

# Craigslist Forum Mining Analytics System

---

CLIENT: ☮ craigslist



# Agenda

1. Background
2. Problem Identification
3. Business Objectives
4. Testing Methodology
5. Validation & Model Reliability
6. Solution & Impact



# BACKGROUND



An "OG" social ads platform allowing users to post ads for jobs, housing, services, and more, enabling anonymous user interactions through discussion forums and private messaging.

Subset for problem  
identification



# User-generated feedback on forums remains unstructured and unanalyzed

bou **Suggestions for 3 Craigslist user enhancements** < **BriGood** > 2025-04-12 16:00 +2  
: . . . **have you used Save Search?** < **HelpfulHedda** > 2025-04-12 17:24 -8+6  
: . . . . . **Safe search - Thanks for the reminder!** < **BriGood** > 2025-04-12 18:36 -1  
: . . . . . **Save search worked - Thanks HelpfulHedda** < **BriGood** > 2025-04-13 05:54 +1

**Same ad, multiple pages** < **JLKSR-1** > 2025-04-08 09:20

How is it that a poster can dominate page after page of the same ad of the same item for sale? To be specific, the ad is under Shipping Containers and the poster is DryBox.

**Why was my listing flagged and deleted** < **Winquiry** > 2025-04-24  
: . . . **you're in the wrong forum.... flagged ads are** < - > 2025-04-24

# User-generated feedback on forums remains unstructured and unanalyzed

Why isn't there a Category for mobile homes < ede365 >

--- Craigslist needs a new sales category:new produc < Shopweasel > 2023-12-06 00:56  
: . . Jesus Christ < Jim203 > 2023-12-06 10:07 +5  
: . . New products already have categories on CL. < - > 2023-12-07 14:00

Listings have image links but no images < LAKnox >  
: . . Weeks or months? < HelpfulHedda > 2025-04-23 09:  
: . . : . . Problem listings... < LAKnox > 2025-04-23 09:31

# User-generated feedback on forums remains unstructured and unanalyzed

Why isn't there a Category for mobile homes < ede365 >

not getting email for my ads < Zolba >

: . . . The anon relay is subject to moderation

: . . . The relay system is glitchy, and has

--- Craigslist needs a new sales category:new product < Shopweasel > 2023-12-06 00:56

: . . . Jesus Christ < Jim203 > 2023-12-06 10:07

: . . . New products already have categories

Listings have image links but no images < LAKnox >

: . . . Weeks or months? < HelpfulHedda > 2025-04-23 09:31

: . . . . . Problem listings... < LAKnox > 2025-04-23 09:31

Why was my listing flagged and deleted < Winquiry > 2025-04-24

: . . . you're in the wrong forum.... flagged ads are < - > 2025-04-24

# User-generated feedback on forums remains unstructured and unanalyzed

bou **Suggestions for 3 Craigslist user enhancements < ede365 >**  
: . . . have you used Save Search? < HelpfulHedda >  
: . . . Safe search. Thanks for the reminder. < Bu

**not getting email for my ads < Zolba >**

: . . **The anon relay is subject to modera**

: . . **The relay system is glitchy, and has**

--- **Craigslist needs a new sales category:new produc < Shopweasel >** 2023-12-06 00:56

: . . **Jesus Christ < Jim203 >** 2023-12-06 10:07

: . . **New products already have categories**

**Listings have image links but no images < LAKnox >**

: . . **Weeks or months? < HelpfulHedda >** 2025-04-23 09:

: . . : . . **Problem listings... < LAKnox >** 2025-04-23 09:31

Same ad, multiple pages < JLKSR-1 > 2025-04-08 09:20

**Why was my listing flagged and deleted < Winquiry >** 2025-04-24

: . . **you're in the wrong forum.... flagged ads are < - >** 2025-04-24

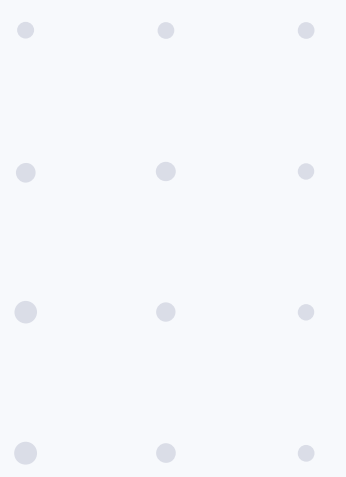
of the same ad of the same item for sale? To be  
the poster is DryBox.



