

# MINI PROJECT CANVAS

Title (preliminary):

Quotes Recommendation System based on posted Text Analysis

Group members: 1. Nikola Srbinski 2. Melany Macias 3. Sabina Zaman

Workshop # :

## MOTIVATION



Which is the target group of our mini-project?  
Who is the end-user?

The target group for our mini-project includes:

- **Social Media Users:** Individuals who actively engage with social platforms and may benefit from personalized emotional support.
- **Mental Health Advocates:** Professionals or volunteers seeking tools to assist others in managing their emotional well-being.
- **General Users:** Anyone interested in self-improvement, motivation, or emotional reflection.

The **end-users** are people who will interact directly with the application to get personalized quotes and insights into their emotions.

What are their objectives? What needs do we need to address with our work?

Their objectives and needs include:

- **Seeking Emotional Support:** Users are looking for motivational, comforting, or inspirational content that resonates with their current feelings.
- **Accurate Emotion Detection:** They need the application to accurately detect their emotions from the text they provide.
- **Relevant Recommendations:** Users expect the system to recommend quotes that are relevant to their detected emotions.
- **Privacy Maintenance:** Especially when social media data is involved, users need assurance that their privacy is protected.
- **Scalability for Future Features:** Users may benefit from additional features and updates, so the system should be scalable.

How will they benefit from this proposed solution?

Users will benefit in several ways:

- **Emotional Support:** They will

## DATA COLLECTION



Which data sources are we planning to use?

We plan to use the following data sources:

- **Emotion-Labeled Text Datasets:** Pre-existing datasets from platforms like Kaggle and GitHub containing text samples labeled with emotions for training our emotion detection model.
- **Motivational Quotes Dataset:** A curated collection of motivational quotes with associated emotion labels to match quotes with detected emotions.
- **User Input Data:** Text inputs provided by users through our web application for real-time emotion analysis.
- **Future Data Sources:** Potentially using social media APIs (with user consent) to collect user posts for more comprehensive analysis.

Mention database tables, API methods, websites to scrape, etc.

Our data infrastructure includes:

### Database Tables:

- **EmotionClassification:** Stores text samples and emotion labels.
- **MotivationalQuotes:** Contains quotes and associated emotion tags.
- **UserTexts:** Records user IDs, usernames, submitted texts, and detected emotions.
- **Users:** Manages user profiles, preferences, and login information.
- **Feedback (Future):** Collects user feedback on quotes for improving recommendations.

### APIs and Scraping:

- Initially, no external APIs or web scraping will be used.
- In future iterations, we may use social media APIs (e.g., Twitter API) to collect user data, adhering strictly to privacy policies and user consent.

Which is the data management plan?

## PREPROCESSING



What are the goals of the preprocessing pipeline?

The goals are:

- **Data Cleaning and Standardization:** To improve data quality by cleaning and standardizing the text data.
- **Improving Model Performance:** Eliminating noise and irrelevant information to enhance the performance of our classification model.
- **Avoiding Overfitting and Underfitting:** Reducing noise in the dataset helps prevent overfitting or underfitting issues.

Give some examples of data preprocessing steps.

Examples include:

- **Text Normalization:** Converting all text to lowercase.
- **Punctuation Removal:** Eliminating punctuation marks and special characters.
- **Tokenization:** Breaking text into individual words or tokens.
- **Stop-word Removal:** Excluding common words that don't contribute to emotional meaning.
- **Stemming/Lemmatization:** Reducing words to their base or root form.
- **Vectorization:** Converting text into numerical vectors using methods like TF-IDF.

What are some possible data cleaning/wrangling methods you're planning to use?

## EXPLORATORY DATA ANALYSIS

(EDA)



Look at the data!

What steps are you planning to take towards exploring and understanding better the data you have?

We plan to:

- **Examine Subsets of Data:** Since it's impractical to review each sample individually, we'll analyze subsets to understand the data structure.
- **Identify Most Frequent Words:** Using techniques like word clouds to visualize common words.
- **Visualize Emotion Distribution:** Using histograms or box plots to understand the distribution of emotions and address any over-sampling issues.
- **Analyze Term Frequencies:** Identifying mostly used but unnecessary words to refine the dataset.
- **Consider Text Length:** To decide if summarization is needed for longer texts.
- **Handle Missing Values:** Clustering emotions to assign labels to texts with missing labels.

