A

PROJECT REPORT

ON

**EXPLORING SENTIMENT ANALYSIS ON SOCIAL MEDIA DATA**

BY

**Team Members**

D. MEGHANA

B. THANUSH

J. ARCHANA REDDY

N. MANASA

**COLLEGE NAME**

SPHOORTHY ENGINEERING COLLEGE

**ABSTRACT**

The abstract of the sentiment analysis project encapsulates its essence, objectives, methodology, and outcomes concisely:

Sentiment analysis, a vital aspect of natural language processing, involves the extraction and interpretation of subjective information from textual and multimodal data. This project aims to develop an advanced sentiment analysis system to address the growing need for accurate, scalable, and real-time sentiment analysis solutions.

The project begins with an exploration of the existing sentiment analysis systems, highlighting their limitations and areas for improvement. Subsequently, a comprehensive solution is proposed, leveraging advanced data collection techniques, sophisticated pre-processing methods, and state-of-the-art sentiment analysis models.

Key features of the proposed system include real-time sentiment monitoring, customizable visualization dashboards, and a sentiment analysis API for seamless integration into various applications. Emphasis is placed on scalability, performance optimization, and user feedback mechanisms to ensure usability and effectiveness.

The project's conclusion underscores the significant advancements made in sentiment analysis, enabling organizations and individuals to gain deeper insights into public opinion, customer feedback, and market trends. The proposed system holds promise for driving informed decision-making and improving various aspects of business operations, marketing strategies, and social media management.

This abstract provides a succinct overview of the sentiment analysis project, highlighting its objectives, methodology, and potential impact.

I

**INDEX**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Contents** | | | | | | **Page No.** |
| Abstract | | | | | | I |
| Index | | | | | | II |
| List of Figures | | | | | | IV |
| **Chapters** | | | | | |  |
| **1** | **Introduction** | | | | | 1 |
|  | 1.1 | | | | Problem Statement | 1 |
|  | 1.2 | | | | Objective | 1 |
|  | 1.3 | | | | Motivation | 1 |
|  | 1.4 | | | | Existing System | 2 |
|  | 1.5 | | | | Proposed System | 2 |
|  | 1.6 | | | | Scope | 3 |
|  | 1.7 | | | | Software Requirements | 4 |
|  | 1.8 | | | | Hardware Requirements | 4 |
| **2** | **Literature Survey** | | | | | 5 |
|  | 2.1 | | Role of Sentiment Analysis on Social media data | | | 5 |
|  | 2.2 | | Machine Learning | | | 5 |
|  | 2.3 | | Types of machine Learning | | | 5 |
|  | 2.4 | | Steps in Machine Learning | | | 6 |
|  | 2.5 | | Sentiment Analysis | | | 7 |
|  | 2.6 | | Feature Extraction in Sentiment Analysis | | | 9 |
|  | 2.7 | | Types of Sentiment Analysis | | | 9 |
|  | 2.8 | | Sentiment Analysis Methods | | | 10 |
| **3** | **System Design** | | | | | 11 |
|  | 3.1 | | | System Architecture | | 11 |
|  | 3.2 | | | Data Flow Diagram | | 12 |
|  | 3.3 | | | Activity Diagram | | 13 |
| **4** | **Implementation** | | | | | 14 |
|  | 4.1 | Methodology | | | | 14 |
|  | 4.2 | Enrolments Phase | | | | 14 |
|  | 4.3 | Environmental Setup | | | | 15 |
|  | 4.4 | Jupyter Notebook | | | | 15 |
|  | 4.5 | Introduction to Python | | | | 16 |
|  | 4.6 | Introduction to NLP | | | | 17 |

II

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **5** | **Libraries and Algorithms** | | | 18 |
|  | 5.1 | | Libraries | 18 |
|  | 5.1.1 | | Python Libraries | 18 |
|  | 5.1.2 | | NLP Libraries | 19 |
|  | 5.2 | | Algorithms | 19 |
| **6** | **Dataset** | | | 21 |
|  | 6.1 | | About Dataset | 21 |
|  | 6.2 | | Key Features | 21 |
|  | 6.3 | | Use of Dataset | 22 |
|  | 6.4 | | Columns | 22 |
|  | 6.5 | | Link of the Dataset | 22 |
| **7** | **Source Code and Conclusion** | | | 23 |
|  | 7.1 | Source Code and Output | | 23 |
|  | 7.2 | Conclusion | | 40 |
| **8** | **Appendix** | | | 41 |

III

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Diagram Name** | **Page No** |
| 1 | Machine Learning Steps | 8 |
| 2 | Sentiment Analysis Process | 9 |
| 3 | Sentiment Analysis Methods | 11 |
| 4 | Architecture of a System | 12 |
| 5 | Data Flow Diagram | 13 |
| 6 | Activity Diagram | 14 |

IV

**CHAPTER-1**

**INTRODUCTION**

**1.1 PROBLEM STATEMENT:**

The problem statement for this project revolves around the need to use sentiment analysis techniques to comprehend and interpret the diverse range of emotions expressed on social media platforms. Social media has become the primary channel for people to express their ideas, opinions, and emotions, generating a wealth of data for analysis. However, the sheer volume and diversity of content pose significant challenges in accurately capturing and interpreting sentiments. The goal of this project is to develop and implement effective sentiment analysis models for social media data. In doing so, we hope to address the complexities of human emotions expressed online, paving the way for the development of emotionally intelligent machines capable of understanding and responding to user sentiments with empathy and accuracy.

**1.2** **OBJECTIVE:**

The goal of this sentiment analysis project on social media data is to investigate and implement effective techniques for understanding and analyzing the various emotions expressed by users across multiple social media platforms. Using advanced machine learning algorithms and natural language processing techniques, we hope to create models that can accurately capture the sentiments, opinions, and attitudes conveyed through text. Furthermore, the project intends to look into the effects of various factors on sentiment analysis performance, such as language, cultural context, and platform-specific characteristics. Our meticulous data collection, preprocessing, and model training aim to uncover meaningful insights and trends in human emotion expression online. Finally, the project aims to contribute to the development of emotionally intelligent machines that can understand and respond to user sentiment with empathy and understanding, thereby enhancing user experiences in the digital realm.

**1.3 MOTIVATION:**

The motivation behind this sentiment analysis project on social media data lies in the profound impact of user sentiments on various aspects of modern life. With the pervasive influence of social media platforms, individuals actively share their thoughts, opinions, and emotions, creating a wealth of data for analysis. Understanding these sentiments is essential for businesses to refine their marketing strategies, governments to gauge public opinion, and researchers to study societal trends in real time. By deciphering the underlying sentiments in social media posts, this project aims to provide actionable insights that can inform decision-making processes across different domains. Additionally, sentiment analysis on social media data holds promises in identifying emerging trends, detecting potential crises, and even predicting future events. Ultimately, the goal is to leverage sentiment analysis to improve user experiences, facilitate more informed decisions, and deepen our understanding of human behavior in the digital era.

1

**1.4 EXISTING SYSTEM:**

Existing systems for sentiment analysis on social media data utilize a variety of techniques and tools to analyze the sentiments expressed in user-generated content. These approaches include lexicon-based methods, machine learning algorithms, deep learning models, hybrid approaches, and commercial sentiment analysis tools.

* Lexicon-based methods rely on sentiment lexicons or dictionaries annotated with sentiment polarities (positive, negative, or neutral). Examples include Sent WordNet and VADER.
* Machine learning algorithms such as Support Vector Machines, Naive Bayes, and Logistic Regression are commonly employed for sentiment classification. These models learn from labelled datasets to predict sentiment in unseen text.
* Deep learning models like Recurrent Neural Networks, Long Short-Term Memory networks, and Transformer architectures (e.g., BERT) excel in capturing complex patterns in text data, making them popular choices for sentiment analysis.
* Hybrid approaches combine multiple techniques, such as using lexicon-based methods for feature extraction and machine learning algorithms for classification, to enhance sentiment analysis accuracy.
* Commercial sentiment analysis tools and APIs like IBM Watson, Google Cloud Natural Language API, and Microsoft Azure Text Analytics offer specialized sentiment analysis capabilities tailored for social media data.
* Despite advancements, challenges persist in accurately capturing linguistic nuances, understanding context, detecting sarcasm and irony, and adapting to evolving language trends. Further research and development efforts are necessary to address these challenges and enhance the performance and reliability of sentiment analysis systems for real-world applications.

**1.5 PROPOSED SYSTEM:**

The proposed system for the sentiment analysis project aims to enhance the existing system by introducing several improvements and additional features. Here's an overview of the proposed system:

**1.** **Advanced Data Collection:** Expand the scope of data collection to include a wider range of sources beyond social media platforms. This could involve gathering data from news articles, blogs, customer reviews, and forums to provide a more comprehensive analysis of sentiment.

**2. Enhanced Data Preprocessing:** Implement more sophisticated data preprocessing techniques to improve the quality of the input data for sentiment analysis. This includes advanced text cleaning methods, entity recognition, and sentiment lexicon enrichment to better handle nuances in language and context.

**3. State-of-the-Art Sentiment Analysis Models:** Utilize state-of-the-art machine learning and deep learning models for sentiment analysis, such as Transformer-based architectures like BERT (Bidirectional Encoder Representations from Transformers) or GPT (Generative Pre-trained Transformer). These models have shown superior performance in understanding context and capturing subtle nuances in text data.

2

**4. Multimodal Sentiment Analysis:** Incorporate multimodal sentiment analysis by combining textual data with other modalities such as images, audio, and video. This allows for a more holistic understanding of sentiment expressed across different media types.