# BUSINESS OBJECTIVE

Our goal is to automate the extraction, structuring, and analysis of Craigslist forum discussions using a text analytics pipeline to assist Craigslist's platform managers in identifying user sentiment, pain points, and systemic issues.

## Why it matters:



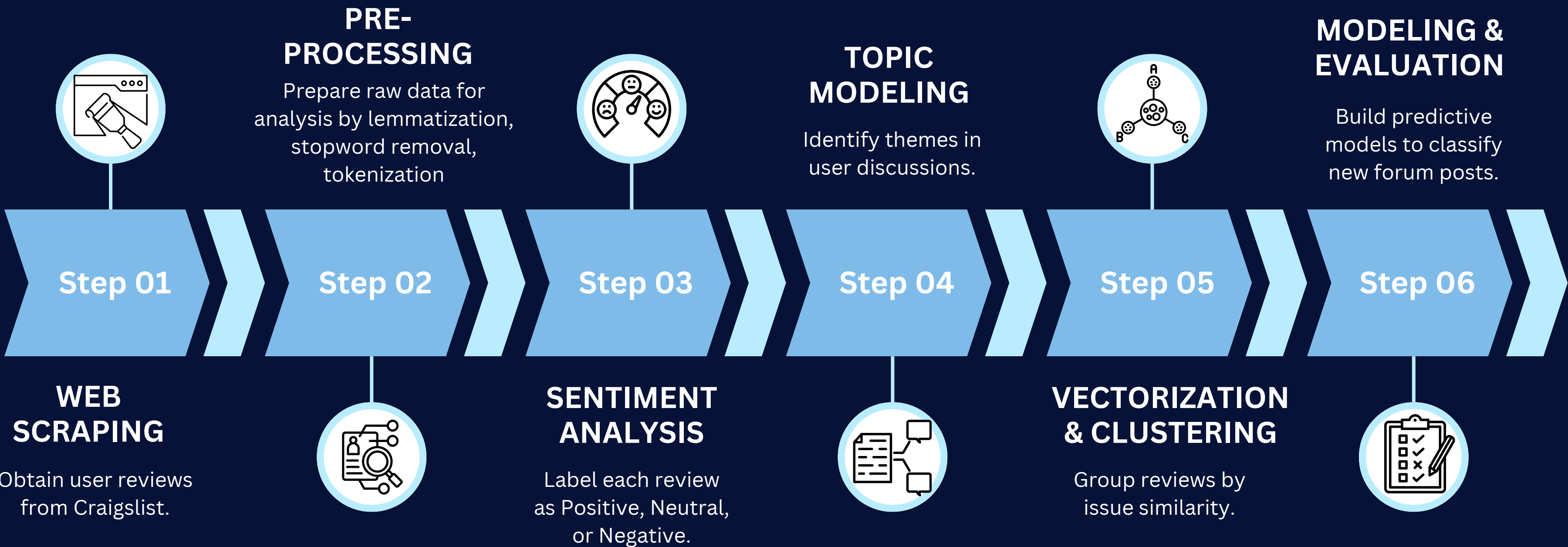
**Reduce manual triage  
for platform managers**

**Identify systemic  
frustrations and  
feature gaps**

**Enhance Craigslist's  
ability to respond to  
user needs**



# MACHINE LEARNING PIPELINE



- • •
- • •
- • •

# WEB SCRAPING & PREPROCESSING

## 1. Data Collection

- Source: Craigslist public forums (Feedback, Flag Help, Help Desk)
- Method: Web scraping using requests and BeautifulSoup
- Volume: ~900 forum posts saved to a DataFrame

## 2. Text Preprocessing

- Lowercased text
- Removed punctuation and numbers
- Tokenized into words
- Removed English stopwords
- Lemmatized tokens

```
class CraigslistSpider(scrapy.Spider):
    name = 'craigslist'
    start_urls = [
        'https://forums.craigslist.org/?forumID=8', # cl-help desk
        'https://forums.craigslist.org/?forumID=9', # cl-flag help
        'https://forums.craigslist.org/?forumID=3'  # cl-feedback
    ]
```

```
comment_text = comment.text.strip()
if comment_text:
    self.data.append({
        'id': 'comment',
        'review': comment_text,
        'type': 'main comment'
    })
```

```
# Preprocessing tools
stop_words = set(stopwords.words('english'))
lemmatizer = WordNetLemmatizer()

# Cleaning function
def clean_text(text):
    text = str(text).lower()
    text = re.sub(r'^\w\s', '', text) # remove punctuation
    tokens = word_tokenize(text)
    tokens = [lemmatizer.lemmatize(t) for t in tokens if t.isalpha() and t not in stop_words]
    return " ".join(tokens)

# Apply cleaning
df["cleaned_review"] = df["review"].apply(clean_text)
df.head()
```

