**Capstone Two - Project Ideas**

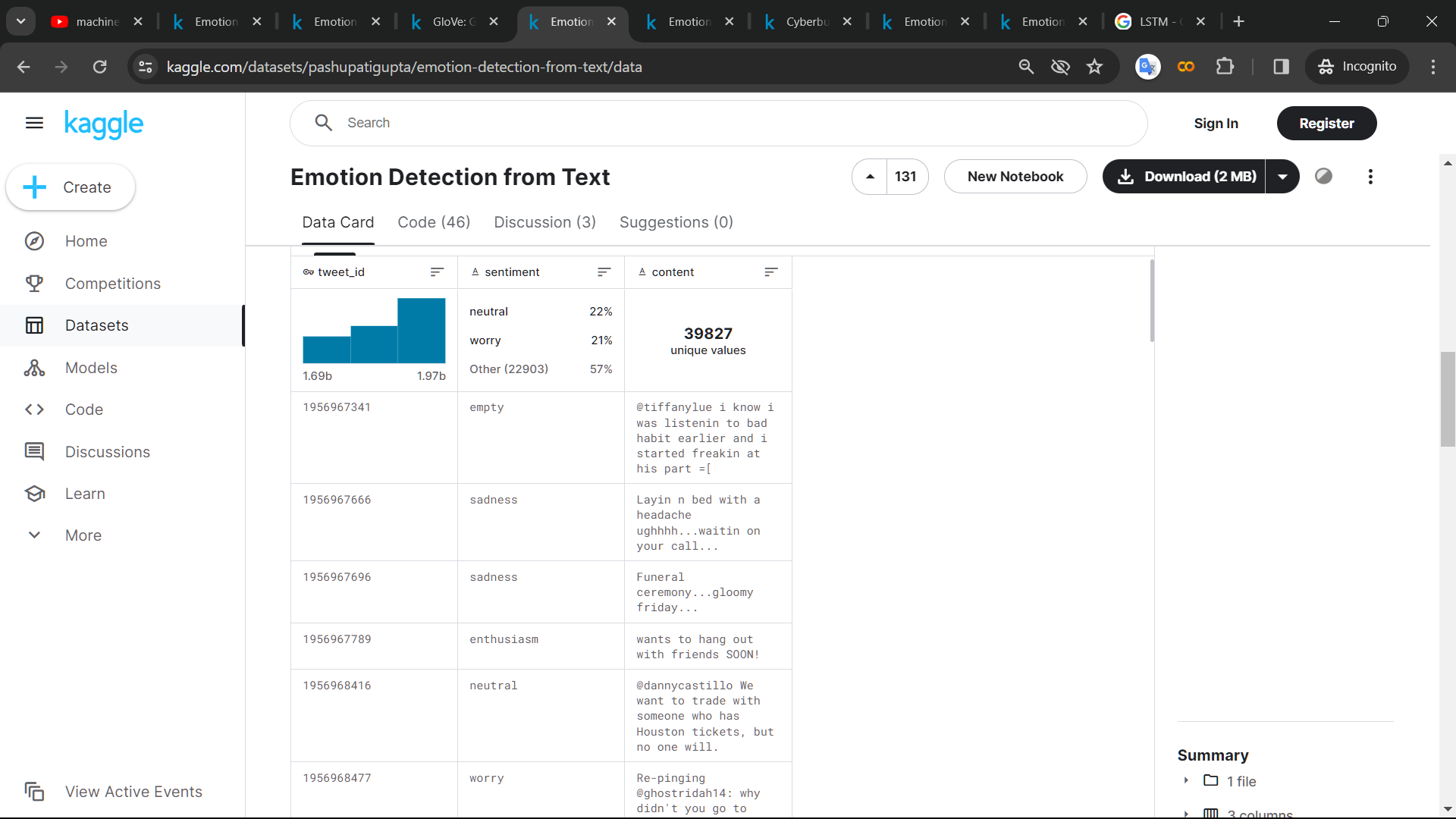
**Project Idea -1 “Emotion Detection from Text’’**

**Context:**

Detecting emotions from text is one of the challenging problems in Natural Language Processing. The reason is the unavailability of labeled datasets and the multi-class nature of the problem. Humans experience a wide range of emotions and it is difficult to collect enough records for each emotion hence the problem of class imbalance arises. Here we have labeled data for emotion detection The objective of this task is to build a model that can interpret text data to detect emotions with a high degree of accuracy.

**Content:** The dataset we are working with consists of a comprehensive collection of tweets, each annotated with the corresponding emotions expressed in them. The dataset has three columns: tweet\_id, sentiment, and content. The "content" column contains the raw tweet text while the "sentiment" column indicates the specific emotion conveyed in each tweet. The emotions can range from positive to neutral to negative and are labeled accordingly.

The data that we have has 13 different emotions 40000 records. So it's challenging to build an efficient multiclass classification model. We may need to logically reduce the number of classes here and use some advanced methods to build an efficient model.



