UC Irvine
Paul Merage
School of Business

Detecting Suspicious and Fake Amazon Reviews Using Text & Sentiment Analysis

Data & Programming for Analytics

Team 18B
Christie Shin
Shivani Vallamdas
Gema Zhu
Vishal Srivastava
Shuai Zhao

December 9, 2025

1. Introduction

1.1 Business Idea

As Amazon and e-commerce platforms continue to grow, identifying inauthentic or suspicious reviews has become increasingly difficult through manual moderation alone. Fraudulent reviews may come from bots, incentivized reviewers, or coordinated networks and often exhibit patterns that are subtle and difficult to detect with rule-based approaches. Therefore, this creates a need for automated techniques capable of detecting suspicious behavior at scale while preserving legitimate customer reviews.

In response to this challenge, our project investigates the use of machine learning techniques to detect suspicious Amazon reviews by leveraging behavioral, sentimental, and linguistic features. By analyzing reviewer posting patterns, sentiment extremity, and textual structure, we aim to distinguish suspicious reviews from legitimate ones and provide insights into the features most indicative of potential review manipulation.

Through this analysis, we aim to address the following questions:

1. Do suspicious reviewers display abnormal posting patterns (frequency, timing)?
2. Do suspicious reviews exhibit more extreme sentiment (overly positive or overly negative) than legitimate reviews?
3. Do suspicious reviews show abnormal linguistic patterns, such as overly generic wording, repeated phrases, or unusually short/long text?

1.2 Data Source

The data used in this study was sourced from Kaggle under the title *"Amazon Products Review"*. The dataset contains approximately 568,000 customer reviews on different Amazon products. It contains 10 columns that include the reviewID, productID, userID, text, time and score rating of the product.

2. Data cleaning

To ensure data consistency, several preprocessing steps were performed:

- Removed duplicates based on ProductId, UserId, Time, and Text (1,208 records deleted).
- Validated helpfulness fields and removed invalid rows where helpful votes exceeded total votes.
- Converted timestamps from Unix to standard calendar dates for easier time-based analysis.
- Checked missing values in Summary, Text, and ProfileName (none found).
- Cleaned text data by creating a CleanedText column — lowercased, trimmed spaces, and standardized content.
- Added behavioral indicators: ReviewLength, UserReviewCount, IsPerfect, AverageScorePerUser, DailyReviewRate, and CountOf5StarReviews to describe user-level activity.

After these steps, the dataset was fully cleaned and ready for feature engineering and modeling.

3. Feature Engineering

3.1 Sentiment Polarity

Using TextBlob, each review received a score from –1 (negative) to +1 (positive).

This detected cases such as:

- Text saying “Amazing!!!”
- Rating given as 1-star

Such inconsistencies signal bot-generated or template-based reviews (*Figure 3.1)*.

```
----Sentiment analysis----
----Summaries----
100%|██████████| 567244/567244 [01:22<00:00, 6908.86it/s]
Summary Sentiment calculated.
----Cleaned Text----
100%|██████████| 567244/567244 [04:32<00:00, 2084.51it/s]
Text Sentiment calculated.

Sample Results:
   Score                Summary  SummarySentiment  \
0      5  Good Quality Dog Food               0.7
1      1      Not as Advertised               0.0
2      4  "Delight" says it all               0.0
3      2          Cough Medicine              0.0
4      5             Great taffy              0.8

                                         CleanedText  TextSentiment
0  i have bought several of the vitality canned d...       0.450000
1  product arrived labeled as jumbo salted peanut...      -0.033333
2  this is a confection that has been around a fe...       0.133571
3  if you are looking for the secret ingredient i...       0.166667
4  great taffy at a great price. there was a wide...       0.483333
```

Figure 3.1

```
Subjectivity calculated.

Sample Results:
   Score  TextSentiment  Subjectivity
0      5       0.450000      0.433333
1      1      -0.033333      0.762963
2      4       0.133571      0.448571
3      2       0.166667      0.533333
4      5       0.483333      0.637500
```

Figure 3.2

3.2 Linguistic Feature

Subjectivity Score from 0 (objective, factual) to 1 (opinion-based, emotional) (*Figure 3.2)*.