What properties would be meaningful to summarize/visualize in this step?

Meaningful properties include:

- **Most Frequent Words:** To understand common themes in the dataset.
- **Emotion Distribution:** To see how emotions are represented and identify any imbalances.
- **Text Length Distribution:** To assess the need for text summarization.
- **Term Frequencies:** To refine the dataset by removing irrelevant words.

## VISUALIZATIONS



List any meaningful visualizations you are planning to produce that will be useful to the end user?

We are planning to create the following meaningful visualizations:

### Word Cloud:

- **Purpose:** To display the most frequently used words in the user's text inputs.
- **Benefit:** Helps users quickly identify common themes or topics in their language, promoting self-awareness.

### Emotion Distribution Chart:

- **Types:** Pie chart or bar chart.
- **Purpose:** To show the frequency of different emotions detected in the user's texts over time.
- **Benefit:** Allows users to understand their emotional patterns and fluctuations.

### Scatter Plot of Text Clusters:

- **Purpose:** To visualize clusters of texts based on similarity or emotional content.
- **Benefit:** Provides insights into how different texts relate to each other, highlighting relationships between emotions and topics.





### Feedback or Quotes List:

- **Format:** Organized list format.
- **Purpose:** To display recommended quotes and collect user feedback.
- **Benefit:** Enhances personalization by allowing users to interact with the content and provide input on its relevance.

Are you planning to produce any interactive visualizations?

Yes, we are planning to incorporate interactive visualizations into our web

|  |   |  |  |  |
|--|---|--|--|--|
| <p>receive personalized motivational quotes that provide comfort and encouragement based on their current emotional state.</p> <ul style="list-style-type: none"> <li>• <b>Self-Reflection:</b> By receiving feedback on their emotions, users can gain insights into their mental state, promoting self-awareness.</li> <li>• <b>Personalization:</b> Recommendations are tailored to their specific emotions and preferences, enhancing the relevance and impact of the content.</li> <li>• <b>Updates on Mental State:</b> Especially when social media data is utilized, users can receive ongoing updates and support based on their posts.</li> <li>• <b>Future Enhancements:</b> With scalability, users can look forward to more customized recommendations and new features based on their feedback.</li> </ul> | <p>Our data management plan involves:</p> <ul style="list-style-type: none"> <li>• <b>Data Collection:</b> Initially, we will use pre-defined datasets from Kaggle and GitHub. User texts will be collected via a Flask web application.</li> <li>• <b>Data Storage:</b> Training data and quote data will be stored in CSV files. User data will be stored in a relational database with appropriate tables as mentioned.</li> <li>• <b>Data Preprocessing:</b> We will apply text preprocessing techniques like converting text to lowercase, removing stop words, and stemming using libraries like NLTK.</li> <li>• <b>Data Security:</b> Currently, as we are not handling sensitive information, data privacy concerns are minimal. However, when we start using social media data, we will handle user data securely and may implement access control mechanisms.</li> </ul> | <p>We plan to:</p> <ul style="list-style-type: none"> <li>• <b>Remove Duplicates:</b> Eliminate repeated entries to prevent bias.</li> <li>• <b>Handle Missing Values:</b> Address any missing labels or text entries appropriately.</li> <li>• <b>Standardize Formats:</b> Ensure consistency in data formats across the dataset.</li> </ul> <p><i>What are some possible data transformations that could be useful?</i></p> <p>Possible transformations include:</p> <ul style="list-style-type: none"> <li>• <b>Word Embeddings:</b> Using Word2Vec or GloVe to capture semantic relationships.</li> <li>• <b>Contextual Embeddings:</b> Implementing BERT for deeper contextual understanding.</li> <li>• <b>POS Tagging:</b> Adding grammatical information to tokens.</li> <li>• <b>Text Summarization:</b> Using algorithms like TextRank for long texts.</li> </ul> <p><i>Any feature engineering necessary?</i></p> <p>Yes, feature engineering steps include:</p> <ul style="list-style-type: none"> <li>• <b>Assigning Multiple Emotions:</b> Considering assigning multiple emotions to a single text to capture nuanced meanings.</li> <li>• <b>Working with Text Length:</b> Using text length as a feature to inform the model or decide when to summarize text.</li> </ul> |  | <p>application to enhance user engagement and provide deeper insights.</p> <p><i>If so, which types of interactivity might be useful to the end user?</i></p> <ul style="list-style-type: none"> <li>• <b>Interactive Word Cloud:</b> Users can click on words within the word cloud to see examples of how they've used those words in their texts. This allows users to delve deeper into their language patterns and understand the themes prevalent in their expressions.</li> <li>• <b>Emotion Distribution Charts:</b> Users can hover over segments of a pie chart or bars in a bar chart to reveal exact counts or percentages of each emotion. Clicking on a particular emotion displays all their texts associated with that emotion, helping them recognize patterns in their emotional states over time.</li> <li>• <b>Interactive Scatter Plot of Text Clusters:</b> Users can zoom in and out, pan across the visualization, or filter data based on time frames or specific emotions. Hovering over data points displays snippets of the text or additional details about the emotional content, providing a nuanced understanding of how different texts relate to each other.</li> <li>• <b>Feedback Mechanism in Quotes List:</b> Users can rate the recommended quotes, mark them as favorites, or provide comments on their relevance. This interaction enhances personalization by adapting future recommendations based on user input and actively involves users in the content curation process.</li> </ul> <p>These interactive elements aim to make the visualizations more engaging and informative, empowering users to explore their emotional data meaningfully and fostering a deeper connection with the insights provided by the application.</p> |
|--|---|--|--|--|

