**PROJECTTITLE**:**DECODING EMOTIONS THROUGH SENTIMENT ANALYSIS OF SOCIAL MEDIA CONVERSATIONS**

**PHASE-2**

**STUDENT NAME : MEENATCHI V**

**REGISTER NUMBER: 620523243035**

**INSTITUTION : CMS COLLEGE OF**

**ENGINEERING ,**

**NAMAKKAL .637003**

**DEPARTMENT : B.TECH- ARTIFICIAL**

**INTELLIGENCE AND**

**DATA SCIENCE**

**DATE OF SUBMISSION: 01/05/2025**

**GITHUB REPOSITORY**

**LINK :** <https://github.com/Meenatchi979/nm-project.git>

1.**Problem Statement**:

The proliferation of social media platforms has resulted in an unprecedented volume of textual data reflecting public opinions, emotions, and attitudes towards various topics, events, products, and services. Understanding these underlying emotions can provide valuable insights for businesses, policymakers, and researchers across diverse domains. However, manually analyzing this vast amount of unstructured text data is time-consuming and impractical. There is a need for automated and efficient methods to accurately identify and categorize the sentiment expressed in social media conversations, enabling a deeper understanding of public sentiment and its dynamics. Inaccurate or superficial sentiment analysis can lead to misinterpretations and flawed decision-making. Therefore, this project aims to develop a robust system for decoding emotions from social media conversations using sentiment analysis techniques

2.**Project Objectives**:

The primary objectives of this project are:

1. Develop a system capable of automatically analyzing the sentiment expressed in social media conversations.

2. Explore and implement various Natural Language Processing (NLP) and Machine Learning (ML)/Deep Learning (DL) techniques for sentiment classification.

3. Identify and analyze relevant features from social media text data that contribute to accurate sentiment detection.

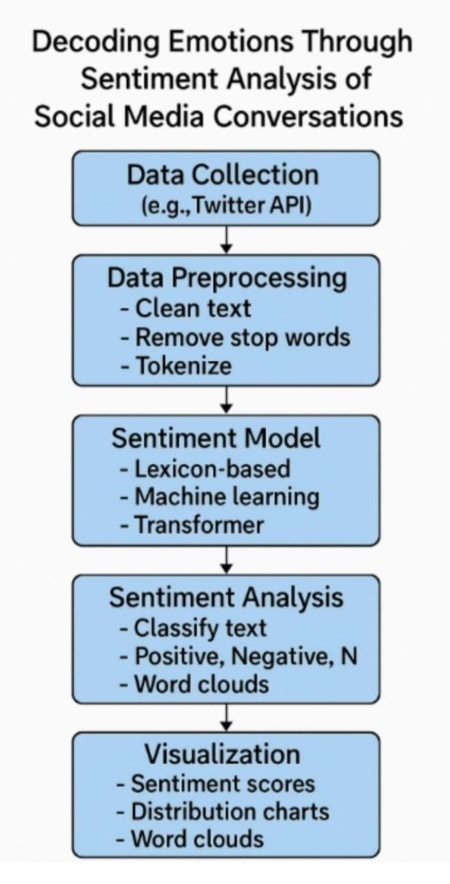
4. Build and evaluate different sentiment analysis models (e.g., lexicon-based approaches, traditional ML classifiers, deep learning models like RNNs and Transformers).

5. Visualize the sentiment distribution and trends within the analyzed social media data.

6. Gain insights into the emotional responses towards specific topics or entities discussed in the social media conversations.

7. Identify potential tools and technologies suitable for building and deploying a sentiment analysis system

**3. Flowchart of the Project Workflow:**



**4.Data Description:**

The dataset for this project will consist of social media conversations, likely from platforms like Twitter, Reddit, Facebook (if publicly available), or specific forums. The data will typically include:

● **Text Content:** The actual text of the social media post or comment.

● **Timestamp:** The date and time of the post.

● **User Information (Optional):** Username, profile details (may be anonymized for privacy).

● **Metadata (Optional):** Likes, retweets, replies, hashtags, mentions.

● **Ground Truth Labels (if available):** Manually annotated sentiment labels (e.g., positive, negative, neutral, or more granular emotions like joy, anger, sadness, fear). If no labeled data is readily available, data labeling will be a crucial initial step.

The dataset will vary in size depending on the chosen platform and the scope of the analysis (e.g., specific keywords, timeframes). We will aim for a dataset that is representative and sufficiently large for training robust models. Considerations regarding data privacy and ethical use will be paramount.

**5. Data Preprocessing:**

Before analysis, the raw social media text data will undergo several preprocessing steps:

1. **Data Cleaning:**

○ Removing irrelevant characters (URLs, special symbols, emojis - with consideration for their sentiment).

○ Handling HTML tags or other platform-specific formatting.

○ Addressing inconsistencies in text encoding.

2. **Text Normalization:**

○ Converting text to lowercase.

○ Tokenization: Splitting the text into individual words or subwords.

**cription:**

○ Stop word removal: Eliminating common words that often don't carry significant sentiment (e.g., "the," "a," "is").

○ Stemming or Lemmatization: Reducing words to their root form to handle variations (e.g., "running," "ran," "runs" to "run").

○ Handling negations (e.g., "not good" might be transformed to "not\_good" to preserve meaning).

