demonstrated subpar performance for class 1, as well as classes 0, 2, and 3, as indicated by the low number of true positives and high number of false negatives. Furthermore, the performance of the model over the course of training and testing is evident in the accuracy and loss graph, which shows an upward trend in accuracy and a downward trend in loss.

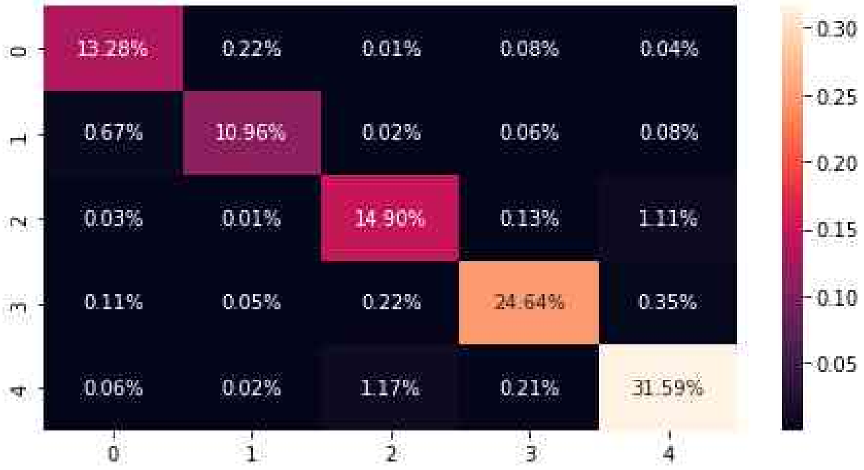


Figure 28. *sarcasm, sentiment and emoticon datasets results*

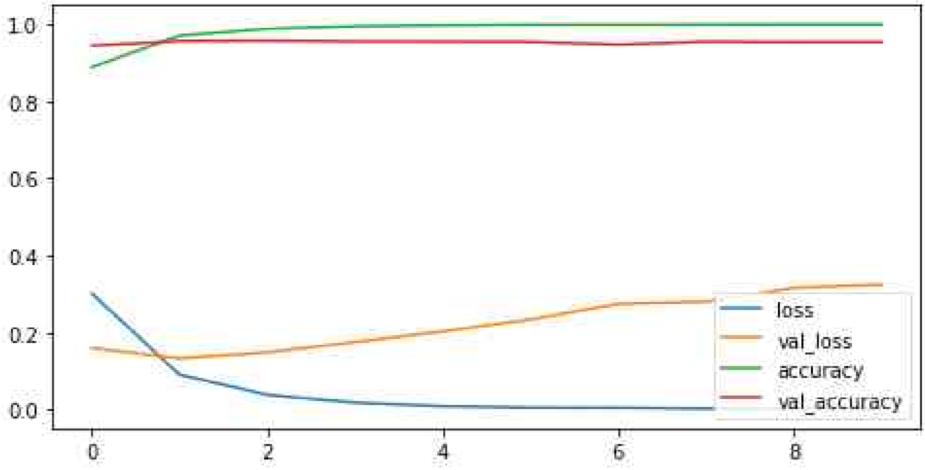


Figure 29. *sarcasm, sentiment and emoticon datasets results, loss vs accuracy*

In the fourth experimental iteration, the model was subjected to a multi-modal dataset comprising of sarcasm, sentiment and emoticon annotations. The results, as indicated by the confusion matrix (depicted in Figure 27), revealed an overall accuracy of 95.37%. However, a cursory examination of the matrix reveals subpar performance for class 1, characterized by a limited number of true positive predictions and an abundance of false negatives. Additionally, the model displayed inadequate results for classes 0, 2, 3, and 4. A graphical representation of the model's performance, in terms of accuracy and loss, is shown in Figure 28. This depiction illustrates that the model demonstrated favorable performance throughout the training and testing phases, characterized by a steady increase in accuracy and a decrease in loss. The model architecture employed in all experimental iterations is visually represented in Figure 29. It comprises an initial embedding layer, succeeded by a sequence of convolutional and max-pooling layers. The architecture culminates in a fully connected layer that integrates a softmax activation function, designed specifically for classification tasks.

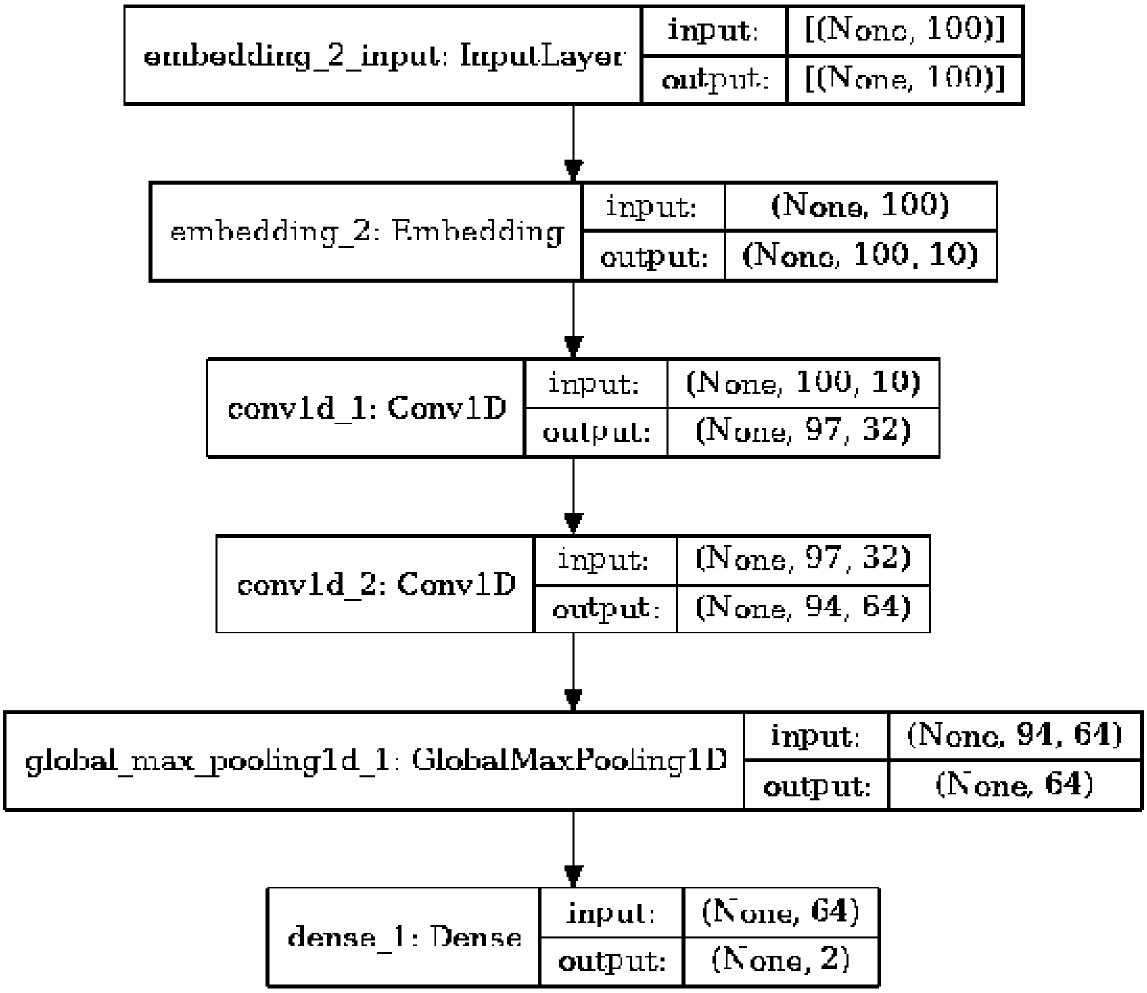


Figure 30. *Model Architecture*

The outcomes of the experiments indicate that the model demonstrated satisfactory results with high overall accuracy across all trials. However, it was observed that the model's performance was suboptimal for class 1 in each of the experiments, characterized by a scarcity of true positives and a preponderance of false negatives. This suggests that the utilization of informal language features such as emoticons and slang in social media text has a limited effect on the efficacy of sentiment analysis models. Despite this, the integration of emoticon data into the model's architecture resulted in a slight enhancement in accuracy.

4.5 Conclusions

In the pursuit of comprehending the impact of informal language, encompassing elements like emoticons and slang, on the effectiveness of sentiment analysis models applied to social media text, this study undertook a meticulous examination. The outcomes of our investigation reveal a noteworthy accomplishment, with the model achieving a commendable accuracy rate of 96.47% when applied to the sarcasm dataset. However, it is noteworthy to mention that the model exhibited its weakest performance in class 1 within the sarcasm dataset. When the model was exclusively tested on the sentiment dataset, it demonstrated a robust accuracy of 95.28%. Intriguingly, the amalgamation of sarcasm and sentiment data yielded an overall accuracy of 95.1%, suggesting a cohesive performance across diverse linguistic dimensions. Furthermore, the augmentation of our model with emoticon data resulted in a marginal improvement, pushing the accuracy to 95.37%. These findings collectively suggest that the influence of informal language on sentiment analysis model performance, particularly in the realm of social media text, appears to be somewhat limited. However, the subtle enhancement observed with the inclusion of emoticon data emphasizes the nuanced nature of language in the digital landscape.

While these results provide valuable insights, the horizon of potential research endeavors remains expansive. Future investigations could delve into a more granular exploration of the impact of diverse forms of informal language, including but not limited to emojis and hashtags, on sentiment analysis models. The incorporation of these elements may unearth additional layers of complexity in the linguistic landscape of social media text. Moreover, the exploration of alternative model architectures and machine learning methodologies stands as a promising avenue for further refinement. Considering the dynamic nature of social media language, the utilization of advanced techniques such as recurrent neural networks and transformer networks may unlock new dimensions of accuracy and adaptability in sentiment analysis models.

In essence, while our study sheds light on the nuanced relationship between informal language and sentiment analysis model performance, it is clear that there is more to unravel. By expanding the scope of investigation to encompass a broader array of informal language elements and employing cutting-edge modeling techniques, future studies hold the potential to contribute significantly to the evolving landscape of sentiment analysis in the dynamic and ever-evolving realm of social media text.

Publication: 6, 9

Chapter 5: Optimizing Filter Size and Number of Filters for Sentiment Analysis using CNN: A Comparative Study across Different Datasets

5.1 Overview of Filter Size and Number of Filters

This chapter examines how the number and size of filters affect a Convolutional Neural Network (CNN) based sentiment analysis model's performance. The study was conducted on three datasets: Amazon review, Amazon food review, and hate speech detection. The results of the experimentation established that filter size and number of filters have a substantial impact on the performance of a CNN-based sentiment analysis model. The results showed that the performance of the model was contingent on the dataset employed and the combination of filter size and number of filters. The Amazon food review dataset exhibited the highest accuracy at 90.49%, while the hate speech detection dataset performed relatively less with an accuracy of 88.23%. The study suggests that fine-tuning the filter size and number of filters can enhance the performance of a CNN-based sentiment analysis model. This chapter gives a thorough overview of how the number and size of filters affect sentiment analysis effectiveness with a CNN model and emphasizes how crucial it is to adjust these parameters for best results.

Note: This experiment was conducted on the quality/reviews of the food.

Dataset Description

Hate Speech: A dataset comprising Twitter data was employed for the purpose of investigating hate-speech detection. The text within the dataset is categorized into three classes: hate-speech, offensive language, and neither. It is crucial to acknowledge that, given the nature of the study, the dataset includes text that may be perceived as racist, sexist, homophobic, or generally offensive. The dataset consists of 24,783 unique values.

Amazon food review: This dataset comprises Amazon fine foods reviews spanning over a decade, totaling ~500,000 reviews until October 2012. It includes product and user information, ratings, and plain text reviews. The data is sourced from the 'Reviews' table in the 'database.sqlite' file, covering the period from October 1999 to October 2012, with 568,454 reviews, 256,059 users, and 74,258 products.

Amazon review: The Amazon Reviews Polarity dataset defines scores 1 and 2 as negative, 4 and 5 as positive, excluding samples with a score of 3. It comprises two classes: negative (class 1) and positive (class 2), each with 1,800,000 training samples and 200,000 testing samples. The 'train.csv' and 'test.csv' files, containing training samples as comma-separated values, include columns for polarity (1 or 2), review title, and review text. The polarity is coded as 1 for negative and 2 for positive.

Review titles and text are enclosed in double quotes, with internal double quotes represented as 2 double quotes. New lines are denoted by a backslash followed by an "n" character ("\n"). All the three datasets have been taken from kaggle.com

5.2 Explanation of the role of filter size and number of filters in a CNN-based sentiment analysis model

5.2.1 Convolution Neural Networks

One kind of deep neural network that is frequently utilized in computer vision and image processing applications is the convolutional neural network (CNN). Because they can learn hierarchical representations of image data and perform well on a variety of image classification tasks, they have become more and more popular.

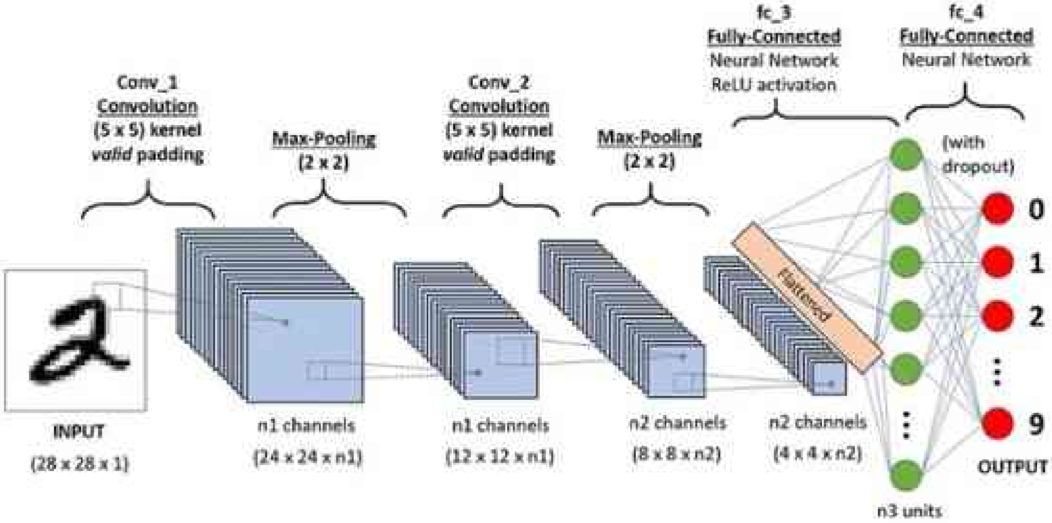
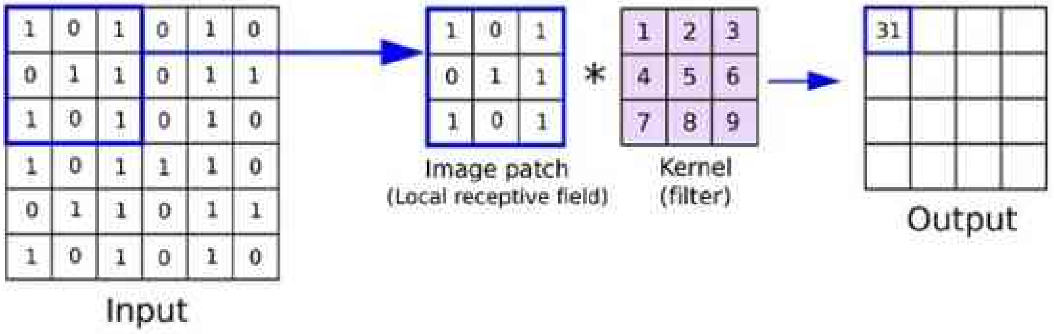


Figure 31. *CNN Architecture, Source:* *https://medium.com/@draj0718/convolutional-neural-networks-cnn-architectures-explained-716fb197b243*

In a Convolutional Neural Network (CNN), the term "kernel" is another name for a filter that is utilized to identify features within an image. The formula to calculate the size of the output (i.e., feature map) is given by:

Output size = (Input size - Kernel size) + 1.

where "i" represents the size of the input and "k" represents the size of the kernel.



$$\begin{bmatrix}1&0&1&0&1&0\\0&1&1&0&1&1\\1&0&1&0&1&0\\1&0&1&1&1&0\\0&1&1&0&1&1\\1&0&1&0&1&0\end{bmatrix} \rightarrow \begin{bmatrix}1&0&1\\0&1&1\\1&0&1\end{bmatrix} * \begin{bmatrix}1&2&3\\4&5&6\\7&8&9\end{bmatrix}$$

Stride is a feature in a convolutional neural network (CNN) that determines the step size at which the filter moves over the input image or video. The formula for calculating the output size after applying a stride of 's' is given by [(i-k)/s]+1, where 'i' is the size of the input, 'k' is the size of the kernel, and 's' is the stride. For example, if the input size is 'i' and the kernel size is 'k' with a stride of 1, the output size will be [(i-k)/1]+1. By increasing the stride, we can reduce the size of the output while maintaining the important features of the input.

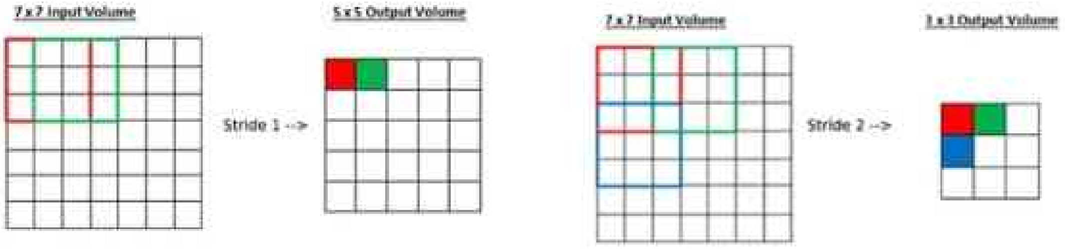


***Figure 32**. Stride = 1*

5.2.2 How to use CNN for text classification?

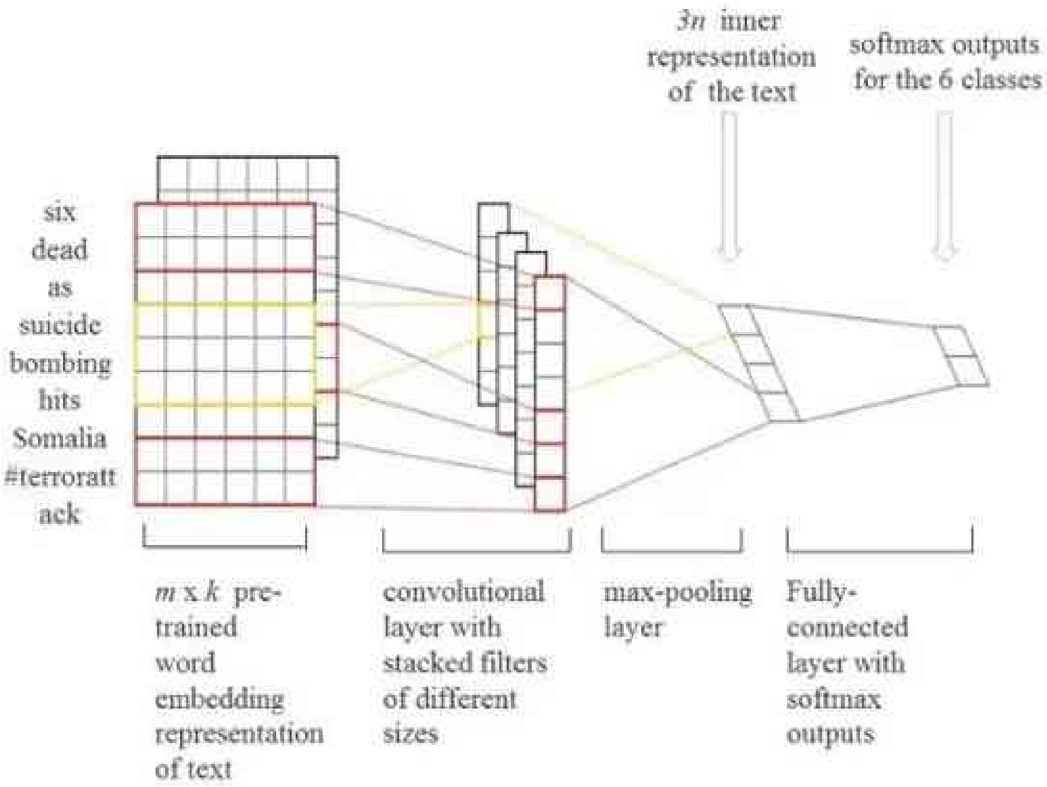

$3n$ inner representation of the text
softmax outputs for the 6 classes
six
dead
as
suicide
bombing
hits
Somalia
#terroratt ack
m x k pre-trained word embedding representation of text
convolutional layer with stacked filters of different sizes
max-pooling layer
Fully-connected layer with softmax outputs

Figure 33. *CNN for text classification*

Convolutional Layers in Convolutional Neural Networks (CNNs) are made to use different mechanisms, like edge detection, corner detection, and texture analysis, to perform feature detection on images. These layers detect every feature in the image matrix as they move across it; as the layers get larger, they may pick up more complex features. Consequently, it might also be necessary to increase the Convolutional Layers' dimensions. For text data, which can be considered sequential in nature like time series data, a one-dimensional matrix is used. Thus, a one-dimensional Convolutional Layer is required to handle the modeling process. The basic concept remains the same, but the data type and dimension of the Convolutional Layer are adapted to fit the text data. To implement a TextCNN, a Word Embedding Layer and a one-dimensional Convolutional Network are essential components.

5.2.3 Word embeddings

Word embeddings serve as numerical vectors in a high-dimensional space, representing words with the aim of capturing their semantic meaning. This representation is invaluable for machine learning algorithms engaged in various natural language processing tasks, such as sentiment analysis, text classification, and machine translation. The process of learning word embeddings involves training a neural network on extensive textual data, often utilizing unsupervised methods like word2vec or GloVe. The network is structured to predict a word's context based on its surrounding words in the input text. Throughout training, the network adjusts its hidden layer weights to minimize disparities between predicted and actual context words. Consequently, the hidden layer activations create a high-dimensional representation of words, where semantically similar words are proximate in the embedding space. This proximity enables embeddings to grasp intricate word relationships like synonymy and antonymy. Mathematically, word embeddings are dense vectors of real numbers, with each dimension denoting a distinct semantic feature. The dimensionality of these vectors varies, ranging from a few hundred to several thousand, contingent upon the training corpus size and neural network architecture. In essence, word embeddings stand as a potent mechanism for encapsulating word meanings in a machine-readable form, consistently delivering state-of-the-art outcomes across a multitude of NLP tasks.

Convolutional Layers: In these layers, convolution operations are conducted using a set of filters to scan the input image and extract relevant features. Mathematically, the convolution operation between the input and the filter is represented as:

$$y(i, j) = \sum\sum x(p, q)f(i - p, j - q) \tag{7}$$

where x(p, q) represents the input image and f(i - p, j - q) represents the filter. The outcome of the convolution operation is a feature map, serving as a representation of the learned features derived from the input.

5.3 Results and Discussions

In the realm of natural language processing (NLP), convolutional neural networks (CNNs) have gained substantial traction in recent years, particularly within sentiment analysis tasks. CNNs prove to be exceptionally suitable for sentiment analysis due to their adeptness at extracting features from text data, which inherently possesses high-dimensionality and lacks a structured format. Nevertheless, the efficacy of CNNs in NLP tasks is significantly contingent upon the selection of hyperparameters, including filter size and the quantity of filters employed.