# SENTIMENT ANALYSIS

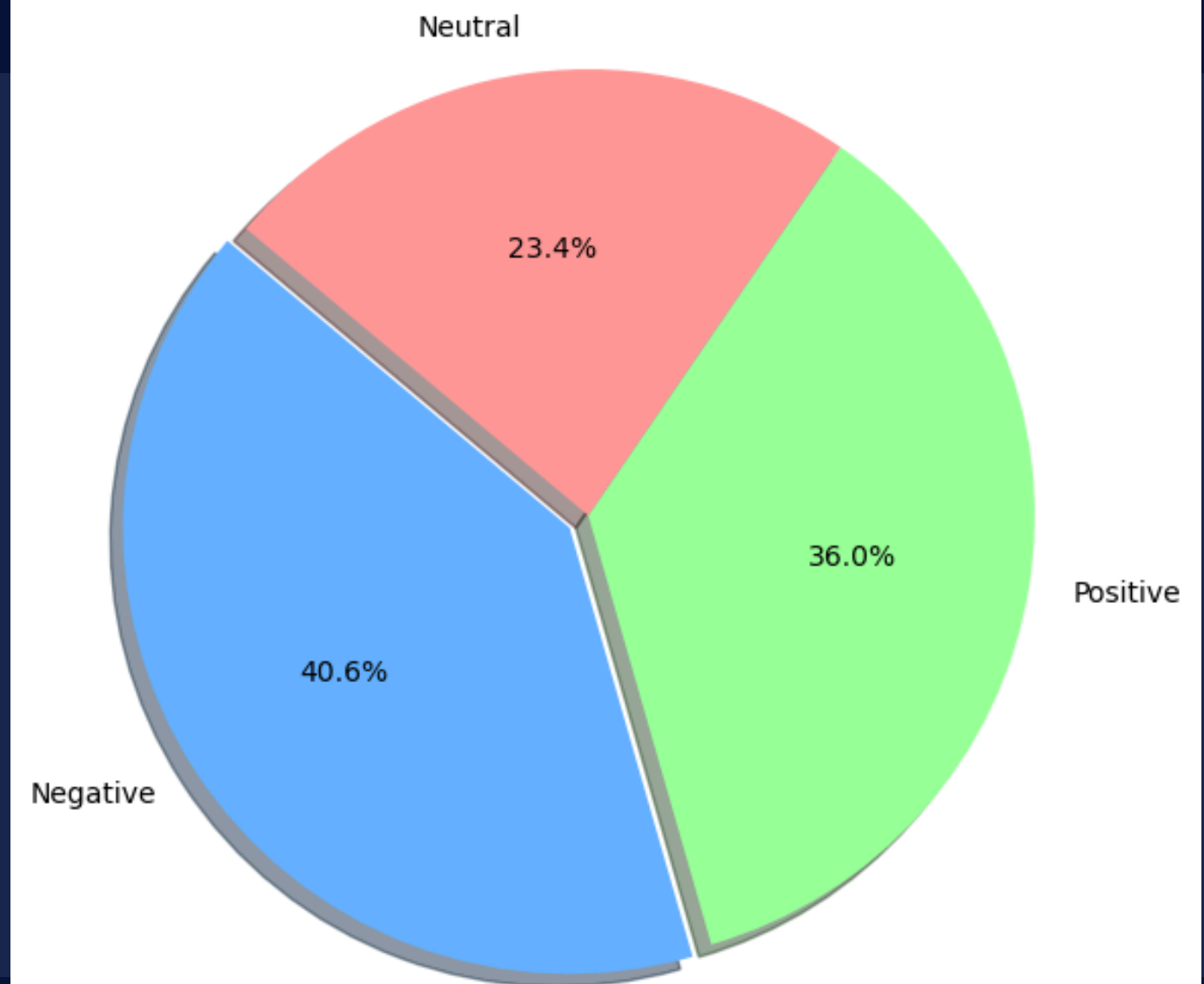
Goal: To label each review as Positive, Neutral, or Negative.

## Method:

- Used VADER sentiment analyzer
- Added a frustration keyword override layer (keywords like "blocked", "issue", "frustrated").
- If a post scored Positive but contained frustration keywords, label was switched to Negative.