**5. Real-time Sentiment Monitoring:** Implement real-time sentiment monitoring capabilities to track sentiment trends and fluctuations as they occur. This involves continuously collecting and analyzing incoming data streams to provide up-to-date insights to users.

**6. Customizable Visualization Dashboard:** Develop a customizable visualization dashboard that allows users to explore sentiment analysis results through interactive charts, graphs, and heatmaps. Users can customize the dashboard based on their specific preferences and requirements.

**7. Sentiment Analysis API:** Create a sentiment analysis API that can be easily integrated into third-party applications and services. This API allows developers to access sentiment analysis capabilities programmatically, enabling the incorporation of sentiment analysis into a wide range of applications

**8. Scalability and Performance Optimization:** Design the system to be scalable and capable of handling large volumes of data efficiently. This involves optimizing the performance of sentiment analysis models, deploying them on cloud infrastructure, and implementing distributed computing techniques if necessary.

**9. User Feedback Mechanism:** Incorporate a user feedback mechanism to gather input from users and stakeholders, allowing them to provide suggestions, report issues, and request new features. This feedback is used to continuously improve and refine the system over time.

**10. Comprehensive Documentation and Support:** Provide comprehensive documentation and support resources to assist users in understanding and utilizing the system effectively. This includes user guides, tutorials, API documentation, and technical support channels.

**1.6 SCOPE:**

The scope of the project encompasses several key components related to sentiment analysis on social media data. Here are some aspects that might fall within the project's scope:

**1. Data Collection:** Gathering data from various social media platforms such as Twitter, Instagram, Facebook, etc. This could involve using APIs provided by these platforms or scraping publicly available data.

**2. Data Preprocessing:** Cleaning and preparing the collected data for analysis. This includes tasks such as removing irrelevant information, handling missing values, standardizing text data, and dealing with duplicates.

**3. Text Analysis:** Applying natural language processing (NLP) techniques to analyze the text data. This involves tokenization, lemmatization, and removing stop words. Additionally, sentiment analysis algorithms are applied to classify the sentiment of each text (positive, negative, or neutral).

**4. Model Development:** Developing machine learning or deep learning models for sentiment analysis. This may include training models using labeled data and evaluating their performance using metrics such as accuracy, precision, recall, and F1-score.

3

**5. Visualization:** Visualizing the results of sentiment analysis to make them interpretable and actionable. This could involve creating charts, graphs, or dashboards to present insights derived from the data.

**6. Interpretation and Insights:** Analyzing the results of sentiment analysis to extract meaningful insights. This may involve identifying trends, patterns, or correlations in the data and drawing conclusions that can inform decision-making.

**7. Deployment:** Integrating the sentiment analysis model into a usable application or platform where stakeholders can interact with it. This could involve building APIs or web interfaces for accessing the sentiment analysis functionality.

**8. Documentation and Reporting:** Documenting the entire process, including data collection methods, preprocessing steps, model development, and evaluation results. This documentation serves as a reference for future use and helps in replicating the project.

**9. Feedback and Iteration:** Gathering feedback from stakeholders and incorporating it into the project to improve its effectiveness. This may involve refining the sentiment analysis model, enhancing data preprocessing techniques, or adding new features based on user requirements.

The scope of the project may vary depending on factors such as available resources, time constraints, and specific objectives. It's essential to define the scope clearly at the outset to ensure that the project stays focused and achievable within the given constraints.

**1.7 SOFTWARE REQUIREMENTS:**

* Windows 7 and above
* Jupyter Notebook
* Kaggle datasets
* Machine Learning Algorithms
* Python Libraries

**1.8 HARDWARE REQUIREMENTS:**

* Processor: i5 and above (64-bit OS)
* Memory: 4GB RAM
* Hard Disk: 6

4

**CHAPTER-2**

**LITERATURE SURVEY**

Literature review is an important aspect of this project since it helps in establishing

Familiarity with the topic. With the help of various literature review we can understand

the current research in the respective field. It makes things clearer and helps in greater

focus to the research problem and understand the findings.

**2.1 The role of Social Media Sentiment Analysis:**

Social media platforms like Twitter, Face book, YouTube, Reddit generate huge an

amounts of big data that can be mined in various ways to understand trends, public

sentiments and opinions. Social media data today has become relevant for branding,

Marketing, and business as a whole. A sentiment analysis learns about various

Sentiments behind a “content piece” through machine learning and predicts the same.

**2.2 Machine Learning:**

* Machine Learning is a subset of artificial intelligence (AI) that focuses on the development of an algorithms and statistical models that enable computers to perform tasks without being explicitly programmed.

Here’s an overview of machine learning: -

**2.3 Types of Machine Learning:**

1. **Supervised Learning: -**

* In supervised learning, the algorithm learns from labelled data, which means each input is associated with corresponding target output.
* Examples include classification and regression.

1. **Unsupervised Learning: -**

* Unsupervised learning deals with unlabelled data, where the algorithm tries to find patterns or structure in the data on its own.
* Clustering (e.g., customer segmentation, image segmentation) and dimensionality reduction (e.g., principal component analysis) are common tasks in unsupervised learning.

1. **Semi-supervised Learning: -**

* Semi-supervised learning combines both labelled and unlabelled data to improve learning accuracy.
* It's useful when labelled data is scarce or expensive to obtain.

5

1. **Reinforcement Learning: -**

* Reinforcement learning involves an agent learning to make decisions by interacting with an environment. The agent receives feedback in the form of rewards or penalties.
* Applications include game playing (e.g., AlphaGo), robotics, and autonomous vehicle control.

**2.4 Steps in Machine Learning: -**

* **Data Collection: -**Gather relevant data for the problem at hand. This can involve various sources such as databases, APIs, or scraping.
* **Data Preprocessing:**-Clean and preprocess the data to handle missing values, outliers, and noise. This step also includes feature scaling, normalization, and encoding categorical variables.
* **Feature Engineering:**-Extract or create meaningful features from the data to represent it effectively for the learning algorithm. Feature engineering can involve transformations, selection, or creation of new features.
* **Model Selection:-** Choose an appropriate machine learning algorithm based on the problem type, data characteristics, and performance requirements. Common algorithms include decision trees, random forests, support vector machines, neural networks, etc.
* **Training:** -Train the selected model on the labelled training data. The model learns to map inputs to outputs based on the provided examples.
* **Evaluation**: - Assess the performance of the trained model on unseen data (the test set). Common evaluation metrics include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).
* **Hyperparameter Tuning:** - Fine-tune the model's hyperparameters to optimize its performance. Techniques like grid search, random search, or Bayesian optimization are used for hyperparameter tuning.
* **Deployment:** -: Deploy the trained model into production to make predictions on new, unseen data. This can involve integrating the model into existing systems or developing APIs for external use.
* **Monitoring and Maintenance:** - Continuously monitor the deployed model's performance and retrain it periodically with new data to ensure its effectiveness over time.

Machine learning is a versatile tool with applications across various domains, including healthcare, finance, e-commerce, natural language processing, computer vision, and more. It's constantly evolving with advancements in algorithms, techniques, and computational resources.

6

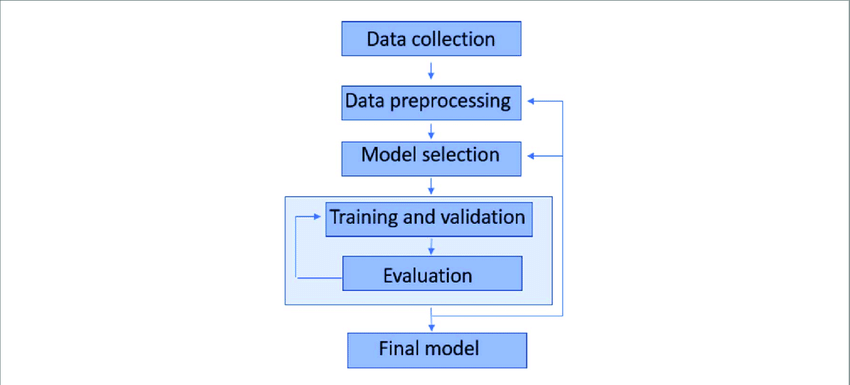


Fig: 1. Machine Learning Steps

**2.5 Sentiment Analysis: -**

The Sentiment analysis (SA) which is commonly known as opinion mining or contextual mining, is used in the Natural Language Processing (NLP), computational linguistics, text analysis which helps in identify, systematically extract and quantify,the subjective information. The sentiment analysis actually works widely in the form of a customer’s voice like reviews or responses on any material or item.

**Example**:

Suppose a customer wants to buy any item online, so before buy that item the customer generally reads reviews about that item or product and this will help to take

the right decision about that item.

Sentiment analysis uses three terms to define sentiment. These are, object about which opinion is given, features of that object, opinion holder who give his opinion about the object. Sentiment analysis handles various challenges such as identification of the object, feature extraction and finds the orientation of opinion. SentimentAnalysis performs the classification task in 3 steps:

* + Document level
  + Sentence level
  + Feature level or Aspect level

7

The **document level** of classification is used where the task is to find the overall polarity of a topic irrespective of opinion holder. Document-level sentiment analysis assumes opinion about the single entity is expressed by the document. This is true incase of product review, movie review, etc., where a document expresses the opinion about a singlemovie or a single product..The sentence is a shorter form of document as collection of sentence makes a document.

**Sentence level** classification assume search sentence holds a single opinion.

Here classification includes two subtasks: **subjectivity detection and opinion detection.**

At **the feature levelor aspect level**, theanalysis of various features of an object is performed.

**Example:**

Suppose a customerbuy a Samsung Mobile Phone, then he observed that the camera quality of the cellphone is fair but the sound quality of the cell phone is not fair. So, to analyze, thevarious aspect of an entity aspect level analysis is performed.