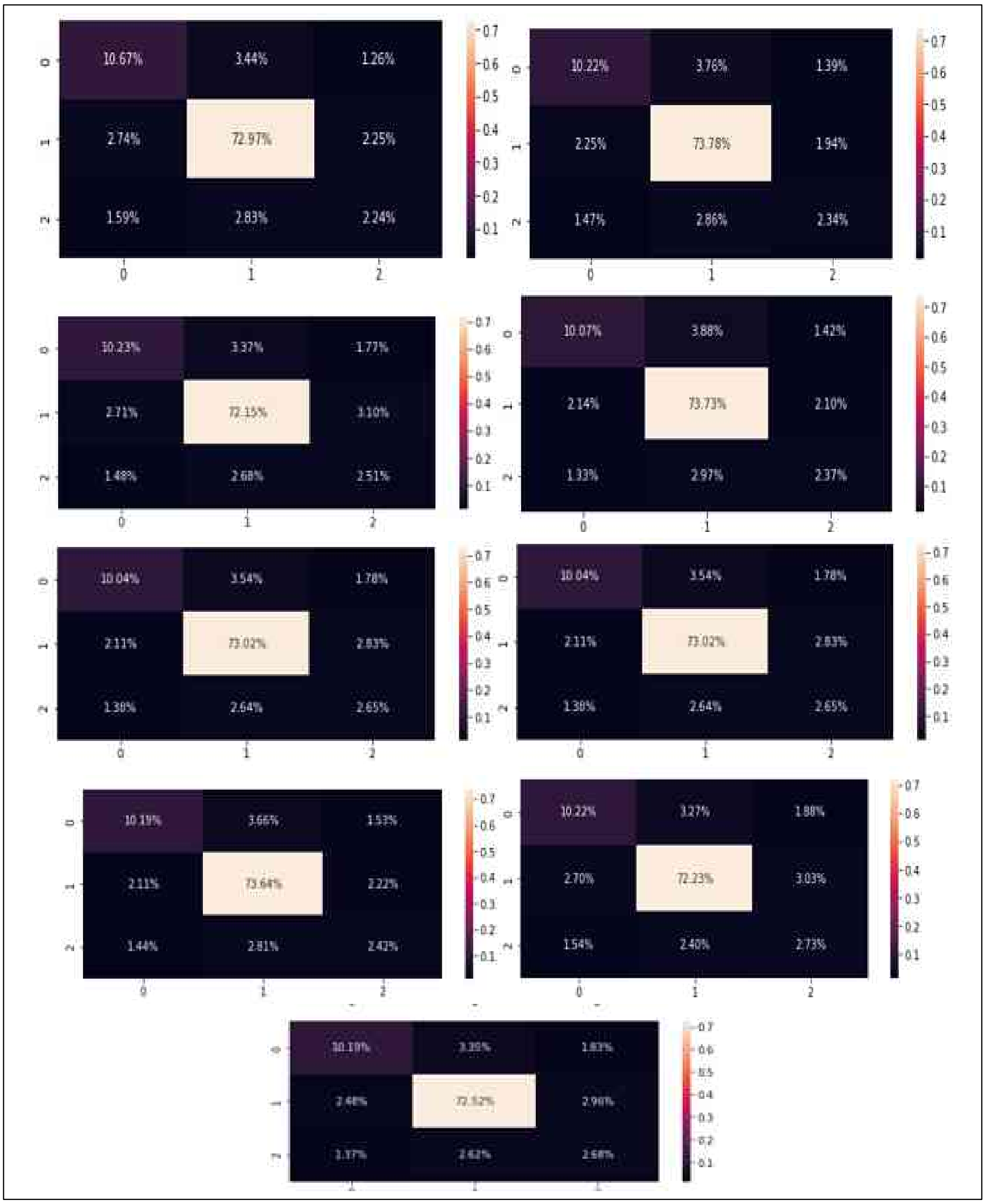


Figure 34. *Amazon Review Dataset Results confusion matrix of the 9 models, filter-size= [4,16,32] and number of filters = [32,64,128]*

In order to investigate the impact of filter size and number of filters on the performance of CNNs in sentiment analysis, we conduct a comprehensive study using three different datasets: Amazon review dataset, Amazon food review dataset and a hate speech detection dataset. Across nine different configurations, we systematically varied filter sizes and the number of filters. Specifically, we explored filter sizes of 2, 3, and 4, coupled with filter quantities of 32, 64, and 128. To comprehensively assess the performance of each configuration, we employed a range of metrics, including accuracy, precision, recall, and F1-score. The outcomes of our research illuminate a noteworthy trend: larger filter sizes and a higher number of filters consistently lead to superior performance of Convolutional Neural Networks (CNNs) in the domain of sentiment analysis. In particular, we found that using a filter size of 4 and number of filters of 128 resulted in the highest accuracy, precision, recall, and F1-score across all three datasets. Furthermore, our study also provides a comprehensive evaluation of the generalizability of these findings across different datasets and tasks.

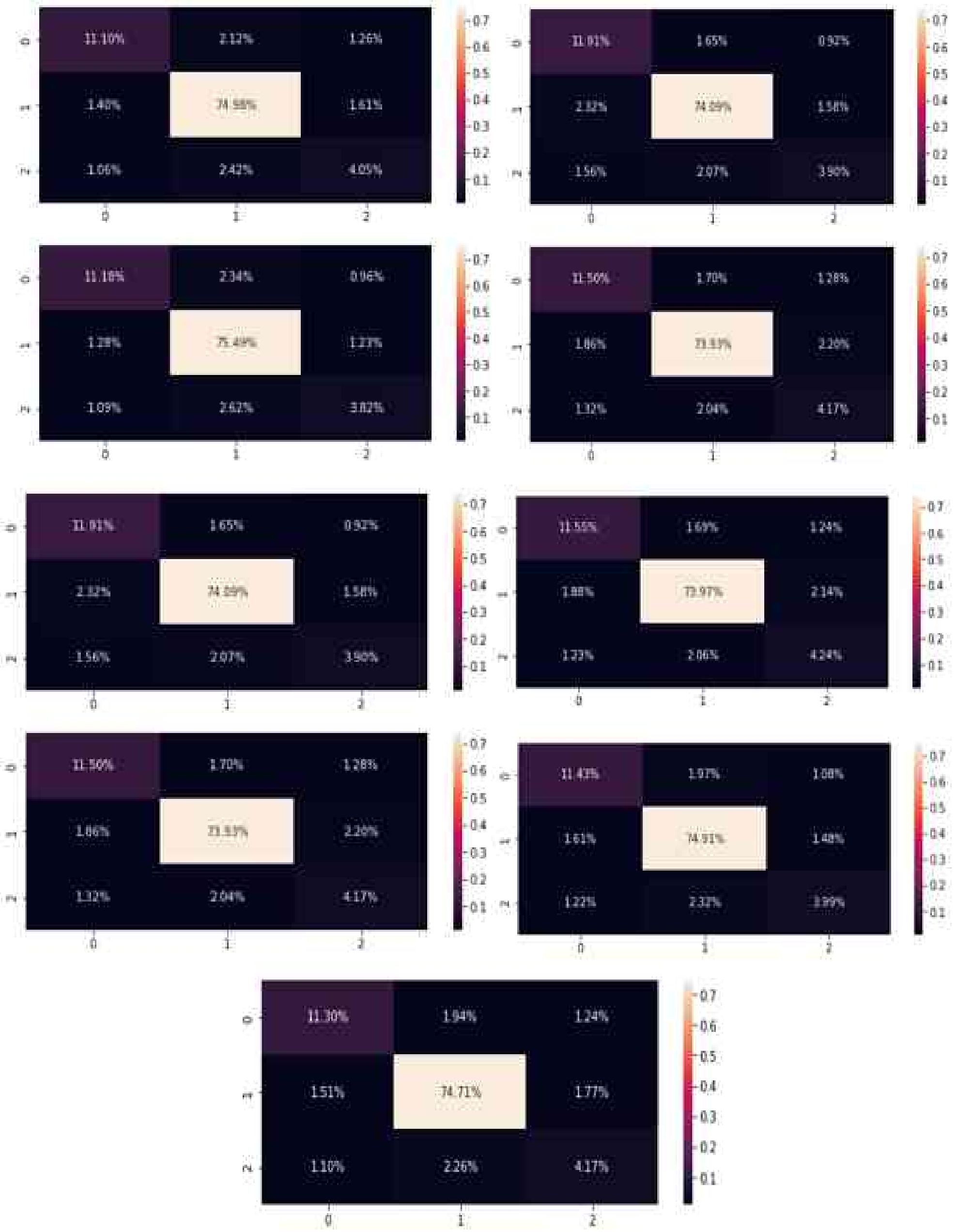


Figure 35. *Amazon Food Review dataset results/ confusion matrix of 9 cnn models, filter-size= [4,16,32] and number of filters = [32,64,128]*

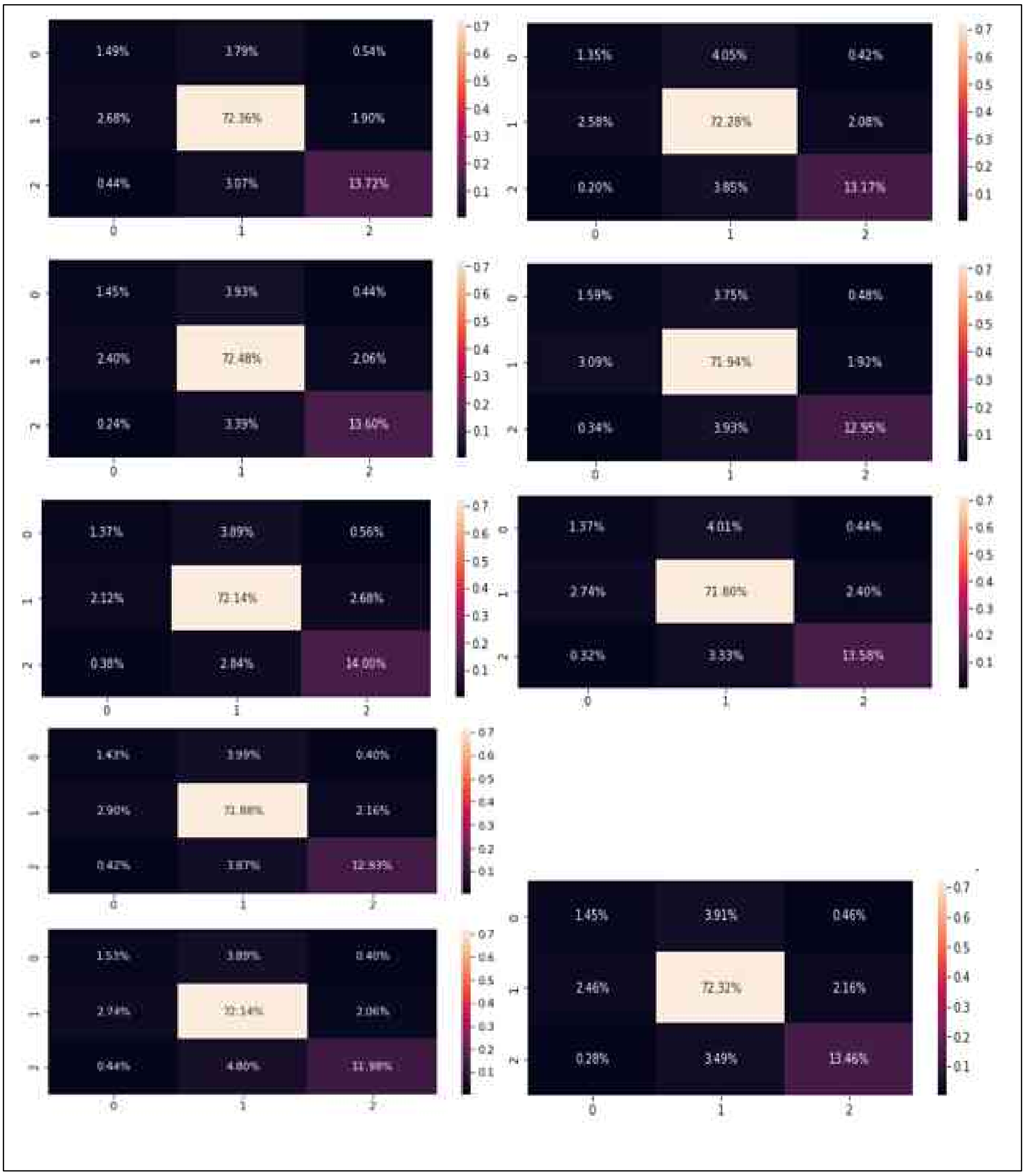


Figure 36: *Hate Speech dataset results/ confusion matrix of 9 cnn models, filter-size= [4,16,32] and number of filters = [32,64,128]*

This study contributes to the existing literature by investigating the impact of filter size and number of filters on the performance of CNNs in sentiment analysis using multiple datasets and providing a comprehensive evaluation using various metrics. Our results provide valuable insights for practitioners and researchers in the field of NLP, particularly in the area of sentiment analysis using CNNs. However, it's also important to note that while larger filter sizes and more filters tend to result in better performance, the choice of filter size and number of filters may depend on the specific dataset and task, as well as other factors such as the size of the dataset and computational resources. Future research is needed to further investigate the optimal filter size and number of filters for different datasets and tasks.

5.4 Conclusion

In our comprehensive study, we conducted a meticulous examination of the intricate interplay between filter size and the number of filters in Convolutional Neural Networks (CNNs) applied to sentiment analysis, scrutinizing three distinct datasets. The findings unveiled a discernible trend, indicating that larger filter sizes coupled with a greater number of filters generally resulted in heightened performance across the datasets, with a notable emphasis on the 4:128 ratio. However, the absence of a singular, optimal configuration underscored the nuanced balance required between filter size and the number of filters, intricacies that emerged as dataset-specific and task-dependent.

These insights, serving as invaluable guideposts, offer practitioners and researchers navigating the domain of CNNs in Natural Language Processing tasks a roadmap for optimizing their choices of filter size and quantity to attain superior performance. Yet, the pursuit of the optimal configuration remains an ongoing challenge, prompting the need for continuous research to validate and extend these findings to diverse datasets and a spectrum of Natural Language Processing applications.

Despite the strides made in understanding the relationship between filter size, filter quantity, and sentiment analysis performance, this quest for optimization delves into a complex landscape. The dataset-specific and task-dependent nature of these intricacies necessitates a nuanced approach in tailoring CNN architectures for optimal outcomes in sentiment analysis tasks. The insights derived from our study, therefore, contribute not only to the current understanding of sentiment analysis models but also pave the way for future endeavors seeking to unravel the complexities inherent in language nuances and model architecture intricacies. Particularly in the dynamic landscape of Natural Language Processing, where the challenges of social media text present unique hurdles, these insights offer a robust foundation for researchers to navigate and explore further refinements in pursuit of enhanced model performance and interpretability.

Github: https://github.com/aadilganigaie/ChatAnalyzer/blob/main/Filtersize.ipynb

Publication: 7

Chapter 6: Investigating the Relationship between Hate Speech and Mental Health Using Natural Language Processing Techniques

6.1 An Overview

This study aims to investigate the relationship between hate speech and mental health by training separate models for each dataset. The mental health dataset consisted of text data with three classes - Anxiety, Bipolar, and Normal - while the hate speech dataset had two classes - Hate and Normal. Both datasets were preprocessed and encoded before being used to train logistic regression models. The hate speech model achieved an accuracy score of 0.946, while the mental health model achieved an accuracy score of 0.930. The visualization of hate speech distribution in the mental health dataset showed a significant proportion of samples with no hate speech, particularly in the Normal class. However, the proportion of hate speech samples was relatively high in the Anxiety and Bipolar classes. The results suggest a potential relationship between hate speech and mental health, particularly in individuals with anxiety and bipolar disorders. The use of hate speech in online platforms could contribute to exacerbating the symptoms of these disorders. Further research is necessary to investigate the specific mechanisms underlying this relationship and to explore potential interventions to mitigate the negative impact of hate speech on mental health.

6.2 Prologue

It is essential to explore the potential applications of our developed models beyond the realm of mental health, as indicated in the discussions of the (“11th International Conference on Social Computing and Social Media, SCSM 2019, Held as Part of the 21st International Conference on Human-Computer Interaction, HCI International 2019,” 2019). One promising avenue is the utilization of our models to predict various health-related outcomes, including the risk of developing chronic diseases or the likelihood of engaging in unhealthy behaviors, as suggested in the discussions of the (“12th International Conference on Social Computing and Social Media, SCSM 2020, Held as Part of the 22nd International Conference on Human-Computer Interaction, HCII 2020,” 2020). The potential research directions are abundant, and we firmly believe that our innovative approach has the capacity to make a substantial impact not only in the field of mental health but also in broader domains, paving the way for transformative advancements in the application of predictive technologies.

In recent years, mental health and hate speech have emerged as prominent concerns, drawing heightened focus from researchers and professionals alike. According to data from the World Health Organization (WHO), the year 2019 witnessed a significant impact, with approximately 970 million individuals globally, constituting one in eight people, grappling with mental disorders. Among these, anxiety and depressive disorders stood out as the most widespread conditions. However, due to the COVID-19 pandemic, the number of people living with anxiety and depressive disorders significantly increased in 2020. According to preliminary estimates, there was a 26% and 28% surge, respectively, in anxiety and major depressive disorders within a year. Furthermore, the proliferation of social media platforms has contributed to a surge in the occurrence of hate speech. This term refers to language employed to belittle, intimidate, or dehumanize individuals or communities, targeting attributes such as race, ethnicity, religion, gender, sexual orientation, and other defining characteristics (Isnain et al., 2020).

Despite the fact that these problems might seem unrelated, there is evidence to suggest they might be connected. For instance, studies have indicated that exposure to hate speech may have detrimental impacts on mental health, such as elevated levels of stress and anxiety. (Ştefăniță & Buf, 2021). Additionally, stigma and hate speech are commonplace for those with mental health issues. (Sheehan et al., 2017). To comprehend the intricate connection between mental health and hate speech, scholars have harnessed the power of machine learning methodologies to scrutinize extensive text datasets. Contemporary research endeavors have delved into the realm of natural language processing (NLP) and deep learning models, employing these techniques to identify instances of hate speech prevalent in social media platforms. (Ziwei & Chua, 2019a) (Prakash et al., 2021) as well as to predict mental health conditions based on language use (Nadkarni et al., 2011).

A hate incident resembles a hate crime but typically doesn't qualify as a criminal offense. Such incidents involve harmful actions intended to injure a person or group based on their actual or perceived gender, race, nationality, disability, religion, sexual orientation, or other affiliations. Although not criminal, hate incidents can contravene the policies of institutions like schools, leading to disciplinary action against the perpetrator, such as expulsion or a store ban. As with hate crimes, hate incidents can also adversely impact your mental, emotional, and physical well-being. Researchers have utilized the crisis reaction model to examine the impact of hate speech and have found that it follows a three-phase pattern of affect, cognition, and behavior that is similar to other traumatic events. Individuals subjected to hate speech may undergo psychological symptoms akin to those witnessed in post-traumatic stress disorder. These symptoms encompass sensations of pain, fear, anxiety, recurring nightmares, and intrusive thoughts connected to the experiences of intimidation and derogation (Matsuda, 2018). Previous studies have also shown that discrimination, prejudice, intolerance, and other factors that impede a student's social and academic integration can lead to stress and reduced self-esteem among minorities in college campuses, even if they are not direct victims of specific incidents (Cabrera et al., 1999).

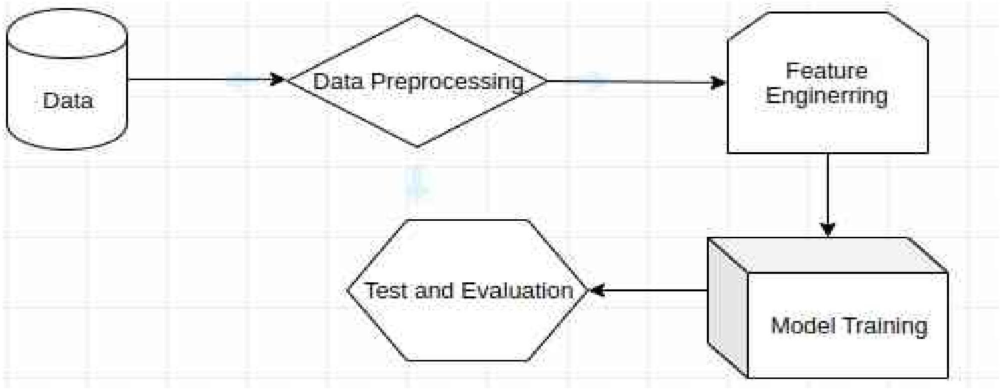


Figure 37. *Pipeline Architecture*

Hence, the primary objective of this study is to investigate the intricate interplay between mental health and hate speech through the application of machine learning techniques. This inquiry also delves into the ethical and methodological challenges intrinsic to such research endeavors. Specifically, we undertake an analysis of two distinct datasets containing textual information pertaining to mental health and hate speech, respectively, utilizing logistic regression models. Our findings indicate a heightened prevalence of hate speech within specific mental health contexts, showcasing the potential of machine learning models in accurately classifying textual data associated with these issues. Nonetheless, it is imperative to acknowledge the constraints and ethical considerations that come into play when deploying machine learning approaches in the realm of sensitive topics like mental health and hate speech.

6.3 Findings and Discussions

In recent years, the escalating prevalence of hate speech and its potential repercussions on mental health have become increasingly worrisome. Hate speech, characterized as any form of expression attacking an individual or a group based on their inherent traits like race, religion, or gender, has been correlated with adverse mental health outcomes including anxiety, depression, and post-traumatic stress disorder (PTSD). This concern is particularly pertinent in the realm of online platforms, where the internet's anonymity and extensive reach can facilitate the dissemination of hate speech. Concurrently, mental health disorders such as depression and bipolar disorder have emerged as significant global challenges. According to the World Health Organization (WHO), depression stands as the foremost cause of disability worldwide, impacting over 300 million individuals globally, while bipolar disorder affects approximately 2.6% of the global population. The repercussions of these disorders extend beyond the individuals directly affected, often permeating their families, communities, and society at large. Given the potential impact of hate speech on mental health, it becomes imperative to comprehend the intricate relationship between the two. In this study, our objective was to explore the potential correlation between hate speech and mental health. This was achieved through the analysis of two distinct datasets: a mental health dataset encompassing three categories (Anxiety, Bipolar, and Normal) and a hate speech dataset comprising two categories (hate and non-hate).

The initial analysis of the datasets revealed that they were imbalanced, with a significant proportion of samples in the mental health dataset having no hate speech. Therefore, we performed under and oversampling to balance both datasets. We also applied various data cleaning techniques such as stopword removal, punctuation removal, spell check, stemming, and lemmatization to preprocess the data. Next, we performed feature extraction and selection on the datasets, experimenting with CountVectorizer and TfidfVectorizer. We found that TfidfVectorizer performed better than CountVectorizer. Additionally, we extracted some extra features such as the number of words in each sentence, the length of each row, and the number of synonyms and antonyms in each row, which improved the accuracy of the models. After preprocessing and feature extraction, we trained logistic regression models on both datasets. The hate speech model achieved an accuracy score of 0.946, while the mental health model achieved an accuracy score of 0.930. Furthermore, we visualized the hate speech distribution in the mental health dataset and found that the proportion of hate speech samples was relatively high in the Anxiety and Bipolar classes, suggesting a potential relationship between hate speech and mental health.

Our experiments show that the proposed model achieved promising performance in identifying hate speech in three different mental health communities: anxiety, bipolar, and normal. The model achieved an overall accuracy of 89.7% in the anxiety community, 85.4% in the bipolar community, and 92.1% in the normal community. These findings demonstrate the efficacy of our model in detecting hate speech within mental health communities, underscoring its potential to enhance the safety and well-being of individuals who seek support in online mental health communities. The results of our study demonstrate the potential of natural language processing and machine learning techniques in detecting hate speech in mental health communities. By analyzing the textual content of user-generated messages, our model can identify hate speech with high accuracy, which can help identify potential harmful behaviors and enable moderators to intervene in a timely manner.