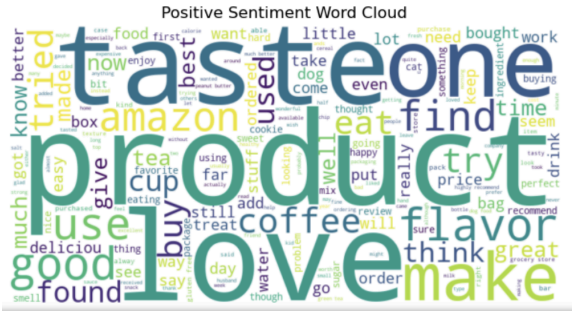
Hypothesis:

- **Real users** describe product specifics (“battery lasted 10hrs”).
- **Fake users** use vague hype (“best item ever!”).

The inclusion of this feature revealed a substantial “Subjectivity Gap”:

- Fake cluster: **0.67**
- Real cluster: **0.48**

3.3 Word Clouds



The word clouds shown above visualizes the most common words used by reviewers in both positive and negative sentiment categories, revealing that while both groups rely

on broad, generic words such as ‘product’, and ‘taste’, the negative reviews contain more complaint-driven language and the positive reviews rely more on repetitive praise terms, which are patterns that support our finding that suspicious reviews often use emotionally charged but nonspecific wording rather than detailed product descriptions.

4. Methodology

Since we lacked labels, we used K-Means Clustering to segment users based on the features defined above and then trained the Decision Tree model based on the results. Furthermore, we ran 2 separate Clustering & Decision Tree models, **"Model A" (Behavioral)** & **"Model B" (Linguistic).**

4.1 Model A Behavioral Approach

4.1.1 K-Means Clustering A

After normalization with StandardScaler, K-Means (k=4) produced four clusters (Figure 4.1.1).

***Cluster 0 emerged as the suspicious group**, defined by:*

- *High SingleDayReviewFrequency (5.6)*
- *High Sentiment (0.48)*
- *Short ReviewLength (236 characters)*

Other clusters represented normal reviewers and long-form writers.

```
----CLUSTER PROFILES----
         SingleDayReviewFrequency  TextSentiment  ReviewLength     Score  \
Cluster
0                        5.604724       0.480741    236.200080  4.796454
1                        3.371013       0.038884    412.425066  1.806346
2                        2.383607       0.188697    397.564530  4.770088
3                        4.424158       0.151587   1666.990326  4.194939

          Count
Cluster
0        162155
1        106654
2        262670
3         35765
```

Figure 4.1.1

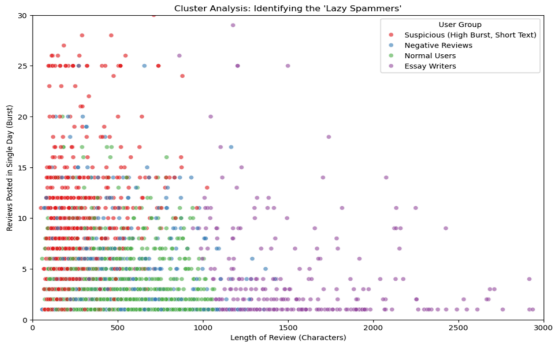


Figure 4.1.2

4.1.2 Decision Tree A

Using cluster labels, we trained a Decision Tree with the following performance:

- Accuracy: 96%
- Fake Recall: 0.90
- Normal Recall: 0.98

Top features :

1. TextSentiment
2. SingleDayReviewFrequency
3. ReviewLength

Confusion matrix observations (Figure 4.1.4 & 4.1.5):

- Correct fake detections: 29,187
- False negatives: 3,229 (fake → normal)
- False positives: 1,836 (normal → fake)

Model A shows that even without linguistic features, behavioral anomalies strongly indicate fraud.

