**Detailed project report on sentiment 140:**

**1. Introduction**

* **Project Title**: Sentiment Analysis on Twitter Data
* **Objective**: To analyze sentiment trends using machine learning models.
* **Scope**: This report focuses on the **Sentiment140 dataset**, explaining its structure, limitations, and preliminary insights.

**2. Data Collection Overview**

* **Source**: Sentiment140 dataset (Preprocessed dataset for Twitter sentiment analysis).
* **Why Sentiment140?**
  + Contains **1.6M labeled tweets** (positive/negative).
  + Eliminates the need for manual scraping.
  + Provides real-world, noisy Twitter data.
* **Dataset Features:**

|  |  |
| --- | --- |
| **Feature** | **Description** |
| polarity | 0 (negative), 2 (neutral), 4 (positive) |
| tweet\_id | Unique ID of the tweet |
| date | Date and time of the tweet |
| query | Query term (mostly "NO\_QUERY") |
| username | Twitter username |
| tweet | The actual tweet text |

**3. Dataset Limitations & Workarounds**

**Limitations:**

* **No emojis, URLs, or mentions** (preprocessed out).
* **Potentially outdated** (collected in 2009).
* **Class imbalance** (neutral class underrepresented).

**Workarounds & Enhancements:**

* **Supplementing with live Twitter data** (via API or scraping).
* **Reintroducing emojis & hashtags for richer sentiment signals** using data augmentation.

**4. Exploratory Data Analysis (EDA) on Sentiment140**

**Data Summary & Distribution**

* **Sentiment class distribution** (how many positive vs. negative tweets).
* **Most common words & bigrams** in positive vs. negative tweets.
* **Word cloud visualization** for different sentiments.

**Challenges in EDA:**

* **Presence of noise** (misspellings, slang).
* **Handling retweets & duplicate content.**

**5. Next Steps & Justification of Time Usage**

* **Reviewing additional datasets** (TweetEval, Kaggle datasets).
* **Developing data augmentation methods** (synthetic tweet generation).
* **Running preliminary ML models** on raw Sentiment140 to compare with refined datasets later.