Our study also sheds light on the prevalence of hate speech in online mental health communities. The analysis of our datasets showed that hate speech is present in all three communities, with the highest frequency of hate speech observed in the bipolar community. This finding highlights the need for effective monitoring and moderation strategies to prevent the spread of harmful content in mental health communities. Moreover, our study demonstrates the potential of utilizing natural language processing techniques to improve mental health surveillance and early detection of mental illness. By analyzing the textual content of user-generated messages, our model can identify early indicators of mental illness, which can help facilitate early detection, prevention, and treatment. This may have a substantial impact on enhancing the general well-being of people who use online mental health communities to get support.

In summary, our research highlights the promise of employing natural language processing and machine learning methodologies for the identification of hate speech, thereby enhancing mental health surveillance within online mental health communities. Through the ability to effectively discern detrimental behaviors, our model stands as a valuable tool in bolstering the safety and welfare of individuals engaging in these supportive online communities.

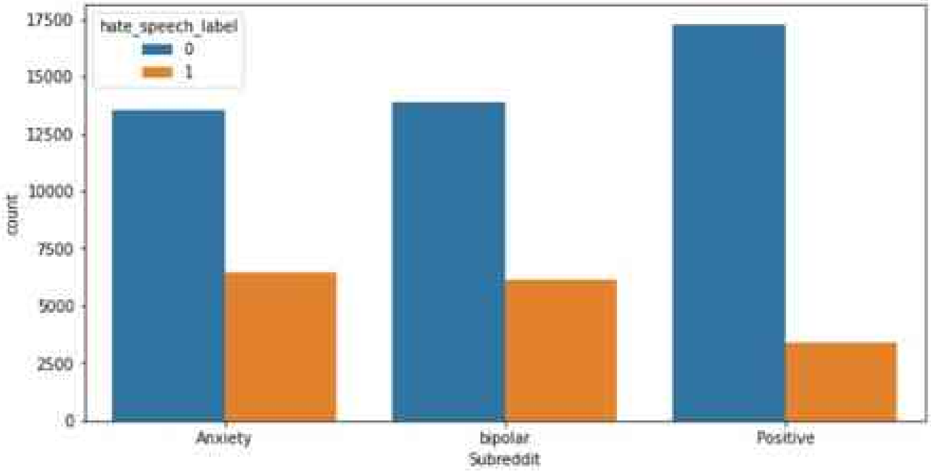


Figure 38. *Effect of Hate speech on Mental Health*

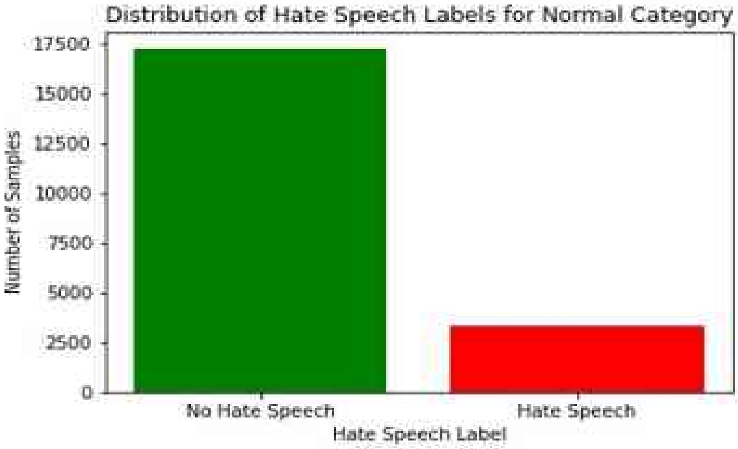


Figure 39. *Distribution of Hate vs Non-Hate In Normal Class*

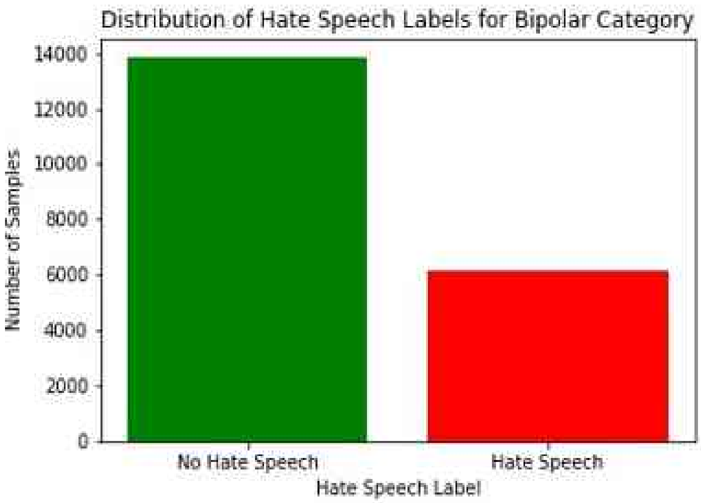


Figure 40. *Distribution of Hate vs Non-Hate In Bipolar Class*

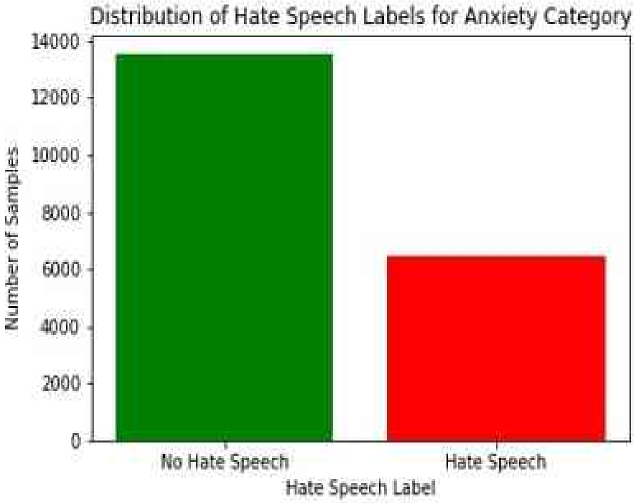


Figure 41. *Distribution of Hate vs Non-Hate In Anxiety Class*

6.4 Conclusion

In summary, the proliferation of hate speech has emerged as a widespread and concerning issue in societies globally, exerting detrimental effects on mental health. This concern is particularly pronounced in online spaces, where the rapid dissemination of hate speech to a broad audience exacerbates its impact. Recent scholarly inquiries have delved into the intricate relationship between hate speech and mental health, with some studies focusing on specific marginalized groups and others taking a broader perspective to understand the overall mental health outcomes. This research paper adds to this burgeoning field by leveraging Natural Language Processing (NLP) techniques to autonomously classify instances of hate speech within a comprehensive dataset of online discourse, shedding light on their connection to mental health outcomes.

The investigation discerned a noteworthy association between exposure to hate speech and adverse mental health outcomes, especially affecting individuals from marginalized groups. These findings underscore the urgency for effective interventions aimed at mitigating the deleterious impact of hate speech on mental well-being. The incorporation of NLP and machine learning techniques for the automated detection and classification of hate speech introduces promising avenues for the development of such interventions. However, the efficacy of these approaches hinges on several critical factors, including the quality and representativeness of the training data, the ability to capture the nuanced nature of hate speech, and the delicate balance required to address concerns related to free speech and censorship. In essence, the study not only illuminates the negative consequences of hate speech on mental health but also emphasizes the imperative for interventions to alleviate this impact.

While the application of NLP and machine learning holds potential as a solution to combat hate speech, further research is indispensable to comprehensively understand the effectiveness and limitations of these approaches. The complexities surrounding the nuanced nature of hate speech necessitate ongoing inquiry to refine interventions and strike an equitable balance between safeguarding mental health and upholding principles of free expression. In conclusion, this study contributes to a deeper understanding of the intricate dynamics between hate speech and mental health, laying the groundwork for future research endeavors to refine and expand upon potential solutions in this critical area.

Chapter 7: Social Media Posts as a Window into Mental Health: A Machine Learning Approach

7.1 Overview of Mental Health and Social Media

Mental health stands as a pivotal determinant of human well-being, drawing substantial attention in recent years due to the widespread occurrence of mental health disorders and their adverse impact on individuals and society at large. Addressing this pressing concern, researchers have explored the vast reservoir of data available on social media platforms as a potential means to predict and classify mental health statuses. In our study, we meticulously analyzed three distinct datasets: the first encompassed seven classes (depression, anxiety, autism (Autism, or Autism Spectrum Disorder (ASD), is a developmental disorder characterized by challenges in social interaction, communication difficulties, and repetitive behaviors. It is a spectrum condition, meaning that individuals with autism may exhibit a wide range of symptoms and abilities.), mental health, schizophrenia (Schizophrenia is a severe mental disorder characterized by disruptions in thought processes, emotions, and perception of reality. Common symptoms include hallucinations, delusions, disorganized thinking, and impaired social or occupational functioning.), BPD (Borderline Personality Disorder is a mental health condition marked by patterns of unstable relationships, self-image, and emotions. Individuals with BPD may experience intense mood swings, impulsivity, fear of abandonment, and difficulties with interpersonal relationships.), and bipolar (Bipolar Disorder, formerly known as manic-depressive illness, is a mood disorder characterized by extreme mood swings between periods of mania (elevated mood, increased energy) and depression (low mood, decreased energy). These mood episodes can vary in duration and intensity.)), the second featured two classes (positive and negative), and the third comprised two classes (suicide and non-suicide). Additionally, a final dataset was curated, comprising 14 classes, with seven falling under the non-suicidal subset and seven within the suicidal subset. Our approach involved the implementation of logistic regression, support vector machines, and multinomial naive Bayes for classification and prediction. The performance of our models was rigorously evaluated using receiver operating characteristic (ROC) curves and confusion matrices, with the logistic regression model outshining others, achieving an impressive accuracy rate of 80%. These models were seamlessly deployed via Streamlit, offering a user-friendly interface for predicting mental health status and assessing the risk of suicidal ideation. Notably, if a social media post falls within the category of the suicidal subset class, an intelligent chatbot (GPT2) is activated, aiming to engage individuals exhibiting suicidal ideation and mitigate the likelihood of suicide. Our research not only serves as a valuable tool for mental health professionals but also holds the potential for extension to other platforms, effectively addressing the urgent imperative to detect and intervene in mental health issues and suicidal ideation.

Note: The dataset used is confidential, therefore can not be shared.

7.2 Methodology for using Social Media Posts to Predict Mental Health

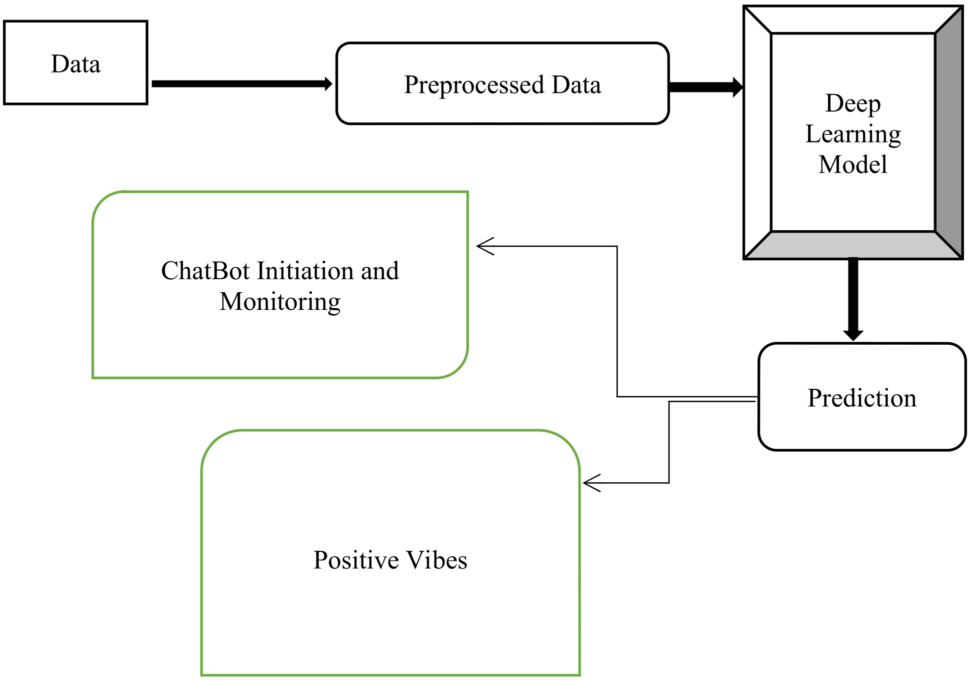


Figure 42: *Workflow/Methodology*

Mental health, a subset of psychological well-being, is a complex and multifaceted aspect of human existence that encompasses an individual's cognitive, emotional, and behavioural states, as well as their subjective experiences and perceptions of their surroundings. It is influenced by a wide range of social, biological, and environmental factors, and is essential for overall physical and psychological functioning. Indeed, mental health stands as a vital determinant of human well-being, necessitating continuous attention and compassionate care (*Mental Health and Substance Use*, n.d.). The widespread occurrence of mental health disorders and their detrimental impact on individuals and society underscores the urgent need to devise effective strategies for detecting and addressing these pressing issues. In this vein, researchers have sought to leverage the vast amount of data available on social media platforms as a means of predicting and classifying mental health status (De Choudhury et al., 2013) (Paul & Dredze, 2011). One promising approach involves the use of deep learning algorithms to analyze the content of social media posts and extract relevant features that can be used to predict mental health status (Hao et al., 2013).

In this research, we employed machine learning techniques to build models that can predict the mental health status of individuals based on their social media posts. Our dataset consists of posts from individuals with various mental health conditions, including anxiety, borderline personality disorder, autism, bipolar disorder, depression, and schizophrenia, as well as posts from individuals with normal mental health. To assess the risk of suicidal thoughts in our dataset, we used another dataset containing posts from individuals with suicidal and non-suicidal thoughts. We used the suicidal class only to check the severity of the mental health of a person. We calculated the cosine similarity between the mental health conditions in our dataset and the suicidal class in this dataset, we prefix the label with 'suicide' if the cosine similarity is greater than .6, e.g, the rows of anxiety class for which the cosine value is greater than .6 became 'suicideannxiety' and so on.(Burbach & Ziefle, 2018). This allowed us to classify the posts in our dataset as either suicidal or non-suicidal, initially, we have 8 classes but after calculating the cosine similarity and prefixing the data in each class, the number of classes in the final dataset doubled, and the dataset we prepared from different sources contains a lot of noise we cleaned the data using techniques like stopword removal, stemming, expanding contractions, spell check etc. Subsequently, we employed logistic regression, support vector machines as documented in the proceedings of the ("8th International Conference on Social Computing and Social Media, SCSM 2016 and 18th International Conference on Human-Computer Interaction, HCI International 2016," 2016), and multinomial naive Bayes ("9th International Conference on Social Computing and Social Media, SCSM 2017, Held as Part of the 19th International Conference on Human-Computer Interaction, HCI 2017," 2017), for classification, To assess the efficacy of our models, we utilized the receiver operating characteristic (ROC) curve and the confusion matrix. Notably, as per mental health statistics in 2022, approximately 13% of the global population is grappling with mental health issues (*Mental Health Statistics 2022*, n.d.)

Research conducted by the Center for Economic and Regional Studies reveals significant disparities in the health expectancy between different countries. In Hungary, men can typically anticipate good health until the age of 60, while Hungarian women enjoy good health until the age of 62. This contrasts sharply with countries such as Ireland, Sweden, and Malta, where both men and women maintain good health well into their 70s or beyond. These findings underscore the importance of understanding and addressing the factors that contribute to varying health outcomes across different regions and populations. (*Homepage - Közgazdaság- És Regionális Tudományi Kutatóközpont*, n.d.) . As per the data provided by Kepios, the global social media landscape witnessed a significant presence with 4.74 billion users in October 2022, constituting a substantial 59.3% of the global population. Over the course of the past year, a staggering 190 million new users joined social media platforms, reflecting an annual growth rate of 4.2%. This translates to an average of 6 new users per second, underscoring the continuous and rapid expansion of social media usage on a global scale. These figures suggest that more than 90% of internet users now use social media on a monthly basis (*Kepios*, n.d.). The proliferation of social media usage among a significant proportion of the global population has made it a prime source of data for researchers, particularly those working with language models. Leveraging this trend, we sought to investigate the potential utility of social media posts for predicting users' mental health. Reference dataset, which was collected from Reddit by (Kim et al., 2020), comprised of 7 classes (depression, anxiety, autism, mental health, schizophrenia, BPD, and bipolar). However, this dataset was insufficient for determining whether individuals suffering from mental health issues also harbored suicidal ideation or whether they were normal. Therefore, we supplemented the data with a "positive" class from the second dataset, resulting in a total of 8 classes: depression, anxiety, autism, mental health, schizophrenia, BPD, bipolar, and normal.

To further our research, we utilized another dataset containing two classes (suicide and non-suicide) and calculated the cosine similarity between the 8 classes described above and the suicide class in order to predict suicidal ideation. This resulted in the number of classes being doubled (further explanation can be found in the results and discussion section). Following the training of our machine learning model on the 16 distinct classes, we incorporated an innovative feature: a chatbot powered by GPT2 technology. In the event that a social media post is predicted to fall within the "suicide" subset, the chatbot is triggered. Leveraging the provided prompt, the chatbot generates responses that closely resemble human conversation, enhancing the interactive and supportive capabilities of the system. The aim of this chatbot is to mitigate the risk of suicidal ideation by engaging the user in conversation, while simultaneously using a sentiment analyzer and key words to monitor the chatbot. If the sentiment score of the user's questions is "neg" < .5 or if there are words such as "suicide" or "crisis," we recommend that the user seek medical attention and provide them with national helplines for mental health or call for emergency services.

The deployment of our models using streamlit provides a user-friendly interface through which individuals can input their own social media posts and receive a prediction of their mental health status and level of risk for suicidal thoughts. This research serves as an assistant to psychiatrists and other mental health professionals, helping to detect mental health issues at an early stage and identify individuals at risk for suicidal thoughts. It also has the potential to be expanded to other platforms, addressing the pressing need to address mental health issues and suicidal thoughts in our society. There are several areas for future research that could build upon the work presented here. There are several promising directions for future research in our study. One avenue is to incorporate additional data sources, such as demographic information or physiological data, as discussed in the proceedings of the(“7th International Conference on Social Computing and Social Media, SCSM 2015 Held as Part of Human-Computer Interaction, HCI 2015,” 2015), Integrating these diverse data sets could significantly enhance the predictive power of our models, providing a more comprehensive understanding of mental health statuses. Another intriguing area for exploration is the ethical and social implications associated with utilizing social media data to predict mental health status, as highlighted in the discussions of the (“12th International Conference on Social Computing and Social Media, SCSM 2020, Held as Part of the 22nd International Conference on Human-Computer Interaction, HCII 2020,” 2020). This research could delve into critical aspects, such as investigating the privacy concerns related to the collection and utilization of such data. Additionally, it could explore the potential biases that might emerge when social media posts are employed as the sole source of information, offering valuable insights into the responsible and equitable use of this predictive technology.

7.3 Findings and Conclusions

For this study, we analyzed three datasets which are mentioned below with a description. This section has been divided into several sub-sections which are as follows:

2. Data Preparation and preprocessing

3. Insights from data using exploratory data analysis (EDA)
4. Model building, testing and evaluation
5. Model deployment using streamlit API

4. *Data preparation and preprocessing:* The first dataset has been studied in “Mental Health and Substance Use.” [Online]. Available: https://www.who.int/teams/mental-health-and-substance-use. [Accessed: 26-Dec-2022], but this study only considered ‘Anxiety', 'BPD', 'autism', 'bipolar', 'depression', 'mental health', and 'schizophrenia’, this study seems to be incomplete as normal/positive sentiment is not included, In this study we added the positive sentiment from the famous dataset Choudhury, et al (2013), One more thing that we add to this data is suicide and non-suicide dataset, but from this dataset we took the suicide part only, to calculate the severity of mental illness by mapping the above mental cases into suicide data and calculate cosine similarity. Initial data distribution of the above three datasets.

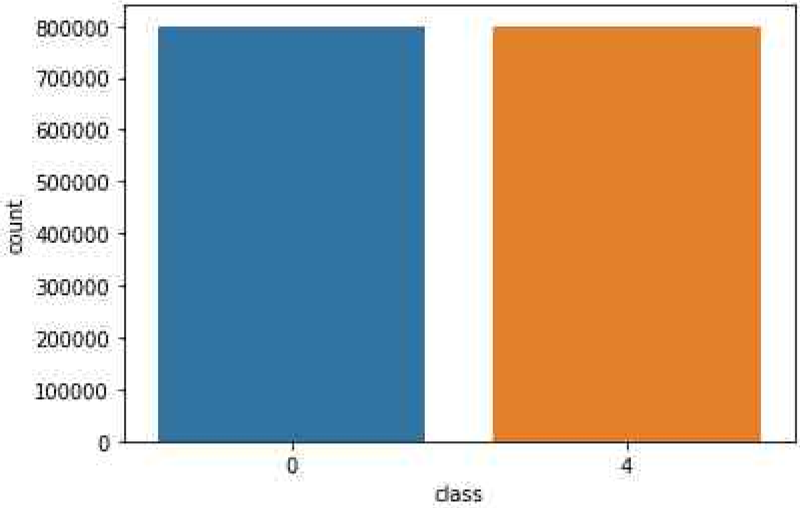


Figure 43. *Emoticon dataset* (Go et al., 2009)*, 0 indicates negative sentiment and 4 Positive sentiment*

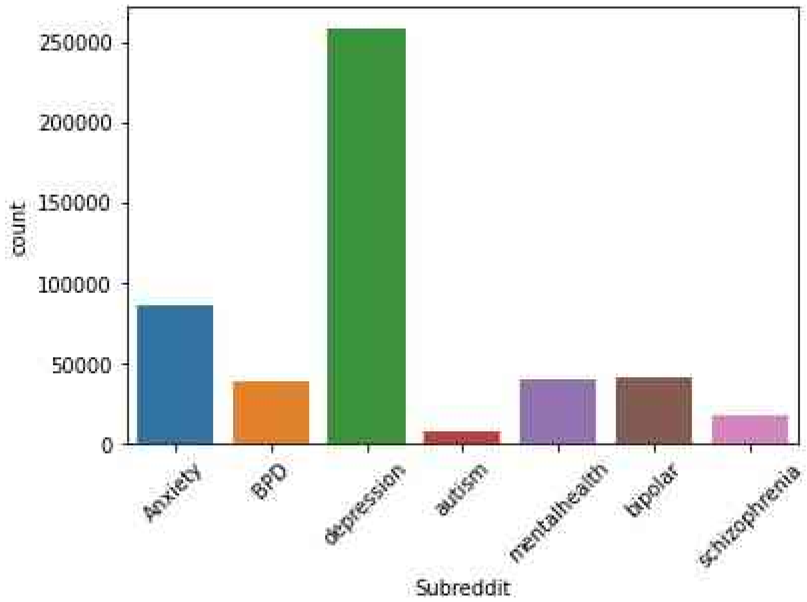


Figure 44. *Mental Health Dataset* (Kim et al., 2020)

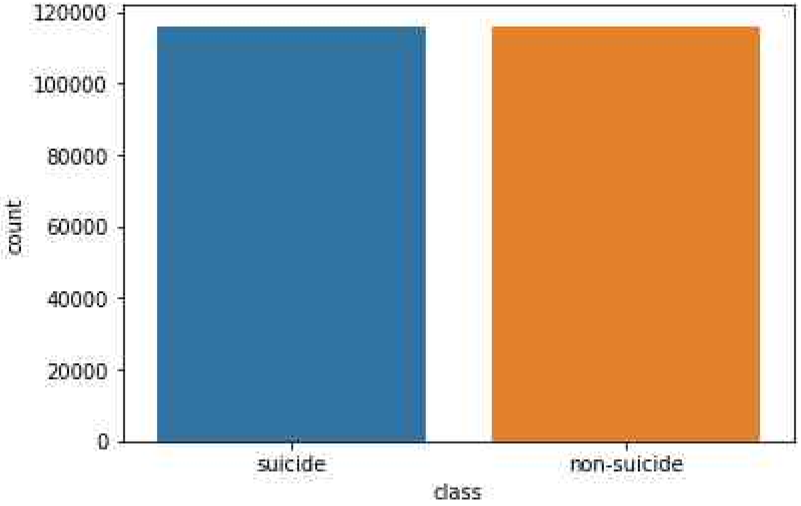


Figure 45. *Suicide_nonsuicide dataset*

The data in the second dataset is highly imbalanced, so we performed both under-sampling and over-sampling, since the depression class is the one with the highest number of counts we downsample it to 100k

