INTERNSHIP REPORT

A report submitted in partial fulfillment of the requirements for the Award of Degree of

BACHELOR OF ENGINEERING

in

COMPUTER ENGINEERING

BY STUDENT

Mr. Aryan Deepak Parishwad

(72292642B)

UNDER SUPERVISION OF

PROF.(MR).AMOL S. KAMBALE

DURATION (JANUARY 2024 - APRIL 2024)

Topic: Twitter Sentiment Analysis using KNN and its Variants

DEPARTMENT OF COMPUTER ENGINEERING

Modern Education Society's Wadia College of Engineering, Pune

APPROVED BY AICTE,

Affiliated to SPPU, Pune

MAHARASHTRA

Department of Computer Engineering

MES's Wadia College Engineering,

19, Bund Garden, V. K. Joag Path, Pune-411 001

Certificate of Internship

This is to certify that **Aryan Deepak Parishwad**, a student of the MES's Wadia College of Engineering, Pune has completed a Research internship in the field of Computer Engineering from **2023** to **2024,** entitled **Twitter sentiment analysis using KNN and its Variants** under the guidance of **PROF. (MR). AMOL S. KAMBALE**.

During the period of her/his internship program with us, she/he had been exposed to different processes and was found diligent, hardworking and inquisitive.
We wish her/his every success in his life and career.

Research Internship Guide Internship Coordinator HOD

ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude towards my internship guide **Prof.(Mr) Amol S. Kamble** for his support, continuous guidance and being so un- derstanding and helpful throughout the Internship.

I furthermore thank Computer Department HOD **Dr. (Mrs). N.F. Shaikh** to encourage me to go ahead and for continuous guidance. I also want to thank **Prof. Mr. G.B. Aoachar** and **Dr. (Mrs). S.P.Deore** for all her assistance and guidance for preparing report.

I would like to thank all those, who have directly or indirectly helped me for the completion of the work during this Internship.

Mr. Aryan Deepak Parishwad

ABSTRACT

Company Information:

The research internship program at Modern Education Society's Wadia College of Engineering, Pune, a unique opportunity for students to engage in hands-on research experiences under the guidance of esteemed faculty members and researchers. This program aims to cultivate research skills, foster intellectual growth, and promote a deeper understanding of the research process.

Programs and Opportunities:

As part of this research internship, we are undertaking a project on Comparison of KNN Variants on Twitter Sentiment Analysis. The project focuses on designing **Twitter Sentiment Analysis using KNN and its Variants** and make this process more efficient than before.

Through this internship, we have been able to apply theoretical concepts learned in the classroom to real-world scenarios, gain practical research skills, and contribute to the ongoing efforts in the management system.

Key parts of Report:

1. Internship Work Identification
2. Technologies used in the Internship
3. Weekly overview of Internship Activities
4. Motivation

Benefits of the Institution from the Report:

1. Resume Building
2. Career Guidance
3. Growth of Skill

Contents

List of Figures **6**

List of Tables **7**

WEEKLY OVERVIEW OF INTERNSHIP **10**

1 INTRODUCTION **12**

2 PROBLEM STATEMENT AND OBJECTIVES **14**

3 Motivation **16**

4 Literature Survey **18**

5 SYSTEM ANALYSIS **19**
5.1 REQUIREMENT ANALYSIS 19
5.1.1 Existing System 21
5.1.2 Proposed System 22

6 SOFTWARE REQUIREMENT SPECIFICATIONS **23**
6.1 System Configuration 23
6.1.1 Software Requirements 23
6.1.2 Hardware Requirements 24

7 METHODOLOGICAL DETAILS **25**

8 RESULT **27**

9 CONCLUSION **30**

10 BIBLIOGRAPHY **31**

List of Figures

7.1 SYSTEM ARCHITECTURE 27
8.1 ACCURACY OF SIMPLE KNN 27
8.2 ACCURACY OF LM-KNN (LOCAL MEAN-KNN). 28
8.3 CONFUSION MATRIX OF KNN AND LM-KNN 28
8.4 PRECISION, RECALL AND F1 SCORE OF KNN AND LM-KNN 29
8.5 COMPARISON OF VARIANTS BASED ON THEIR ACCURACY 29

List of Tables

8.1 ACCURACY TABLE OF KNN AND ITS VARIANTS USING ‘K’ VALUE…. 28

- **GENERAL GUIDELINES AND INSTRUCTIONS:**

1. Internships are educational and career development opportunities, providing practical experience in a field or discipline. Internships are far more important as the employers are looking for employees who are properly skilled and having awareness about industry environment, practices and culture. Internship is structured, short-term, supervised training often focused around particulartasks or projects with defined time scales.
2. Core objective is to expose technical students to the industrial environment, which cannot be simulated/experienced in the classroom and hence creating competent professionals in the industryand to understand the social, economic and administrative considerations that influence the workingenvironment of industrial organizations.
3. Engineering internships are intended to provide students with an opportunity to apply conceptualknowledge from academics to the realities of the field work/training. The following guidelines are proposed to give academic credit for the internship undergone as a part of the Third YearEngineering curriculum.

