University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

ADS-509 – Team Project Status Update Form

Fill out this form and submit it by the end of Module 4 in Blackboard.

Team Number: 2

Team Leader/Representative: Tarane Javaherpour

Full names of team members:

1. Tarane Javaherpour
2. Davood Aein

Title of your Applied Text Mining project: Disaster Tweet Classification & Topic Modeling

Short description of your project and objectives:

We will acquire a corpus of tweets (via rehydrating Kaggle's "Disaster Tweets" IDs through the Twitter API), clean and tokenize the text, and then build a supervised classification model that distinguishes "real-disaster" tweets from "non-disaster" tweets. After measuring classification accuracy on a hold-out set, we will ignore the original labels and apply an unsupervised topic model (e.g., LDA or NMF) to the same cleaned corpus. Finally, we will compare how well the discovered topics align with the a priori "disaster vs. non-disaster" labeling. This dual approach demonstrates both classification and topic-modeling workflows that are commonly used in real-world text-mining.

Name of your selected dataset: Kaggle "Disaster Tweets" (tweet IDs rehydrated via Twitter API)

Description of your selected dataset (data source, number of variables, size of dataset, etc.):

- **Data source:** Kaggle provides a CSV containing approximately 10,000 Twitter IDs labeled with `target` = 1 (real disaster) or 0 (non-disaster).

- **Rehydration step:** We will use the Twitter API v2 (via Tweepy) to fetch the full tweet text and metadata (e.g., `created_at`, `lang`, `public_metrics`).

- **Post-rehydration schema:** After rehydration, each record will have:
    - `id` (tweet ID)
    - `text` (full tweet content)
    - `created_at` (timestamp)
    - `lang` (language code)
    - `retweet_count`, `reply_count`, `like_count`, `quote_count` (public metrics)
    - `target` (original Kaggle label 0/1)

- **Size of dataset (post-rehydration):** Approximately 9,200 – 10,000 tweets (some IDs may be missing if deleted), each with the fields above.

For this project, use GitHub as a code hosting platform for version control and collaboration.
Provide the link here: https://github.com/Tarane2028/ADS-509-Project/tree/main
How many times have your members met in the last two weeks? 2

List the specific contributions that each team member is providing for the Final Team Project in the table below.

- **NOTE:** ALL students on the team should contribute equally to the Final Team Project.

Team Member 1 (Tarane)	Team Member 2 (Davood)	Team Member 3 (if applicable) (Name)
• Data acquisition (Twitter API rehydration) • Data cleaning & tokenization (NLTK/regex) • Descriptive statistics (vocabulary size, top unigrams/bigrams) • Topic modeling (LDA/NMF) & topic-label comparison	• Vectorization (TF-IDF/CountVectorizer) • Classification model (Logistic Regression, evaluation metrics) • Feature importance (top tokens) • Preparation of slide decks & video recordings	List of contributions

Comments/Roadblocks:

- Obtaining and configuring valid Twitter API v2 credentials (Bearer Token) took extra time, and we must ensure we stay within rate-limit windows when rehydrating.

- A small percentage (~5–8%) of tweet IDs have been deleted or made private, so our final post-rehydration corpus is slightly smaller than the original ~10,000 IDs.

- We are coordinating across two schedules but have agreed on three standing meetings per week (via Zoom or Slack huddles) to keep progress aligned.