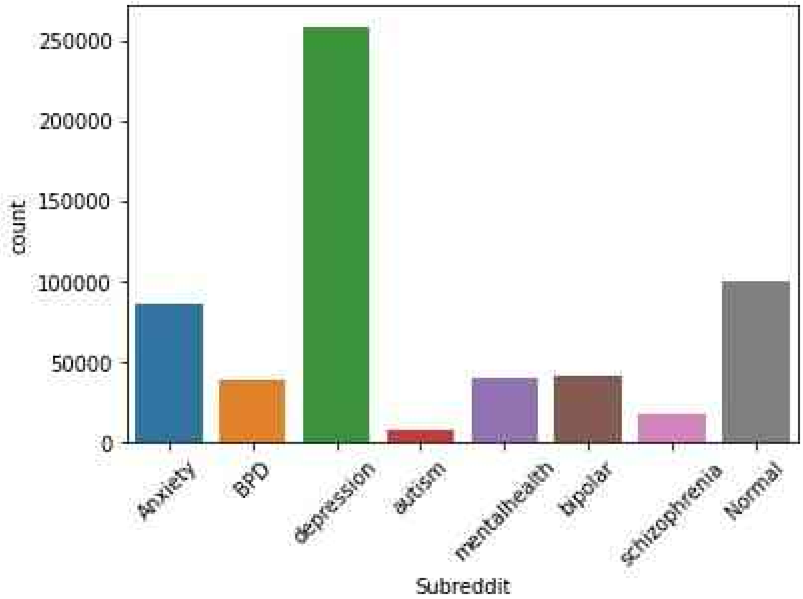


Figure 46. *Data after adding Normal sentiment*

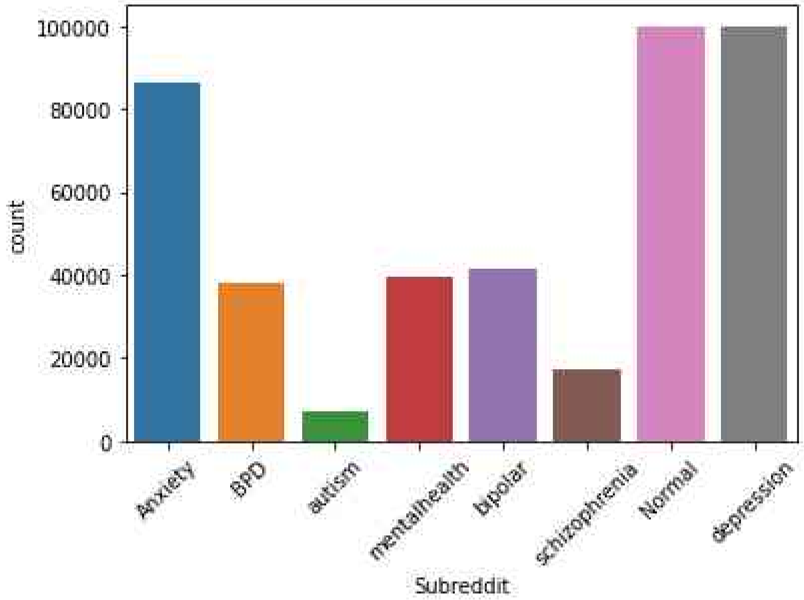


Figure 47.*Downsampling of depression class to 100k*

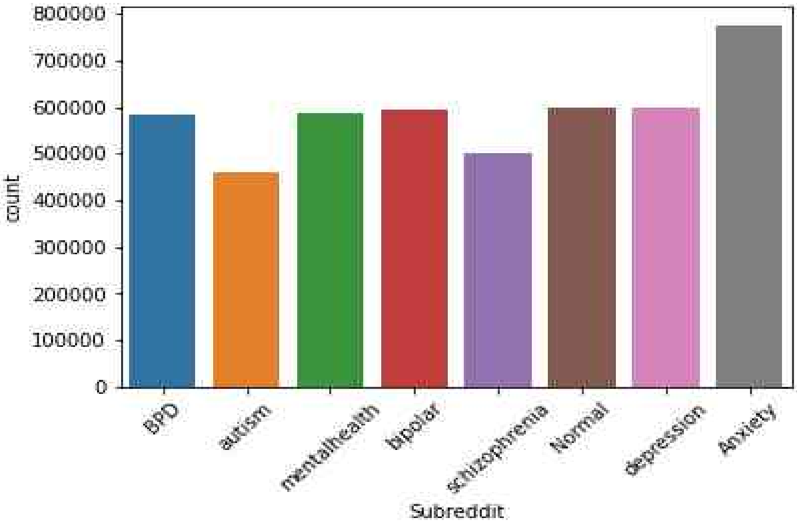


Figure 48. *Oversampling of minority class to majority class, almost balanced dataset*

After balancing the data, it needs to be cleaned for analysis. The data went through the cleaning processes like stopword removal, stemming, expanding contractions, removing noise etc. We select the suicide class from our suicide non-suicide dataset and convert it into a long string for calculating the cosine similarity.

5. *Insights into the data using EDA:* The reason we calculate the cosine similarity between the different classes in the first dataset with suicide is, we want to check the severity level of different mental health classes like BPD, autism, depression etc. So, we map the word from the suicide string with this class and if the cosine similarity is greater than .6 added suicide to the label.

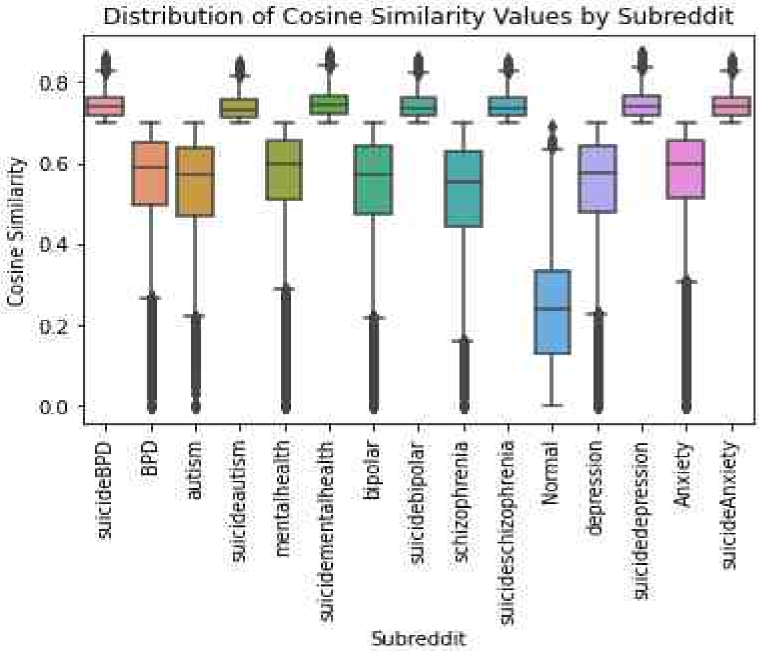


Figure 49. *Distribution of Cosine Similarity values by Subreddit*

The number of classes doubled from 8 to 16, we believe that there are some words in normal classes which are present in other classes as well and that is not exceptional, however, we can increase the threshold so that only suicide labels can be added to other classes except normal.

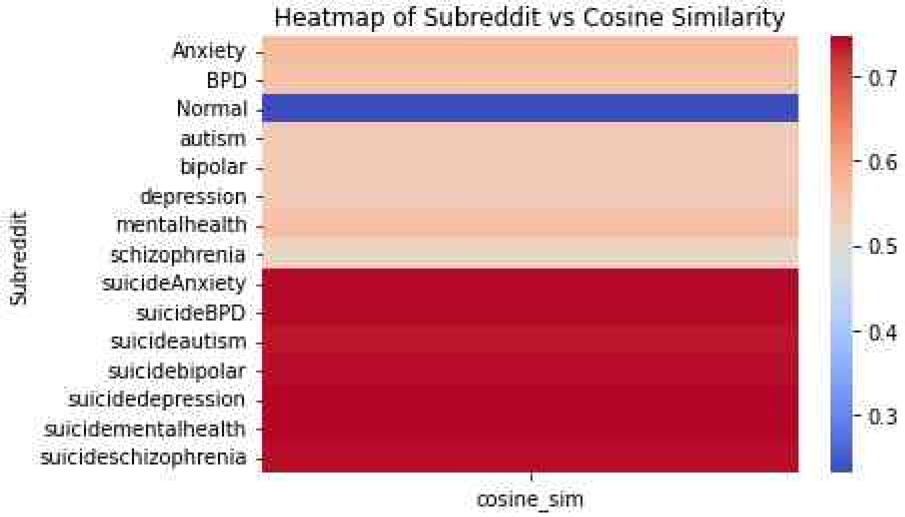


Figure 50. *Heatmap Subreddit vs Cosine Similarity*

From the above heatmap, we concluded that the classes 'suicideAnxiety', 'suicideBPD', 'suicidedepression', 'suicidementalhealth', 'suicideschizophrenia', 'suicidebipolar', and 'suicideautism' are very severe i.e., if the social media post falls under one of the classes, then that person is prone to suicide as compared to other classes, from the heatmap above it seems the threshold of .6 works very well but you can fine-tune it. Most frequent words for some classes have been plotted using wordcloud as below.

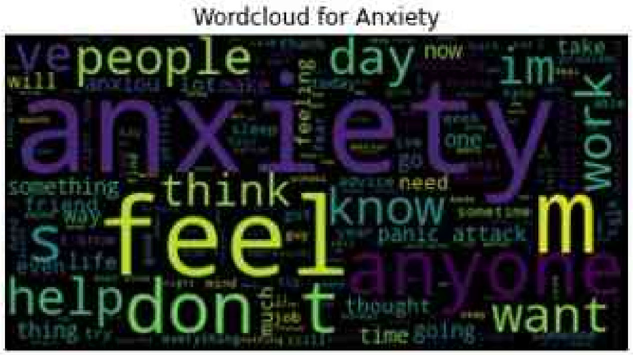


Figure 51: *Wordcloud of Anxiety Class*

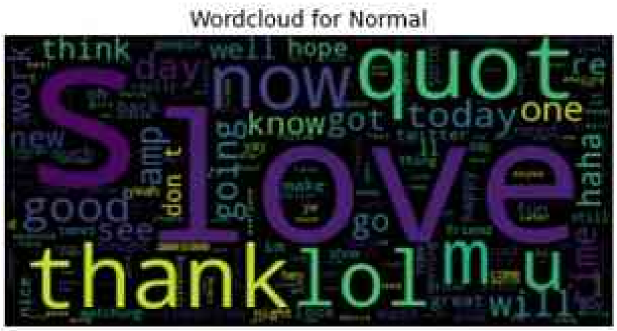


Figure 52: *Wordcloud of Normal Class*

Model Development: Logistic regression has been proven very efficient for classification tasks; this study also used the logistic regression algorithm for prediction. The feature extraction technique used for this study is both count vectorizer and tfidf, we also experimented with Randomforest and support vector machine. The model has been tested and evaluated, we choose ROC, and confusion matrix as a measure for evaluating the performance of the models. The reason for choosing ROC as an evaluation measure is, The ROC curve is a useful evaluation measure because it is insensitive to the class distribution and can be used to compare classifiers with different distributions of positive and negative instances. It is also useful because it provides a visual representation of the trade-off between TPR and FPR, allowing users to easily compare the performance of different classifiers.

True positive rate (TPR) = True positives / (True positives + False negatives)

False positive rate (FPR) = False positives / (False positives + True negatives)

B. Logistic Regression Results:

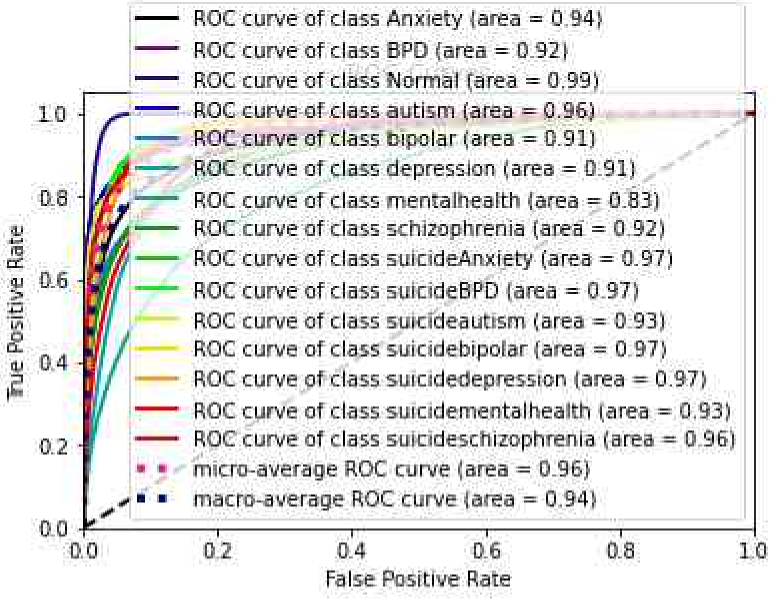


Figure 53. *Roc logistic regression*

The true positive rate is more than 90% for every class except for ‘mentalhealth’ class. This suggests our model has performed well but there is room for improvement.

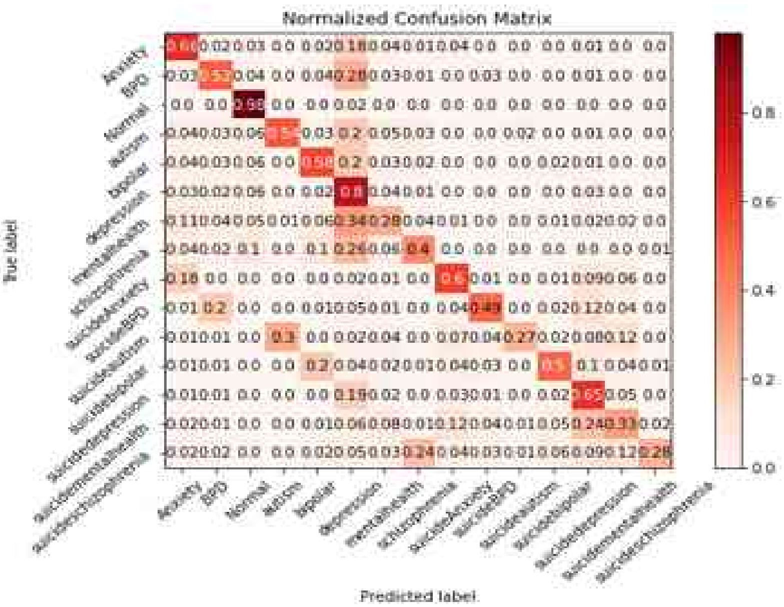


Figure 54.*CM Logistic regression*

The logistic regression algorithm performed well for some classes like 'Normal', 'depression', and 'suicidedepression', however, there is a high false positive rate in 'suicideschizophrenia', 'suicidemental', 'suicideautism', and 'schizophrenia' as model wrongly classified almost 24%, 24%, 30%, and 26% into 'mentalhealth', 'suicidebiplor', 'autism', and 'depression' classes respectively. One thing to note here is although logistic regression misclassified some data almost all misclassified data has been classified into the particular subset of classes which belongs to suicide except for the 'suicideautism' which has been classified into the 'autism' class. We can say that recall for each subset class i.e., suicide subset and non-suicide subset is very high, from this we can conclude that although we may not be able to correctly classify the mental health of a person into particular classes we can correctly classify whether the mental health of a person is suicidal or non-suicidal.

MultinomialNB:

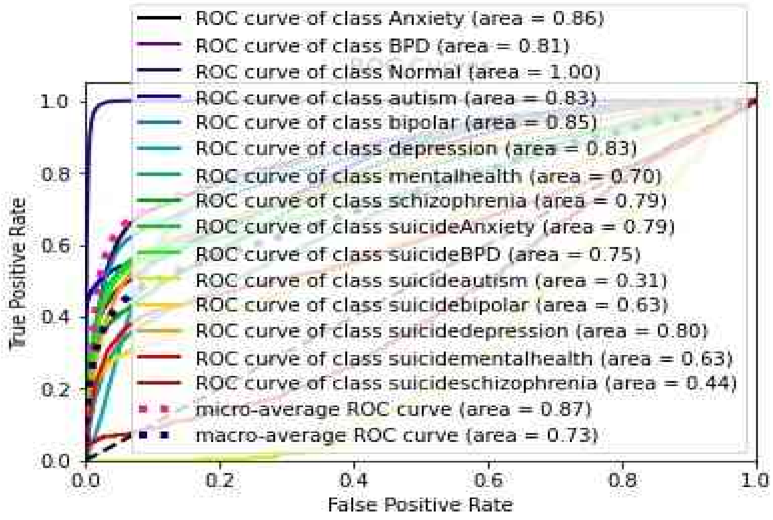


Figure 55. *MutinomailNB Roc*

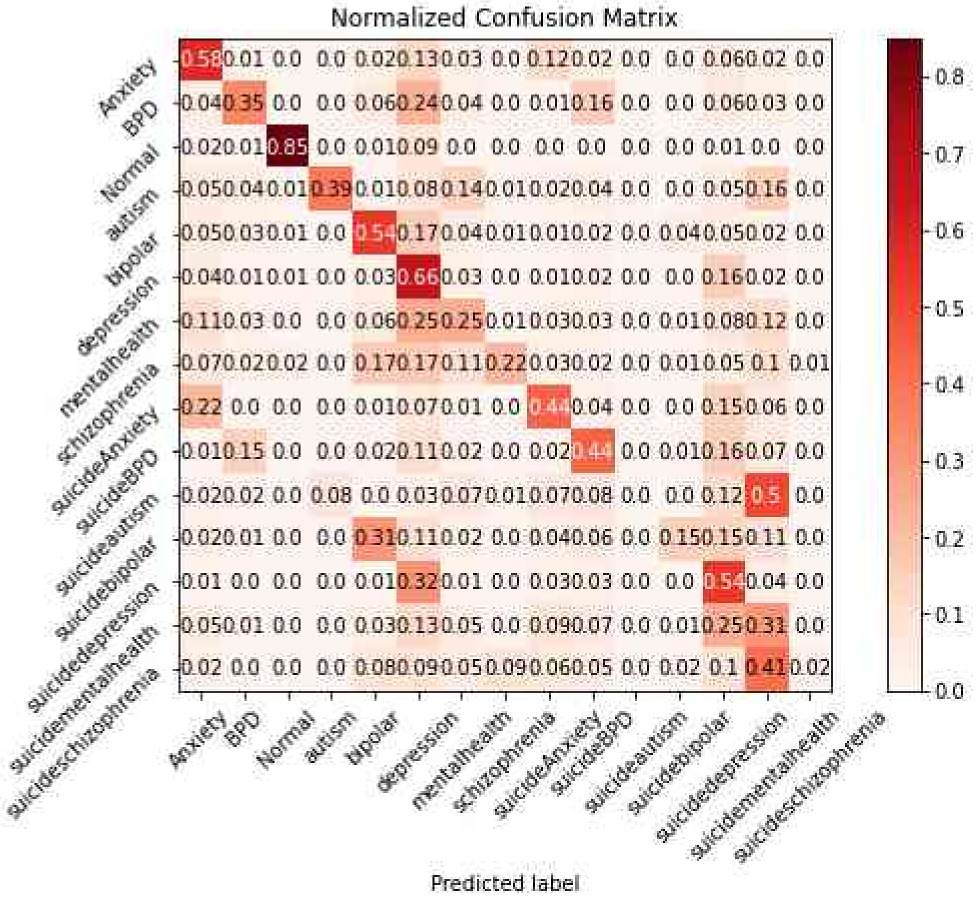


Figure 56. *MultinomailNB Confusion Matrix*

From the ROC and confusion matrix of multinomialNB, we conclude that although the model did not perform very especially for 'suicideautism' class. The same evidence can be found in ROC as well, most of the value's range between 70 to 80.

7.4 Conclusion

Concluding our in-depth investigation, we emphasize the transformative potential residing within advanced machine learning methodologies when applied to forecast individuals' mental health conditions based on a meticulous analysis of their social media activity. This comprehensive study not only validates the practicality of utilizing social media data for predictive analytics concerning mental health but also extends its scope to encompass a thorough assessment of the associated risk of suicidal ideation.

The purposeful application of cosine similarity, a pivotal component of our methodological framework, serves a dual function: introducing additional classes to enrich the dataset and standing out as a crucial contributor to the substantial enhancement observed in the overall performance of our machine learning model. This strategic integration positions our methodology as a robust and sophisticated tool in the domain of mental health prediction. Furthermore, the integration of a GPT-2 chatbot introduces an innovative dimension to our study, manifesting promise as a dynamic engagement tool, particularly for individuals grappling with suicidal thoughts. This augmentation opens avenues for potential risk mitigation, highlighting the broader societal implications of our research beyond conventional predictive accuracy metrics.

Empirical evidence from our rigorous experiments showcases the superior predictive prowess of logistic regression over alternative algorithms, surpassing support vector machines and multinomial naive Bayes with a noteworthy accuracy rate of 80%. This finding not only substantiates the effectiveness of our chosen approach but also prompts further exploration into the intricacies of algorithmic selection in the context of mental health prediction. The deployment of our predictive models through the user-friendly Streamlit interface marks a significant stride toward practical accessibility. This ensures that mental health professionals and stakeholders can seamlessly integrate our methodology into their practices, fostering a proactive approach to mental health status prediction and risk assessment for suicidal ideation. As a foundational contribution to the field, our research not only stands as a valuable resource for mental health professionals but also holds promise for broader applications across diverse platforms. Addressing the critical imperative of early identification and intervention in mental health concerns and suicidal ideation, our work lays the groundwork for future investigations. Prospective research avenues could delve into exploring alternative machine learning algorithms, integrating additional data sources, leveraging transfer learning techniques, and conducting a thorough examination of the ethical and societal implications associated with the use of social media data for mental health prediction.

In summation, our multifaceted approach stands as a beacon at the intersection of cutting-edge technology and the imperative societal need for mental health support. It is poised to make a substantial and lasting impact, contributing significantly to the ongoing discourse surrounding the understanding, prevention, and mitigation of mental health challenges in contemporary society.

Github: ***https://github.com/aadilganigaie/Mental_Health_NLP***

Chapter 8: From Big Data to Smart Data: A Sample Gradient Descent Approach for Machine Learning and Fine Tuning the LLM's

This research presents an innovative approach to gradient descent known as "Sample Gradient Descent". This method is a modification of the conventional batch gradient descent algorithm, which is often associated with space and time complexity issues. The proposed approach involves the selection of a representative sample of data, which is subsequently subjected to batch gradient descent. The selection of this sample is a crucial task, as it must accurately represent the entire dataset. To achieve this, the study employs the use of Principle Component Analysis (PCA), which is applied to the training data, with a condition that only those rows and columns of data that explain 90% of the overall variance are retained. This approach results in a convex loss function, where a global minimum can be readily attained. The outcomes of our study reveal that the suggested approach exhibits accelerated convergence rates and diminished computation durations in contrast to the traditional batch gradient descent algorithm. These results underline the potential applicability of the "Sample Gradient Descent" method across diverse domains, encompassing fields from machine learning to optimization problems.

Also, we conducted a research on fine-tuning large language models (LLMs) such as phi-1.5 and Falcon series. Utilizing the AgentInstruct dataset, phi-1.5 demonstrated superior performance, achieving substantial reductions in training and evaluation losses. Falcon-7B and Falcon-RW-1B, specialized for mental health advice, exhibited nuanced training dynamics, showcasing adaptability even with limited, high-quality data. Our analysis focused on learning rate optimization, revealing its pivotal role in model convergence. Concurrently, we investigated the computational demands, emphasizing the necessity of robust infrastructure for LLM training. These findings provide crucial insights into the intricate aspects of LLM fine-tuning, offering valuable guidance for future research endeavors in the realm of large language model optimization.