Fig.2. Sentiment Analysis Process

Sentiment Analysis includes Data Pre-processing, Feature Selection, and classification then find the polarity of data as shown in **Fig. 2**.

Data pre-processing includes tokenization, stop word removal, stemming, lemmatization, etc.

* **Tokenization** is a task of breaking a sequence of words into individual words called tokens.
* **Stop words** are the words (is, am, are, in, to, etc.) which do not hold any opinion, so it is beneficial to remove them.
* **Stemming** is a task of converting word’s variant forms to its base form like helping to help.

8

**2.6 Features Extraction in Sentiment Analysis: -**

Feature extraction in sentiment analysis involves converting raw text data into numerical or categorical features that capture relevant information for sentiment classification. Common techniques include:

* **Bag-of-Words (BoW):** Representing text as word frequency vectors.
* **Term Frequency-Inverse Document Frequency (TF-IDF):** Adjusting word frequencies based on importance.
* **Word Embeddings:** Converting words into dense vectors capturing semantic relationships.
* **N-grams:** Extracting sequences of n words for richer context.
* **Part-of-Speech (POS) Tagging:** Assigning grammatical categories to words.
* **Sentiment Lexicons:** Using annotated word lists for sentiment analysis.
* **Syntactic and Semantic Features:** Capturing structural and conceptual information.
* **Emotion Features:** Identifying emotional content in text.

Each technique provides different perspectives on the text data, enabling sentiment analysis models to effectively classify sentiments.

**2.7 Types of Sentiment Analysis:**

Sentiment analysis, also known as opinion mining, involves determining the sentiment expressed in text data. Here are the main types of sentiment analysis:

* **Document-Level Sentiment Analysis**: Analyses the sentiment of an entire document, such as a review or a tweet, and classifies it as positive, negative, or neutral.
* **Sentence-Level Sentiment Analysis**: Determines the sentiment of individual sentences within a document. Useful for identifying nuanced opinions within longer texts.
* **Aspect-Based Sentiment Analysis**: Focuses on identifying the sentiment towards specific aspects or entities mentioned in the text. For example, in a product review, it might analyze sentiments towards various features of the product.
* **Multimodal Sentiment Analysis**: Incorporates multiple modalities such as text, images, audio, and video to analyze sentiment. This can provide richer insights by considering non-textual cues.
* **Domain-Specific Sentiment Analysis**: Tailors sentiment analysis models to specific domains or industries, such as healthcare, finance, or hospitality, to capture domain-specific sentiment expressions accurately.
* **Fine-Grained Sentiment Analysis**: Goes beyond binary sentiment classification (positive/negative) and categorizes sentiments into multiple classes, such as very positive, positive, neutral, negative, very negative.

9

* **Temporal Sentiment Analysis**: Tracks changes in sentiment over time, allowing for the analysis of trends, fluctuations, or sentiment shifts in response to events or topics.

Each type of sentiment analysis serves different purposes and may require different approaches and techniques for accurate analysis, depending on the specific use case and requirements.

**2.8 Sentiment Analysis Methods: -**

Sentiment analysis methods are machine learning-based, lexicon-based and hybridmethod. In machine learning method labelled dataset is used where the polarity ofa sentence is already mentioned. From that dataset, we extract the feature and thatfeatures help to classify the polarity of the unknown input sentence. Machine learningmethods divided into supervised learning and unsupervised learning (Fig. 3).



Fig: 3. Sentiment Analysis Methods

10

**CHAPTER 3**

**SYSTEM DESIGN**

**3.1 System Architecture: -**

****

Fig 3.1.1 Architecture of a system

This architecture outlines the flow of data and processing steps involved in sentiment analysis on social media data, from data collection to sentiment analysis. Each component plays a crucial role in extracting insights from social media content.

* **Social Site:** Refers to the platform where data is collected from, such as Twitter, Facebook, Reddit, etc.
* Data Collection: Involves gathering text data from the social media site using APIs or web scraping techniques. This data includes posts, comments, tweets, etc., which will be used for sentiment analysis.
* **Data Processing:** Cleans and preprocesses the collected data to remove noise, such as URLs, emojis, and special characters. Tokenization, removing stopwords, and lemmatization or stemming are commonly applied preprocessing steps.
* **Data Filtering**: Filters the processed data to focus on relevant content for sentiment analysis. This may involve selecting specific topics, keywords, or user-generated content based on the analysis objectives.
* **Feature Extraction:** Extracts meaningful features from the filtered data to represent text numerically or categorically. Techniques such as Bag-of-Words (BoW), TF-IDF, word embeddings, or n-grams can be used to capture important information for sentiment analysis.

11

* **Classification Algorithm:** Utilizes machine learning or deep learning algorithms to classify text into sentiment categories, such as positive, negative, or neutral. Common classification algorithms include Support Vector Machines (SVM), Naive Bayes, Logistic Regression, Random Forests, or neural network-based models.
* **Sentiment Analysis:** Analyses the sentiment expressed in the social media data based on the classification results. It provides insights into the overall sentiment trends, sentiment distribution, and key findings, which can be visualized and reported to stakeholders.

**3.2 Data Flow Diagram: -**

****

Fig 3.2.1 Data Flow Diagram

This streamlined process outlines how a web application employs sentiment analysis with the Naïve Bayes algorithm to assess user comments' sentiment. The system receives user comments, processes them, trains a sentiment analysis model with labelled data, classifies the comments using Naïve Bayes or similar methods, and displays the sentiment polarity to the user. Through this iterative cycle of data processing and analysis, the application delivers insightful feedback on the sentiment expressed in the comments, enhancing user engagement and understanding.

* **Enter Comment:** Users input comments on the web application interface.
* **Comment Retrieval:** The web application retrieves the entered comments.
* **Request for Data:** The system requests relevant data from the social media platform or database.

12

* **Extract Data:** Data containing comments or posts related to the request is extracted from the social media platform or database.
* **Web Application:** The web application serves as the user interface for interacting with the system.
* **User:** The user interacts with the web application by entering comments.
* **Process Comment:** The system processes each comment to prepare it for sentiment analysis.
* **Training Data:** Utilizing labelled data, the system trains its sentiment analysis model.
* **Get Feature Vector**: Each comment is converted into a feature vector representing its characteristics.
* **Extract Feature**: Features relevant to sentiment analysis are extracted from the feature vector.
* **Get Training Sets:** Training sets containing feature vectors and their corresponding sentiment labels are prepared.
* **Display Pattern and Polarity:** The system analyses the training data to identify patterns and polarity (positive, negative, neutral).
* **Naïve Bayes Algorithm:** The sentiment analysis model utilizes the Naïve Bayes algorithm (or another classification algorithm) to classify the comments based on their features.
* **Calculate Polarity:** The system calculates the polarity (sentiment) of each comment based on the classification results.

**3.3 Activity Diagram: -**



Fig 3.3.1 Activity Diagram

13

**CHAPTER 4**

**IMPLEMENTATION**

**4.1 Methodology: -**

To conduct a sentiment analysis, social media monitoring tools use natural language processing (NLP) and machine learning (ML) to analyze and classify the sentiment expressed in social media posts and comments, and decipher if a message is positive, negative, or neutral.

* **Define objectives:** Determine the specific goals and objectives of the sentiment analysis project. This could include understanding customer sentiment towards a product, tracking brand perception, or identifying emerging trends.
* **Data collection:** Gather social media data from platforms such as Twitter, Facebook, Instagram, etc. Utilize APIs or web scraping techniques to extract relevant data.
* **Data preprocessing:** Clean the data by removing noise, irrelevant information, and formatting it into a suitable structure for analysis. This may involve tasks such as tokenization, removing stop words, and handling emojis and slang.
* **Sentiment analysis model selection:** Choose an appropriate sentiment analysis model based on the nature of the data and project requirements. This could involve using pre-trained models like BERT or developing custom models using machine learning algorithms such as Naive Bayes or Support Vector Machines.
* **Model training and evaluation:** Train the selected model using labelled data (if available) and evaluate its performance using metrics such as accuracy, precision, recall, and F1-score. Iterate on the model and fine-tune parameters as needed**.**

**4.2 Enrollment Phase:**

* **Stakeholder engagement:** Engage with stakeholders to understand their requirements, expectations, and constraints regarding the sentiment analysis project. This could involve conducting meetings, surveys, or interviews.
* **Resource allocation:** Allocate necessary resources including budget, personnel, and technology infrastructure for the implementation of sentiment analysis.
* **Team formation:** Form a multidisciplinary team comprising data scientists, software engineers, domain experts, and project managers to collaborate on the implementation process.
* **Establish timelines:** Define a timeline with milestones and deliverables to track progress and ensure timely completion of the project.

14

**4.3 Environmental Setup:**

* **Infrastructure setup:** Configure computing resources such as servers or cloud platforms for data storage, processing, and analysis.
* **Software installation**: Install and configure required software tools and libraries for data collection, preprocessing, analysis, and visualization.
* **Data integration:** Establish connections with social media APIs or set up web scraping tools to collect data from relevant sources.
* **Security measures:** Implement security protocols to protect sensitive data and prevent unauthorized access to the sentiment analysis system.
* **Testing environment:** Set up a testing environment to validate the functionality and performance of sentiment analysis algorithms before deployment in production.
  1. **Jupyter Notebook: -**

Jupyter Notebook is an open-source web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text. It supports various programming languages, including Python, R, Julia, and more. Originally, it was developed as part of the I Python project but has since evolved to support multiple languages.

