The Sleep-Deprived Internet:
Night vs Day Posting and Sentiment Analysis

DSA210 – Introduction to Data Science
Fall 2025–2026

1 Introduction

Many people believe that posts shared late at night feel different from those shared during the day. In this project, I examine whether this idea can be observed in real data by analyzing tweets posted at different times.

I compare tweets posted at night and during the day using sentiment information that already exists in the dataset. To enrich the analysis, I also connect night posting behavior with country-level Quality of Life indicators.

2 Research Questions

The main questions of this project are:

- Do tweets posted at night differ from day tweets in sentiment behavior?
- Is posting time related to sentiment categories (negative, neutral, positive)?
- Can we predict whether a tweet was posted at night using only its text?
- Does late-night posting vary across countries with different quality of life levels?

3 Datasets

3.1 Tweets Dataset https://www.kaggle.com/datasets/tango911/airline-sentiment-tweets

The main dataset consists of airline-related tweets collected from Twitter.

3.2 Quality of Life Dataset https://www.kaggle.com/datasets/marcelobatalhah/quality-of-life-index-by-country

The second dataset provides country-level Quality of Life indicators.

4 Data Preparation

I converted tweet timestamps into hours and defined two time groups:

- **Night:** 00:00–06:00
- **Day:** 08:00–22:00

Tweets outside these ranges were removed to make the comparison clearer. A binary variable `is_night` was created (1 = night, 0 = day).

5 Exploratory Analysis

I explored:

- How many tweets were posted at night vs day
- Sentiment distributions for each time group
- Differences in sentiment confidence
- Tweet length differences

Overall, night tweets were fewer but showed slightly different sentiment patterns.

6 Statistical Tests

6.1 Sentiment Confidence (t-test)

I tested whether the average sentiment confidence differs between night and day tweets.

- Night mean ≈ 0.906
- Day mean ≈ 0.899

The Welch t-test produced:

$$t \approx 1.95, \quad p \approx 0.051$$

This result is very close to the 0.05 threshold, but technically I fail to reject the null hypothesis. The difference exists, but it is small.

6.2 Sentiment Categories (Chi-square test)

I also tested whether sentiment categories depend on posting time.

The Chi-square test result was:

$$\chi^2 \approx 9.13, \quad p \approx 0.010$$

Since the p-value is below 0.05, I reject the null hypothesis. This means that sentiment category distributions differ between night and day tweets.

7 Dataset Enrichment

To enrich the analysis, I used the `user_timezone` column to assign tweets to approximate countries. This mapping is not exact, so results are interpreted cautiously.

For each country, I calculated:

- Night posting ratio
- Number of tweets

These values were merged with the Quality of Life dataset.

8 Country-Level Findings

I computed correlations between night posting ratio and quality of life indicators:

- Night ratio vs Quality of Life Index: $\approx +0.35$
- Night ratio vs Pollution Index: ≈ -0.33

These are moderate correlations and are meant to be descriptive rather than causal.

9 Machine Learning Analysis

I trained text-based models to predict whether a tweet was posted at night.

9.1 Models Used

- Logistic Regression
- Random Forest
- Linear SVM
- Multinomial Naive Bayes

TF-IDF was used to convert text into numerical features.

9.2 Class Imbalance Problem

Initially, about 85% of tweets were day tweets and only 15% were night tweets. Because of this imbalance, models achieved high accuracy but very low F1 and recall for night tweets.

9.3 Oversampling Fix

I applied simple oversampling to balance the dataset. After this step, model performance on night tweets improved significantly.

10 Results Summary

After fixing class imbalance:

- Models could successfully identify night tweets
- Random Forest achieved the best overall performance
- The earlier poor results were caused by imbalance, not lack of signal

11 Conclusion

This project shows that posting behavior at night is different from daytime posting in several ways. While sentiment confidence differences are small, sentiment categories and text patterns differ clearly. Machine learning models confirm that posting time can be inferred from text when data imbalance is handled properly.

12 Limitations

- Country mapping is approximate
- Results are correlational, not causal
- Oversampling may exaggerate model performance

13 Future Work

Possible extensions include:

- Better location information
- Topic modeling of night vs day tweets
- Adding engagement metrics such as retweets