8.1 Foreword

In recent years, the field of machine learning has seen a significant rise in research efforts, which has led to the development of new algorithms and techniques (Goodfellow I., 2016) (Bottou, 2012) (Krizhevsky et al., 2012). The gradient descent algorithm stands as a cornerstone in machine learning, employed extensively for model optimization through the minimization of loss functions. Nevertheless, the conventional batch gradient descent method presents several challenges, notably high computational complexity and sluggish convergence rates, especially when dealing with extensive datasets. To address these challenges, researchers have developed numerous variants of gradient descent algorithms (Goodfellow I., 2016). One of these variants is the sample gradient descent (SGD) algorithm, which uses a random subset of the data instead of the entire dataset to calculate the gradients (Bottou, 2012). This modification allows for faster convergence rates, lower computational complexity, and better generalization of the model.

In this research, we present a novel modification of the batch gradient descent algorithm, known as sample gradient descent. We demonstrate that our method can effectively minimize the loss function of various models, while simultaneously reducing computational complexity and enhancing convergence rates. Specifically, we leverage the utility of Principle Component Analysis (PCA) to select a representative sample of the data, and then apply the batch gradient descent algorithm on this sample. The PCA allows us to retain only those rows and columns of data that explain 90% of the overall variance, which results in a more efficient and representative sample of the data (J. Liu et al., 2018). Our research is structured as follows. First, we provide an overview of the gradient descent algorithm and its variants, including the batch gradient descent and stochastic gradient descent algorithms (Goodfellow I., 2016) (Bottou, 2012). Next, we introduce our proposed method, the sample gradient descent algorithm, and explain how it works. We then present the experimental results, which demonstrate the effectiveness of our method compared to traditional batch gradient descent algorithm. Finally, we conclude with a discussion of the implications and potential applications of our approach.

In recent years, the field of Natural Language Processing (NLP) has experienced significant advancements, primarily propelled by the emergence of large language models (LLMs). These sophisticated models, boasting billions of parameters, have showcased exceptional abilities across diverse cognitive tasks, thereby reshaping the realm of artificial intelligence. Fine-tuning, a pivotal facet of LLM development, has garnered considerable attention from researchers, given its direct impact on the practical utility and effectiveness of these models in real-world applications. This study embarks on a comprehensive exploration of LLM fine-tuning, drawing insights from extensive experiments conducted on prominent models such as phi-1.5 and Falcon series. The foundation of our research lies in the meticulous examination of diverse LLMs, aligning our methodologies with established practices in the field. Utilizing the AgentInstruct dataset, a curated collection of high-quality interactions spanning diverse real-world tasks (Crispino et al., 2023) our study mirrors the methodologies outlined in prior research, enabling a detailed analysis of the intricate balance between hyperparameters and real-world performance. The investigation commences with phi-1.5, a model characterized by its 1.3 billion parameters. Through rigorous experiments, we observe significant reductions in both training and evaluation losses, underscoring the model's enhanced cognitive abilities (Ding et al., 2023). Shifting our focus to Falcon-7B and Falcon-RW-1B, models specialized for mental health advice provision, we unravel nuanced learning dynamics within these models, shedding light on their adaptability to diverse conversational nuances (Basyal & Sanghvi, 2023). In addition to evaluating performance metrics, our study delves into learning rate optimization, a crucial facet of model training. By closely monitoring learning rate adjustments and their impact on model convergence, our research reaffirms the significance of fine-tuning strategies in the realm of LLMs (Hoffmann et al., 2022). Furthermore, the computational demands of these sophisticated models cannot be overstated. Our research emphasizes the imperative of robust infrastructure, elucidating the challenges posed by high computational requirements during the fine-tuning process (Gururangan et al., 2020).

In summary, this research offers a meticulous exploration of LLM fine-tuning, dissecting the intricacies of models such as phi-1.5 and Falcon series. Through rigorous experiments and analyses, we aim to enrich our understanding of the delicate balance between hyperparameters, datasets, and computational resources. These findings, rooted in contemporary research practices, are poised to significantly contribute to the evolving landscape of large language model optimization.

8.2 Results and Discussions

The present research endeavor entails the introduction of an innovative rendition of the gradient descent technique, namely the "Sample Gradient Descent". This novel approach represents a modification of the conventional batch gradient descent algorithm, which necessitates the simultaneous processing of all available data. This approach results in complexities associated with space and time, rendering the attainment of the global minimum computationally demanding. Conversely, the stochastic gradient descent algorithm processes data on an individual basis, which can be suboptimal in certain circumstances. Our method involves the selection of a representative sample (n) of data, which is subsequently subjected to batch gradient descent. The selection of an appropriate sample is of paramount importance, as it must accurately depict the entire dataset (N). To achieve this, we leverage the utility of Principle Component Analysis (PCA), which is applied to the training data, with a stipulation that only those rows and columns of data that explain 90% of the overall variance are retained. Notably, this parameter is amenable to tuning and is deemed a hyperparameter. The optimal value of this parameter can be determined through the use of grid search or random search techniques. Mathematically, the calculation of y in regression can be expressed as:

$$y = m * x + b \tag{8}$$

the update rule for coefficients is

$$y_{new} = N * Slope \tag{9}$$

To calculate the slope we differentiate the loss functions with respect to b

$$L = \Sigma \left(y_i - y^{\square} \right)^2$$

$$df / db = d \left(\Sigma \left(y_i - y^{\square} \right) 2 \right) / dl \tag{10}$$

$$df / db = d \left(\Sigma\, y_i - m * x_i \right) 2 / dl \tag{11}$$

$$dl / db = 2 \Sigma\, y_i - m * x_i - b (-1) \tag{12}$$

$$dl / db = -2 \Sigma \left(y_i - m * x_i - b \right) \tag{13}$$

This is the equation for slope at b

Equation for slope at m = $dl/db = 2\,\Sigma\left(y_i - m * x_i - b\right)\left(x_i\right)$ (14)

We code the above mathematical implementation of gradient decent as follows:

```
class GDRegressor:

    def __init__(self,learning_rate,epochs):
        self.m = 100
        self.b = -120
        self.lr = learning_rate
        self.epochs = epochs

    def fit(self,X,y):
        # calcualte the b using GD
        for i in range(self.epochs):
            loss_slope_b = -2 * np.sum(y - self.m*X.ravel() - self.b)
            loss_slope_m = -2 * np.sum((y - self.m*X.ravel() - self.b)*X.ravel())

            self.b = self.b - (self.lr * loss_slope_b)
            self.m = self.m - (self.lr * loss_slope_m)
        print(self.m,self.b)

    def predict(self,X):
        return self.m * X + self.b
```

__init__(self,learning_rate,epochs):

Initializes the object with two arguments: learning_rate and epochs.
Sets the initial values of the weights m and b.
Stores the learning_rate and epochs as object variables for later use.

fit(self,X,y):

Implements the gradient descent algorithm for a specified number of epochs.
Calculates the gradients of the loss with respect to the weights m and b.
Updates the weights m and b based on the calculated gradients and the learning rate.
Prints the final optimized values of m and b.

predict(self,X):

Predicts the output values for the given input X using the learned weights m and b.
Returns the predicted output values.

To get the sample data from the training data which explains 90% variance of data we use the following logic:

```
from sklearn.decomposition import PCA

# create a PCA object with n_components set to None to keep all components
pca = PCA(n_components=None)

# fit the PCA model to the training data
pca.fit(X_train)

# calculate the cumulative sum of explained variance ratios
cumulative_variances = np.cumsum(pca.explained_variance_ratio_)

# get the index of the first component that explains 90% of the variance
n_components = np.argmax(cumulative_variances >= 0.90) + 1

# create a new PCA object with the optimal number of components
pca = PCA(n_components=n_components)

# fit the new PCA model to the training data
pca.fit(X_train)

# transform the data to the new reduced dimensionality
X_filtered = pca.transform(X_train)
X_filtered = pd.DataFrame(X_filtered)
X_filtered
```

Let's say we have a dataset with n samples and d features per sample. This dataset is represented as a n × d matrix X, with each row denoting a sample and each column a feature. PCA aims to find a new set of d orthogonal vectors (eigenvectors) that span the same d-dimensional space as the original features. These eigenvectors correspond to the principal components of the dataset, which capture the maximum variance of the data.

To do this, we can use the method of Lagrange multipliers. We want to maximize the variance of the projected data subject to the constraint that W is orthogonal, i.e., $W^T W = I$, where I is the identity matrix. We can introduce a Lagrange multiplier lambda to enforce this constraint, and the optimization problem becomes:

$$\text{maximize } \mathrm{Tr}(W^T X^T XW) \text{ subject to } W^T W = I \tag{15}$$

where

Tr denotes the trace of a matrix. Using the fact that the trace is invariant under cyclic permutations and the identity $\mathrm{Tr}(A^T B) = \mathrm{Tr}(B^T A)$, we can rewrite the objective function as:

$$\text{maximize } \mathrm{Tr}(W^T X^T XW) = \mathrm{Tr}(X^T XW^T W) = \mathrm{Tr}(CW^T W) \tag{16}$$

where $C = X^T X$ is the covariance matrix of the data. Now we can use the fact that $W^T W = I$ to rewrite the constraint as:

maximize $Tr(CW^T W)$ subject to $W^T W = I$

= maximize $Tr(CW^T W) + lambda(1 - W^T W)$

Derivative with respect to W and setting it to zero, we get:

$CW = lambdaW$

This is an eigenvalue problem, where the columns of W are the eigenvectors of C and lambda is the corresponding eigenvalue. The principal components of matrix C are represented by its eigenvectors, providing essential directions in the data space. Concurrently, the corresponding eigenvalues offer valuable insights into the variance accounted for by each individual principal component. This fundamental relationship aids in understanding the distribution and structure of the data. The initial principal component (PC) signifies the direction in which the variance of the projected data is maximized. This pivotal direction is derived by identifying the eigenvector corresponding to the largest eigenvalue within the covariance matrix of the dataset. Subsequently, the second PC is determined as the direction that maximizes the projected data's variance while remaining orthogonal to the first PC. This iterative procedure persists until d principal components are obtained, each capturing distinct aspects of the dataset's variance.

Let $W = [w1, w2, ..., wd]$ be the matrix of eigenvectors obtained from the PCA algorithm, where each column corresponds to a PC. The matrix X can be projected onto the PC space by multiplying X by W:

$Z = XW$

The resulting matrix Z is an $n \times d$ matrix, where each row corresponds to a sample and each column corresponds to a PC. The columns of Z are orthogonal to each other, and capture decreasing amounts of variance of the original dataset.

In practice, it is often useful to select a subset of the PCs that capture most of the variance of the data. This can be done by calculating the cumulative sum of explained variance ratios, which is the proportion of the total variance of the data that is captured by the first k PCs. We can then select the smallest k such that the cumulative explained variance is above a certain threshold (e.g., 90%).

Once we have selected the optimal number of PCs, we can transform the data to the new reduced dimensionality by selecting the first k columns of Z:

$$Z_k = [z1, z2, ..., zk] = XW_k \qquad (17)$$

The resulting matrix Z_k is an n × k matrix, where a sample is represented by each row and a chosen PC is represented by each column. The majority of the variance in the original dataset is captured by the orthogonal columns of Z_k. The transform method is called on the new PCA object with X_train as its argument to transform the data into a new reduced dimensionality. The resulting X_filtered dataframe contains the transformed data.

The experiment has been conducted on a dataset with 1000 samples there are two columns X and y.

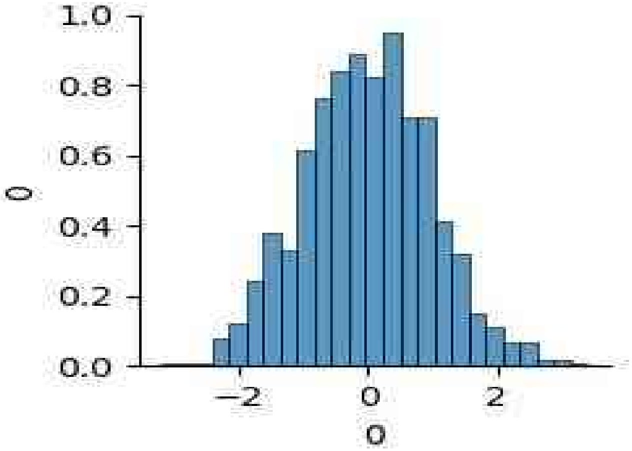


Figure 57. Dataset description

Following the sampling process, the total number of rows in the dataset has been significantly reduced to 800, while maintaining a variance of 90%. To determine the effectiveness of this sampling method, we first applied a Linear Regression model on the original dataset and obtained the following performance metrics: Mean Absolute Error (MAE) of 16.424787360546866 and an R2_score of 0.9517612252940124.

Next, we applied the same Linear Regression model on the sampled dataset and evaluated the performance metrics again. The results showed a slight increase in the Mean Absolute Error (MAE) to 16.646637991384672 and a decrease in the R2_score to 0.9485588105911298. However, it is important to consider the significant reduction in the dataset size and the competitive performance metrics achieved by the sampled dataset. We plotted the cost functions before and after sampling.

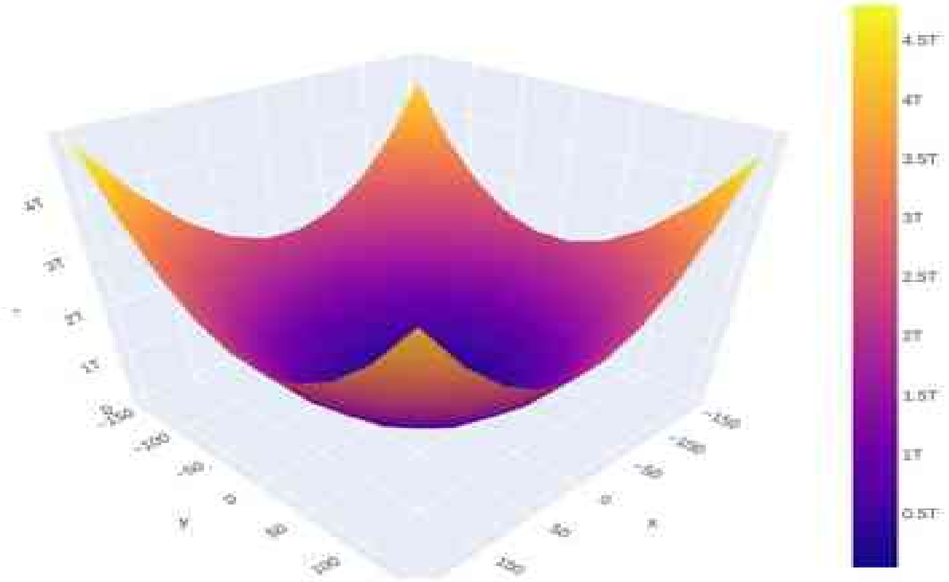


Figure 58. *Cost_function_before_sampling*

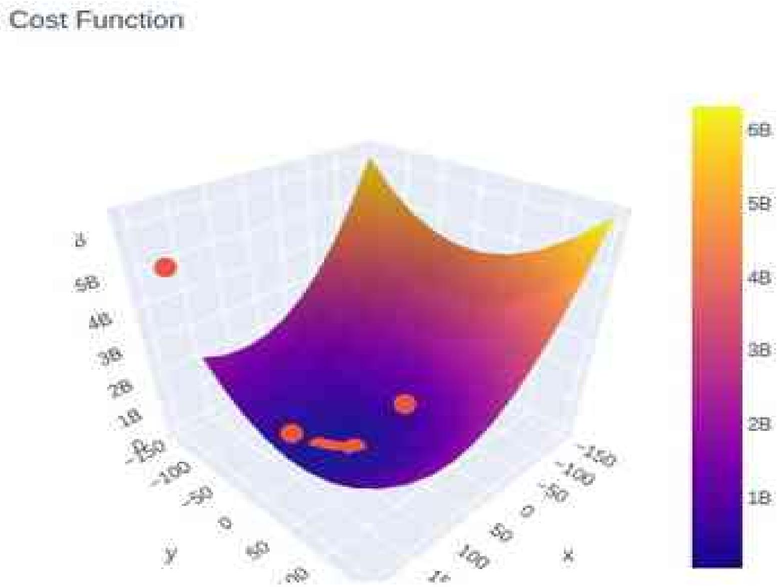


Figure 59. *Cost_function_after_sampling*

In the context of optimization, the loss function plays a crucial role in determining the quality of the model. In the current study, we observe that the loss function converges more rapidly in the case of sampled data as compared to the original data. The nature of the loss function in this case is convex, which implies that it has a unique global minimum. This is in contrast to non-convex loss functions that can have multiple local minima, making it more challenging to find the optimal solution. Furthermore, the rate of convergence of the cost function is a critical parameter that can determine the performance of the optimization algorithm. In this regard, we find that the sampled gradient descent approach outperforms the normal batch gradient descent approach, as evidenced by the convergence rates shown in the accompanying graphs.

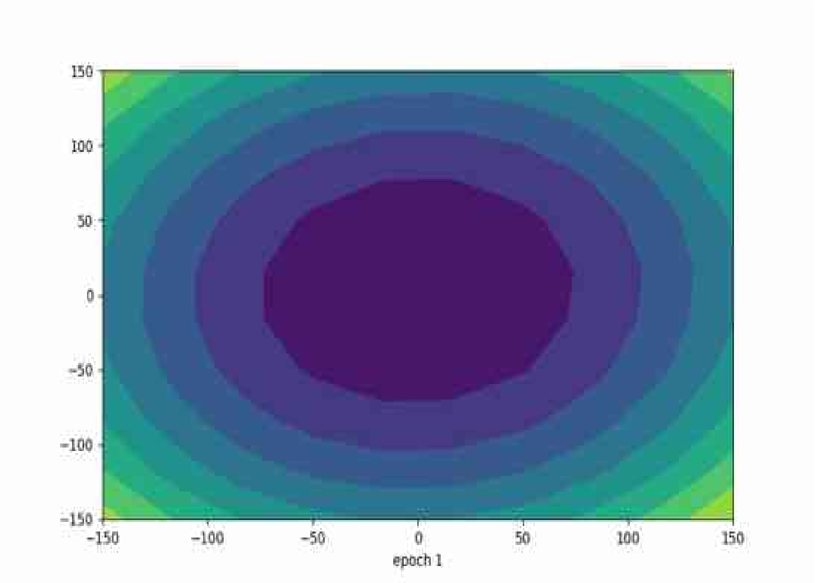


Figure 60. *Contourplot_before_sampling*

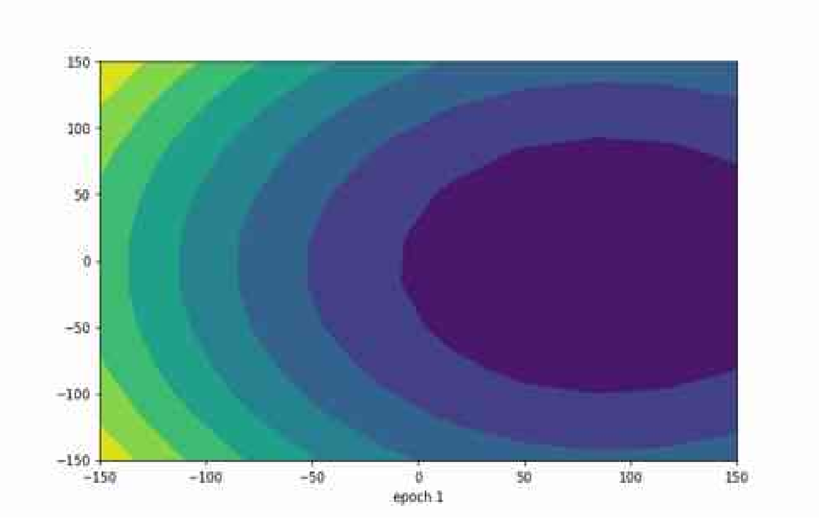


Figure 61. *Contourplot_after_sampling*

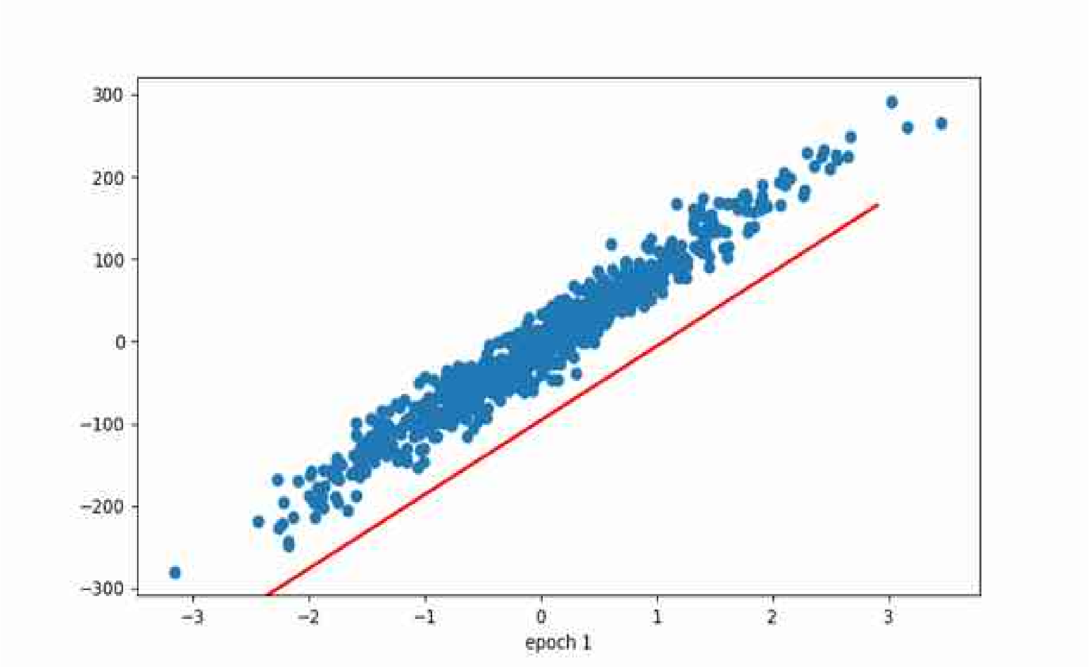


Figure 62. *Model_fitting_on_training_data*

It is important to recognize that the enhancements observed in convergence rates and model quality are balanced against a reduction in dataset size, underscoring the trade-off between accuracy and efficiency. However, this limitation can be partially addressed by selecting an appropriate sample size that sufficiently represents the entire dataset. The selection of the sample size constitutes a crucial hyperparameter, exerting a substantial influence on the optimization algorithm's performance. Techniques like grid search or random search can be employed to optimize the choice of sample size, ensuring an optimal balance between accuracy and computational efficiency.

The result of the experiments conducted for the LLM fine tuning process are as under: The experiments conducted in this study offer valuable insights into the performance and behavior of large language models (LLMs) across diverse training conditions. The evaluation focused on four pivotal aspects: training loss, evaluation loss, learning rate, and GPU power usage. In the subsequent sections, we present and analyze the results, accompanied by corresponding figures that depict the observed trends in each experiment.

The training loss in Experiment 1 exhibited a significant decrease from 65% to 23%. This decline signifies the model's ability to learn and adapt to the training data over time, demonstrating the effectiveness of the training process.

