Student Name: A. ANAL

Register Number: 612823205005

Institution: VARUVAN VADIVELAN INSTITUTE OF TECHNOLOGY

Department: INFORMATION TECHNOLOGY

Date of Submission: 03.05.2025

Github Repository Link:

Problem Statement

Social media platforms have become a primary medium for human interaction, hosting a vast amount of user-generated content reflecting emotions, opinions, and sentiments. However, interpreting emotions accurately from textual conversations poses significant challenges due to sarcasm, ambiguity, and varying linguistic styles.

Traditional sentiment analysis methods often struggle to distinguish nuanced emotional expressions, limiting their effectiveness in applications such as mental health monitoring, customer feedback analysis, and social trend prediction.

 Sentiment **analysis** traditionally focuses on polarity: positive, negative, or neutral.

 Emotion **detection** adds granularity—decoding feelings like fear, surprise, happiness.

 Social media platforms provide large-scale, real-time emotional data.

 Combining **classification** (categorical emotion) and **regression** (intensity scoring) offers a deeper understanding.

Project Objectives

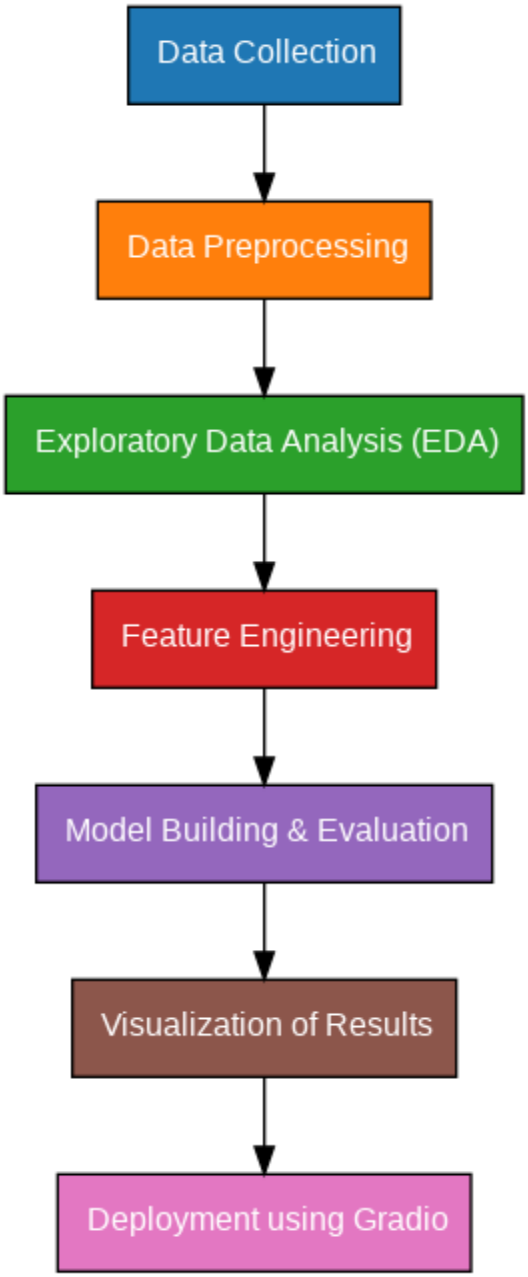
1. **To collect and preprocess social media conversation data**
   Gather real-world text data from platforms like Twitter, Reddit, or Facebook and apply natural language preprocessing techniques (tokenization, normalization, etc.).

2. **To classify text into emotional categories using machine learning models**
   Develop and evaluate models that can accurately classify text into discrete emotions such as happiness, anger, sadness, fear, etc.

3. **To predict the intensity of emotions using regression models**
   Apply regression techniques to estimate how strongly an emotion is expressed within a given text (e.g., on a scale from 0 to 1).

4. **To compare the performance of traditional ML and deep learning models**
   Benchmark models like Logistic Regression, SVM, Random Forest, LSTM, and BERT to determine which provides the best performance for emotion decoding tasks.