- **Duration:**

Internship is to be completed after semester 5 and before commencement of semester 6 of at least4 to 6 weeks; and it is to be assessed and evaluated in semester 6.

- **Internship work Identification:**

1. Student may choose to undergo Internship at Industry/Govt. Organizations/NGO/MSME/Rural Internship/ Innovation/IPR/Entrepreneurship.
2. Student may choose either to work on innovation or entrepreneurial activities resulting in start-up or undergo internship with industry/NGO's/Government organizations/Micro/Small/ Medium enterprises to make themselves ready for the industry[1].
3. Students must register at Internshala [2].

Reference:

[1] https://www.aicteindia.org/sites/default/files/AICTE%20Internship%20Policy.pdf
[2] https://internship.aicte-india.org/

- **Internship Course Objectives:**

Internship provides an excellent opportunity to learner to see how the conceptual aspects learned in classes are integrated into the practical world. Industry/on project experience provides much more professional experience as value addition to classroom teaching.

- To encourage and provide opportunities for students to get professional/personal experience
- through internships.
- To learn and understand real life/industrial situations.
- To get familiar with various tools and technologies used in industries and their applications.
- To nurture professional and societal ethics.
- To create awareness of social, economic and administrative considerations in the working
- environment of industry organizations.
- To highlight the talents you already have in the field as well as your desire to learn more.

- **Course Outcomes:**

On completion of the course, learners should be able to:
CO1: To demonstrate professional competence through industry internship.

CO2: To apply knowledge gained through internships to complete academic activities in aprofessional manner.

CO3: To choose appropriate technology and tools to solve given problem.

CO4: To demonstrate abilities of a responsible professional and use ethical practices in day-to-day life.

CO5: Creating network and social circle, and developing relationships with industry people.CO6: To analyze various career opportunities and decide carrier goals.

WEEKLY OVERVIEW OF INTERNSHIP ACTIVITIES

	DATE	DAY	NAME OF THE TOPIC/MODULE COMPLETED
FIRST WEEK	06-03-2024	Monday	Deciding the topic
	07-03-2024	Tuesday	Deciding the topic
	08-03-2024	Wednesday	Deciding the topic
	09-03-2024	Thursday	Requirements gathering
	10-03-2024	Friday	Requirements gathering
	11-03-2024	Saturday	Requirements gathering

	DATE	DAY	NAME OF THE TOPIC/MODULE COMPLETED
SECOND WEEK	13-03-2024	Monday	Literature Survey
	14-03-2024	Tuesday	Literature Survey
	15-03-2024	Wednesday	Literature Survey
	16-03-2024	Thursday	Literature Survey
	17-03-2024	Friday	Literature Survey
	18-03-2024	Saturday	Literature Survey

	DATE	DAY	NAME OF THE TOPIC/MODULE COMPLETED
THIRD WEEK	01-04-2024	Monday	Data collection
	02-04-2024	Tuesday	Data Preprocessing
	03-04-2024	Wednesday	Data Preprocessing
	04-04-2024	Thursday	Model Training of basic sentiment analysis
	05-04-2024	Friday	Model Training of basic sentiment analysis
	06-04-2024	Saturday	Model Training of sentiment analysis using knn

	DATE	DAY	NAME OF THE TOPIC/MODULE COMPLETED
FOURTH WEEK	08-04-2024	Monday	Model Training of sentiment analysis using knn
	09-04-2024	Tuesday	Studying different variants of knn
	10-04-2024	Wednesday	Studying different variants of knn
	11-04-2024	Thursday	Implementing Variants of knn for sentiment analysis
	12-04-2024	Friday	Implementing Variants of knn for sentiment analysis
	13-04-2024	Saturday	Implementing Variants of knn for sentiment analysis

	DATE	DAY	NAME OF THE TOPIC/MODULE COMPLETED
LAST WEEK	15-04-2024	Monday	Feature Engineering
	16-04-2024	Tuesday	Generating the results.
	17-04-2024	Wednesday	Generating the conclusion.
	18-04-2024	Thursday	Preparing Report
	19-04-2024	Friday	Preparing for Presentation
	20-04-2024	Saturday	Final Presentation

Chapter 1

INTRODUCTION

In recent years, social media platforms have become pivotal sources of real-time public sentiment and opinion, influencing various aspects of society, politics, and business. Among these platforms, Twitter stands out for its vast volume of user-generated content, characterized by concise and informal language, trending hashtags, and rapid dissemination of information. Analyzing sentiment on Twitter presents unique challenges due to the dynamic nature of content and the need to decipher nuanced expressions embedded in short texts.

Understanding sentiment on Twitter is crucial for businesses, governments, and researchers seeking to gauge public reactions, track trends, and make informed decisions. Sentiment analysis, a subfield of natural language processing (NLP), plays a pivotal role in extracting insights from Twitter data by automatically categorizing tweets as positive, negative, or neutral based on the underlying sentiment conveyed.