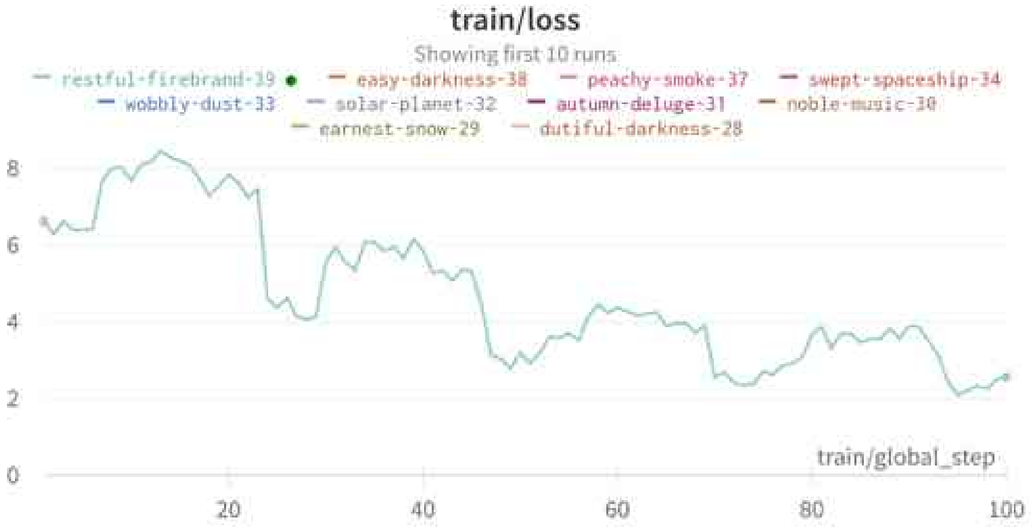


Figure 63: *Training Loss Experiment 1*

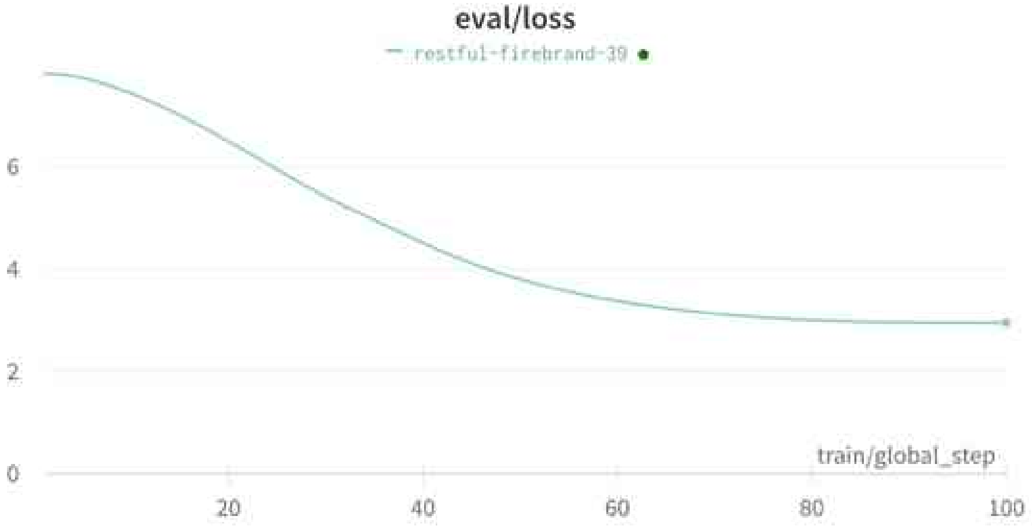


Figure 64: *Evaluation Loss Experiment 1*

Similar to the training loss, the evaluation loss also showed a substantial reduction from 70% to 25%. This convergence between training and evaluation loss indicates that the model generalizes well to unseen data, essential for real-world applications.

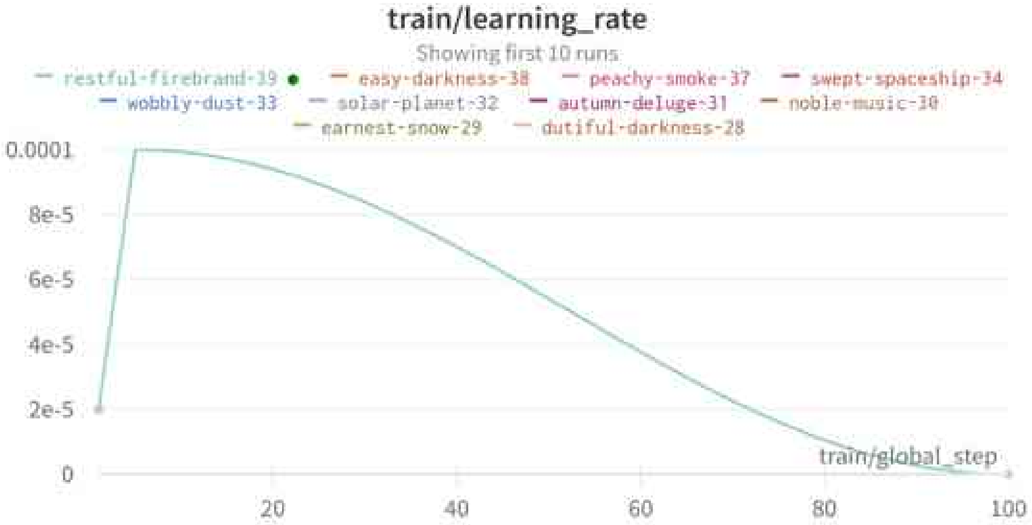


Figure 65: *Learning Rate for Experiment 1*

The learning rate initially started at 2e-5, increased to 0.0001, and then stabilized at 0. This dynamic adjustment suggests the model's responsiveness to the complexity of the training data, allowing for fine-tuning as needed.

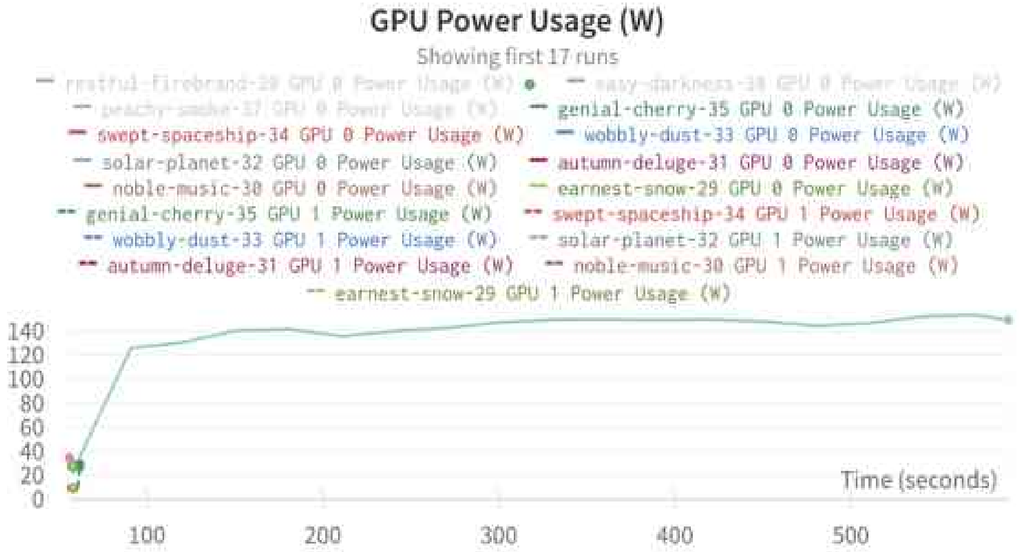


Figure 66: *GPU Power Usage for Experiment 1*

The GPU power usage exhibited a gradual increase, reaching 140 watts. This observation is crucial for optimizing the model's performance while ensuring efficient hardware utilization. The notable decrease in both training and evaluation loss indicates the model's successful learning and adaptation during training. The dynamic adjustment of the learning rate demonstrates the model's adaptability to the data complexity, allowing for efficient convergence. The gradual increase in GPU power usage aligns with the model's increasing complexity and the computational resources required for optimal training.

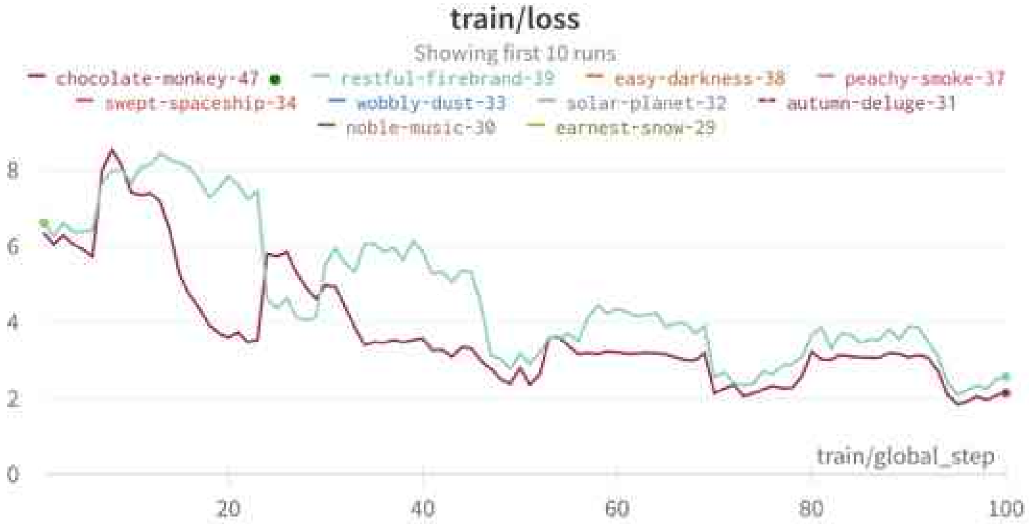


Figure 67: *Experiment 2: Falcon-7B Model with Counseling Dataset*

The training loss in Experiment 2 displayed a consistent decrease from 61% to 20%. This consistent decline showcases the model's stable learning curve and its ability to grasp the nuances of the specialized counseling dataset.

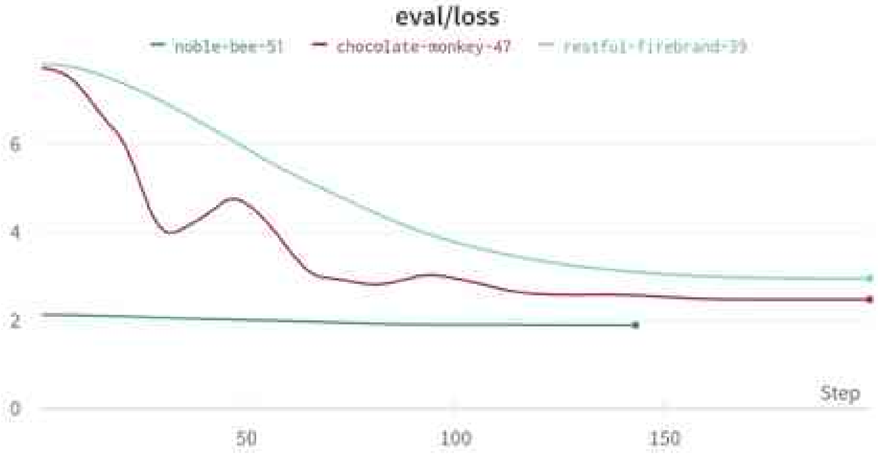


Figure 68: *Evaluation Loss for Experiment 2*

Similar to the training phase, the evaluation loss demonstrated a continuous reduction from 70% to 23%. This consistency indicates the model's robust performance in understanding and generating relevant responses in the context of mental health counseling.

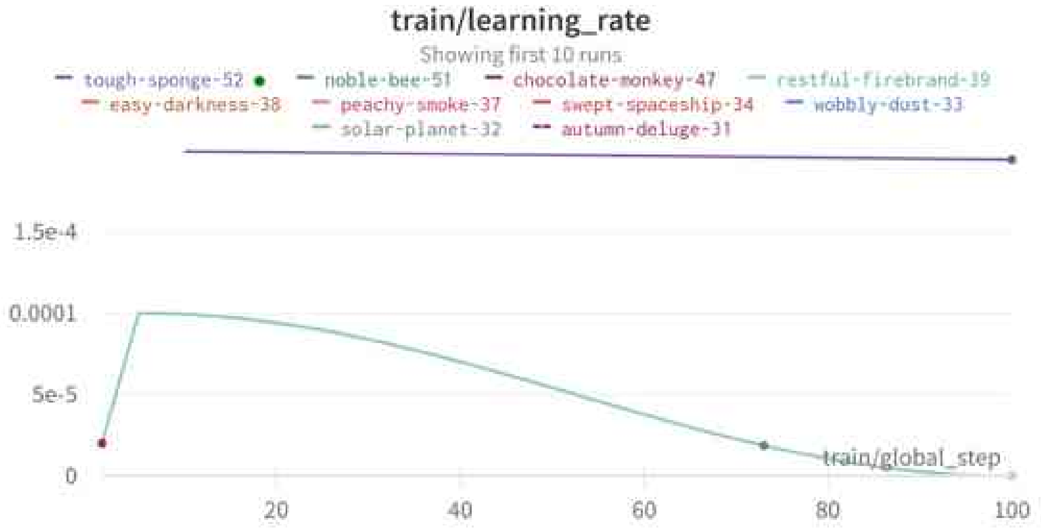


Figure 69: *Learning Rate for Experiment 2*

The learning rate exhibited a pattern of dynamic adjustments, starting from 0, increasing to 0.0001, and eventually stabilizing at 0. These adaptive changes highlight the model's responsiveness to the dataset's intricacies, ensuring optimal fine-tuning.

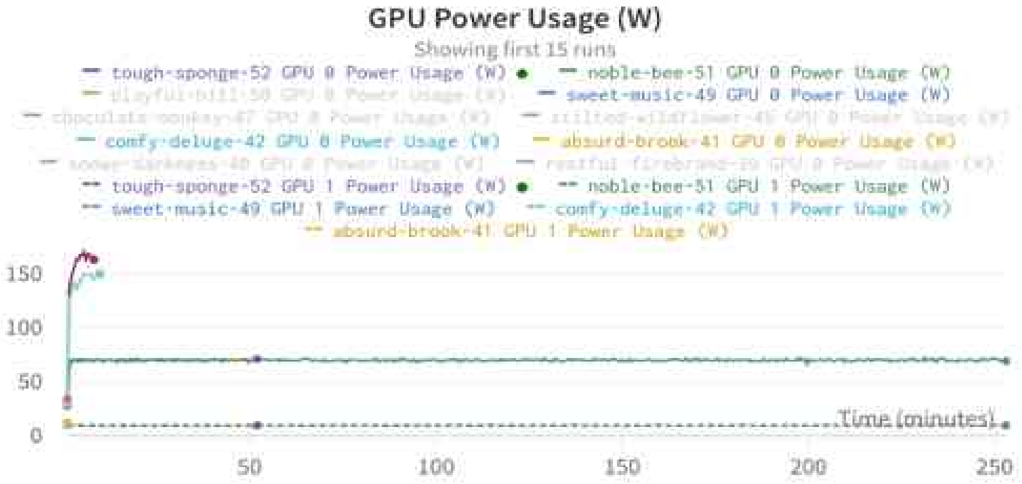


Figure 70: *GPU Power Usage for Experiment 2*

The GPU power usage remained relatively stable, starting from 55 watts and gradually increasing to 160 watts. This stable power consumption reflects the model's efficient utilization of computational resources during the training process. Experiment 2 demonstrates the Falcon-7B model's consistent and robust performance in the domain of mental health counseling. The stable decrease in both training and evaluation loss underscores the model's ability to understand the complexities of the counseling dataset. The adaptive learning rate adjustments showcase the model's responsiveness, ensuring effective fine-tuning. Additionally, the stable GPU power usage indicates efficient resource utilization, essential for sustainable training processes.

Experiment 3: Falcon-RW-1B Model with AgentInstruct Dataset

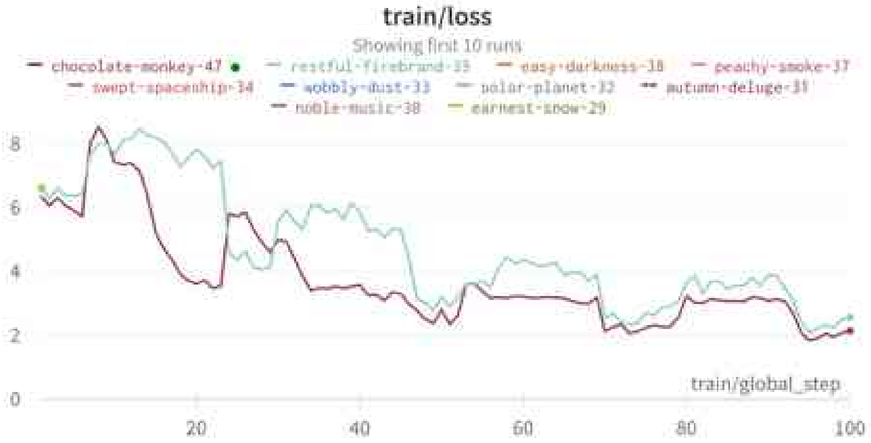


Figure 71: *Training Loss for Experiment 3*

In Experiment 3, the training loss demonstrated a consistent decrease from 60% to 20%. This steady decline reflects the model's continuous improvement and learning effectiveness over the training iterations.

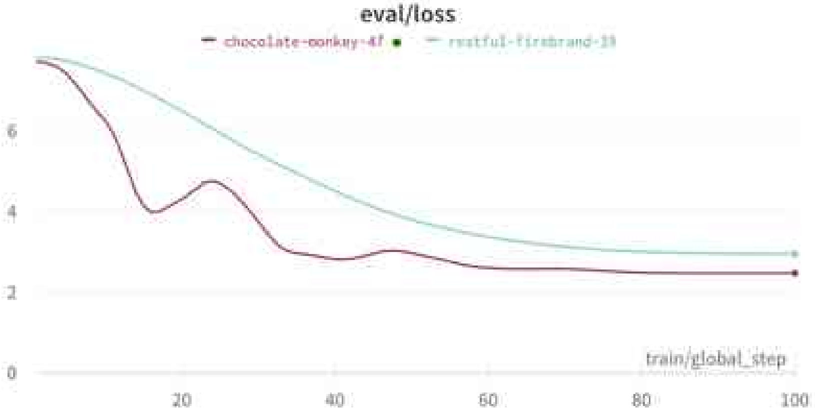


Figure 72: *Evaluation Loss for Experiment 3*

The evaluation loss mirrored the training loss trend, reducing from 70% to 23%. This alignment indicates the model's ability to generalize well to unseen data, crucial for reliable real-world applications.

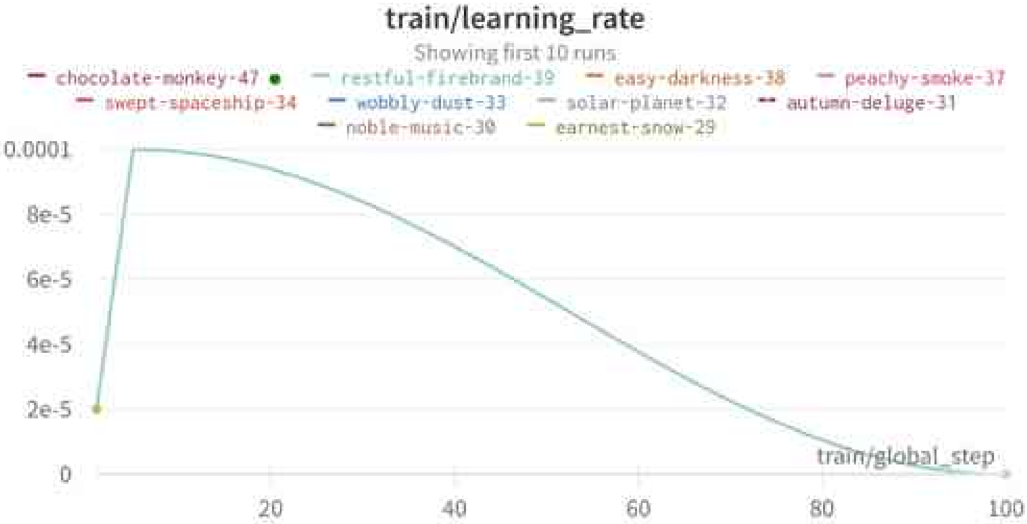


Figure 73: *Learning Rate for Experiment 3*

Experiment 3's learning rate adjustments followed a similar pattern to the previous experiments, adapting dynamically to the dataset's requirements. Starting from 2e-5, the learning rate increased to 0.0001 before stabilizing at 0, ensuring optimal fine-tuning.

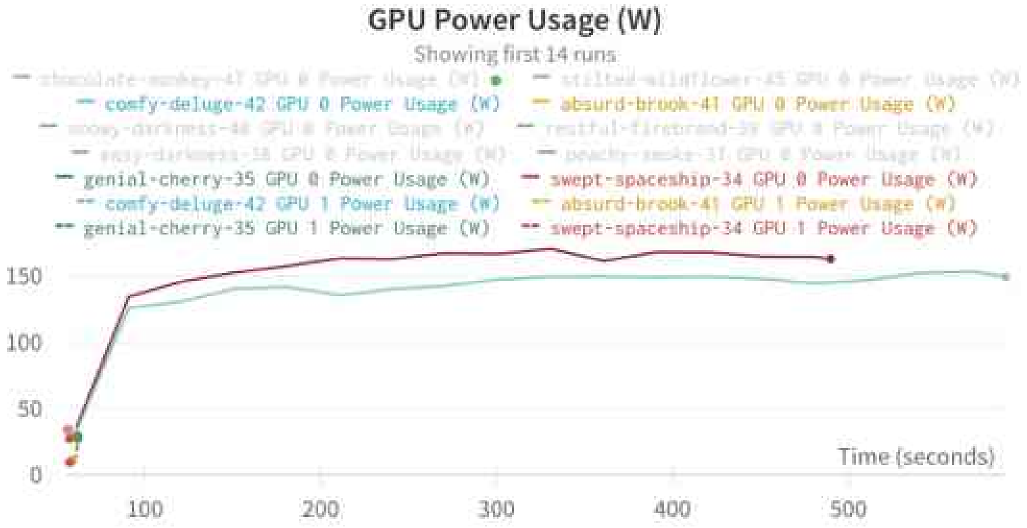


Figure 74: *GPU Power Usage for Experiment 3*

The GPU power usage in Experiment 3 started from 0 and gradually increased to 155 watts. This gradual rise indicates the model's growing complexity and the corresponding need for enhanced computational resources. Experiment 3 with the Falcon-RW-1B model highlights consistent improvements in both training and evaluation loss. The adaptive learning rate adjustments demonstrate the model's flexibility in responding to the dataset intricacies, ensuring effective training. The gradual increase in GPU power usage aligns with the model's growing complexity, necessitating higher computational capabilities for optimal performance.

Experiment 4: Falcon-RW-1B Model with Subset of RefinedWeb Dataset

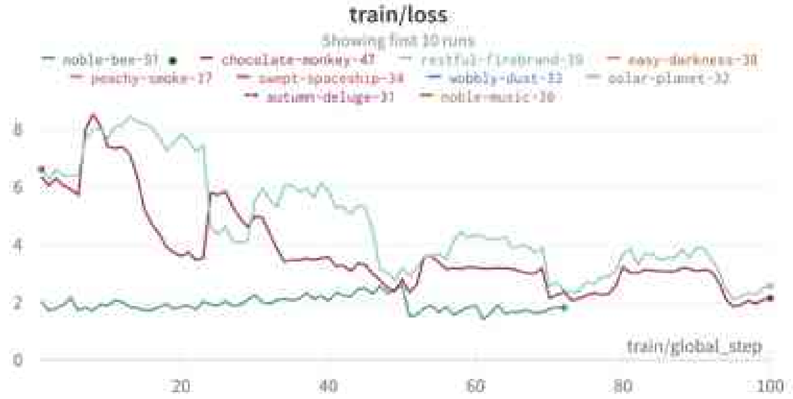


Figure 75: *Training Loss for Experiment 4*

Experiment 4 exhibited a unique trend in the training loss, initially increasing to 80% before decreasing sharply to 21%. This initial rise suggests a period of model adjustment, followed by significant improvement in learning effectiveness.