**Source:** [Dataset.csv](https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text)

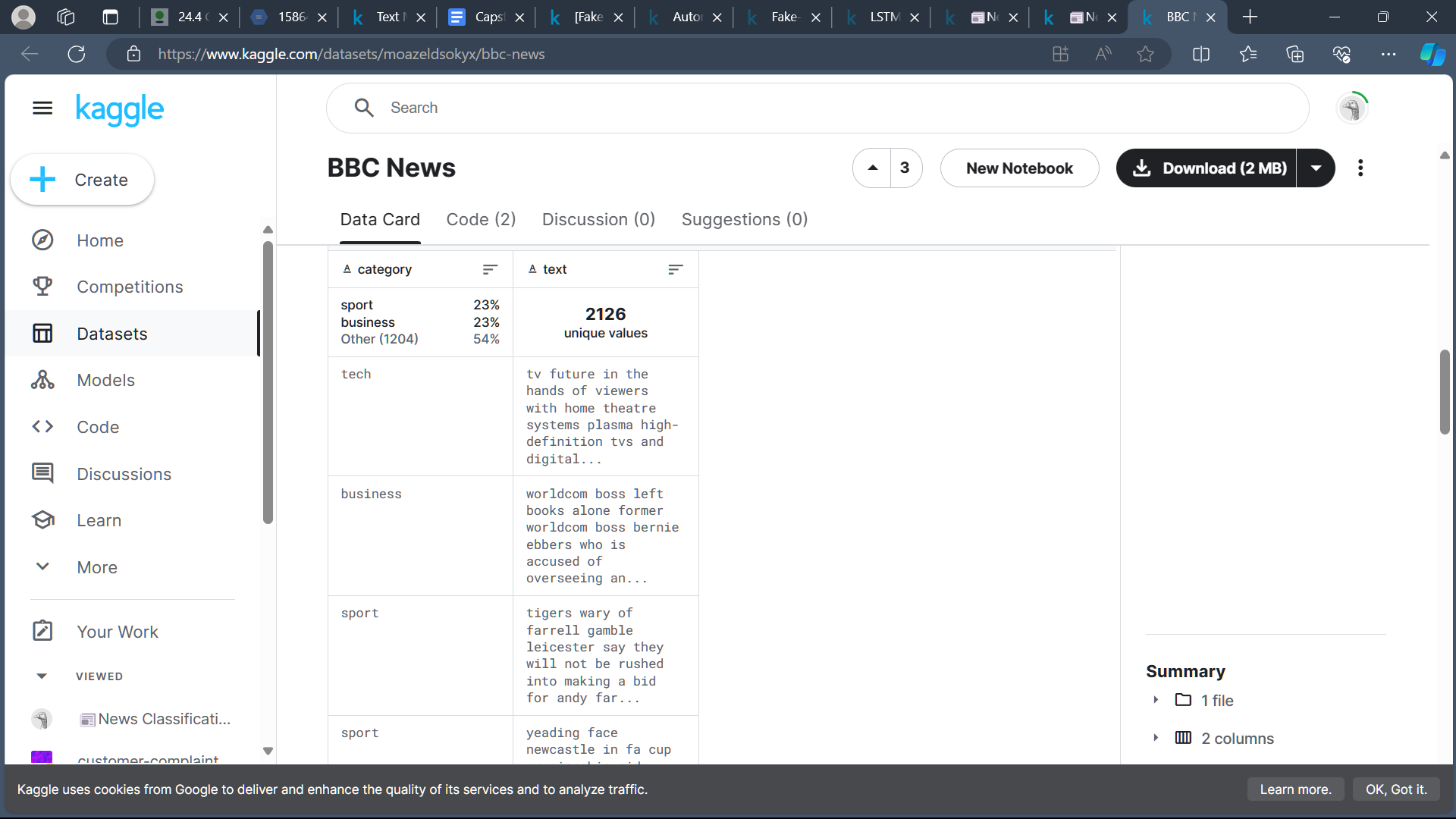
# **Project Idea -2 “Twitter Financial News”**

**Context :**

The BBC News Article Dataset is a comprehensive compilation of 2,225 news articles published by the renowned news organization, BBC News. The articles cover diverse categories, including but not limited to Sport, Business, Politics, Tech, and Entertainment, and are accompanied by corresponding category labels. This enables easy categorization and analysis of the news content, making it an invaluable resource for researchers, journalists, anddata scientists alike. The dataset offers a broad selection of news articles, each with unique perspectives and insights, making it a valuable tool for gaining a deeper understanding of current affairs and trends across variousdomains.

**Content:**-

1. Category: The category or topic of the news article. This dataset includes five categories:  
   [Sport, Business, Politics, Tech, and Entertainment].
2. Text: The actual textual content of the news article, providing detailed information on the topic covered in the article.

****

**Source:** [Dataset](https://www.kaggle.com/datasets/moazeldsokyx/bbc-news/data).csv

# **Project Idea -3 “Automatic Ticket Classification”**

# **Context:**

# A financial company aims to efficiently handle customer complaints by classifying them into relevant categories based on their products and services. To improve customer service, the company wants to build a model that can classify customer complaints into five clusters. Each cluster corresponds to a different product or service offered by the company, such as credit cards, banking, and mortgages or loans. The ultimate aim is to automate the process of assigning tickets to relevant departments, thereby streamlining customer support operations and enhancing overall efficiency.