Here are some key features and aspects of Jupyter Notebook:

* **Interactive Computing**: Jupyter Notebooks allow for interactive computing, meaning you can write and execute code in individual cells. This enables iterative development and easy experimentation.
* **Mixing Code with Text**: You can include formatted text, equations (using LaTeX), images, videos, links, and HTML in Jupyter Notebooks, making them suitable for documenting and explaining your code and analysis.
* **Kernel Support:** Jupyter Notebooks are powered by kernels, which are separate computing processes responsible for executing the code contained within a notebook document. Each kernel supports a specific programming language. For example, the IPython kernel runs Python code, while there are kernels available for other languages like R and Julia.
* **Rich Output:** Jupyter Notebooks can display a wide range of outputs, including plain text, images, HTML, LaTeX, JSON, audio, video, and custom MIME types. This makes it convenient for visualizing data and presenting results directly within the notebook.
* **Collaboration and Sharing**: Notebooks can be easily shared with others via email, Dropbox, GitHub, and the Jupyter Notebook Viewer. They can also be converted into various formats, such as HTML, PDF, and slideshows, for easier sharing and presentation.

15

* **Extensions and Widgets**: Jupyter supports extensions and widgets that enhance its functionality. Extensions provide additional features like code folding, table of contents, and spell-checking, while widgets enable interactive user interfaces within notebooks.
* **Support for Data Science and Education:** Jupyter Notebooks are widely used in data science, machine learning, scientific computing, and education due to their versatility, interactivity, and ease of use. They are commonly used for tasks such as data exploration, prototyping, visualization, and teaching.

Overall, Jupyter Notebook is a powerful tool for interactive computing and communication, making it popular among researchers, data scientists, educators, and anyone who wants to write code, document their work, and share their findings effectively.

* 1. **Introduction to Python: -**

Python is a high-level, versatile, and popular programming language known for its simplicity and readability. It's widely used in various domains, including web development, data analysis, machine learning, scientific computing, automation, and more. It was created by Guido van Rossum in 1991 and further developed by the Python Software Foundation. It was designed with an emphasis on code readability, and its syntax allows programmers to express their concepts in fewer lines of code.

It was created by Guido van Rossum in 1991 and further developed by the Python Software Foundation. It was designed with an emphasis on code readability, and its syntax allows programmers to express their concepts in fewer lines of code.

**Feature of Python: -**

1. **Interpreted**

• There are no separate compilation and execution steps like C and C++.

• Directly run the program from the source code.

• Internally, Python converts the source code into an intermediate form called bytecode which is then translated into native language of specific computer to run it.

• No need to worry about linking and loading with libraries, etc.

1. **Platform Independent**

* Python programs can be developed and executed on multiple operating system platforms.
* Python can be used on Linux, Windows, Macintosh, Solaris and many more.

1. **High-level Language**

* In Python, no need to take care about low-level details such as managing the memory used by the program.

16

1. **Simple**

* Closer to English language;
* Easy to Learn
* More emphasis on the solution to the problem rather than the syntax Embeddable
* Python can be used within C/C++ program to give scripting capabilities for the program’s users.

1. **Robust:**

* Exceptional handling features
* Memory management techniques in built.

1. **Rich Library Support:**

* The Python Standard Library is very vast.
* Known as the **“batteries included”** philosophy of Python;
* It can help do various things involving regular expressions, documentation generation, unit testing, threading, databases, web browsers, CGI, email, XML, HTML, WAV files, cryptography, GUI and many more.
* Besides the standard library, there are various other high-quality libraries such as the Python Imaging Library which is an amazingly simple image manipulation library**.**
  1. **Introduction to Natural Language Processing (NLP): -**

NLP, or Natural Language Processing, is a field of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. It involves tasks such as text processing, language understanding, and language generation using techniques like machine learning and deep learning. NLP finds applications in search engines, virtual assistants, sentiment analysis, machine translation, text summarization, and more.

17

**CHAPTER 5**

**LIBRARIES AND ALGORITHMS**

**5.1 Libraries: -**

**5.1.1 Python Libraries: -**

**Pandas:** A library for data manipulation and analysis. It provides data structures like Data Frames and Series for handling structured data.

**NumPy:** A library for numerical computing in Python. It provides support for arrays, matrices, and various mathematical operations.

**matplotlib:** A plotting library for creating static, interactive, and animated visualizations in Python.

**seaborn:** A statistical data visualization library based on matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics.

**colorama:** A library for generating console text with colored output on Windows and Linux/OS X terminals.

**plotly.express:** A high-level library for creating interactive visualizations based on the plotly.js library.

**re:** A library for regular expressions in Python, which is used for text processing and manipulation.

**nit**: The Natural Language Toolkit, a comprehensive library for natural language processing in Python. It includes tools for tokenization, stemming, sentiment analysis, and more.

**string:** A built-in Python library for string manipulation and formatting.

warnings: A built-in Python library for controlling and filtering warning messages during program execution.

**sklearn:** Scikit-learn, a machine learning library for Python. It includes tools for data preprocessing, model training, model evaluation, and hyperparameter tuning.

**tfidfVectorizer**: A class from sklearn. feature\_extraction.text for converting text data into a matrix of TF-IDF features.

**PassiveAggressiveClassifier, LogisticRegression, RandomForestClassifier, SVC,and MultinomialNB**: Classes from sklearn.linear\_model, sklearn.ensemble, sklearn.svm, and sklearn.naive\_bayes for training machine learning models.

**train\_test\_split:** A function from sklearn.model\_selection for splitting data into training and testing sets.

**accuracy\_score and classification\_report**: Functions from sklearn.metrics for evaluating the performance of machine learning models.

**RandomizedSearchCV**: A class from sklearn.model\_selection for performing randomized hyperparameter tuning.

These libraries are used to perform data preprocessing, exploratory data analysis, visualization, and machine learning tasks on the sentiment dataset.

"Pip install twython" is a command to install Twython, a Python wrapper for the Twitter API. Twython is an open source library that supports Python 3 and can be used to access Twitter data. It can be used for:

* Querying data for user information, Twitter lists, timelines, direct messages, and anything found in the Twitter API docs
* Uploading images Updating user status with an image, Changing user avatar, Changing user background image, and Changing user banner image.

pip install vaderSentiment. VADER Sentiment Analysis: VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

18

**5.1.2 NLP Libraries: -**

**nltk (Natural Language Toolkit):** A comprehensive library for NLP tasks, including tokenization, stemming, sentiment analysis, and more.

**nltk.corpus:** A module in the NLTK library that provides access to various text corpora and linguistic data.

**nltk.tokenize:** A module in the NLTK library that provides functions for tokenizing text into words, sentences, and other units.

**nltk.stem:** A module in the NLTK library that provides classes for stemming words, such as the Porter Stemmer used in the code.

**nltk.sentiment:** A module in the NLTK library that provides functions for sentiment analysis, including the VADER (Valence Aware Dictionary and sentiment Reasoner) sentiment analysis tool used in the code.

**nltk.download:** A function in the NLTK library that allows you to download various linguistic data and corpora, such as the VADER lexicon and the Punkt tokenizer.

**collections.Counter:** A class in the Python collections library that provides a simple way to count the frequency of elements in a list or other iterable.

These libraries are used for various NLP tasks, such as text cleaning, tokenization, stemming, and sentiment analysis, in the given code.

**5.2 Algorithms: -**

**Passive Aggressive Classifier (PAC):**

* Passive Aggressive algorithms are online learning algorithms.
* They are well suited for tasks where data streams in continuously or where there are resource constraints.
* In the context of text classification, Passive Aggressive classifiers aim to correctly classify instances while minimizing the loss.
* The algorithm is called "passive" because it does not significantly update the model for correct classifications, and it's called "aggressive" because it updates aggressively for misclassified instances.
* It's often used in scenarios where rapid adaptation to changing data distributions is necessary.

**Logistic Regression:**

* Logistic Regression is a popular classification algorithm used when the target variable is binary.
* It models the probability that a given input belongs to a particular category.
* Despite its name, it's a classification algorithm rather than a regression one.
* It works by fitting a sigmoid function to the input data, which maps inputs to probabilities.

19

**Random Forest Classifier:**

* Random Forest is an ensemble learning method that constructs a multitude of decision trees during training.
* Each tree in the forest predicts the target variable, and the final prediction is determined by aggregating the predictions of all the trees (e.g., by averaging or taking a majority vote).
* Random Forest is known for its robustness, scalability, and ability to handle high-dimensional data with complex interactions.

**Support Vector Machine (SVM):**

* SVM is a powerful supervised learning algorithm used for classification and regression tasks.
* It works by finding the hyperplane that best separates the classes in the feature space.
* SVM aims to maximize the margin between the classes, which leads to better generalization performance.
* It's effective in high-dimensional spaces and is particularly useful when the number of dimensions exceeds the number of samples.

**Multinomial Naive Bayes:**

* Naive Bayes classifiers are based on Bayes' theorem and assume that features are conditionally independent given the class.
* Multinomial Naive Bayes is specifically designed for classification with discrete features (e.g., word counts in text classification).
* It's commonly used in text classification tasks like spam filtering, sentiment analysis, and document categorization.

These algorithms are applied to perform sentiment analysis on textual data, where the goal is to classify the sentiment of the text as positive, negative, or neutral. Each algorithm learns patterns from the text features (e.g., word frequencies or TF-IDF scores) and predicts the sentiment label for new unseen text data. The performance of each algorithm is evaluated using metrics like accuracy, precision, recall, and F1-score.

20

**CHAPTER 6**

**DATASET**

**6.1 About Dataset: -**

The **Social Media Sentiments Analysis Dataset** captures a vibrant tapestry of emotions, trends, and interactions across various social media platforms. This dataset provides a snapshot of user-generated content, encompassing text, timestamps, hashtags, countries, likes, and retweets. Each entry unveils unique stories—moments of surprise, excitement, admiration, thrill, contentment, and more—shared by individuals worldwide.