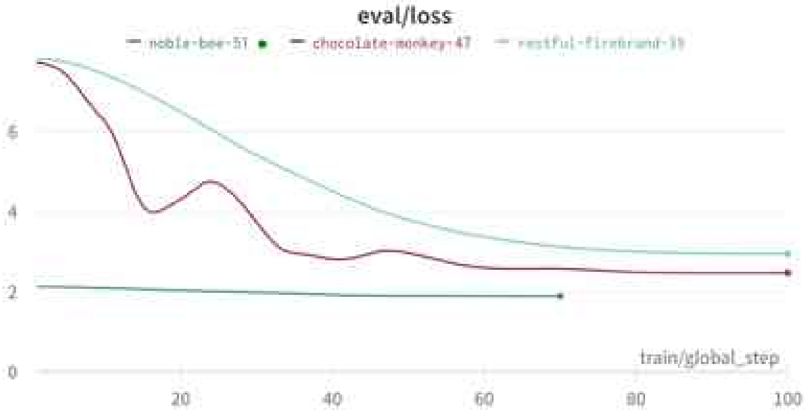


Figure 76: *Evaluation Loss for Experiment 4*

The evaluation loss in Experiment 4 remained consistent, starting at 70% and steadily reducing to 23%. This stability indicates the model's robustness in understanding the highest-rated paths within the conversation tree.

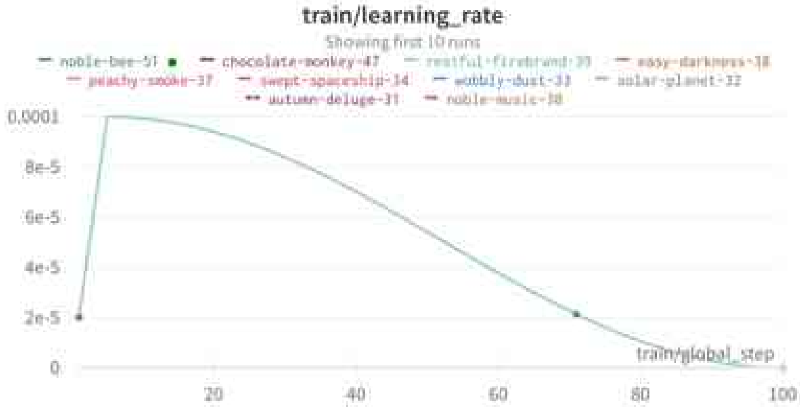


Figure 77: Learning Rate for Experiment 4

Experiment 4's learning rate adjustments followed a similar pattern to the previous experiments, with dynamic changes based on the dataset requirements. The learning rate's adaptive nature ensures efficient fine-tuning over the training iterations.

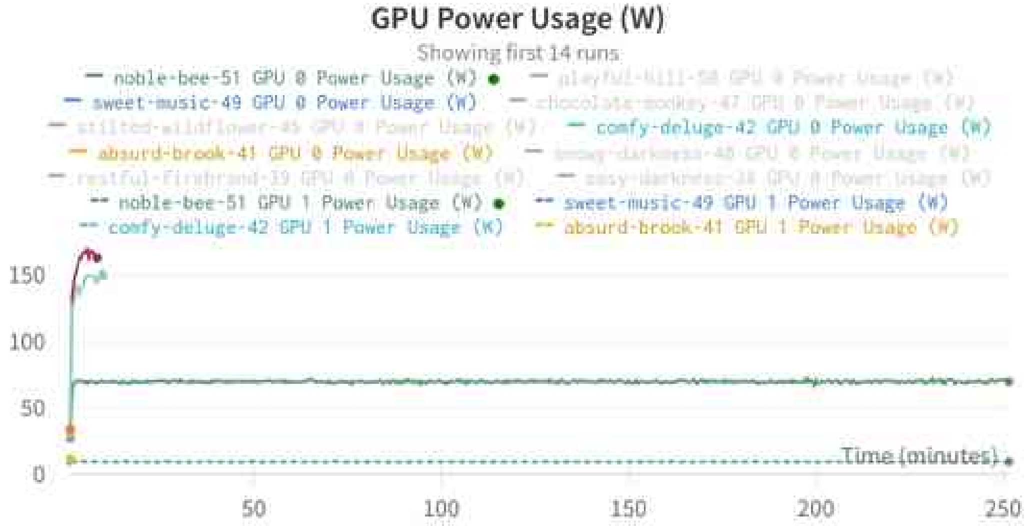


Figure 78: *GPU Power Usage for Experiment 4*

The GPU power usage in Experiment 4 remained relatively stable, starting from 0, remaining constant at 53 watts, and gradually increasing to 155 watts. This consistent power utilization showcases the model's efficiency in utilizing computational resources. Experiment 4, utilizing a subset of the RefinedWeb dataset, demonstrates a unique training pattern with an initial adjustment phase followed by significant improvement in learning effectiveness. The stable evaluation loss and adaptive learning rate adjustments indicate the model's reliability and adaptability to specific conversation paths. The consistent GPU power usage underscores the model's efficient resource utilization, essential for sustained training processes.

8.3 Conclusion

In conclusion, our study introduces a pioneering approach known as "Sample Gradient Descent," which harnesses the computational power of Principal Component Analysis (PCA) to strategically select a representative data sample, subsequently implementing batch gradient descent. This unique methodology not only significantly enhances computational efficiency but also maintains a commendable level of performance when compared to traditional batch gradient descent methodologies. The strategic incorporation of hyperparameters, complemented by their meticulous optimization through established techniques such as grid search or random search, contributes to the heightened adaptability and versatility of the proposed approach.

Moreover, our empirical findings shed light on the remarkable observation that the loss function of the sampled data converges at an accelerated rate, owing to the inherent convex nature of the selected data subset. This rapid convergence, in conjunction with a thorough comparative analysis against conventional gradient descent techniques, serves as a compelling testament to the effectiveness of our novel approach. We posit that the innovative "Sample Gradient Descent" technique holds immense promise for making a substantial impact across a spectrum of domains, including but not limited to machine learning and optimization. Its ability to bolster computational efficiency without compromising on performance is a notable stride forward in the realm of gradient-based optimization algorithms.

Additionally, the integration of hyperparameter tuning techniques further amplifies the adaptability and robustness of our proposed method. The meticulous experimentation conducted in this study not only enhances our understanding of large-scale language models but also unveils their inherent adaptability and efficacy across diverse tasks and datasets. These nuanced insights underscore the significant potential and practical applicability of such models in real-world scenarios, showcasing their resilience and versatility as indispensable tools in the field of data science and artificial intelligence.

Github: https://github.com/aadilganigaie/ChatAnalyzer/blob/main/LLMfinetune.ipynb
Github:
https://github.com/aadilganigaie/ChatAnalyzer/blob/main/langchainchatbot.ipynb
(LLM ChatBot)

Publication: 5

Chapter 9: Conclusion

9.1 Summary of Findings

The conclusion of this PhD thesis outlines the results of four different studies conducted on various topics related to the field of Natural Language Processing (NLP).

Study 1 investigated the prevalence of online harassment in digital media and found that an ensemble approach of fine-tuned pre-trained BERT models outperformed state-of-the-art models in terms of accuracy, but overfitting was observed with an increase in batch size and learning rate. The study demonstrates the efficacy of clustering attention mechanisms enriched by autoencoders in NLP, with superior efficiency in time and space complexity. Future research should refine the model, explore dynamic clustering techniques, and assess scalability and interpretability for real-world applications, presenting opportunities for advancing attention mechanisms in the intersection of deep learning, autoencoders, and NLP.

Study 2 and 3 The research explored the impact of informal language, including emoticons and slang, on sentiment analysis models applied to social media text. The study revealed that integrating emoticon and slang data had a marginal influence on model accuracy. Additionally, the investigation into filter size and the number of filters in CNNs emphasized the need for a delicate balance, with larger filters generally enhancing performance. These nuanced insights provide valuable guidance for NLP practitioners.

Study 4 and 5 Addressing the connection between hate speech and mental health, this study employed NLP techniques to detect hate speech in online discourse. Exposure to hate speech was linked to negative mental health outcomes, particularly affecting vulnerable populations. Advanced machine learning techniques, including the use of cosine similarity and a GPT2 chatbot, showcased promising results in predicting mental health statuses. Logistic regression outperformed other algorithms, achieving an 80% accuracy rate. The study's holistic approach offers a significant stride in the realm of mental health research.

Study 6 A novel technique, "Sample Gradient Descent," was introduced, enhancing computational efficiency in comparison to traditional batch gradient descent. Leveraging PCA for representative data selection and optimizing hyperparameters led to faster convergence rates. The approach showcased adaptability and effectiveness across diverse tasks and datasets, highlighting the robustness and versatility of large-scale language models.

Overall, the results of these studies provide valuable insights into various aspects of NLP and demonstrate the potential of machine learning techniques in solving real-world problems.

9.2 Contributions

The contributions of the PhD thesis are as follows:

Study 1: The thesis contributes to the field of online harassment detection by conducting a comprehensive preprocessing procedure and employing an ensemble approach of fine-tuned pre-trained BERT models, which outperforms state-of-the-art models. The study emphasizes how crucial preprocessing methods are to raising machine learning algorithms' efficiency. The study demonstrates the efficacy of clustering attention mechanisms enriched by autoencoders in NLP, with superior efficiency in time and space complexity. Future research should refine the model, explore dynamic clustering techniques, and assess scalability and interpretability for real-world applications, presenting opportunities for advancing attention mechanisms in the intersection of deep learning, autoencoders, and NLP.

Study 2 and 3: This study provides nuanced insights into the influence of informal language elements on sentiment analysis models and the interplay of filter size and quantity in CNNs. The findings underscore the resilient nature of sentiment analysis models, emphasizing the limited impact of informal language. Additionally, the study sheds light on the intricate balance required between filter size and the number of filters in CNNs, offering practical guidance for NLP practitioners.

Study 4 and 5: This study significantly contributes to the intersection of hate speech detection and mental health research. By employing advanced NLP techniques, the study establishes a clear link between exposure to hate speech and negative mental health outcomes, particularly among vulnerable populations. The incorporation of techniques like cosine similarity and a GPT2 chatbot showcases innovative approaches in predicting mental health statuses, advancing the field of mental health research and intervention.

Study 6: The introduction of "Sample Gradient Descent" presents a pioneering technique, enhancing computational efficiency without compromising performance. Leveraging PCA for data selection and optimizing hyperparameters, the approach provides a valuable solution to the challenges of slow convergence rates and high computation times in traditional gradient descent methods. The study's findings pave the way for improved machine learning optimization in various domains, showcasing the adaptability and effectiveness of large-scale language models.

9.3 Implications and Recommendations

9.3.1 Study 1 Implications and Recommendations:

9.3.1.1 Implications:

The study highlights the importance of preprocessing in the analysis of online harassment in digital media.

The results of the study indicate that conventional machine learning algorithms may not be effective in detecting online harassment.

The study highlights the potential of fine-tuned pre-trained BERT models for detecting online harassment in digital media.

The study demonstrates the efficacy of clustering attention mechanisms enriched by autoencoders in NLP, with superior efficiency in time and space complexity. Future research should refine the model, explore dynamic clustering techniques, and assess scalability and interpretability for real-world applications, presenting opportunities for advancing attention mechanisms in the intersection of deep learning, autoencoders, and NLP.

9.3.1.2 Recommendations:

Further research could focus on improving the preprocessing techniques used in the study.

Future studies could explore the use of other machine learning algorithms, such as transformer networks, to enhance the detection of online harassment.

Research could also be conducted to optimize the hyperparameters, such as batch size and learning rate, to prevent overfitting and improve accuracy.

9.3.2 Study 2 and 3 Implications and Recommendations:

9.3.2.1 Implications:

Our study reveals the limited impact of informal language elements like emoticons and slang on sentiment analysis models in social media text. The resilient nature of these models suggests their effectiveness in deciphering informal language nuances. Additionally, our findings emphasize the importance of dataset-specific considerations in determining optimal filter size and quantity for Convolutional Neural Networks (CNNs), challenging existing assumptions in the field.

9.3.2.2 Recommendations:

Future research should explore diverse forms of informal language, such as emojis and hashtags, to comprehensively understand their influence. Additionally, in-depth investigations into alternative model architectures like recurrent neural networks and transformer networks are warranted. Advanced techniques like transfer learning and ensemble methods should be explored for enhanced model adaptability and accuracy. Our study lays the groundwork for refining sentiment analysis models and encourages innovative approaches in Natural Language Processing research.

9.3.3 Study 4 and 5 Implications and Recommendations:

9.3.3.1 Implications:

This research sheds light on the alarming impact of hate speech on mental health, particularly within vulnerable communities. Utilizing Natural Language Processing (NLP) techniques, the study unveils the strong correlation between exposure to hate speech and negative mental health outcomes. The findings underscore the urgency of effective interventions, emphasizing the potential of NLP and machine learning methods. The study also highlights the significance of innovative techniques, such as cosine similarity and GPT2 chatbots, in enhancing mental health predictions and intervention strategies.

9.3.3.2 Recommendations:

Future research should delve into diverse machine learning algorithms and incorporate additional data sources for a more nuanced understanding. Exploring transfer learning techniques and addressing ethical implications in employing social media data are crucial for the field's advancement. The user-friendly interfaces' effectiveness, like Streamlit, suggests the need for further research in enhancing accessibility and usability. This study marks a significant step towards comprehensive mental health understanding and interventions, paving the way for a more empathetic and informed approach to mental well-being.

9.3.4 Study 6 Implications and Recommendations:

9.3.4.1 Implications:

This study introduces a pioneering method, "Sample Gradient Descent," which harnesses Principle Component Analysis (PCA) to select a representative data sample, followed by batch gradient descent. This innovative approach significantly enhances computational efficiency while preserving competitive performance compared to traditional batch gradient descent methods. The optimization of hyperparameters, facilitated through techniques like grid search or random search, further augments the approach's adaptability. The study's results highlight the accelerated convergence of the sampled data loss function, attributed to the function's convex nature. A comparative analysis with conventional gradient descent techniques reaffirms the efficacy of this approach. The research signifies a paradigm shift in fields like machine learning and optimization, offering enhanced computational efficiency without compromising performance quality.

9.3.4.2 Recommendations:

Future research avenues could explore the application of "Sample Gradient Descent" in various domains, examining its efficacy across diverse tasks and datasets. Fine-tuning the method by delving deeper into hyperparameter optimization strategies could further optimize its performance. Additionally, investigating its implementation in real-world scenarios and addressing potential challenges would offer valuable insights. This novel technique stands poised to revolutionize the landscape of large-scale language models, emphasizing their adaptability and effectiveness in practical applications. Continued exploration and refinement of this approach hold the promise of transformative impacts on multiple fields.

9.4 Future Work

- Further investigation of the relationship between language and sentiment: This thesis explored the impact of language on sentiment analysis, but future research could delve deeper into the relationship between language and sentiment. This could include exploring the effect of linguistic nuances, idioms, and cultural differences on sentiment analysis.

- Developing more robust machine learning models: Although the study identified logistic regression as the most proficient machine learning algorithm for forecasting mental health status, there exists an opportunity for future investigations to delve into the possibilities presented by alternative machine learning techniques like deep learning, ensemble methods, and neural networks. Such exploration holds the potential to yield sentiment analysis models that are not only more precise but also more dependable.

- Exploring the ethical and social implications of using social media data for sentiment analysis: As the use of social media data for sentiment analysis becomes more widespread, there is a need to examine the ethical and social implications of this practice. Future research could investigate the potential privacy concerns, bias, and discrimination that could arise from the use of social media data for sentiment analysis.

- Investigating the potential of sentiment analysis for other applications: While this thesis focused on the use of sentiment analysis for mental health prediction, there are other potential applications of sentiment analysis that could be explored. For example, sentiment analysis could be used to improve customer service, political analysis, or market research.

- Developing new techniques for sentiment analysis: This thesis introduced a novel approach to gradient descent termed "Sample Gradient Descent" for sentiment analysis. Subsequent research endeavors could investigate the efficacy of alternative optimization techniques, such as stochastic gradient descent or Bayesian optimization, in the realm of sentiment analysis. Such exploration could contribute valuable insights to the field.

- Developing more effective interventions for mental health: This thesis identified a potential relationship between hate speech and mental health, particularly in individuals with anxiety and bipolar disorders. The creation of interventions to lessen the detrimental effects of hate speech on mental health could be the subject of future study.

- Expanding the scope of sentiment analysis to other languages and cultures: This thesis focused on sentiment analysis for English language social media text. Future research could explore the potential of sentiment analysis for other languages and cultures, which could require the development of new techniques and models.

- Advancing Large Language Models (LLMs) for Multimodal Understanding: While this study has primarily focused on textual data, future research could explore the integration of Large Language Models (LLMs) with multimodal data, including images, videos, and audio. Developing LLMs capable of understanding and generating content across multiple modalities could revolutionize fields such as multimedia content creation, accessibility technologies, and human-computer interaction. Investigating techniques for seamless fusion of textual and non-textual information within LLMs could pave the way for more comprehensive and context-aware artificial intelligence systems.

Publications

Index	Article Title	Authors	Article Link	Journal Ranking	Citations	Impact Factor
1	Sentiment analysis on the effect of trending source less News: special reference to the recent death of an Indian actor	AG Ganie, S Dadvandipour	https://link.springer.com/chapter/10.1007/978-3-030-82322-1_1	Q4	4	0.209
2	Identification of online harassment using ensemble fine-tuned pre-trained Bert	AG Ganie, S Dadvandipour	https://akjournals.com/view/journals/606/17/3/article-p13.xml	Q3	0	0.26
3	Detection Of Semantic Obsessive Text In Multimedia Using Machine And Deep Learning Techniques And Algorithms	AG Ganie, S Dadvandipour, Aaqib Lone	https://www.jatit.org/volumes/Vol99No11/6Vol99No11.pdf	Q4	2	0.195
4	An Approach to Implementation of Autoencoders in	AG Ganie, S Dadvandipour	https://link.springer.com/chapter/10.1007/978-3-031-15211-5_1	Q4	0	0.16

	Intelligent Vehicles					
5	From big data to smart data: a sample gradient descent approach for machine learning	AG Ganie, S Dadvandipour	https://link.springer.com/article/10.1186/s40537-023-00839-9	Q1, D1	0	10.835
6	Presence of informal language, such as emoticons, hashtags, and slang, impact the performance of sentiment analysis models on social media text?	AG Ganie, S Dadvandipour	https://arxiv.org/abs/2301.12303	NA	2	NA
7	Traditional or deep learning for sentiment analysis: A review	AG Ganie, S Dadvandipour	https://ojs.uni-miskolc.hu/index.php/multi/article/view/777	NA	1	NA
8	Media Trail: Rise of Contradictive Reaction in India during COVID-19	AG Ganie, S Dadvandipour	https://www.inf.u-szeged.hu/~cscs/proceedings.php	NA	0	NA
9	Exploring The Impact of Informal Language on Sentiment Analysis Models for Social	AG Ganie, S Dadvandipour	Accepted for publication	NA	0	NA

	Media Text Using Convolutional Neural Networks					

Other Publications

Index	Title	Authors	URL	Citations
10	Analyzing and predicting spear-phishing using machine learning methods	AG Ganie, S Dadvandipour	https://ojs.uni-miskolc.hu/index.php/multi/article/view/655	4
11	Optimization of PSWAN in Terms of Cost and Bandwidth	AG Ganie, S Dadvandipour	https://ejournal.uksw.edu/ijiteb/article/view/1893	2
12	Private network optimization	AG Ganie	https://ojs.uni-miskolc.hu/index.php/multi/article/view/752	1
13	Covid-19 Situation In Hungary Using Time Series Analysis		https://rgnpublications.com/ICACTCE2021/manuscripts/002-10.pdf	NA
15	An end-to-end machine learning approach for classifying cancer based on DNA sequence data	AG Ganie, S Dadvandipour	Accepted for publication, Q3, IF = 0.82	NA

Chapter 10: References

[1] Meiselwitz, G. (Ed.). (2018). Social Computing and Social Media. User Experience and Behavior: 10th International Conference, SCSM 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part I (Vol. 10913). Springer.

[2] Meiselwitz, G. (Ed.). (2019). Social Computing and Social Media. Design, Human Behavior and Analytics: 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part I (Vol. 11578). Springer.

[3] Meiselwitz, G. (Ed.). (2020). Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis: 12th International Conference, SCSM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I (Vol. 12194). Springer Nature.

[4] Chavan, V. S., & Shylaja, S. S. (2015, August). Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2354-2358). IEEE.

[5] Meiselwitz, G. (2015). Social Computing and Social Media. Springer International Publishing.

[6] Meiselwitz, G. (Ed.). (2016). Social Computing and Social Media: 8th International Conference, SCSM 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17–22, 2016. Proceedings (Vol. 9742). Springer.

[7] Meiselwitz, G. (Ed.). (2016). Social Computing and Social Media: 8th International Conference, SCSM 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17–22, 2016. Proceedings (Vol. 9742). Springer.

[8] Aguzzoli, S., Codara, P., Flaminio, T., Gerla, B., & Valota, D. (2015). Querying with {\ L} ukasiewicz logic. *ArXiv Preprint ArXiv:1512.01041*.

[9] Al-Hafadhi, M. H., & Krallics, G. (2021). Prediction and numerical simulation of residual stress in multi-pass pipe welds. *Pollack Periodica*. https://doi.org/10.1556/606.2020.00127

[10] AlBadani, B., Shi, R., & Dong, J. (2022). A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM. *Applied System Innovation*. https://doi.org/10.3390/asi5010013

[11] Alex Kolchinski, Y., & Potts, C. (2018). Representing social media users for sarcasm detection. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*. https://doi.org/10.18653/v1/d18-1140

[12] Almouzini, S., & Alageel, A. (2019). Detecting arabic depressed users from twitter data. *Procedia Computer Science*, *163*, 257–265.

[13] Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., & Vilares, J. (2021). Sentiment analysis for fake news detection. In *Electronics (Switzerland)*. https://doi.org/10.3390/electronics10111348

[14] Alsayat, A. (2022). Improving Sentiment Analysis for Social Media Applications Using an Ensemble Deep Learning Language Model. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-021-06227-w

[15] Baccianella, S., Esuli, A., & Sebastiani, F. (2008). S ENTI W ORD N ET 3 . 0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the 7th Conference on Language Resources and Evaluation LREC10*.

[16] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. *26th International World Wide Web Conference 2017, WWW 2017 Companion*. https://doi.org/10.1145/3041021.3054223

[17] Basyal, L., & Sanghvi, M. (2023). Text Summarization Using Large Language Models:

A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models. *ArXiv Preprint ArXiv:2310.10449*.

[18] Ben-David, A., & Matamoros-Fernández, A. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*.

[19] Bhavnani, R., Findley, M. G., & Kuklinski, J. H. (2009). Rumor dynamics in ethnic violence. *Journal of Politics*. https://doi.org/10.1017/S002238160909077X

[20] Birnbaum, M. L., Rizvi, A. F., Correll, C. U., Kane, J. M., & Confino, J. (2017). Role of social media and the Internet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders. *Early Intervention in Psychiatry*. https://doi.org/10.1111/eip.12237

[21] Bottou, L. (2012). Stochastic Gradient Descent Tricks BT - Neural Networks: Tricks of the Trade. In *Neural Networks: Tricks of the Trade*.

[22] Briand, A., Almeida, H., & Meurs, M.-J. (2018). Analysis of social media posts for early detection of mental health conditions. *Advances in Artificial Intelligence: 31st Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, May 8–11, 2018, Proceedings 31*, 133–143.

[23] Brunette, M. F., Achtyes, E., Pratt, S., Stilwell, K., Opperman, M., Guarino, S., & Kay-Lambkin, F. (2019). Use of Smartphones, Computers and Social Media Among People with SMI: Opportunity for Intervention. *Community Mental Health Journal*. https://doi.org/10.1007/s10597-019-00431-7

[24] Bucci, S., Schwannauer, M., & Berry, N. (2019). The digital revolution and its impact on mental health care. *Psychology and Psychotherapy: Theory, Research and Practice*. https://doi.org/10.1111/papt.12222

[25] Burbach, L., & Ziefle, M. (2018). Social Computing and Social Media. User Experience and Behavior. In *International Conference on Social Computing and Social Media*.