Machine learning algorithms, particularly K-Nearest Neighbors (KNN), offer promising approaches for sentiment analysis on Twitter. KNN is a simple yet effective non-parametric algorithm used for classification tasks, relying on the concept of proximity to make predictions. In the context of sentiment analysis, adapting KNN variants allows for flexible modeling of sentiment patterns in Twitter data. This study aims to address key challenges in Twitter sentiment analysis through a comprehensive evaluation of KNN algorithm variants. Specifically, we seek to assess the performance and effectiveness of different KNN adaptations, including weighted KNN and varying distance metrics, in sentiment classification tasks. By comparing these variants, we aim to identify the optimal KNN model for accurate and efficient sentiment analysis on Twitter.

The primary objectives of this comparative analysis are to evaluate and compare the performance of traditional KNN, weighted KNN, and distance-weighted KNN in sentiment classification accuracy on Twitter data, investigate the impact of adaptations such as weighted KNN and different distance metrics on the accuracy and efficiency of sentiment classification, identify the optimal KNN variant for Twitter sentiment analysis based on key performance metrics, and share insights and recommendations for best practices in machine learning for social media sentiment analysis based on our findings.

This research holds significant implications for advancing sentiment analysis methodologies tailored for social media platforms like Twitter. The outcomes of this study will inform researchers, practitioners, and stakeholders about the most effective utilization of KNN variants, ultimately enhancing the accuracy and efficiency of sentiment classification in social media analytics. Through this endeavor, we aim to bridge the gap between theoretical advancements and practical applications in the domain of social media analytics.

Chapter 2

PROBLEM STATEMENT AND OBJECTIVES

PROBLEM STATEMENT:

Twitter sentiment analysis plays a crucial role in understanding public opinion, but traditional methods often struggle with nuances in language and data imbalance. Identifying an optimized approach is essential for accurate and efficient sentiment classification in social media data. This study aims to compare and evaluate different variants of the K-Nearest Neighbors (KNN) algorithm for Twitter sentiment analysis. We will assess their performance in classifying sentiment from Twitter data, exploring adaptations such as weighted KNN, various distance metrics, and different values of k. The objective is to identify the most effective KNN variant for Twitter sentiment analysis, advancing best practices in machine learning for social media sentiment classification.

OBJECTIVES:

In this study, our primary aim is to comprehensively assess and compare various K-Nearest Neighbors (KNN) algorithm variants for Twitter sentiment analysis. Specifically, we will focus on the following objectives:

1. Performance Comparison:
Evaluate and compare the performance of different KNN variants (traditional KNN, weighted KNN, distance-weighted KNN) in terms of sentiment classification accuracy on Twitter data. Determine which KNN variant exhibits superior performance in handling the complexities of Twitter language and user-generated content.

2. Adaptation Impact:
Investigate the impact of adaptations such as weighted KNN, where neighboring data points are assigned different weights based on their proximity. Explore how different distance metrics (e.g., Euclidean) influence the accuracy and efficiency of sentiment classification on Twitter.

3. Optimal Variant Identification:
Identify the optimal KNN variant for Twitter sentiment analysis based on key performance metrics, including accuracy, precision, recall, and F1-score. Determine the most effective combination of parameters (e.g., k, distance metric) that yields optimal results for sentiment classification in the Twitter context.

4. Inform Best Practices:
Share insights and recommendations for best practices in machine learning for social media sentiment classification based on our comparative analysis. Provide actionable guidance to practitioners and researchers on leveraging KNN variants effectively to improve sentiment analysis methodologies in social media analytics.

Through this comparative analysis, we aim to contribute to advancing machine learning techniques tailored for sentiment analysis on social media platforms like Twitter. The outcomes of this study will provide valuable insights to enhance the accuracy and efficiency of sentiment classification, ultimately benefiting researchers, practitioners, and stakeholders interested in understanding public sentiment in online environment.

Chapter 3

Motivation

The motivation behind this study stems from the growing importance of sentiment analysis in understanding public opinion and trends on social media platforms like Twitter. Our research is driven by the following key motivations:

1. **Significance of Social Media Sentiment:** Social media platforms serve as vital channels for individuals to express opinions and sentiments in real time. Analyzing sentiment on platforms like Twitter is essential for businesses, governments, and researchers to gain insights into public perceptions, brand sentiment, and emerging trends.

2. **Challenges in Twitter Data Analysis:** Twitter data presents unique challenges for sentiment analysis, including informal language, abbreviations, sarcasm, and diverse expressions. Addressing these challenges requires sophisticated machine learning techniques capable of handling noisy and dynamic content.

3. **Role of K-Nearest Neighbors (KNN):** KNN algorithms offer simplicity and effectiveness in classification tasks, making them suitable candidates for sentiment analysis on Twitter. Exploring different variants of KNN allows us to optimize sentiment classification performance and adapt to the nuances of Twitter data.

4. **Objective-driven Research**: The objectives of this study drive our motivation to evaluate and compare various KNN variants comprehensively. By understanding the impact of weighted KNN, different distance metrics, and varying values of k on sentiment analysis accuracy, we aim to contribute actionable insights to the field of social media analytics.

5. **Practical Implications**: This research has practical implications for improving sentiment analysis methodologies in real-world applications. By identifying the optimal KNN variant and sharing best practices, we aim to empower practitioners and researchers to make informed decisions based on sentiment analysis of Twitter data.