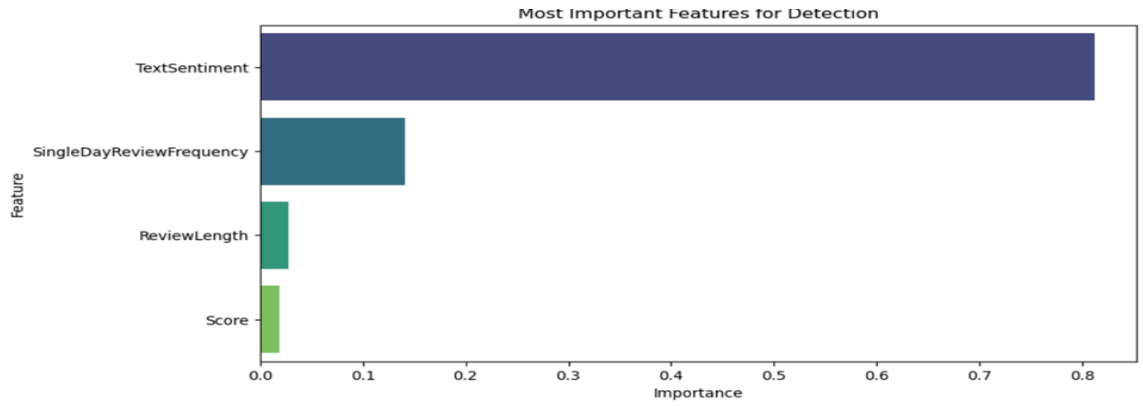


Figure 4.1.3

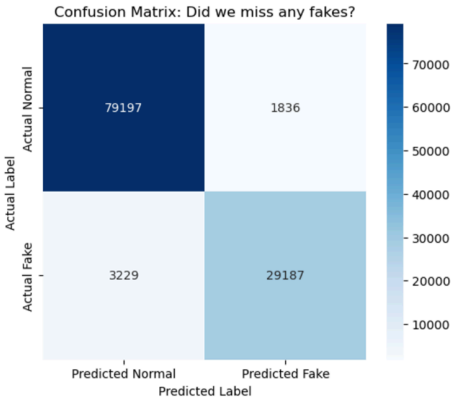


Figure 4.1.4

	precision	recall	f1-score	support
Normal	0.96	0.98	0.97	81033
Fake	0.94	0.90	0.92	32416
accuracy			0.96	113449
macro avg	0.95	0.94	0.94	113449
weighted avg	0.96	0.96	0.96	113449

Figure 4.1.5

4.2 Model B Linguistic Approach

4.2.1 K-Means Clustering B

With Subjectivity added, cluster boundaries shifted significantly (Figure 4.2.1). The new fraud-prone cluster displayed:

- Subjectivity: 0.67
- Sentiment: 0.45
- ReviewLength: 251 characters

This confirmed fraudsters' tendency toward subjective overstatement instead of factual detail.

Cluster	SingleDayReviewFrequency	TextSentiment	ReviewLength	Score	Subjectivity
0	3.988581	0.451217	251.180708	4.784355	0.669484
1	3.390369	0.035752	415.112800	1.795435	0.513767
2	4.779998	0.153440	1638.122535	4.201916	0.514196
3	3.255005	0.184925	395.760229	4.764862	0.478281

Figure 4.2.1

4.2.2 Decision Tree B

After adding Subjectivity, the Decision Tree redistributed its feature importance:

- Subjectivity became the second most influential feature (27%)
- SingleDayReviewFrequency dropped to 0% (Fig 4.2.2)

This demonstrates a paradigm shift:
from catching fast reviewers → to catching emotionally inflated writing.

Performance:

- Accuracy: 95%
- Fake Recall: 0.92 (better than Model A)
- False negatives reduced (2,881 vs 3,229)

Normal false positives increased slightly (1,836 → 2,997), as the model became more sensitive to emotional tone. Overall precision remained high (0.96).(Figure 4.2.3 & 4.2.4)

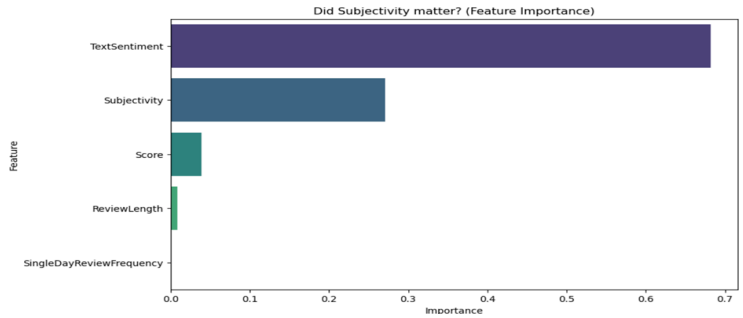


Figure 4.2.2

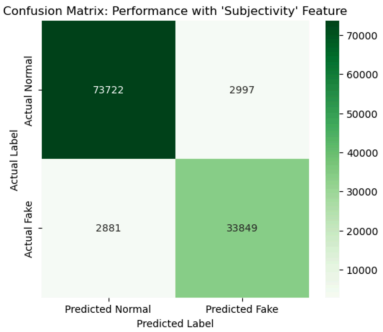


Figure 4.2.3

	precision	recall	f1-score	support
Normal	0.96	0.96	0.96	76719
Fake	0.92	0.92	0.92	36730
accuracy			0.95	113449
macro avg	0.94	0.94	0.94	113449
weighted avg	0.95	0.95	0.95	113449