Flowchart of the Project Workflow



Data Description

1. Data Source

The dataset(s) for this project can be obtained from:

- **Public Datasets**:
  - **GoEmotions** (Google): Annotated Reddit comments with 27 emotion labels + neutral.
  - **Sentiment140**: Tweets labelled as positive, neutral, or negative.
  - **Emotion Dataset (Kaggle)**: Includes text labelled with emotions such as anger, joy, fear, sadness, etc.
  -
- **Custom Scraping**:
  - Tweets collected using **Twitter API** (Tweepy or snscrape)
  - Reddit data collected using **Pushshift API**

2. Features (Input Variables)

Feature Name	Description
text	The raw social media post or comment
cleaned text	Text after preprocessing (cleaning, tokenization, etc.)
tokens	List of words or tokens from the processed text
embedding	Vector representation of text (TF-IDF, BERT, etc.)

3. Dataset Statistics (Typical)

- **Total samples**: 50,000 – 200,000 (varies by source)
- **Number of emotion classes**: 6 to 28
- **Average sentence length**: ~12–20 words
- **Data balance**: Often imbalanced (e.g., joy and anger are more frequent than disgust or surprise)

Data Preprocessing

This step prepares raw social media text for feature extraction and modelling.

Step	Description
Lowercasing	Convert all text to lowercase to maintain consistency.
Removing noise	Strip URLs, mentions (@user), hashtags, numbers, emojis, and punctuation.
Tokenization	Split text into words or subwords.
Stopword removal	Remove common words like “and”, “the”, “is” that don't add meaning.
Lemmatization/Stemming	Reduce words to their root forms (e.g., “running” → “run”).
Handling contractions	Expand contractions (e.g., "can't" → "cannot").
Spelling correction (optional)	Use libraries like TextBlob or SymSpell for misspellings.

Removing Noise

Remove non-textual elements that do not contribute to sentiment/emotion understanding.

Element	Example	Tools
URLs	http://..., www...	Regex
Mentions	@username	Regex
Hashtags	#happy → happy (optional)	Regex
Emojis (optional)	😊 😢	emoji lib or regex
Punctuation	!, ?, .	string.punctuation
Numbers	123, 2020	Regex or isnumeric()

Exploratory Data Analysis (EDA)

- **Univariate Analysis:**
    - Distribution of features using histograms, boxplots, countplots, etc.

- **Bivariate/Multivariate Analysis**:
    - Correlation matrix, pairplots, scatterplots, grouped bar plots, etc.

    ○ Analysis of relationship between features and the target variable.

- **Insights Summary:**
    - Highlight patterns, trends, and interesting observations.

    ○ Mention which features may influence the model and why.]

Feature Engineering

1. Bag of Words (BoW)

- Represents text as a frequency count of words.
- Simple but ignores word order and context.

2. TF-IDF (Term Frequency – Inverse Document Frequency)

- Weighs words by importance (common words get lower weights).
- Better than BoW for most sentiment tasks.

3. Word Embeddings

- Words are represented as dense vectors capturing semantic meaning.
- Better at capturing context and relationships than BoW/TF-IDF.

a. Pre-trained Embeddings

- **Word2Vec**, **GloVe**, or **FastText**
- Convert each word into a vector and average/sum them.

Method	Captures Context?	Dimensionality	Suitable for
BoW	❌	High	Baselines
TF-IDF	❌	High	ML models
Word2Vec	✅	Medium	LSTM, CNN
BERT	✅	High	Transformers
Custom	❌	Low	Add-on boosts

Text-Based Features

- Bag of Words (BoW): Represents text as a collection of words for frequency-based analysis.
- TF-IDF (Term Frequency-Inverse Document Frequency): Helps identify important words by weighting their significance.

- Word Embeddings (Word2Vec, GloVe, BERT): Captures word relationships and contextual meanings.

- N-Grams: Extracts sequences of words to analyze sentiment patterns in phrases.

2. Lexicon-Based Features

- Sentiment Lexicons: Predefined dictionaries like VADER, SentiWordNet, and AFINN provide sentiment scores for words.

- Polarity Scores: Measures the positivity, negativity, or neutrality of words.

3. Audio-Based Features

(For Speech Sentiment Analysis)

- Pitch & Tempo Analysis: Detects emotional cues in speech.

- MFCCs (Mel-Frequency Cepstral Coefficients): Extracts essential features from voice signals.

- Prosodic Features: Analyzes speech rhythm and intonation.

4. Visual-Based Features (For Facial Emotion Recognition)

- Facial Landmarks: Tracks facial expressions.
- Histogram of Oriented Gradients (HOG): Helps detect patterns in facial features.
- Deep Learning-Based Features: CNNs extract hierarchical features from images for emotion recognition.