**6.2 Key Features**

|  |  |
| --- | --- |
| **Feature** | **Description** |
| Text | User-generated content showcasing sentiments |
| Sentiment | Categorized Emotions |
| Timestamp | Date and time information |
| User | Unique identifiers of users contributing |
| Platform | Social media platform where the content originated |
| Hashtags | Identifies trending topics and themes |
| Likes | Quantifies user engagement (likes) |
| Retweets | Reflects content popularity (retweets) |
| Country | Geographical origin of each post |
| Year | Year of the post |
| Month | Month of the post |
| Day | Day of the post |
| Hour | Hour of the post |

**6.3 Use of Dataset**

The Social Media Sentiments Analysis Dataset is a rich source of information that can be leveraged for various analytical purposes. Below are key ways to make the most of this dataset:

**Sentiment Analysis:**

* Explore the emotional landscape by conducting sentiment analysis on the "Text"

Column.

* Classify user-generated content into categories such as surprise, excitement, admiration, thrill, contentment, and more.

**Temporal Analysis:**

* Investigate trends over time using the "Timestamp" column.
* Identify patterns, fluctuations, or recurring themes in social media content.

21

**User Behavior Insights:**

* Analyze user engagement through the "Likes" and "Retweets" columns.
* Discover popular content and user preferences.

**Platform-Specific Analysis:**

* Examine variations in content across different social media platforms using the "Platform" column.
* Understand how sentiments vary across platforms.

**Hashtag Trends:**

* Identify trending topics and themes by analyzing the "Hashtags" column.
* Uncover popular or recurring hashtags

**Geographical Analysis:**

* Explore content distribution based on the "Country" column.
* Understand regional variations in sentiment and topic preferences.

**User Identification:**

* Use the "User" column to track specific users and their contributions.
* Analyze the impact of influential users on sentiment trends.

**Cross-Analysis:**

* Combine multiple features for in-depth insights.
* For example, analyze sentiment trends over time or across different platforms and countries.

**6.4 Columns: -**

* **Text:** Text content of the post.
* **Sentiment:** Sentiment of the post (positive, negative, or neutral).
* **Timestamp:** Date and time of the post.
* **User:** Username of the poster.
* **Platform:** Social media platform where the post was made (Twitter, Instagram, Facebook).
* **Hashtags:** Hashtags used in the post.
* **Retweets:** Number of retweets for the post.
* **Likes:** Number of likes for the post.
* **Country:** Country of origin for the poster.
* **Year:** Year in which the post was made.
* **Month:** Month in which the post was made.
* **Day:** Day of the month on which the post was made.
* **Hour:** Hour of the day at which the post was made.

**6.5 Link of the Dataset: -**

[Social Media Sentiments Analysis Dataset 📊 (kaggle.com)](https://www.kaggle.com/datasets/kashishparmar02/social-media-sentiments-analysis-dataset)