Figure 4.2.4

5. Model Validation

5.1 Cross-Validation

Model A

Average Accuracy: 95.69%

Standard Deviation: 0.176

(Figure 5.1)

→ High model stability based on behavioral patterns alone.

```
Results:
Scores for each run: [0.95807808 0.95556594 0.95828081 0.95416443 0.95858896]
Average Accuracy:    95.69%
Standard Deviation:  0.176
```

Figure 5.1

Model B

- Average Accuracy: 94.80%
- Standard Deviation: 0.079

  (Figure 5.2)

  → Linguistic features introduce complexity but remain stable across folds.

```
Results:
Scores for each run: [0.94773863 0.94839972 0.94783559 0.94683955 0.94923665]
Average Accuracy:    94.80%
Standard Deviation:  0.079
```

Figure 5.2

5.2 Forensic Validation

To confirm the existence of real fraudulent activity, we conducted a text-duplication forensic audit.

- Identified reviews with **identical strings > 50 characters** from different UserIDs
- Found **141 verified duplicate reviews**

6. Research Question Results

1. Do suspicious reviewers display abnormal posting patterns (frequency, timing)?

Yes. Suspicious reviewers displayed clear posting anomalies. While normal users posted 1–2 reviews per day, the suspicious cluster frequently produced **20–70 reviews in a single day**, with an average of **5.6/day**. The Decision Tree confirmed **SingleDayReviewFrequency > 6.5** as a key fraud indicator, showing that “burst posting” is a defining behavioral signature of fake accounts.

2. Do suspicious reviews exhibit more extreme sentiment (overly positive or overly negative) than legitimate reviews?

Yes. Suspicious reviews were heavily polarized, clustering near **+1.0 or -1.0** sentiment scores, unlike legitimate reviews which showed moderate, nuanced polarity. The model also detected mismatches, such as highly positive text paired with low star ratings, indicating automated or error-prone generation. This consistent emotional extremity distinguishes fake reviews from genuine customer feedback.

3. Do suspicious reviews show abnormal linguistic patterns, such as overly generic wording, repeated phrases, or unusually short/long text?

Yes. Fraudulent reviews were noticeably **shorter and more generic**, averaging **~236 characters** compared to ~397 for normal users, >1,600 for detailed reviewers & high subjectivity. Forensic analysis further identified **141 duplicated reviews** originating from repeated scripted templates, confirming coordinated posting behavior.

7. Conclusion

7.1 Recommendations

Based on the findings of our analysis, we propose two primary recommendations for Amazon.

Recommendation 1: Implement Fraud Detection and Sentiment-Mismatch Controls

To improve the reliability of user reviews, Amazon can introduce automated systems that detect suspicious activity and inconsistencies in reviewer behavior. This can be achieved through the following:

- Flag reviews where the emotional tone of the text contradicts the numeric rating
    - E.g.,overly negative language paired with a 5-star store
- Temporarily freeze posting ability for users submitting 7+ reviews per day

- Require CAPTCHA or identity verification for repeated high-volume or inconsistent sentiment activity

Recommendation 2: Develop a Reviewer Risk Scoring System

To further enhance review integrity, Amazon should establish a scoring system that identifies potentially unreliable reviewers through behavioral and text-based analysis. This approach may include the following:

- Profile reviewer behavior to detect unusual patterns
- Track key metrics such as posting frequency, sentiment consistency, and text-length variation
- Generate a Reviewer Credibility Score to prioritize which reviews require manual verification

7.2 Limitations

While the project provides useful insights into review fraud patterns, several limitations should be acknowledged.

- The reviews in our dataset were collected in the early 2010s, meaning the model primarily captures older bot behaviors and does not account for modern threats such as LLM-generated text. As review fraud has evolved significantly, our results may not fully reflect today's more sophisticated patterns.
- The dataset lacked verified "fake" or "real" tags from Amazon. Without ground-truth labels, the model identifies mathematical anomalies and high-risk behavior rather than confirmed fraudulent activity, which limits the certainty of our predictions.
- Our detection approach is optimized for high-volume burst attacks and low-effort bots. While effective for these categories, the model may be less sensitive to low-frequency or more subtle forms of manipulation.
- The strict threshold we applied, which flags users who post more than 6.5 reviews per day, may classify all high-activity users as suspicious. This could lead to false positives among legitimate power reviewers or product testers.

References

Arhamrumi. *Amazon product reviews*. Kaggle. https://www.kaggle.com/datasets/arhamrumi/amazon-product-reviews