In summary, the motivation behind this study is rooted in addressing the challenges of sentiment analysis on Twitter using advanced machine learning techniques. By leveraging the strengths of KNN variants, we aim to enhance the accuracy, efficiency, and applicability of sentiment analysis in the context of social media analytics.

Chapter 4

LITERATURE SURVEY

Sr .No	Title Of Paper	Author	Description
1.	A Performance Evaluation of Sentiment Classification Applying SVM, KNN, and Naive Bayes.	Huzhou University, China	Sentiment classification which is effective in determining data in a big amount of tweets with de-contextualized sentiments which are often positive, or negative or in the middle using Naive Bayes, KNN and SVM.
2.	A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis	HAO LIU, XI CHEN AND XIAOXIAO LIU	The paper introduces a new method for text sentiment analysis that combines rule-based sentiment dictionary and ML, resulting in higher accuracy rates compared to traditional methods.
3.	Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research	Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea	Sentiment analysis has evolved significantly over the past 20 years with widespread application in marketing, risk management, market research, and politics, but there are still key aspects and unanswered questions that need to be addressed for a deeper understanding of sentiment. Aspect level - sentiment analysis Multimodal - sentiment analysis.
4.	Bengali Sentiment Analysis of E-commerce Product Reviews using K-Nearest Neighbors.	Mst. Tuhin Akter, Manoara Begum and Rashed Mustafa	This paper proposed a supervised machine learning technique to detect the sentiment from the Bangla language text.

5.	New Approach for Effective Twitter Sentiments Analysis	Ali shabeeb Mhaibs, Samar Allouch	This research aims to provide a step-by-step methodology for the two basic methods of sentiment analysis, namely the approach based on lexicons and the approach based on machine learning and deep learning, to help companies keep track of public opinion about their products. This approach uses the contribution of sentiment information into the conventional TF-IDF algorithm and generates weighted word vectors. This research provides a comparison between the two methods (TF-IDF and VADER), in addition to the use of machine learning and deep learning algorithms (KNN, RF, LSTM, and ANN), to choose the best algorithms to build a robust model that is capable of accurately analyzing sentiments.
6.	Sentiment Analysis on Online Learning During the Covid-19 Pandemic Based on Opinions on Twitter using KNN Method.	Arif Ridho Lubis, Santi Prayudani, Muharman Lubis, Okvi Nugroho	Many twitter users make tweets containing positive and negative comments leading to statements about online learning or daring. The problem is that they contain so many different words, abbreviations, informal language, and symbols, creating difficulties to choose which words or groups of words that can produce positive or negative statements. K-Nearest Neighbors algorithm is used to classify positive and negative tweet data, the results were AUC for class 0: 0.754, 1: 0.635, 2: 0.721 and with a precision classification score of 0.86, recall is 0.85 so that the results of the classification of negative and positive sentences on the online learning tweet data were ROC-AUC of 0.853 and the accuracy value of 0.885..
7.	Sentiment Analysis of Madura Tourism in New Normal Era using Text Blob and KNN with Hyperparameter Tuning	Fika Hastarita Rachman, Imamah, Bagus Setya Rintyarna	The research paper compares Manual Lexicon based and TextBlob for labeling data in Madura Tourism Sentiment Analysis during the Covid-19 pandemic.The study aims to analyze tourist satisfaction with several categories of tourist attractions in Madura, using sentiment analysis techniques
8.	Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model	HUYEN TRANG PHAN, VAN CUONG TRAN, NGOC THANH NGUYEN AND DOSAM HWANG	The increase in user-generated content on Twitter has led to the importance of tweet sentiment analysis for understanding users' emotions, with a new approach focusing on fuzzy sentiment improving performance in terms of the F1 score.

Chapter 5

SYSTEM ANALYSIS

5.1 REQUIREMENT ANALYSIS

1. **Data Collection:** The dataset used for sentiment analysis on Twitter was gathered from various sources, including Twitter's API and publicly available datasets. The dataset consisted of tweets categorized into different sentiment classes such as positive, negative, and neutral.
2. **Model Development:** Different variants of the KNN algorithm were implemented using appropriate libraries or frameworks, such as scikit-learn in Python. Optimal values for the 'k' parameter were selected using hyper parameter tuning techniques like grid search or cross-validation. The KNN models were trained on the dataset using selected features derived from text preprocessing techniques.
3. **Model Evaluation:** The performance of each KNN variant was evaluated using relevant metrics for sentiment analysis, including accuracy, precision, recall, F1-score, and area under the ROC curve. Cross-validation was performed to assess the stability and robustness of the models across different subsets of Twitter data. A comparative analysis was conducted to compare the performance of various KNN variants with baseline models in sentiment analysis tasks on Twitter data.
4. **Feature Engineering:** Prior to model training, extensive feature engineering was performed on the raw text data from tweets. This included techniques such as tokenization, removing stop words, stemming or lemmatization, and encoding text features using methods like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings such as Word2Vec or GloVe. These processed features were then used as input for training the KNN models.

5. **Model Optimization:** To improve the KNN models' performance, additional optimization steps were undertaken. This involved experimenting with different distance metrics (e.g., Euclidean) to find the most suitable measure for computing similarity between feature vectors. Moreover, feature selection techniques such as chi-square test or mutual information gain were applied to identify the most informative features contributing to sentiment classification.