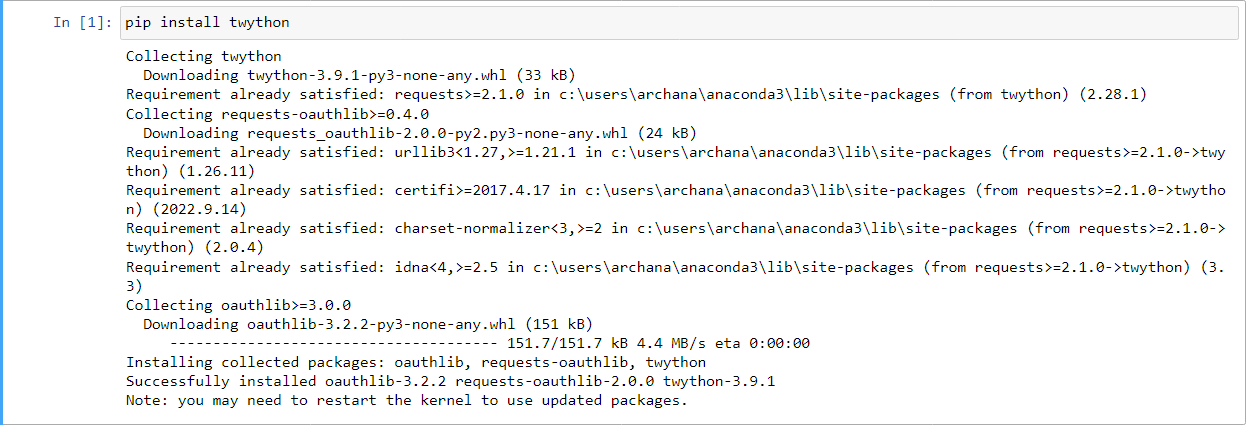
22

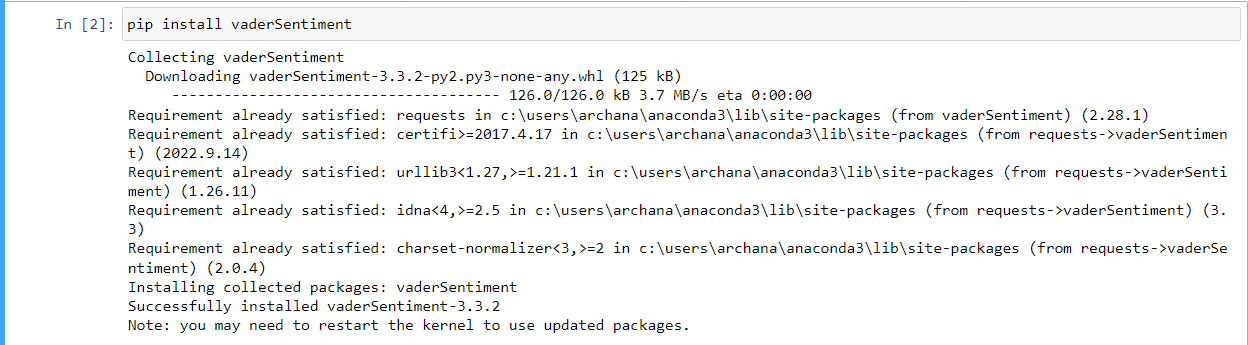
**CHAPTER 7**

**SOURCE CODE AND CONCLUSION**

**7.1 Source code and Output**

**Import Libraries**

****

****

23

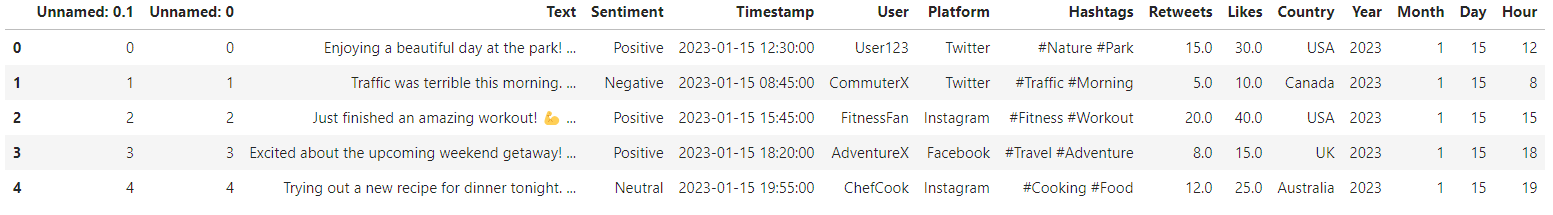


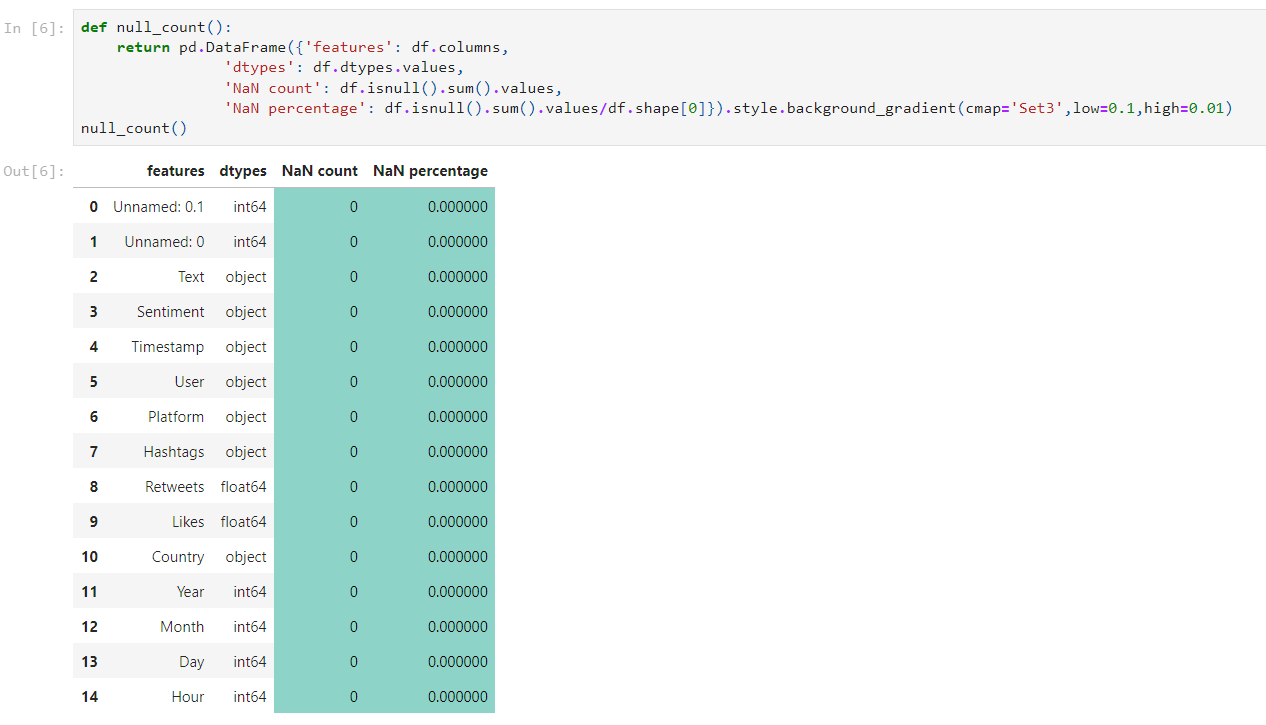
24

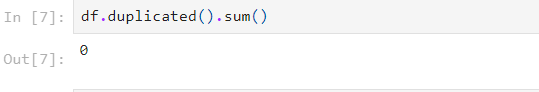
**Load Data**

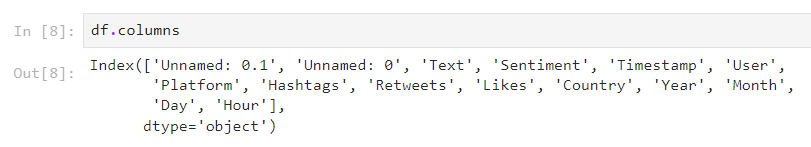
****

****

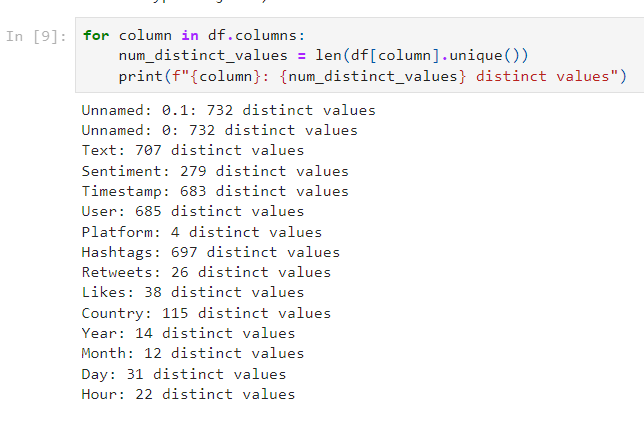
****

****

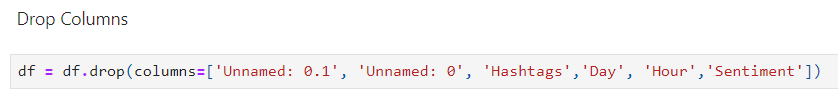
****

****

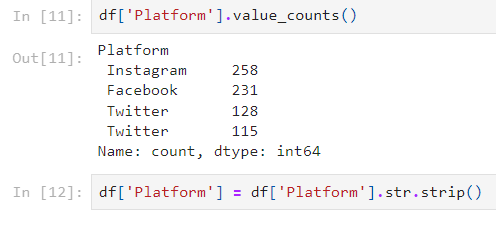
25



**Feature Engineering: -**

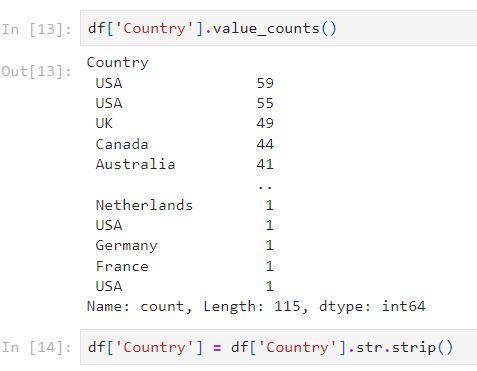
****

**Platform: -**

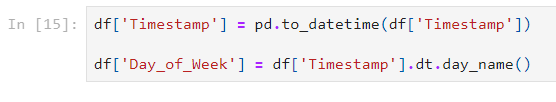
****

26

**Country**

****

**Timestamp**

****

**Month**

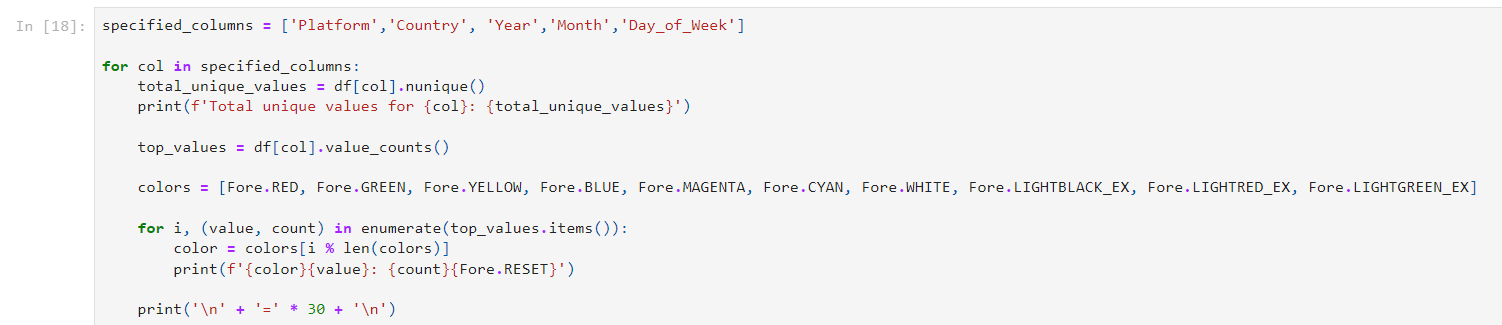
****

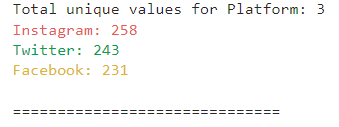
27

**Text**

****

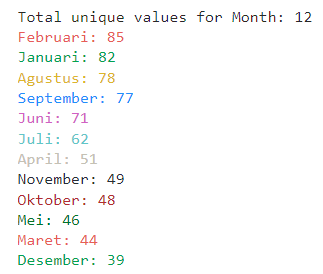
**Unique Columns**

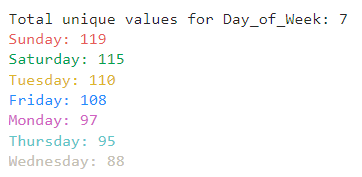
****

****

28

****

****

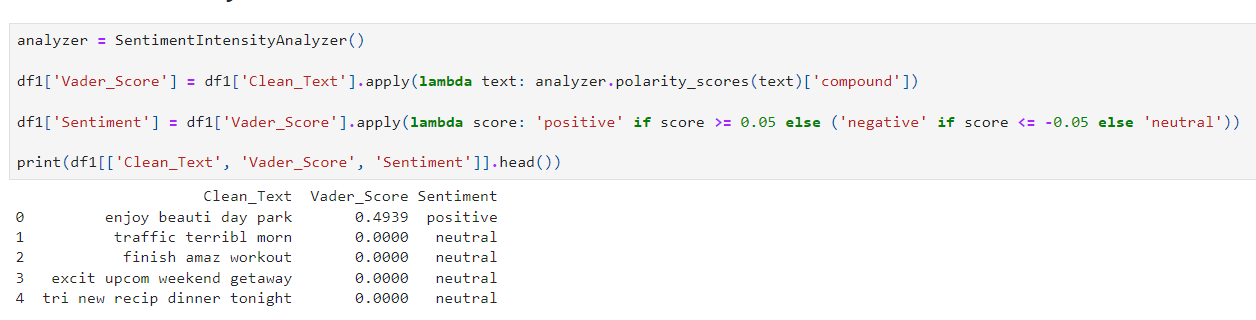
****

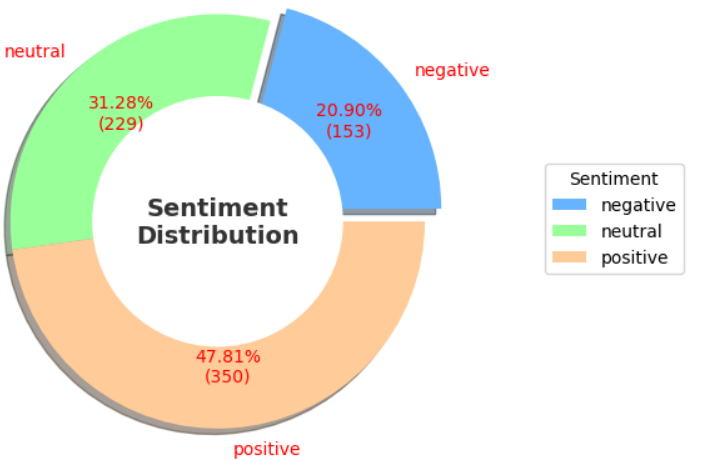
29

**E D A**

****

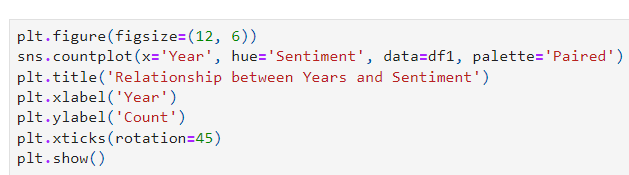
**Sentiment Analysis**

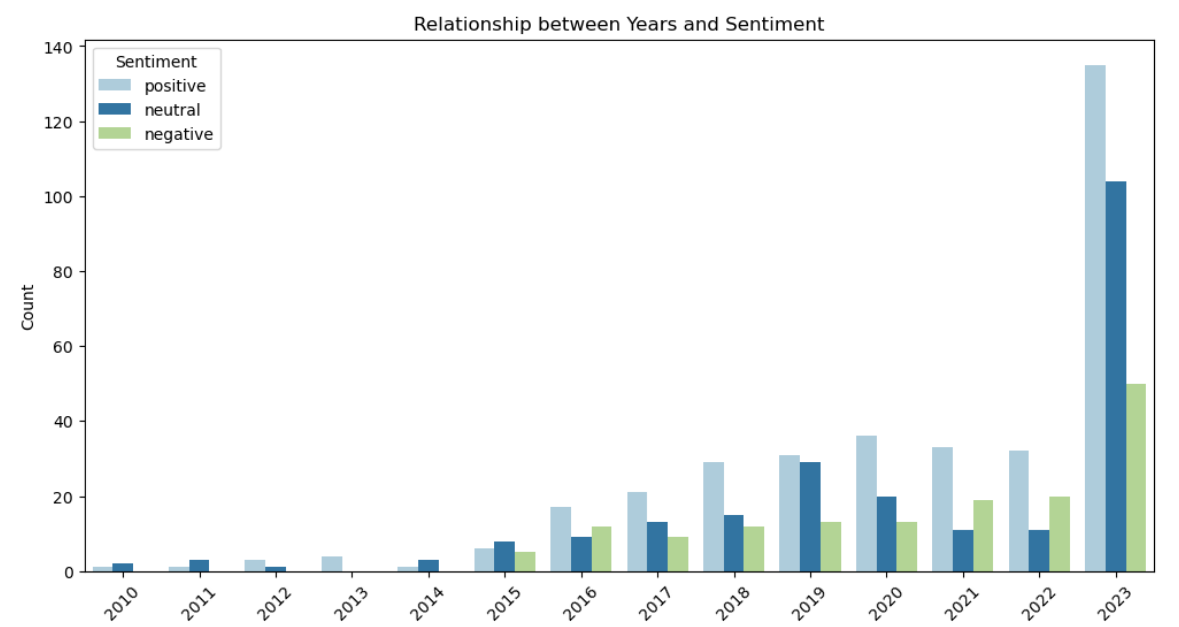
****

****

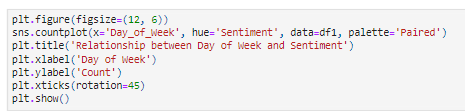
30

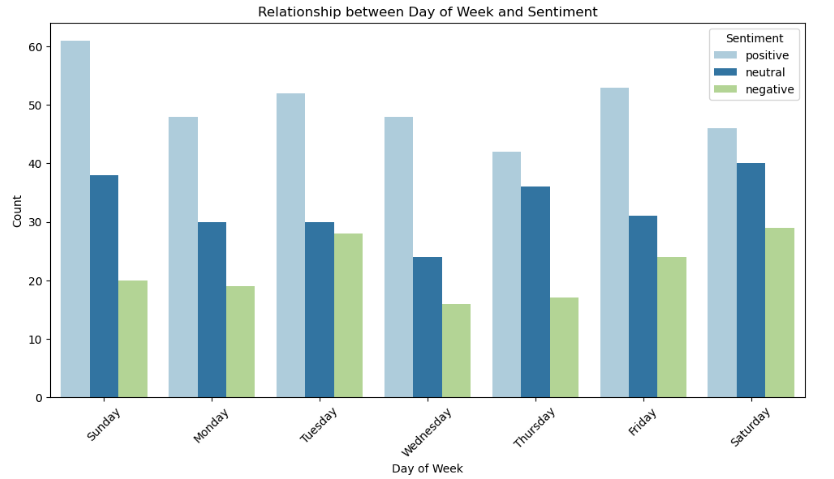
**Year**

****

****

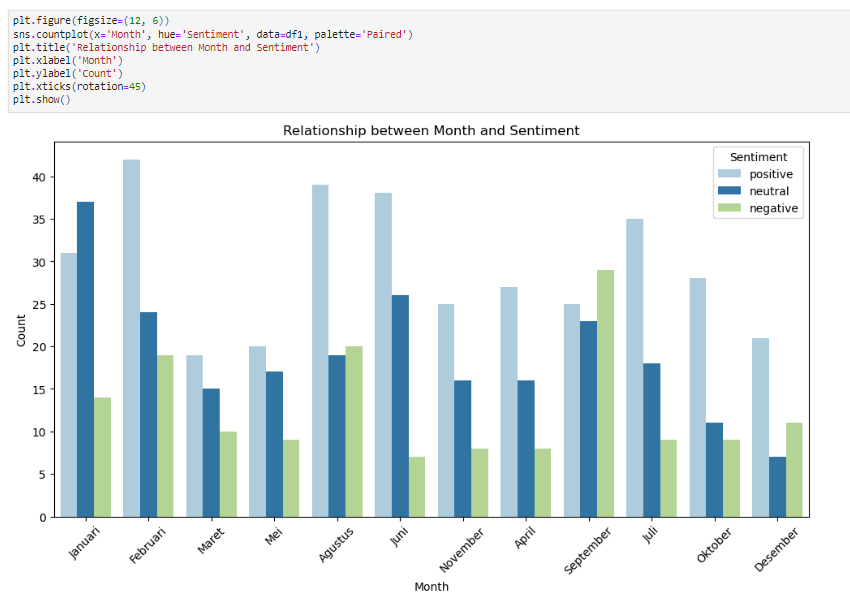
**Day of Week**

****

****

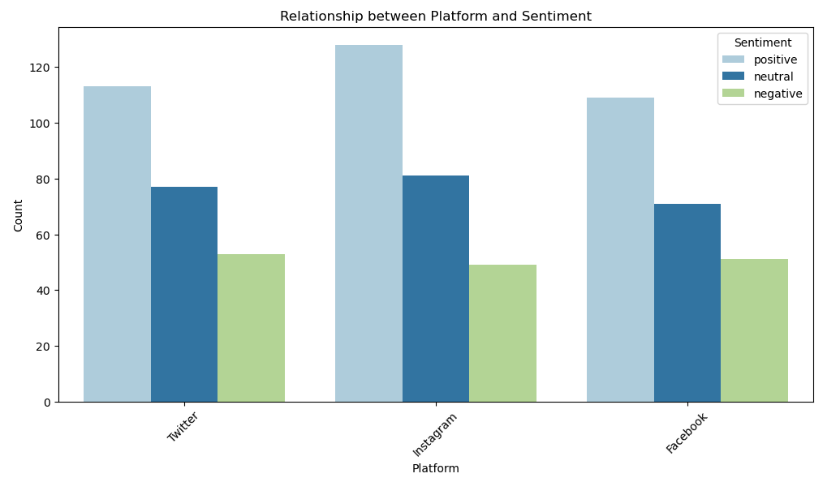
31

**Month**

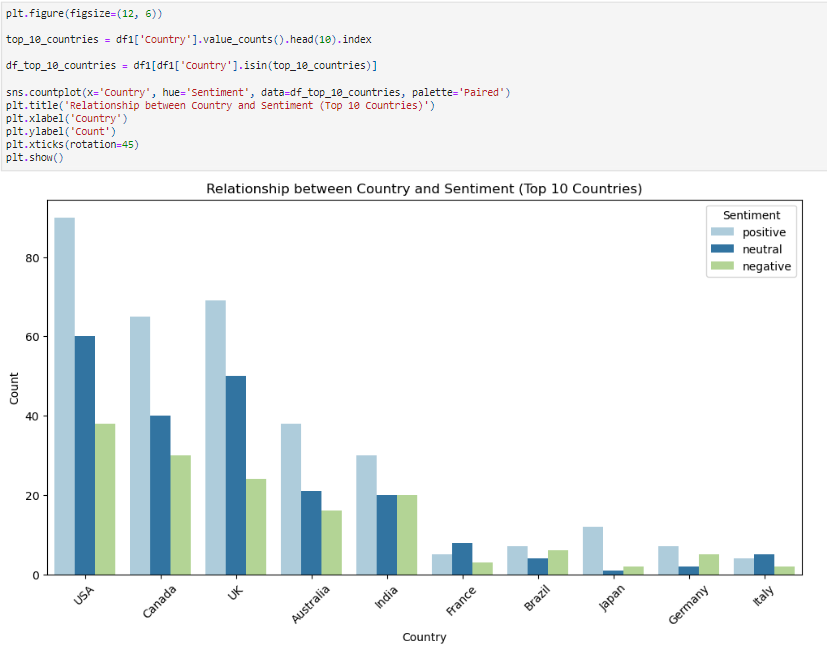
****

**Platform**

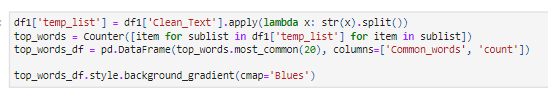
****

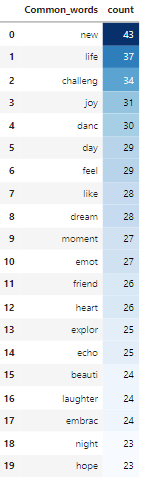
****

32

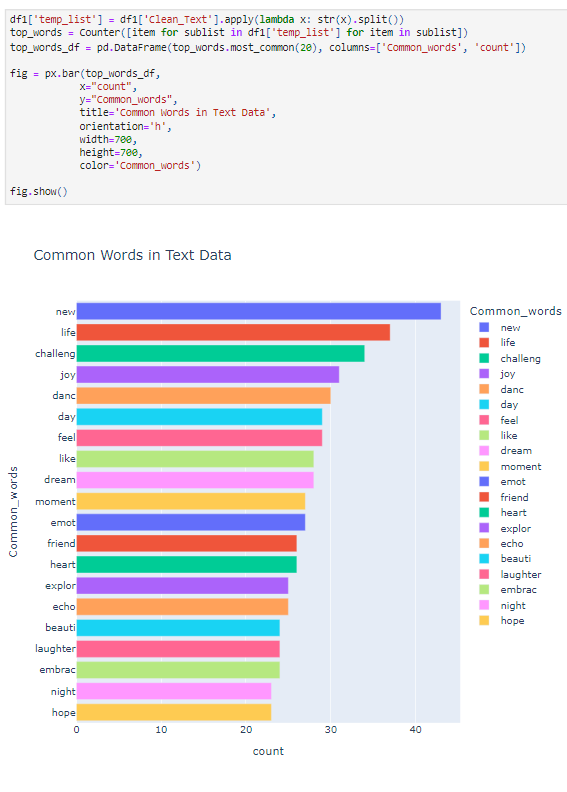


**Common Words**

****

****

33





34

**Positive Common Words**

****

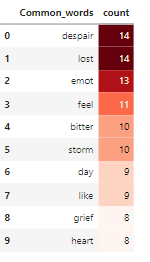
**Neutral Common Words**

****

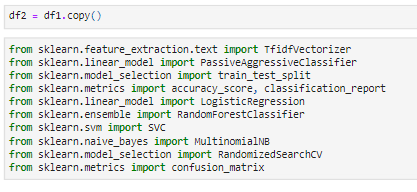
**Negative Common Words**

****

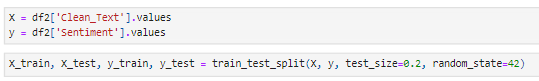
35



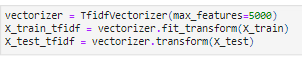
**Data Preparation**

****

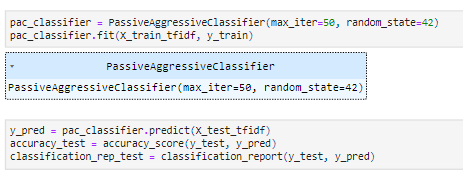
**Split Data**

****

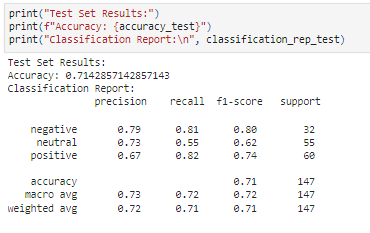
**Modelling**

****

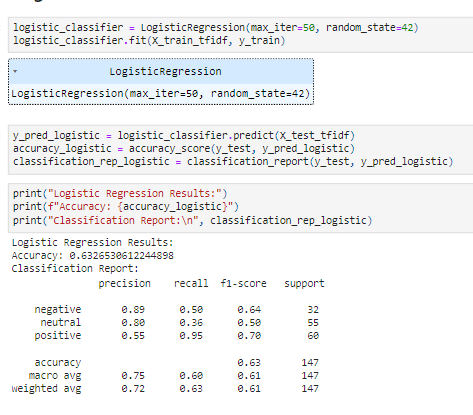