3. **Handling Social Media Specifics:**

○ Processing mentions (@usernames).

○ Handling hashtags (#topics).

○ Addressing slang, abbreviations, and internet jargon.

**6. Exploratory Data Analysis (EDA):**

EDA will be conducted to understand the characteristics of the social media data and the distribution of sentiments:

1. **Descriptive Statistics:** Analyzing the length of posts, frequency of words, and distribution of users (if available).

2. **Sentiment Distribution Analysis:** Examining the overall distribution of sentiment labels (if available).

3. **Word Frequency Analysis:** Identifying the most frequent words and their association with different sentiments.

4. **Visualization:**

○ Histograms and bar charts to show sentiment distribution.

○ Word clouds to visualize the most frequent words for each sentiment category.

○ Time series plots to observe sentiment trends over time.

○ Box plots to compare the length of posts across different sentiment categories.

5. **Analysis of Hashtags and Mentions:** Exploring the relationship between specific hashtags or mentioned users and the expressed sentiment.

**7. Feature Engineering:**

To train effective sentiment analysis models, relevant features will be extracted from the preprocessed text data:

1. **Text-based Features:**

○ **Bag-of-Words (BoW):** Representing text as a collection of its words, ignoring grammar and word order.

○ **TF-IDF (Term Frequency-Inverse Document Frequency):** Weighting words based on their frequency in a document and their rarity across the entire corpus.

○ **N-grams:** Sequences of N words, capturing some contextual information.

○ **Word Embeddings (e.g., Word2Vec, GloVe, FastText):** Dense vector representations of words that capture semantic relationships.

○ **Document Embeddings (e.g., Doc2Vec):** Vector representations of entire documents.

2. **Linguistic Features:**

○ Presence of specific sentiment-laden words or phrases (using lexicons).

○ Use of intensifiers (e.g., "very," "extremely").

○ Presence of negations.

○ Punctuation and capitalization patterns.

○ Syntactic features (e.g., part-of-speech tags).

3. **Social Media Specific Features (if available):**

○ Number of likes, retweets, or replies.

○ Presence of specific emojis (if not removed).

○ Sentiment of related posts or comments in a thread.

**8. Model Building:**

Various machine learning and deep learning models will be explored and built for sentiment classification:

1. **Lexicon-based Approaches:**

Using predefined dictionaries of words associated with positive and negative sentiments (e.g., VADER, SentiWordNet).

2. **Traditional Machine Learning Classifiers:**

○ Naive Bayes

○ Support Vector Machines (SVM)

○ Logistic Regression

○ Random Forest

○ Gradient Boosting (e.g., XGBoost, LightGBM)

3. **Deep Learning Models:**

○ Recurrent Neural Networks (RNNs), specifically LSTMs and GRUs, to capture sequential dependencies in text.

○ Convolutional Neural Networks (CNNs) for text classification.

○ Transformer-based models (e.g., BERT, RoBERTa, DistilBERT) that leverage attention mechanisms for contextual understanding.

For each model, the process will involve:

● Splitting the data into training, validation, and testing sets.

● Training the model on the training data using the engineered features.

● Tuning hyperparameters using the validation set to optimize performance.

**9. Visualization of Results & Model Insights:**

Visualizations will be used to present the results of the sentiment analysis and provide insights into how the models are making predictions:

● **Sentiment Distribution Charts:** Pie charts or bar charts showing the overall proportion of different sentiments in the analyzed data.

● **Sentiment Trends Over Time:** Line graphs showing how sentiment towards a specific topic evolves over time.

● **Geographical Sentiment Maps (if location data is available):** Visualizing sentiment distribution across different regions.

● **Comparison of Model Performance:** Bar charts comparing the evaluation metrics (accuracy, precision, recall, F1-score) of different models.

● **Confusion Matrices:** Showing the types of errors made by the best-performing model.

● **Feature Importance Visualizations:** Highlighting the features that have the most significant impact on the model's predictions (e.g., using coefficients from linear models or feature importance scores from tree-based models).

● **Attention Maps (for Transformer models):** Visualizing which words the model is paying the most attention to when making a sentiment prediction for a given text.

● **Examples of Correct and Incorrect Predictions:** Displaying sample social media posts along with the model's predicted sentiment and the actual sentiment (if available) to understand the model's strengths and weaknesses.

**10. Tools and Technologies Used:**

● **Programming Language:** Python

● **NLP Libraries:** NLTK, spaCy, Transformers (Hugging Face)

● **Machine Learning Libraries:** scikit-learn

● **Deep Learning Libraries:** TensorFlow (with Keras), PyTorch

● **Data Manipulation and Analysis Libraries:** Pandas, NumPy

● **Data Visualization Libraries:** Matplotlib, Seaborn, Plotly

● **Text Annotation Tools (if manual labeling is required):** Label Studio, Doccano

● **Cloud Computing Platforms (Optional, for large datasets and complex models):** Google Cloud Platform (GCP), Amazon Web Services (AWS), Microsoft Azure.

**11. Team Members and Contributions:**

● **MEENATCHI V:** Project Lead and coordinating tasks.

● **DEEPAK V:** Data Engineer and setting up data pipelines.

● **ABINESH N:** NLP/ML Engineer and evaluation.

● **GOWSICK S:** Data Analyst/Visualizer and deriving insights.