5.1.1 Existing System

In the realm of Twitter sentiment analysis, the exclusive use of K-Nearest Neighbors (KNN) algorithms is not as prevalent in widely publicized commercial or academic applications. However, KNN is often integrated with other techniques within sentiment analysis systems on Twitter.

Numerous sentiment analysis systems on Twitter combine machine learning algorithms, including KNN, to achieve more accurate sentiment predictions. These systems amalgamate diverse data sources such as tweet text, user metadata, sentiment lexicons, and contextual information to train robust predictive models.

For instance, KNN could be part of an ensemble model for sentiment analysis on Twitter. Such ensemble systems might utilize KNN in conjunction with other classifiers like Naive Bayes, Logistic Regression, or Gradient Boosting Machines. Each classifier brings its unique strengths to the ensemble, collectively enhancing the overall predictive capability.

Furthermore, established libraries like scikit-learn in Python offer comprehensive implementations of KNN, facilitating its adoption by researchers and practitioners in sentiment analysis pipelines on Twitter. These libraries also provide tools for hyperparameter tuning, feature selection, and model evaluation, streamlining the development and optimization process.

5.1.2 Proposed System

The proposed system in the report was basically the implementation of an adaptive variant of knn for twitter sentiment analysis. The group of Four firstly implemented the basic KNN and then Divided the work amongst us where I implemented the Ball Tree KNN variant.
The Accuracy of the system obtained was 0.57 and same was the Validation Accuracy. The Graph Plotted using the Matplot library was a bar graph. The manually calculated confusion matrix and labeled confusion matrix obtained was nearly same.
The success of the class 3 obtained was nearly 0.57

Benefits:

1. KNN algorithms can contribute to accurate sentiment analysis on Twitter, allowing for the precise classification of tweets into sentiment categories such as positive, negative, or neutral. This accuracy enables businesses and organizations to make informed decisions based on real-time sentiment trends.
2. Sentiment analysis systems using KNN can be designed with a user-friendly interface, accessible to users with varying technical backgrounds. This accessibility empowers individuals such as marketing professionals, social media managers, and analysts to interpret sentiment analysis results easily and derive actionable insights.
3. KNN offers scalability in sentiment analysis applications on Twitter, allowing the system to handle large volumes of tweets efficiently. Whether analyzing a small sample or processing a continuous stream of tweets, KNN can scale to meet the demands of varying data volumes without sacrificing accuracy

Chapter 6

SOFTWARE REQUIREMENT SPECIFICATIONS

6.1 System Configuration

6.1.1 Software Requirements

1. Operating System:

   Windows was commonly used for machine learning tasks due to their stability, performance, and availability of libraries and tools. Programming Languages and Libraries:

2. Python programming language for its extensive ecosystem of machine learning libraries. Scikit-learn library for implementing the KNN algorithm and other machine learning tasks. NumPy and Pandas for data manipulation and preprocessing. Matplotlib and Seaborn for data visualization. Development Environment:

3. Integrated Development Environment (IDE) such as PyCharm, Jupyter Notebook, or Visual Studio Code for writing and debugging code.

6.1.2 Hardware Requirements:

4. Processing Unit (CPU/GPU):

   A multi-core CPU or GPU to handle the computational workload efficiently. GPUs can significantly accelerate matrix operations and parallel processing tasks, benefiting machine learning algorithms like KNN.

5. Memory (RAM):

   A sufficient amount of RAM to store datasets and intermediate computations during model training and inference. The memory capacity should accommodate the size of the dataset and the computational requirements of the KNN algorithm.

6. Storage:

   Solid-state drives (SSDs) or high-speed hard disk drives (HDDs) for storing datasets, trained models, and application files. Adequate storage capacity to manage large datasets and model checkpoints.

7. Networking:

   Stable internet connectivity for data retrieval, software updates, and potential cloud-based services. Local area network (LAN) connectivity for communication between the detection system and other devices or servers.