| <div> <div> <div>LEARNING TASK</div> <div>(focus on problem definition)</div> </div> <div>  </div> </div> <div> <div>Define the problem setting.</div> <div>Our project involves:</div> <div> <div>1. <b>Emotion Classification:</b> A supervised learning task where we predict the emotion label of a given text.</div> <div>2. <b>Recommendation System:</b> An unsupervised learning task using clustering to suggest motivational quotes based on detected emotions.</div> <div>3. <b>Web Application Integration:</b> Implementing these models within a user-friendly web app using Flask.</div> </div> <div> <div>Is this supervised / unsupervised / other...?</div> <div> <ul style="list-style-type: none"> <li><b>Emotion Classification: Supervised Learning</b> (Classification).</li> <li><b>Recommendation System: Unsupervised Learning</b> (Clustering).</li> <li><b>Web App Development:</b> Not a learning task but involves software engineering.</li> </ul> </div> <div> <div>Classification / regression / other...?</div> <div> <ul style="list-style-type: none"> <li><b>Emotion Detection: Classification</b> problem.</li> <li><b>Recommendation System: Clustering</b> and potentially <b>Collaborative Filtering</b> in future iterations.</li> </ul> </div> </div> <div> <div>What are we planning to learn? E.g. What is the target variable / learning outcome?</div> <div>We plan to learn:</div> <div> <ul style="list-style-type: none"> <li><b>How to Classify Emotions:</b> Developing a model to accurately detect emotions from text.</li> <li><b>How to Cluster Texts for Recommendations:</b> Grouping similar texts or emotions to recommend relevant quotes.</li> <li><b>Web Application Development Skills:</b> Using HTML, CSS, JavaScript, and Flask to build an interactive application.</li> </ul> </div> <div>The target variable is the emotion label associated with each text input.</div> </div> </div> </div> | <div> <div> <div>LEARNING APPROACH</div> <div>(focus on solution implementation)</div> </div> <div>  </div> </div> <div> <div>Which ML/statistical methods seem more relevant for the defined problem setting and why?</div> <div> <ul style="list-style-type: none"> <li><b>Naive Bayes Classifier:</b> Effective for text classification due to its simplicity and performance with high-dimensional data.</li> <li><b>Support Vector Machines (SVM):</b> Suitable for handling complex relationships in data.</li> <li><b>k-Nearest Neighbors (k-NN) Clustering:</b> Useful for grouping similar texts or quotes without prior labels.</li> <li><b>Collaborative Filtering</b> (Future): For personalized recommendations based on user feedback.</li> </ul> </div> <div> <div>Which evaluation metrics could be relevant?</div> <div>For evaluating our models:</div> <div> <ul style="list-style-type: none"> <li><b>Accuracy Rate:</b> To measure the overall correctness of the classification model.</li> <li><b>Confusion Matrix:</b> To understand the performance of the classification model across different emotion classes.</li> <li><b>Elbow Method:</b> For determining the optimal number of clusters in the clustering model.</li> <li><b>K-Fold Cross-Validation:</b> To assess the model's ability to generalize to unseen data and prevent overfitting.</li> </ul> </div> <div> <div>Is any special treatment relevant regarding how we choose to split the data or how we cross-validate?</div> <div>Yes:</div> <div> <ul style="list-style-type: none"> <li><b>Data Splitting:</b> We will split our dataset into training and testing subsets to evaluate model performance.</li> <li><b>Cross-Validation:</b> Implementing K-fold cross-validation to ensure the model's robustness and to mitigate overfitting.</li> <li><b>Handling Imbalanced Data:</b> Visualizing emotion distribution during EDA helps us address any class imbalance issues.