Contextual Features

- User Sentiment Over Time: Tracks emotional trends across conversations.
- Engagement Metrics: Measures interaction levels, such as response length and frequency.
- Multimodal Fusion: Combines text, speech, and visual data for comprehensive emotion analysis.

Model Building

Model building involves training machine learning or deep learning algorithms to predict either **emotion classes** (classification) or **emotion intensity scores** (regression) from processed text data.

🏵 1. Task Definitions

- **Classification**: Predict emotion labels (e.g., joy, anger, sadness)
- **Regression**: Predict emotion intensity (e.g., 0.0 to 1.0)

⚙ 2. Model Selection

◈ A. Traditional Machine Learning Models

For Classification:

Model	Notes
Logistic Regression	Simple, interpretable
Support Vector Machine (SVM)	Works well with TF-IDF
Random Forest	Handles non-linearity
Naive Bayes	Good for text classification

Visualization of Results & Model Insights

Emotion decoding and sentiment analysis are crucial for understanding how users interact with a model, especially in AI-driven conversations. These techniques can help refine responses, improve user experience, and identify emotional cues that guide meaningful interactions.

 Emotion Recognition: Goes deeper by identifying specific emotions like joy, sadness, anger, or surprise.

 Context Awareness: Analyses how sentiment shifts based on previous interactions, helping models adapt to users' moods.

 Text-Based Insights: Examines linguistic features, word choices, and punctuation to infer emotional tone.

 Improving Model Responses: Adjusts replies to be more empathetic, supportive, or engaging based on detected sentiment.

It looks like my visualization tool wasn't triggered properly! Sometimes that happens. I can definitely describe how sentiment analysis results can be visualized, though.

Here are a few common ways to represent emotion decoding and sentiment analysis:

- Bar Charts: Show sentiment distribution across different emotions (e.g., happiness, anger, sadness).

- Word Clouds: Highlight frequently used words associated with specific sentiments.

- Heatmaps: Display intensity of emotions over time or across different categories.

- Pie Charts: Illustrate proportions of positive, negative, and neutral sentiments.

- Time-Series Graphs: Track sentiment trends over a conversation or dataset.

Tools and Technologies Used

1. Machine Learning-Based Sentiment Analysis

- Natural Language Processing (NLP): Uses models like BERT, GPT, and LSTMs to analyze text sentiment.

- Supervised Learning: Trains models using labeled datasets to classify emotions.

- Unsupervised Learning: Uses clustering techniques to detect underlying emotional patterns in data.

2. Deep Learning & Neural Networks

- Convolutional Neural Networks (CNNs): Often applied to facial emotion recognition using images or videos.

- Recurrent Neural Networks (RNNs): Helps capture sentiment trends over time in conversational AI.

- Transformers (e.g., BERT, GPT): Improves contextual understanding of emotions in text

.

3. Speech & Audio Emotion Recognition

- Spectrogram Analysis: Converts voice signals into visual representations for emotion detection.

- Mel-Frequency Cepstral Coefficients (MFCCs): Extracts audio features from speech to analyze tone and emotion.
- Voice Pitch & Tempo Analysis: Determines emotional state based on tone variations and speaking speed.
-

4. Facial Emotion Recognition

- OpenCV & DeepFace: Detects facial expressions through AI-driven image processing.

- EmotionNet & Affectiva: Advanced tools for tracking microexpressions in real-time.

5. Multimodal Emotion Analysis

Fusion Models: Combine text, speech, and facial recognition for deeper emotional insights.

AI-driven Chatbots: Utilize sentiment tracking across multiple modalities to improve human interaction.

Sentiment Analysis Methods: Traditional approaches include rule-based and lexicon-based methods, while modern techniques leverage machine learning and deep learning

.

- **Top Sentiment Analysis Tools**: Various AI-powered tools help businesses track customer sentiment, analyze brand perception, and monitor social media conversations.

- **Applications Across Industries**: Sentiment analysis is widely used in marketing, customer service, finance, and healthcare to understand public opinion and improve decision-making

TOOLS AND TEAM MEMBERS

S.NO	NAME	RESPONSIBILITIES
1.	A ANAL	DATA COLLECTION AND CLEANING
2.	A MATHUSRI	EXPLORATORY DATA ANALYSIS(EDA)
3.	M.ABINAYA	FEATURE ENGINEERING MODEL
4.	P.BHUVANESHWARI	DOCUMENTATON AND REPORTING