**Passive Aggressive Classifier**

****

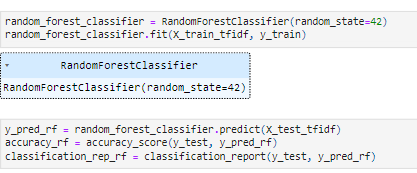
36



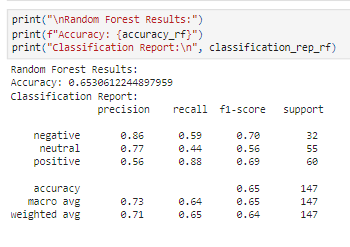
**Logistic Classifier**

****

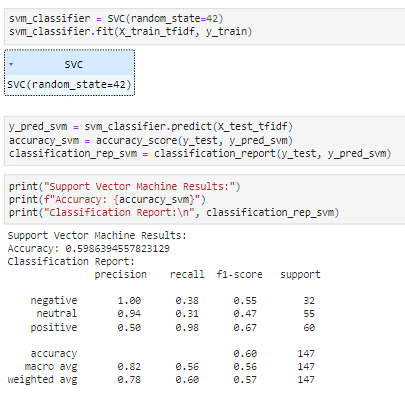
**Random Forest Classifier**

****

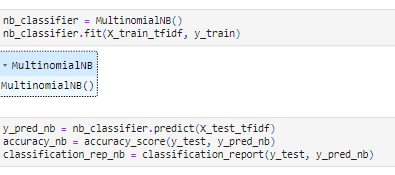
37

****

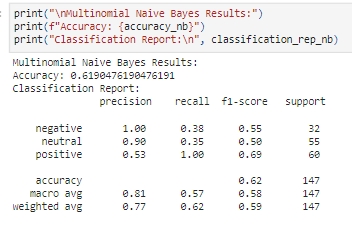
**SVM Classifier**

****

**Multinomial NB**

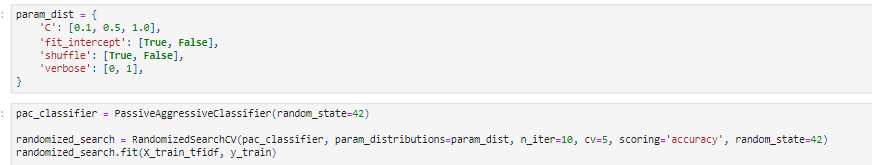
****

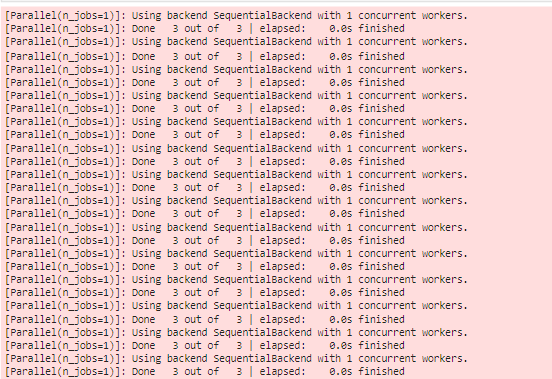
38



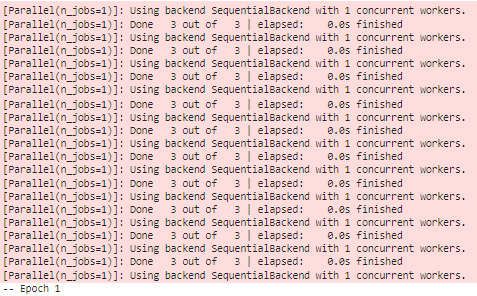
**Best Modelling: Passive Aggressive Classifier**

1. **Hyperparameters: -**

****

****

39

****

**7.2 Conclusion:-**

Sentiment analysis has proven to be a valuable tool in gauging public opinion in various disciplines, including financial market prediction, health issues, customer analytics, commercial valuation assessment, brand marketing, politics, crime prediction, and emergency management. The majority of studies have focused on sentiment analysis of Twitter messages due to the large and diverse population expressing opinions on this platform daily. This paper provides a comprehensive review of sentiment analysis in social networks, exploring methods, temporal dynamics, causal relationships, and applications in industry. The review emphasizes the importance of temporal characterization and causal effects in sentiment analysis in social networks, and their applications in different contexts such as stock market value, politics, and cyberbullying in educational centers. There is a strong interest from industry in this discipline, with over 8,000 patents issued on the topic in five years, and over 2,300 articles published in 15 years. However, there is still room for research opportunities in domains, techniques, and practical applications.

40

**CHAPTER 8**

**APPENDIX**

1. **Sentiment Analysis**: The process of determining the sentiment expressed in text data, usually categorized as positive, negative, or neutral.
2. **VADER Sentiment:** A lexicon and rule-based sentiment analysis tool specifically designed for social media text, developed by researchers at Georgia Tech.
3. **Data Preprocessing:** The initial step in data analysis where raw data is cleaned, transformed, and organized to prepare it for further analysis.
