Social Media Sentiment Analysis Report (Kaggle Dataset)

1. Introduction

This report details a social media sentiment analysis project using a real-world dataset from Kaggle, specifically the Sentiment140 dataset containing 1.6 million tweets. The goal is to understand public sentiment towards various topics by applying Natural Language Processing (NLP) techniques, including data preprocessing, sentiment scoring, and visualization of sentiment trends.

2. Methodology

2.1 Data Source

The dataset used is the Sentiment140 dataset from Kaggle, which comprises 1.6 million tweets. Each tweet is annotated with a polarity (0 = negative, 2 = neutral, 4 = positive). For this analysis, a random sample of 10,000 tweets was selected to facilitate faster processing and demonstration.

2.2 Data Preprocessing

The raw tweet text data underwent the following preprocessing steps:

- **Lowercasing:** All text was converted to lowercase.
- **Punctuation and Number Removal:** Non-alphabetic characters were removed.
- **Tokenization:** Text was broken down into individual words.
- **Stop Word Removal:** Common English stop words were removed.

2.3 Sentiment Analysis

Sentiment analysis was performed using two Python libraries:

- **TextBlob:** Provides a polarity score ranging from -1 (negative) to 1 (positive).
- **VADER (Valence Aware Dictionary and sEntiment Reasoner):** A lexicon and rule-based sentiment analysis tool specifically designed for social media text. It provides a compound score from -1 (most negative) to 1 (most positive).

2.4 Data Visualization

Visualizations were generated to illustrate sentiment trends and distributions:

- **Overall Sentiment Trend Over Time (VADER):** A line plot showing the average VADER sentiment score over time, aggregated daily.
- **Overall Sentiment Distribution (VADER):** A pie chart categorizing sentiments as Positive (compound score >= 0.05), Negative (compound score <= -0.05), or Neutral (otherwise).

3. Results and Analysis

3.1 Overall Sentiment Trend Over Time

The overall sentiment trend over time (overall_sentiment_trend_kaggle.png) for the sampled Kaggle dataset shows daily average VADER sentiment scores. The plot indicates fluctuations in sentiment, with periods of higher positive and negative averages. The data points are from May-June 2009, reflecting the historical nature of the Sentiment140 dataset. The trends appear somewhat volatile, which is typical for social media data where public opinion can shift rapidly.

3.2 Overall Sentiment Distribution

The overall sentiment distribution pie chart (overall_sentiment_distribution_kaggle.png) provides a clear breakdown of sentiment categories within the sampled dataset. The analysis reveals that approximately 48.0% of the tweets are classified as Positive, 28.0% as Neutral, and 24.0% as Negative. This distribution suggests a relatively balanced mix of sentiments, with a slight leaning towards positive, which is common in large social media datasets even when negative events occur.

4. Conclusion

This project successfully applied NLP techniques to perform sentiment analysis on a real-world social media dataset from Kaggle. The process involved data preprocessing, sentiment scoring using TextBlob and VADER, and visualization of sentiment trends and distributions. The analysis of the Sentiment140 dataset provided insights into historical public sentiment, demonstrating the applicability of these techniques to large-scale social media data. Future work could involve exploring more advanced machine learning models for sentiment classification, conducting topic modeling to identify key discussion areas, and performing more in-depth time-series analysis to predict sentiment shifts.