</li> </ul> </div> </div> </div> </div> |  | <div> <div> <div>COMMUNICATION OF RESULTS</div> <div></div> </div> <div>  </div> </div> <div> <div>Which type of deliverable will benefit most the end-user? Do we choose to write a blog post, create a website, an app, or other..?</div> <div>We have chosen to develop an interactive <b>web application</b>. It provides immediate accessibility without the need for downloads and allows seamless integration of our models with an engaging user interface.</div> <div> <div>How do we communicate best our results to the predefined target group?</div> <div> <ul style="list-style-type: none"> <li><b>User-Friendly Interface:</b> Simplify interaction with clear instructions and intuitive design.</li> <li><b>Immediate Feedback:</b> Display emotion analysis and recommendations promptly after input submission.</li> <li><b>Visual Aids:</b> Use visualizations like emotion icons or graphs to enhance understanding.</li> <li><b>Accessibility:</b> Ensure the app is accessible on various devices and compliant with accessibility standards.</li> </ul> </div> </div> <div> <div>Short description of your interface/workflow (if applicable).</div> <div> <div>Interface and Workflow:</div> <div> <div>1. <b>Homepage:</b></div> <div>Text input field for users to enter their message or status. "Submit" button to analyze the text.</div> </div> <div> <div>2. <b>Emotion Detection:</b></div> <div>The system analyzes the text using pre-trained models like SVM or Naïve Bayes.</div> </div> <div> <div>3. <b>Results Display:</b></div> <div>Detected emotion is presented to the user. Relevant motivational quotes are displayed.</div> </div> <div> <div>4. <b>Additional Features</b> (Future):</div> <div>Option to provide feedback on quotes. Visualizations of user's emotional trends over time. Profile creation for personalized experiences.</div> </div> </div> </div> </div> | <div> <div> <div>DATA PRIVACY AND ETHICAL CONSIDERATIONS</div> <div>(if applicable)</div> </div> <div>  </div> </div> <div> <div>Are there any fairness constraints that apply to our proposed pipeline?</div> <div>Yes, since we are working with pre-built datasets, there is a possibility of bias towards specific groups based on age, gender, or ethnicity. This could affect the fairness of our emotion classification model. To address this, we plan to:</div> <div> <ul style="list-style-type: none"> <li><b>Provide Equal Treatment:</b> Ensure that all users receive the same level of service and that the model does not favor any group over another.</li> <li><b>Future Dataset Improvements:</b> In future work, we aim to use more diverse and representative datasets to reduce bias.</li> </ul> </div> <div> <div>Is there a need to ask for consent during the data collection process?</div> <div>Initially, as we are collecting single text inputs directly from users, explicit consent may not be required beyond informing them how their data will be used. However, when we start working with social media data:</div> <div> <ul style="list-style-type: none"> <li><b>User Consent:</b> We will obtain explicit confirmation from users before accessing and analyzing their social media posts.</li> <li><b>Transparency:</b> Clearly communicate the purpose of data collection and how it will be used.</li> </ul> </div> </div> <div> <div>Is there a need for data pseudonymization/anonymization?</div> <div>Yes, to protect user privacy:</div> <div> <ul style="list-style-type: none"> <li><b>Data Anonymization:</b> We will store user data as hashed values rather than in plain text to prevent identification.</li> <li><b>Secure Data Handling:</b> Implementing access control mechanisms and secure storage solutions to protect user information.</li> </ul> </div> <div> <div>Any other privacy considerations that come to mind?</div> <div>Additional privacy considerations include:</div> <div> <ul style="list-style-type: none"> <li><b>Data Security:</b> Ensuring that user data is protected from</li> </ul> </div> </div> </div> </div> |
|--|---|--|---|--|
|--|---|--|---|--|

