Twitter Sentiment Analysis

Phuoc-Lanh Ho – 2102114

School of Information Technology, Tan Tao University

CS441/19: Data Visualization

Xuan-Ha Nguyen, Ph.D.

May 31, 2025

I. Abstract

This report explores public sentiment toward tech and gaming entities using over 69,000 labeled tweets from Twitter. Through tools such as pandas, matplotlib, seaborn, NLTK, and Streamlit, I created a visual analytics dashboard to present sentiment distribution, tweet length analysis, and word patterns. Statistical methods including the two-proportion Z-Test, ANOVA, and Kruskal-Wallis test were applied. Key insights include a dominance of negative sentiment, significant variation in tweet length across sentiment categories, and limitations in sentiment detection due to sarcasm and domain-specific slang. My results provide practical implications for brand monitoring and customer engagement in the digital space.

II. Introduction

Twitter is a prominent platform where users express opinions, particularly around tech and gaming topics. Sentiment analysis enables companies to gauge public perception and respond accordingly.

In this report, I:

- Describe the dataset and preprocessing steps
- Explore the sentiment trends visually
- Apply statistical tests to examine sentiment differences
- Discuss implications and recommendations based on my findings

III. Dataset Description

1. Data Source

I used the dataset from the Kaggle Twitter Entity-Level Sentiment Analysis competition. Tweets were manually labeled for sentiment toward specific

2. Structure

- Rows: 69,247
- Columns: tweet_id, entity, sentiment, text
- Sentiment classes: Positive, Negative, Neutral, Irrelevant

3. Preprocessing Summary

- Removed nulls and placeholder tokens like <unk> (2,414 rows)
- Eliminated duplicate rows
- Added tweet length column

Note: For analysis, Irrelevant and Neutral sentiments are considered separately.

IV. Data Cleaning & Preprocessing

To ensure clean input for analysis, I applied the following preprocessing steps:

- Dropped rows with null values in text or sentiment
- Removed tweets containing placeholder tokens like <unk>
- Dropped duplicate rows based on tweet ID and content
- Standardized text formatting (e.g., lowercasing, punctuation removal)
- Calculated tweet length and added it as a new column for analysis

These steps ensured cleaner input for visualization and statistical testing.

I identified several rows with missing values in text columns. These were visualized using a null heatmap. Although removing them reduces the dataset size slightly (by 2,414 rows), it ensures the reliability of further analyses. Incomplete records could skew sentiment classification or word-based frequency calculations.

V. Exploratory Data Analysis

1. Sentiment Distribution

I visualized the count of tweets per sentiment label. Negative tweets were the most frequent, followed by positive, neutral, and irrelevant tweets.

2. Tweet Length Overview

- Mean tweet length: 110 characters
- 99th percentile threshold: 319 characters
- Tweets exceeding this threshold were considered outliers and analyzed separately.

3. Outlier Insights

Outlier tweets often contained marketing content, long hashtags, or embedded links. Identifying and isolating these tweets helped prevent skewed results in subsequent statistical tests, especially those sensitive to distribution and variance.

VI. Visual Insights

1. Word Clouds by Entity & Sentiment

Using word clouds filtered by entity and sentiment, I observed:

- Positive (TomClancysRainbowSix): Words such as "love", "amazing", and "thank you" were prominent.
- Negative: Frequent terms included "fix", "can't play", and "stuck loading screen".

2. N-Gram Analysis

- Top Bigrams (Positive): "cant wait", "looking forward"
- Top Trigrams (Positive): "rainbow six siege", "help new players"

These indicate user anticipation and a strong community aspect in positive tweets.

VII. Statistical Analysis

Test 1: Z-Test: Comparing Proportions of Positive vs. Negative Sentiment

Purpose: To determine if the proportion of negative tweets significantly exceeds that of positive tweets for a selected entity.

Hypotheses:

- H0 (Null): The proportion of negative and positive tweets are equal.
- H1 (Alternative): The proportion of negative tweets is greater than that of positive tweets.

Why this test? The two-proportion Z-Test is appropriate for comparing proportions in large samples where responses fall into binary categories (e.g., Negative vs. Positive). With sample sizes in the thousands, assumptions for the Z-Test (normal approximation of binomial distribution) are satisfied.

Results (TomClancysRainbowSix) :$Z = 18.38$, $p < 0.001$

Interpretation: I reject the null hypothesis. There is a statistically significant higher proportion of negative tweets compared to positive ones for this entity.

Test 2: ANOVA: Tweet Length by Sentiment Group

Initial Intention: To examine whether tweet length significantly differs across sentiment categories (Positive, Negative, Neutral, Irrelevant), one might consider using a one-way ANOVA.

Why ANOVA was reconsidered: ANOVA requires assumptions of normality and homogeneity of variances. However, my data violates these assumptions significantly:

- Shapiro-Wilk test for normality: $p < 0.001$ for all sentiment groups, indicating strong non-normality.
- Levene’s test for equal variances: stat = 11.39, $p < 0.001$, indicating heteroscedasticity.
- Distribution of tweet lengths: highly right-skewed and discrete in nature (integer character counts), making the use of ANOVA inappropriate.

Conclusion: Given these violations, I opted not to use one-way ANOVA for inference. Instead, I applied a non-parametric alternative that better fits the distributional properties of my dataset: the Kruskal-Wallis H-Test.

Test 3: Kruskal-Wallis H-Test: Non-Parametric Alternative

Purpose: To validate ANOVA results without assuming normality or equal variance.

Hypotheses:

- H0: All sentiment groups have equal distributions of tweet length.
- H1: At least one group differs in distribution.

Why this test? The Kruskal-Wallis test is a rank-based non-parametric alternative to ANOVA. It is appropriate when normality or equal variance assumptions are violated, as in my case.

Results: $H = 192.76$, $p < 0.001$

Interpretation: I reject the null hypothesis. There is a statistically significant difference in tweet length distributions across sentiment groups.

VIII. Key Findings

Negative sentiment dominates discussions across entities.

Mean tweet length varies significantly by sentiment, with negative tweets being longer on average.

Word usage patterns (n-grams, word clouds) show distinct linguistic signals tied to sentiment.

Sarcasm and informal gaming language pose challenges to lexicon-based sentiment tools such as VADER.

IX. Limitations

Sarcasm and slang may lead to mislabeling of sentiment.

The dataset includes only English-language tweets.

Sentiment classes are imbalanced and may affect modeling.

Only the top 10 entities were analyzed, limiting broader generalizability.

X. Conclusion

This project demonstrates how sentiment analysis, paired with statistical testing and visualization, can uncover meaningful trends in public discourse on Twitter. My findings confirm a predominance of negative sentiment in the tech and gaming space and highlight how tweet length and language structure vary by tone. With refinements in modeling and analysis, dashboards like this can become powerful tools for real-time brand monitoring and audience engagement.

XI. References

1. Kaggle Dataset: https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis
2. NLTK: https://www.nltk.org/

3. VADER Sentiment: https://github.com/cjhutto/vaderSentiment

4. Streamlit: https://streamlit.io/

5. Seaborn, Matplotlib, Pandas