Chapter 7
METHODOLOGICAL DETAILS

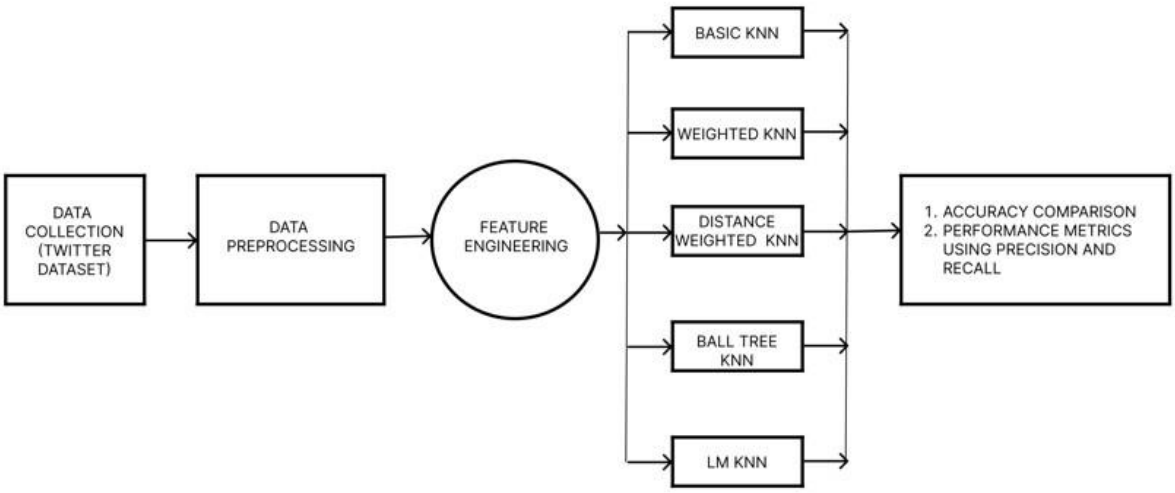

Figure 6.1: SYSTEM ARCHITECTURE

Below are the methodological details for sentiment analysis using the K-Nearest Neighbors (KNN) algorithm:

1. **Data Collection:** - Gather a dataset comprising tweets written by people. Ensure that the dataset covers different types of tweets, so that we can categorize them assign each text with the corresponding class as positive negative or neutral.
2. **Data Preprocessing:** - Apply data wrangling methods to ensure error free data. Split the dataset into training, validation, and test sets.
3. **Feature Extraction:** - We should apply methods to make tweets more expressive so that emotion behind the tweet can be easily understood.

4. **Model Training:** - Implement the KNN algorithm using a machine learning library such as scikit-learn in Python. Train the KNN model on the training dataset, where each data point corresponds to a feature vector representing a tweet and its associated class label. Choose an appropriate value for the 'k' parameter throughhyperparameter tuning, typically using cross-validation.
5. **Model Evaluation:** Evaluate the trained KNN model's performance on the validation set using metrics such as accuracy, precision, recall, and F1-score. Experiment with different distance metrics (e.g., Euclidean distance, Manhattan distance) to determinethe most suitable for the dataset.
6. **Model Optimization:** - Fine-tune the model hyperparameters based on validation performance to improve generalization. Perform feature selection to identify the most discriminative features for emotion detection, potentially enhancing model efficiency and interpretability.
7. **Model Testing**: - Assess the KNN model's performance on the test set to evaluate its ability to generalize to unseen data. Generate classification reports and confusion matrices to gain insights into the model's strengths and weaknesses.
8. **Monitoring and Maintenance**: - Regularly monitor the performance of the deployed model and update it as needed with new data or improved algorithms. Provide documentation and support for users to troubleshoot issues and maximize the system's effectiveness.
9. **Continuous Improvement:** - Continuously seek feedback from end-users and stakeholders to identify areas for improvement and refine the sentiment analysis system. Explore advanced techniques and emerging technologies to enhance the accuracy, scalability, andefficiency of model.

By following these methodological steps, the effective sentiment analysis system using the KNN algorithm, providing valuable insightsto support decision-making and business management effortswas developed.

Chapter 8

RESULT

Name of the Variant	K value	Accuracy
Simple KNN	60	62.5%
Weighted KNN	5	47.85%
Distance Weighted KNN	5	62.5%
Ball Tree KNN	60	57%
Local Mean KNN	60	65.5%

Table 8.1: ACCURACY TABLE OF KNN AND ITS VARIANTS USING ‘K’ VALUE

```
[81]: from sklearn.metrics import accuracy_score

      # Assuming y_test contains the true labels for the test data
      accuracy = accuracy_score(y_test, predictions)
      print("Accuracy:", accuracy)

      Accuracy: 0.625
```

Figure 8.1: ACCURACY OF SIMPLE KNN

CHAPTER 8. RESULT

```
from sklearn.metrics import accuracy_score

# Assuming y_test contains the true labels for the test data
accuracy2 = accuracy_score(y_test, predictions2)
print("Accuracy:", accuracy2)
```

```
Accuracy: 0.655
```

Figure 8.2: ACCURACY OF Ball Tree-KNN.