|  |   |  |   |  |
|--|---|--|---|--|
| <p>What variables are we using as input?</p> <p>Input variables include:</p> <ul style="list-style-type: none"><li>• <b>Tokens:</b> Words extracted from the text after preprocessing.</li><li>• <b>Text Length:</b> The length of the input text.</li><li>• <b>Sentiment Labels:</b> Labels from our training dataset.</li><li>• <b>User Text Posts:</b> Text inputs provided by the users.</li><li>• <b>Choice of Recommendation:</b> User preferences, possibly including historical data like previous posts and feedback.</li></ul> |   |  | <p>5. <b>Navigation:</b></p> <p>Easy access to different sections like "About Us", "Contact", and "Feedback".</p> | <p>unauthorized access or breaches.</p> <ul style="list-style-type: none"><li>• <b>Compliance with Regulations:</b> Adhering to data protection laws and regulations relevant to the regions where our users reside.</li><li>• <b>Ethical Data Use:</b> Using collected data solely for the purposes stated and not sharing it with third parties without consent.</li><li>• <b>User Control Over Data:</b> Allowing users to request deletion of their data and providing options to opt-out of data collection.</li><li>• <b>Handling Sensitive Information:</b> Being cautious with any sensitive data that may be inferred from user inputs and providing resources if needed.</li></ul> |
|  | <p><b>ADDED VALUE</b> </p> <p><i>Is there a possibility for added value from the data we're planning to use?</i></p> <p>Yes, there is a possibility for added value from the data we are using. By predicting emotions from user-provided text, we can offer personalized motivational quotes that directly address the user's current mental state.</p> <p><i>What is the added value?</i></p> <p>The added value includes:</p> <ul style="list-style-type: none"><li>• <b>Improved Mental Health:</b> Providing users with relevant motivational content can enhance their emotional well-being.</li><li>• <b>Personalization:</b> Tailoring recommendations based on individual emotions and preferences increases the relevance and impact of the content.</li><li>• <b>Emotional Support:</b> Offering timely support and comfort through the application.</li><li>• <b>Scalability for Future Features:</b> The system can be expanded with additional functionalities like feedback mechanisms and social media analysis.</li></ul> | <p><i>How are predictions turned into added value for the end-user?</i></p> <p>By accurately predicting the user's emotions from their text input, we can:</p> <ul style="list-style-type: none"><li>• <b>Deliver Personalized Quotes:</b> Recommendations are more meaningful when they align with how the user is feeling.</li><li>• <b>Provide Immediate Support:</b> Users receive instant feedback and motivation, which can positively affect their mood.</li><li>• <b>Enhance User Engagement:</b> Personalized content encourages users to interact more with the application.</li><li>• <b>Foster Self-Reflection:</b> Users gain insights into their emotions, promoting self-awareness and personal growth.</li></ul> |   | <p><b>LEGEND</b></p> <p><b>WEEK 1:</b> Data collection/preprocessing</p> <p><b>WEEK 2:</b> EDA &amp; visualizations</p> <p><b>WEEKS 3-4:</b> Machine/deep learning</p> <p><b>WEEK 5:</b> Fairness &amp; data privacy</p>   |