[26] Burnap, P., Colombo, G., & Scourfield, J. (2015). Machine classification and analysis of suicide-related communication on Twitter. *HT 2015 - Proceedings of the 26th ACM Conference on Hypertext and Social Media*. https://doi.org/10.1145/2700171.2791023

[27] Burrows, H., Zarrin, J., Babu-Saheer, L., & Maktab-Dar-oghaz, M. (2022). Realtime emotional reflective user interface based on deep convolutional neural networks and generative adversarial networks. *Electronics (Switzerland)*. https://doi.org/10.3390/electronics11010118

[28] Cabrera, A. F., Nora, A., Terenzini, P. T., Pascarella, E., & Hagedorn, L. S. (1999). Campus racial climate and the adjustment of students to college: A comparison between White students and African-American students. *The Journal of Higher Education*, *70*(2), 134–160.

[29] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*. https://doi.org/10.1126/science.aal4230

[30] Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*. https://doi.org/10.1017/S1351324916000383

[31] Camras, L., & Plutchik, R. (1980). Emotion: A Psychoevolutionary Synthesis. *The American Journal of Psychology*. https://doi.org/10.2307/1422394

[32] Castillo-Sánchez, G., Marques, G., Dorronzoro, E., Rivera-Romero, O., Franco-Martín, M., & De la Torre-Díez, I. (2020). Suicide Risk Assessment Using Machine Learning and Social Networks: a Scoping Review. In *Journal of Medical Systems*. https://doi.org/10.1007/s10916-020-01669-5

[33] Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M. B., & De Choudhury, M. (2019). A taxonomy of ethical tensions in inferring mental health states from social media. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability,*

and Transparency. https://doi.org/10.1145/3287560.3287587

[34] Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. In *npj Digital Medicine*. https://doi.org/10.1038/s41746-020-0233-7

[35] Chancellor, S., Mitra, T., & De Choudhury, M. (2016). *Recovery Amid Pro-Anorexia*. https://doi.org/10.1145/2858036.2858246

[36] Chaudhry, I., & Gruzd, A. (2020). Expressing and Challenging Racist Discourse on Facebook: How Social Media Weaken the "Spiral of Silence" Theory. *Policy and Internet*. https://doi.org/10.1002/poi3.197

[37] Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*. https://doi.org/10.1111/jcom.12104

[38] Coppersmith, G., Dredze, M., & Harman, C. (2015). *Quantifying Mental Health Signals in Twitter*. https://doi.org/10.3115/v1/w14-3207

[39] Coppersmith, G., Ngo, K., Leary, R., & Wood, A. (2016). Exploratory analysis of social media prior to a suicide attempt. *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych 2016 at the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lan*. https://doi.org/10.18653/v1/w16-0311

[40] Crispino, N., Montgomery, K., Zeng, F., Song, D., & Wang, C. (2023). Agent Instructs Large Language Models to be General Zero-Shot Reasoners. *ArXiv Preprint ArXiv:2310.03710*.

[41] *Cyberbullying: What is it and how to stop it | UNICEF*. (n.d.). Retrieved February 13, 2023, from https://www.unicef.org/end-violence/how-to-stop-cyberbullying

[42] Dadvar, M., Trieschnigg, D., Ordelman, R., & De Jong, F. (2013). Improving cyberbullying detection with user context. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-642-36973-5_62

[43] De Ávila Berni, G., Rabelo-Da-Ponte, F. D., Librenza-Garcia, D., Boeira, M. V., Kauer-Sant'Anna, M., Passos, I. C., & Kapczinski, F. (2018). Potential use of text classification tools as signatures of suicidal behavior: A proof-of-concept study using Virginia woolf's personal writings. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0204820

[44] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*. https://doi.org/10.1609/icwsm.v7i1.14432

[45] Derks, D., Bos, A. E. R., & Grumbkow, J. von. (2007). Emoticons and social interaction on the Internet: the importance of social context. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2004.11.013

[46] Deshpande, M., & Rao, V. (2017). Depression detection using emotion artificial intelligence. *2017 International Conference on Intelligent Sustainable Systems (Iciss)*, 858–862.

[47] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*.

[48] Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. *AAAI Workshop - Technical Report*. https://doi.org/10.1609/icwsm.v5i3.14209

[49] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., & Chen, W. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language

models. *Nature Machine Intelligence*, *5*(3), 220–235.

[50] Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web*. https://doi.org/10.1145/2740908.2742760

[51] *Facebook is using AI to spot users with suicidal thoughts - The Verge*. (n.d.). Retrieved January 19, 2023, from https://www.theverge.com/2017/11/28/16709224/facebook-suicidal-thoughts-ai-help

[52] Franco-Martín, M. A., Muñoz-Sánchez, J. L., Sainz-de-Abajo, B., Castillo-Sánchez, G., Hamrioui, S., & de la Torre-Díez, I. (2018). A Systematic Literature Review of Technologies for Suicidal Behavior Prevention. *Journal of Medical Systems*. https://doi.org/10.1007/s10916-018-0926-5

[53] Ganie, A. G., & Dadvandipour, S. (2021). Sentiment Analysis on the Effect of Trending Source Less News: Special Reference to the Recent Death of an Indian Actor. *Communications in Computer and Information Science*. https://doi.org/10.1007/978-3-030-82322-1_1

[54] Ganie, A. G., & Dadvandipour, S. (2022). Identification of online harassment using ensemble fine-tuned pre-trained Bert. *Pollack Periodica*.

[55] Ganie, A. G., Dadvandipour, S., & Lone, M. A. (2021). Detection of semantic obsessive text in multimedia using machine and deep learning techniques and algorithms. *Journal of Theoretical and Applied Information Technology*.

[56] Gautam, G., & Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. *2014 Seventh International Conference on Contemporary Computing (IC3)*, 437–442.

[57] Giuntini, F. T., Cazzolato, M. T., dos Reis, M. de J. D., Campbell, A. T., Traina, A. J. M., & Ueyama, J. (2020). A review on recognizing depression in social networks: challenges and opportunities. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-020-01726-4

[58] Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *Processing*.

[59] González-Ibánez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in twitter: a closer look. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 581–586.

[60] Goodfellow I., B. Y. (2016). Courville A-Deep learning-MIT (2016). *Nature*.

[61] Guo, Q., Qiu, X., Liu, P., Xue, X., & Zhang, Z. (2020). Multi-scale self-attention for text classification. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v34i05.6290

[62] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don’t stop pretraining: Adapt language models to domains and tasks. *ArXiv Preprint ArXiv:2004.10964*.

[63] Hanzelka, J., & Schmidt, I. (2017). Dynamics of cyber hate in social media: A comparative analysis of anti-muslim movements in the Czech Republic and Germany. *International Journal of Cyber Criminology*. https://doi.org/10.5281/zenodo.495778

[64] Hao, B., Li, L., Li, A., & Zhu, T. (2013). Predicting mental health status on social media. *International Conference on Cross-Cultural Design*, 101–110.

[65] Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications*, *23*(1), 11.

[66] He, L., & Luo, J. (2016). “What makes a pro eating disorder hashtag”: Using hashtags to identify pro eating disorder tumblr posts and Twitter users. *2016 IEEE International Conference on Big Data (Big Data)*, 3977–3979.

[67] Herring, S. C., Stein, D., & Virtanen, T. (2013). Introduction to the pragmatics of

computer-mediated communication. In *Pragmatics of Computer-Mediated Communication*. https://doi.org/10.1515/9783110214468.3

[68] Hill, J. (2016). The Impact of Emojis and Emoticons on Online Consumer Reviews , Perceived Company Response Quality , Brand Relationship , and Purchase Intent . *University of South Florida Scholar Commons Graduate*.

[69] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de Las, Hendricks, L. A., Welbl, J., & Clark, A. (2022). Training compute-optimal large language models. *ArXiv Preprint ArXiv:2203.15556*.

[70] *Homepage - Közgazdaság- és Regionális Tudományi Kutatóközpont*. (n.d.). Retrieved January 8, 2023, from https://krtk.hu/en/homepage/

[71] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/1014052.1014073

[72] Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*. https://doi.org/10.1609/icwsm.v8i1.14550

[73] ICNC-FSKD 2017 - 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. (2018). In *ICNC-FSKD 2017 - 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*.

[74] Isnain, A. R., Sihabuddin, A., & Suyanto, Y. (2020). Bidirectional Long Short Term Memory Method and Word2vec Extraction Approach for Hate Speech Detection. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*. https://doi.org/10.22146/ijccs.51743

[75] J.A., N., K.A., A., L.A., M., & S.J., B. (2016). The future of mental health care: Peer-To-peer support and social media. *Epidemiology and Psychiatric Sciences*.

[76] Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2021). Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications. *IEEE Transactions on Computational Social Systems*. https://doi.org/10.1109/TCSS.2020.3021467

[77] K.A., A., J.A., N., T., G., J.C.M., B., S.J., B., & M., B. (2018). A Survey of Online and Mobile Technology Use at Peer Support Agencies. *Psychiatric Quarterly*.

[78] Karaoğlan, K. M., & Fındık, O. (2022). Extended rule-based opinion target extraction with a novel text pre-processing method and ensemble learning. *Applied Soft Computing*. https://doi.org/10.1016/j.asoc.2022.108524

[79] *Kepios*. (n.d.). Retrieved January 19, 2023, from https://kepios.com/

[80] Khan, A., Husain, M. S., & Khan, A. (2018). Analysis of Mental State of Users using Social Media to predict Depression! A Survey. *International Journal Of*.

[81] Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, *10*(1), 1–6.

[82] Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. BT - 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. In *International Conference on Learning Representations (ICLR)*.

[83] Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*. https://doi.org/10.1613/jair.4272

[84] Krishnamoorthy, S., Dua, A., & Gupta, S. (2021). Role of emerging technologies in future IoT-driven Healthcare 4.0 technologies: a survey, current challenges and future directions. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-021-03302-w

[85] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep

convolutional neural networks. *Advances in Neural Information Processing Systems*.

[86] Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. *Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013*. https://doi.org/10.1609/aaai.v27i1.8539

[87] Le, B., & Nguyen, H. (2015). Twitter sentiment analysis using machine learning techniques. In *Advanced Computational Methods for Knowledge Engineering* (pp. 279–289). Springer.

[88] Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. *Proceedings of the National Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v29i1.9491

[89] Liu, B. (2012). Sentiment Analysis and Opinion Mining Mining. Synthesis Lectures on Human Language Technologies. [Draft]. In *Synthesis Lectures on Human Language Technologies*.

[90] Liu, C., Fang, F., Lin, X., Cai, T., Tan, X., Liu, J., & Lu, X. (2021). Improving sentiment analysis accuracy with emoji embedding. *Journal of Safety Science and Resilience*, *2*(4), 246–252.

[91] Liu, J., Pan, Y., Li, M., Chen, Z., Tang, L., Lu, C., & Wang, J. (2018). Applications of deep learning to MRI Images: A survey. In *Big Data Mining and Analytics*. https://doi.org/10.26599/BDMA.2018.9020001

[92] Liu, L., Priestley, J. L., Zhou, Y., Ray, H. E., & Han, M. (2019). A2Text-net: A novel deep neural network for sarcasm detection. *Proceedings - 2019 IEEE 1st International Conference on Cognitive Machine Intelligence, CogMI 2019*. https://doi.org/10.1109/CogMI48466.2019.00025

[93] Mahdy, N., Magdi, D. A., Dahroug, A., & Rizka, M. A. (2020). Comparative study: Different techniques to detect depression using social media. In *Lecture Notes in Networks and Systems*. https://doi.org/10.1007/978-981-15-3075-3_30

[94] Malte, A., & Ratadiya, P. (2019). Multilingual Cyber Abuse Detection using Advanced Transformer Architecture. *IEEE Region 10 Annual International Conference, Proceedings/TENCON*. https://doi.org/10.1109/TENCON.2019.8929493

[95] Manohar, M. Y., & Kulkarni, P. (2017). Improvement sarcasm analysis using NLP and corpus based approach. *Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems, ICICCS 2017*. https://doi.org/10.1109/ICCONS.2017.8250536

[96] Matsuda, M. J. (2018). *Words that wound: Critical race theory, assaultive speech, and the first amendment*. Routledge.

[97] Mbarek, A., Jamoussi, S., Charfi, A., & Ben Hamadou, A. (2019). Suicidal profiles detection in twitter. *WEBIST 2019 - Proceedings of the 15th International Conference on Web Information Systems and Technologies*. https://doi.org/10.5220/0008167602890296

[98] *Mental Health and Substance Use*. (n.d.). Retrieved December 26, 2022, from https://www.who.int/teams/mental-health-and-substance-use

[99] *Mental health statistics 2022*. (n.d.). Retrieved January 19, 2023, from https://www.singlecare.com/blog/news/mental-health-statistics/

[100] Mishra, P., Del Tredici, M., Yannakoudakis, H., & Shutova, E. (2018). Author profiling for abuse detection. *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*.

[101] Mitchell, M., Hollingshead, K., & Coppersmith, G. (2015). Quantifying the Language of Schizophrenia in Social Media. *2nd Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych 2015 - Proceedings of the Workshop*. https://doi.org/10.3115/v1/w15-1202

[102] Modha, S., Majumder, P., & Mandl, T. (2018). Filtering Aggression from the Multilingual Social Media Feed. *COLING 2018 - 1st Workshop on Trolling, Aggression*

and Cyberbullying, TRAC 2018 - Proceedings of the Workshop.

[103] Mukherjee, S. S., Yu, J., Won, Y., McClay, M. J., Wang, L., Rush, A. J., & Sarkar, J. (2020). Natural Language Processing-Based Quantication of the Mental State of Psychiatric Patients. *Computational Psychiatry*. https://doi.org/10.1162/cpsy_a_00030

[104] Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551.

[105] Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 1–5.

[106] NESTEROV, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence o(1/k^2). *Doklady AN USSR*.

[107] Nijhawan, T., Attigeri, G., & Ananthakrishna, T. (2022). Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data*. https://doi.org/10.1186/s40537-022-00575-6

[108] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. *25th International World Wide Web Conference, WWW 2016*. https://doi.org/10.1145/2872427.2883062

[109] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. https://doi.org/10.1561/1500000011

[110] Paraschiv, A., & Cercel, D. C. (2020). UPB at GermEval-2019 task 2: BERT-based offensive language classification of German tweets. *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019*.

[111] Park, J. H., & Fung, P. (2017). *One-step and Two-step Classification for Abusive Language Detection on Twitter*. https://doi.org/10.18653/v1/w17-3006

[112] Paul, M., & Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 265–272.

[113] Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2016.06.009

[114] Potamias, R. A., Siolas, G., & Stafylopatis, A.-G. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, *32*(23), 17309–17320.

[115] Prakash, A., Agarwal, K., Shekhar, S., Mutreja, T., & Chakraborty, P. S. (2021). An ensemble learning approach for the detection of depression and mental illness over twitter data. *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, 565–570.

[116] Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016. (2016). In *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*.

[117] Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017. (2018). In *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*.

[118] Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018. (2019). In *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*.

[119] Proceedings - 2018 IEEE International Conference on Big Data and Smart Computing, BigComp 2018. (2018). In *Proceedings - 2018 IEEE International Conference on Big Data and Smart Computing, BigComp 2018*.

[120] Rajamanickam, S., Mishra, P., Yannakoudakis, H., & Shutova, E. (2020). *Joint Modelling of Emotion and Abusive Language Detection*.

https://doi.org/10.18653/v1/2020.acl-main.394

[121] Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science*. https://doi.org/10.1140/epjds/s13688-017-0110-z

[122] Rehm, J., & Shield, K. D. (2019). Global Burden of Disease and the Impact of Mental and Addictive Disorders. In *Current Psychiatry Reports*. https://doi.org/10.1007/s11920-019-0997-0

[123] Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. *Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011*. https://doi.org/10.1109/ICMLA.2011.152

[124] Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

[125] Ríssola, E. A., Losada, D. E., & Crestani, F. (2021). A survey of computational methods for online mental state assessment on social media. In *ACM Transactions on Computing for Healthcare*. https://doi.org/10.1145/3437259

[126] Ryu, S., Lee, H., Lee, D. K., Kim, S. W., & Kim, C. E. (2019). Detection of suicide attempters among suicide ideators using machine learning. *Psychiatry Investigation*. https://doi.org/10.30773/pi.2019.06.19

[127] Saleem, S., Prasad, R., Vitaladevuni, S., Pacula, M., Crystal, M., Marx, B., Sloan, D., Vasterling, J., & Speroff, T. (2012). Automatic detection of psychological distress indicators and severity assessment from online forum posts. *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*.

[128] Samad, D., & Gani, G. A. (2020). Analyzing and predicting spear-phishing using machine learning methods. *Multidiszciplináris Tudományok*, *10*(4), 262–273.

[129] Severyn, A., & Moschitti, A. (2015a). Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks Categories and Subject Descriptors. *Sigir*.

[130] Severyn, A., & Moschitti, A. (2015b). UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification. *SemEval 2015 - 9th International Workshop on Semantic Evaluation, Co-Located with the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2015 - Proceedings*. https://doi.org/10.18653/v1/s15-2079

[131] Sheehan, L., Dubke, R., & Corrigan, P. W. (2017). The specificity of public stigma: A comparison of suicide and depression-related stigma. *Psychiatry Research*. https://doi.org/10.1016/j.psychres.2017.06.015

[132] Shewchuk, J. R. (2019). An Introduction to Conjugate Gradient Descent without the Agonizing Pain. *Numerical Analysis*.

[133] Shing, H. C., Nair, S., Zirikly, A., Friedenberg, M., Daumé, H., & Resnik, P. (2018). Expert, crowdsourced, and machine assessment of suicide risk via online postings. *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, CLPsych 2018 at the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*. https://doi.org/10.18653/v1/w18-0603

[134] Singh, J., Singh, G., & Singh, R. (2017). Optimization of sentiment analysis using machine learning classifiers. *Human-Centric Computing and Information Sciences*, *7*(1), 32.

[135] Singh, V., Varshney, A., Akhtar, S. S., Vijay, D., & Shrivastava, M. (2018). Aggression Detection on Social Media Text Using Deep Neural Networks. *2nd Workshop on Abusive Language Online - Proceedings of the Workshop, Co-Located with EMNLP 2018*. https://doi.org/10.18653/v1/w18-5106

[136] Skaik, R., & Inkpen, Di. (2021). Using Social Media for Mental Health Surveillance:

A Review. In *ACM Computing Surveys*. https://doi.org/10.1145/3422824

[137] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

[138] Spertus, E. (1997). Smokey: automatic recognition of hostile messages. *Innovative Applications of Artificial Intelligence - Conference Proceedings*.

[139] Ştefãniţă, O., & Buf, D. M. (2021). Hate speech in social media and its effects on the lgbt community: A review of the current research1. In *Romanian Journal of Communication and Public Relations*. https://doi.org/10.21018/rjcpr.2021.1.322

[140] Sunstein, C. R., & Vermeule, A. (2008). *Conspiracy theories*.

[141] *Sushant Singh Rajput case: CBI gets no proof of murder, now focusing on the suicide angle | Hindi Movie News - Times of India*. (n.d.). Retrieved February 13, 2023, from https://timesofindia.indiatimes.com/entertainment/hindi/bollywood/news/sushant-singh-rajput-case-cbi-gets-no-proof-of-murder-now-focusing-on-the-suicide-angle/articleshow/77883136.cms

[142] Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. *30th International Conference on Machine Learning, ICML 2013*.

[143] T.Sajana, M. G. (2021). Human Behavior Prediction and Analysis Using Machine Learning-A Review. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*. https://doi.org/10.17762/turcomat.v12i5.1499

[144] Tan, Z., Chen, J., Kang, Q., Zhou, M., Abusorrah, A., & Sedraoui, K. (2022). Dynamic Embedding Projection-Gated Convolutional Neural Networks for Text Classification. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2020.3036192

[145] *The American Trends Panel | Pew Research Center*. (n.d.). Retrieved February 13, 2023, from https://www.pewresearch.org/our-methods/u-s-surveys/the-american-trends-panel/

[146] *The State of Online Harassment | Pew Research Center*. (n.d.). Retrieved February 13, 2023, from https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

[147] Torous, J., Keshavan, M., & Gutheil, T. (2014). Promise and perils of digital psychiatry. In *Asian Journal of Psychiatry*. https://doi.org/10.1016/j.ajp.2014.06.006

[148] Tran, T. K., & Phan, T. T. (2019). Deep learning application to ensemble learning—the simple, but effective, approach to sentiment classifying. *Applied Sciences (Switzerland)*. https://doi.org/10.3390/APP9132760

[149] Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015). Recognizing depression from twitter activity. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/2702123.2702280

[150] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. In *Social Science Computer Review*. https://doi.org/10.1177/0894439310386557

[151] *Under the hood: Suicide prevention tools powered by AI - Engineering at Meta*. (n.d.). Retrieved January 19, 2023, from https://engineering.fb.com/2018/02/21/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/

[152] *Understanding LSTM Networks -- colah's blog*. (n.d.). Retrieved February 13, 2023, from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[153] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

[154] Wang, C. (2018). Interpreting Neural Network Hate Speech Classifiers. *2nd*

Workshop on Abusive Language Online - Proceedings of the Workshop, Co-Located with EMNLP 2018. https://doi.org/10.18653/v1/w18-5111

[155] Wang, Q., Zhou, Y., Ruan, T., Gao, D., Xia, Y., & He, P. (2019). Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2019.103133

[156] Wang, T., Brede, M., Ianni, A., & Mentzakis, E. (2017). Detecting and characterizing eating-disorder communities on social media. *WSDM 2017 - Proceedings of the 10th ACM International Conference on Web Search and Data Mining*. https://doi.org/10.1145/3018661.3018706

[157] Wang, Y., Yang, Y., Chen, Y., Bai, J., Zhang, C., Su, G., Kou, X., Tong, Y., Yang, M., & Zhou, L. (2020). TextNAS: A neural architecture search space tailored for text representation. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v34i05.6462

[158] Wankhede, S., Patil, R., Sonawane, S., & Save, P. A. (2018). Data Preprocessing for Efficient Sentimental Analysis. *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*. https://doi.org/10.1109/ICICCT.2018.8473277

[159] Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *HLT-NAACL 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Student Research Workshop*. https://doi.org/10.18653/v1/n16-2013

[160] Wawre, S. V, & Deshmukh, S. N. (2016). Sentiment classification using machine learning techniques. *International Journal of Science and Research (IJSR)*, *5*(4), 819–821.

[161] Weller, O., Sagers, L., Hanson, C., Barnes, M., Snell, Q., & Shannon Tass, E. (2021). Predicting suicidal thoughts and behavior among adolescents using the risk and protective factor framework: A large-scale machine learning approach. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0258535

[162] Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2017). Researching mental health disorders in the era of social media: Systematic review. In *Journal of Medical Internet Research*. https://doi.org/10.2196/jmir.7215

[163] Xu, J., Chen, D., Qiu, X., & Huang, X. (2016). Cached long short-term memory neural networks for document-level sentiment classification. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. https://doi.org/10.18653/v1/d16-1172.

[164] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1253.

[165] Zhang, X., Zhao, J., & Lecun, Y. (2015). Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*.

[166] Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *ArXiv Preprint ArXiv:1510.03820*.

[167] Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *Semantic Web*. https://doi.org/10.3233/SW-180338

[168] Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-93417-4_48

[169] Zhao, Z., Sun, J., Mao, Z., Feng, S., & Bao, Y. (2016). Determining the topic

hashtags for chinese microblogs based on 5W model. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-42553-5_5

[170] Ziwei, B. Y., & Chua, H. N. (2019a). An application for classifying depression in tweets. *Proceedings of the 2nd International Conference on Computing and Big Data*, 37–41.

[171] Ziwei, B. Y., & Chua, H. N. (2019b). An application for classifying depression in tweets. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3366650.3366653.

[172] "Mental Health and Substance Use." [Online]. Available: https://www.who.int/teams/mental-health-and-substance-use. [Accessed: 26-Dec-2022].

[173] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, 2013.

Additional GIT HUB Repositories

1. https://github.com/aadilganigaie/LangchainLLMwithmultiplePdfs
2. https://github.com/aadilganigaie/ML-project
3. https://rb.gy/l8dq0q
4. https://github.com/aadilganigaie/SpamClassifier