4. **Tokenization:** The process of breaking down text data into individual words or tokens for analysis.
5. **Stemming:** A text normalization technique that removes affixes from words to obtain their root form, reducing inflectional forms to their base or root word.
6. **TF-IDF (Term Frequency-Inverse Document Frequency):** A statistical measure used to evaluate the importance of a word in a document relative to a collection of documents, commonly used in text mining and information retrieval.
7. **Machine Learning Models:** Algorithms and statistical models that enable computers to learn from data and make predictions or decisions without being explicitly programmed.
8. **Passive Aggressive Classifier:** A type of online learning algorithm used for classification tasks that updates its model only when it makes a mistake, making it suitable for large-scale and real-time applications.
9. **Logistic Regression:** A statistical method used for binary classification that predicts the probability of an outcome based on one or more predictor variables.
10. **Random Forest Classifier:** An ensemble learning method that builds multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.
11. **Support Vector Machine (SVM):** A supervised learning algorithm used for classification and regression tasks that finds the hyperplane that best separates the classes in the feature space.
12. **Naive Bayes Classifier:** A probabilistic machine learning algorithm based on Bayes' theorem that assumes independence between features, often used for text classification tasks.
13. **Hyperparameter Tuning:** The process of selecting the best set of hyperparameters for a machine learning model to optimize its performance on unseen data.
14. **RandomizedSearchCV:** A technique used for hyperparameter tuning that randomly searches the hyperparameter space to find the best parameter values for a given machine learning model.
15. **Classification Report:** A summary of the performance of a classification model that includes metrics such as accuracy, precision, recall, and F1 score for each class.

41