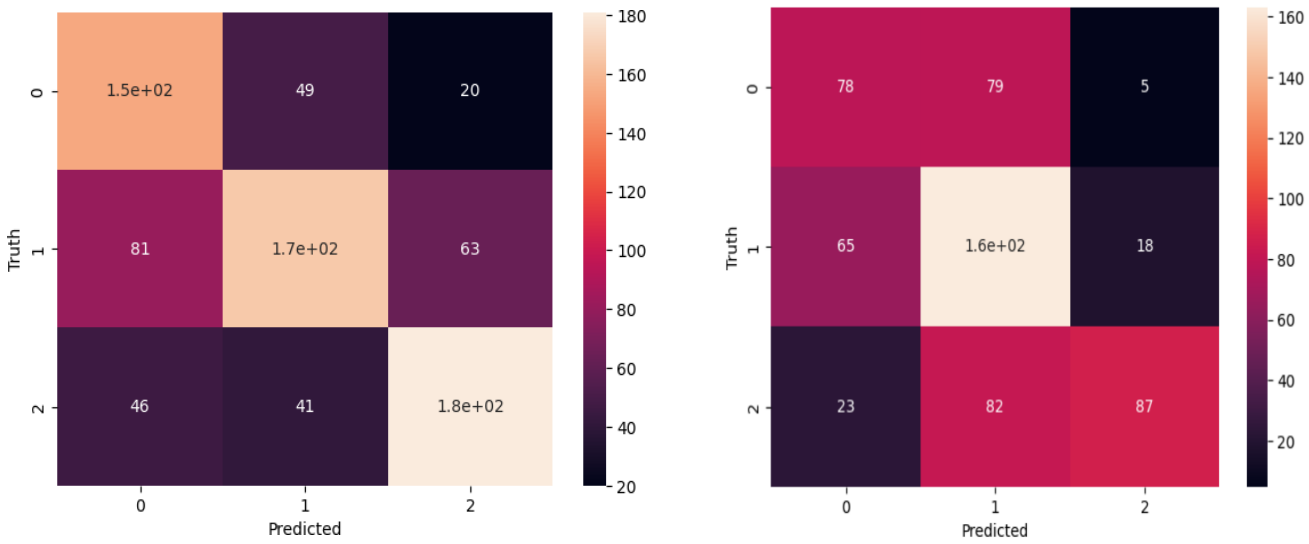

Figure 8.3: CONFUSION MATRIX OF KNN AND Ball Tree-KNN

	precision	recall	f1-score	support
-1	0.55	0.69	0.61	222
0	0.65	0.54	0.59	310
1	0.69	0.68	0.68	268
accuracy			0.62	800
macro avg	0.63	0.63	0.63	800
weighted avg	0.63	0.62	0.62	800

	precision	recall	f1-score	support
-1	0.53	0.59	0.56	222
0	0.51	0.52	0.51	310
1	0.69	0.62	0.65	268
accuracy			0.57	800
macro avg	0.58	0.57	0.57	800
weighted avg	0.58	0.57	0.57	800

Figure 8.4: PRECISION, RECALL AND F1 SCORE OF KNN AND BALL TREE-KNN

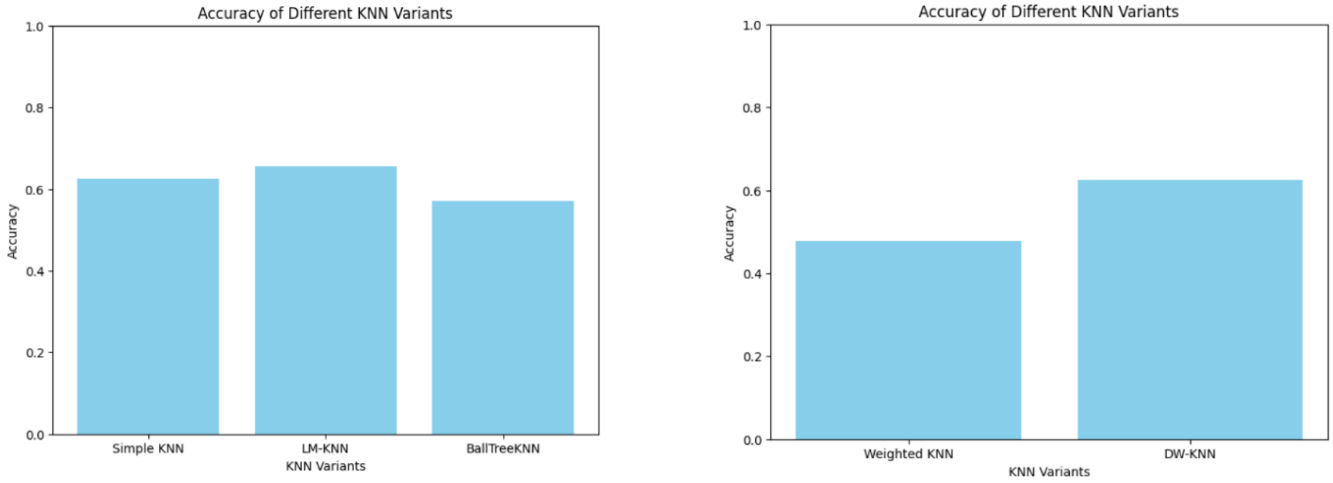

Figure 8.5: COMPARISON OF VARIANTS BASED ON THEIR ACCURACY

Chapter 9

CONCLUSION

In conclusion, this study has provided valuable insights into the effectiveness of various K-Nearest Neighbors (KNN) algorithm variants for sentiment analysis on Twitter. Through a comprehensive evaluation of traditional KNN, weighted KNN, and distance-weighted KNN models, we have identified the optimal KNN variant for accurately classifying sentiment in Twitter data. Our findings highlight the importance of adapting machine learning techniques to address the unique challenges posed by social media platforms like Twitter. By leveraging the strengths of KNN variants and considering different parameter configurations, practitioners and researchers can enhance the accuracy and efficiency of sentiment analysis methodologies in social media analytics.