**Content**:- The dataset consists of 78,313 customer complaints in the .json format. The dataset has 22 features that need to be processed to extract meaningful insights. However, before start analyzing the data, we need to convert it into a data frame format, which will allow us to perform various operations such as sorting, filtering, and grouping of data. Once the conversion is complete, we can begin processing the complaints to identify patterns, trends, and other relevant information that can be used to improve customer satisfaction and address issues proactively.

**Source:** [Dataset.csv](https://www.kaggle.com/datasets/venkatasubramanian/automatic-ticket-classification)

Certainly! Let’s delve into the details of applying **Non-Negative Matrix Factorization (NMF)** for topic modeling on the unlabeled .json data. Our goal is to uncover underlying patterns and group the tickets into distinct clusters based on their products or services.

Here’s a step-by-step breakdown of how we can proceed:

1. **Data Preprocessing**:
   * Load the .json data and extract relevant information.
   * Clean the text data by removing special characters, stopwords, and performing tokenization.
   * Convert the text data into a document-term matrix (DTM).
2. **NMF Model Creation**:
   * Initialize the NMF model with a specified number of topics (in our case, five).
   * Fit the model to the DTM.
   * Extract the topic-word and document-topic matrices.
3. **Interpreting Topics**:
   * Examine the top words associated with each topic. These words represent the key terms for that topic.
   * Assign meaningful labels to each topic based on the prominent terms.
4. **Assigning Tickets to Clusters**:
   * Transform the original ticket data into the topic space using the document-topic matrix.
   * Assign each ticket to the cluster/topic with the highest weight.
5. **Visualization and Analysis**:
   * Visualize the clusters using techniques like t-SNE or UMAP.
   * Explore the characteristics of each cluster to understand the underlying patterns.
6. **Refinement and Iteration**:
   * Fine-tune the model by adjusting hyperparameters (e.g., number of topics, regularization).
   * Evaluate the quality of the clusters using metrics like coherence or silhouette score.
   * Iterate as needed to improve the results.

Remember that NMF assumes non-negativity in the data, making it suitable for text data where word frequencies are non-negative. Additionally, interpretability of the topics is crucial, so we’ll focus on meaningful labels for each cluster.

Feel free to share more details about the .json data (such as its structure, fields, and any specific challenges), and we can proceed with the implementation! 🚀

What visualization tools can we use to explore clusters visually?

Sent by you: What visualization tools can we use to explore clusters visually?

Certainly! When exploring clusters visually, there are several effective techniques to consider. Let’s dive into some of them:

1. **Scatter Plots**:
   * Scatter plots are commonly used to visualize clusters in **two- or three-dimensional data**.
   * Each data point is represented as a dot on the plot, and the position of the dot corresponds to the feature values.
   * Clusters often appear as groups of closely packed points.
   * You can create scatter plots for different pairs of features to explore relationships within and between clusters.
2. **Dendrograms**:
   * Dendrograms are tree diagrams that show how clusters are merged in hierarchical clustering.
   * They are useful for visualizing hierarchies and relationships among data points.
   * Each branch in the dendrogram represents a cluster, and the height of the branches indicates the similarity between clusters.
   * [You can use libraries like scipy or seaborn to create dendrograms](https://www.handsonmentor.com/post/a-comprehensive-guide-to-visualizing-hierarchies-and-clustering-in-data-science)[1](https://www.handsonmentor.com/post/a-comprehensive-guide-to-visualizing-hierarchies-and-clustering-in-data-science).
3. **Heatmaps**:
   * Heatmaps display a matrix of values using colors.
   * They are particularly useful for visualizing the similarity or dissimilarity between data points.
   * Rows and columns represent data points, and the color intensity represents the similarity (e.g., correlation) between them.
   * Heatmaps can reveal patterns and clusters in high-dimensional data.
4. **t-SNE (t-Distributed Stochastic Neighbor Embedding)**:
   * t-SNE is a dimensionality reduction technique that maps high-dimensional data to a lower-dimensional space (usually 2D or 3D).
   * It preserves pairwise similarities between data points.
   * Visualizing t-SNE results can help reveal clusters and patterns.
   * Libraries like scikit-learn provide t-SNE implementations.
5. **PCA (Principal Component Analysis)**:
   * Although PCA is primarily used for dimensionality reduction, it can also aid in cluster visualization.
   * By projecting data onto the first few principal components, you can create scatter plots that highlight clusters.
   * PCA captures the most important features, making it easier to visualize patterns.
6. **Parallel Coordinates Plot**:
   * Parallel coordinates plots display multiple numerical variables along parallel axes.
   * Each data point is represented as a line connecting the values on each axis.
   * Clusters often appear as groups of similar lines.
   * This technique is useful for exploring high-dimensional data.

Remember that the choice of visualization technique depends on your specific data, the number of dimensions, and the insights you want to extract. Feel free to experiment with different methods to find the one that best suits your needs! 📊🔍