Text Mining Course: Final Project Outline

# Dataset Description

Dataset Chosen:

Parenting Advice Posts from Reddit

Data Card:

This dataset comprises posts from various subreddits on Reddit where users seek or offer parenting advice. It includes text posts (questions needing advice) and comment posts (advice), along with metadata such as the subreddit, post title, post content, comment content, comment score, comment date, and provider.

# Explanation of Dataset Choice

The dataset was chosen because it provides a rich source of unstructured text data related to parenting, covering a wide range of topics, concerns, and advice sought by parents and caregivers. Analyzing these posts can offer valuable insights into common parenting challenges, trends, and popular strategies.

# Text Mining Methods

## Method 1: NLP Pipeline (Preprocessing & Normalization)

Implementing a Natural Language Processing pipeline to preprocess and normalize text data. This will involve steps such as tokenization, removing stopwords, lemmatization, and removing special characters.

## Method 2: Sentiment Analysis

Conducting a sentiment analysis to assess the emotional tone of the parenting advice posts. This will help identify whether posts express positive, negative, or neutral sentiments towards parenting issues and advice.

## Method 3: Text Similarity

Text similarity analysis measures the degree of similarity between pairs of text documents. This method will help identify parenting advice posts that are closely related or similar in content, allowing users to explore similar discussions or solutions to common parenting challenges.

## Method 4: Question-Answering System

A question-answering system enables users to pose questions in natural language and receive relevant answers extracted from parenting advice posts. This method adds interactivity and allows users to directly engage with the corpus of advice posts.

# Final Product Proposal

Proposed End Product:

The final product will be a comprehensive dashboard for analyzing and accessing parenting advice extracted from Reddit posts. The dashboard will provide users with various functionalities to explore, analyze, and interact with the corpus of parenting advice posts. Here is a breakdown of the key features and functionalities.

1. Sentiment Analysis:
   1. Sentiment analysis scores are assigned to each post to evaluate the emotional tone towards parenting topics.
   2. Posts can be filtered or sorted based on sentiment polarity (positive, negative, neutral) to identify posts with specific emotional contexts.
2. Text Similarity
   1. Users can discover related or similar parenting advice using similarity analysis.
   2. Similar posts are presented to users to explore alternative perspectives or solutions to common parenting challenges.
3. Question-Answering System
   1. Users can ask specific questions related to parenting advice and receive relevant information from the corpus of Reddit posts.
   2. The question-answering leverages natural language processing techniques to understand and provide accurate responses to user queries.
4. Interactive Visualization
   1. Visualizations such as word clouds, bar charts, and network graphs are used to present insights and patterns derived from the text mining analysis.
   2. Users can interact with visualizations to gain deeper insights into parenting topics and trends.

**Presentation Form**

For the presentation, a professional slide deck will be prepared to demonstrate the functionality and features of the final project dashboard. The slide deck will include:

* Overview of the project objectives and dataset
* Description of the methodologies and text mining techniques employed.
* Demonstration of user interactions such as sentiment analysis, text similarity, and question-answering.
* Discussion of the value and potential applications of the final product in analyzing and accessing parenting advice from Reddit posts.

The presentation will be delivered in a clear and organized manner, highlighting the effectiveness and usability of the final product in providing valuable insights and support for parents and caregivers seeking advice and guidance.

By integrating the four methods into the final product dashboard, users will have a comprehensive tool for exploring, analyzing, and accessing parenting advice from a diverse corpus of Reddit posts.