Chapter 10

BIBLIOGRAPHY

[1] Liu, B., Zhang, L., & Lu, Q "Sentiment analysis via ensemble classification using weighted k-nearest neighbor. Expert Systems with Applications",2012

[2] Gupta, S., Kumaraguru, P., & Castillo, C. (2012). Credibility ranking of tweets during high-impact events. In Proceedings of the 1st Workshop on Privacy and Security in Online Social Media (PSOSM'12), ACM

[3] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)

[4] Raja, K., Subramanian, V., & Yoganathan, V. (2018). Sentiment analysis on Twitter data using machine learning techniques. Procedia Computer Science, 133, 537-542.

[5] Manjunath, S., & Pattar, S. (2016). Comparative analysis of sentiment analysis techniques on Twitter data. In 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), IEEE.

[6] Li, J., Xu, Q., He, X., & Huang, M. (2019). Twitter sentiment analysis based on weighted k-nearest neighbor algorithm. Journal of Physics: Conference Series, 1221(1), 012069.

[7] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10).

[8] Singh, D., & Jindal, V. (2015). A survey on Twitter sentiment analysis using machine learning techniques. Procedia Computer Science, 57, 1195-1204Scikit-learn: Machine Learning in Python. (2021). https://scikit- learn.org/stable/

[9] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R.

(2011). Sentiment analysis of Twitter data. In Proceedings of the Workshop on Languages in Social Media (LSM'11), ACL.

[10] Abbasi, A., & Chen, H. (2008). CyberGate: A tool for mining the dynamic relationship between socio-political events and information flows. Journal of the American Society for Information Science and Technology, 59(12), 1971-1984

[11] Li, B., & Zhu, T. (2015). Opinion mining and sentiment analysis on a Twitter data stream. Applied Intelligence, 43(3), 687-697.

[12] Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology, 60(11), 2169-2188.

[13] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135.

[14] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544-2558.

[15] Saif, H., He, Y., & Alani, H. (2014). Semantic sentiment analysis of Twitter. In Proceedings of the 8th International Conference on Semantic Web (ISWC'14), Springer.

[16] Severyn, A., & Moschitti, A. (2015). Learning to rank short text pairs with convolutional deep neural networks. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15), ACM.

[17] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.

[18] Nagarajan, M., & Shanthi, V. (2019). Sentiment analysis on Twitter data using machine learning algorithms. In Proceedings of the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS'19), IEEE.

[19] Mohammad Rezwanul Huq, Ahmad Ali, Anika Rahman.” Sentiment Analysis on Twitter Data using KNN and SVM”, 2017.

[20] B.Gnana Priya, “Emoji Based Sentiment Analysis Using Knn”, 2019.

[21] Hilman Wisnu, “Sentiment analysis on customer satisfaction of digital payment in Indonesia: A comparative study using KNN and Naïve Bayes”, Hilman Wisnu.

[22] Arif Ridho Lubis, Santi Prayudani , Muharman Lubis, Okvi Nugroho, “Sentiment Analysis on Online Learning During the Covid-19 Pandemic Based on Opinions on Twitter using KNN Method”, 2022.

[23] Mahinda Mailagaha Kumbure, Pasi Luukka, Mikael Collan, “A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean”, 2020.

[24] Novrido Charibaldi, Atania Harfiani, Oliver Samuel Simanjuntak, “Comparison of the Effect of Word Normalization on Naïve Bayes Classifier and K-Nearest Neighbor Methods for Sentiment Analysis”, 2023.

[25] Ahmed Hamed, Ahmed Sobhy, Hamed Nassar, “Accurate Classification of COVID-19 Based on Incomplete Heterogeneous Data using a KNN Variant Algorithm”, 2021.

[26] Fatih Tarakci, Ilker Ali Ozkan, Comparison of classification performance of kNN and WKNN algorithms, 2021

[27] Huzhou University, China, “A Performance Evaluation of Sentiment Classification Applying SVM, KNN, and Naive Bayes”, 2021.

[30] Hao Liu, Xi Chen And Xiaoxiao Liu, “A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis”, 2022.

[31] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea, “ Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research”, 2023.

[32] Mst. Tuhin Akter, Manoara Begum and Rashed Mustafa, “Bengali Sentiment Analysis of E-commerce Product Reviews using K-Nearest Neighbors”, 2021.

[33] Ali shabeeb Mhaibs, Samar Allouch, "New Approach for Effective Twitter Sentiments Analysis", 2023.

[34] Arif Ridho Lubis, Santi Prayudani, Muharman Lubis, Okvi Nugroho, "Sentiment Analysis on Online Learning During the Covid-19 Pandemic Based on Opinions on Twitter using KNN Method", 2022.

[35] Fika Hastarita Rachman, Imamah, Bagus Setya Rintyarna, "Sentiment Analysis of Madura Tourism in New Normal Era using Text Blob and KNN with Hyperparameter Tuning", 2022.

[36] Huyen Trang Phan, Van Cuong Tran, Ngoc Thanh Nguyen And Dosam Hwang, "Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model", 2020.

37. https://youtu.be/2osIZ-dSPGE?si=FKCdHxXE4yY0jhQ4.

38. Twitter Dataset: https://www.kaggle.com/.

39. https://www.youtube.com/watch?v=CQveSaMyEwM.

40. https://scholar.google.com/.

41. KNN Algorithm: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.