**Insight:** This enhancement significantly improved the detection of falsely optimistic posts that were actually complaints.

Sentiment Distribution (Post-VADER + Frustration Override)



# TOPIC MODELING

Goal: To identify themes in user discussions.

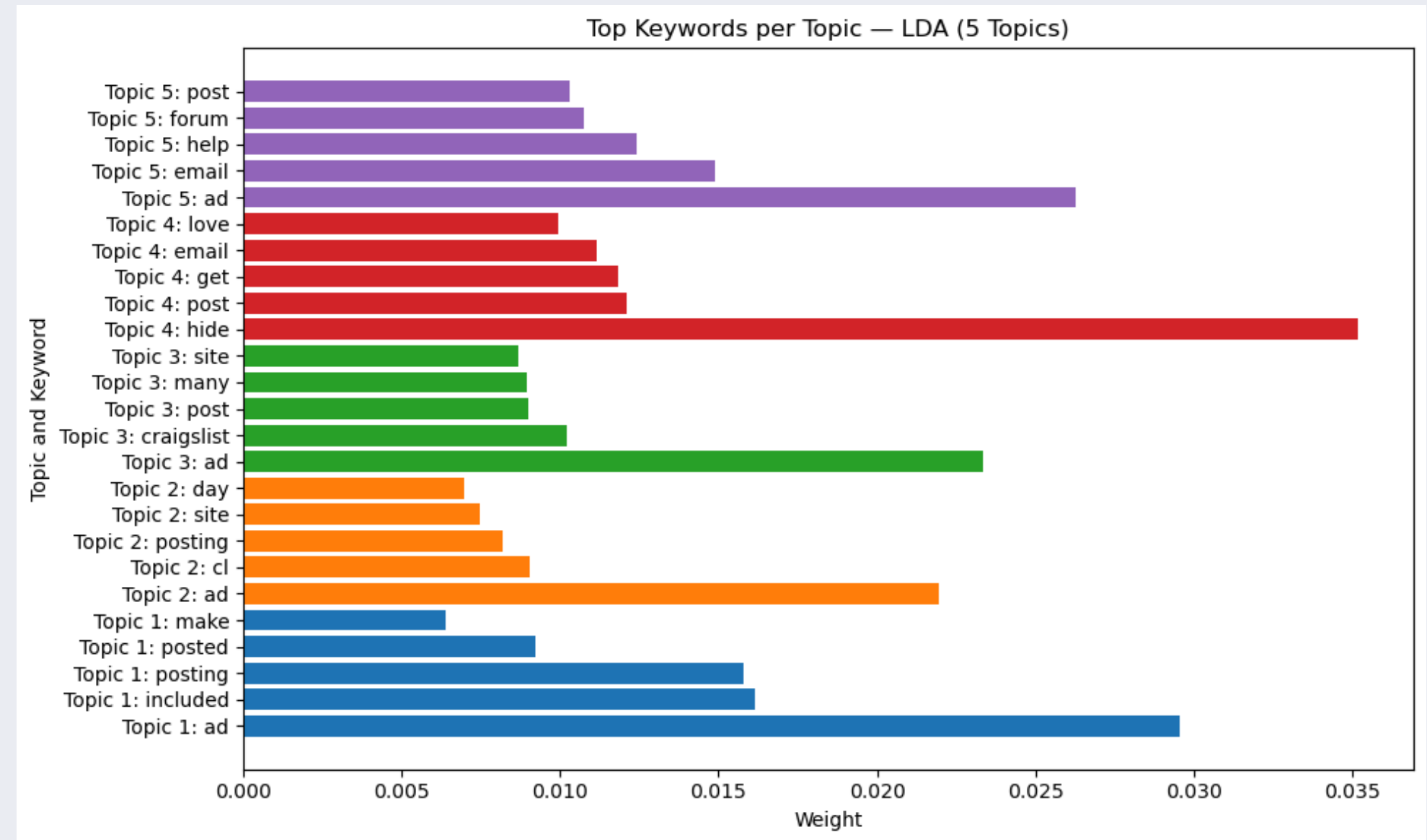
## Technique:

LDA via gensim on tokenized cleaned reviews.

- Number of topics: 5 (selected based on coherence & interpretability).

## Key Themes Identified:

- Posting mechanics and ad visibility
- Navigation and general UX
- Formatting/repetition complaints
- Listing descriptions (items/space)
- User outcomes and emotional feedback



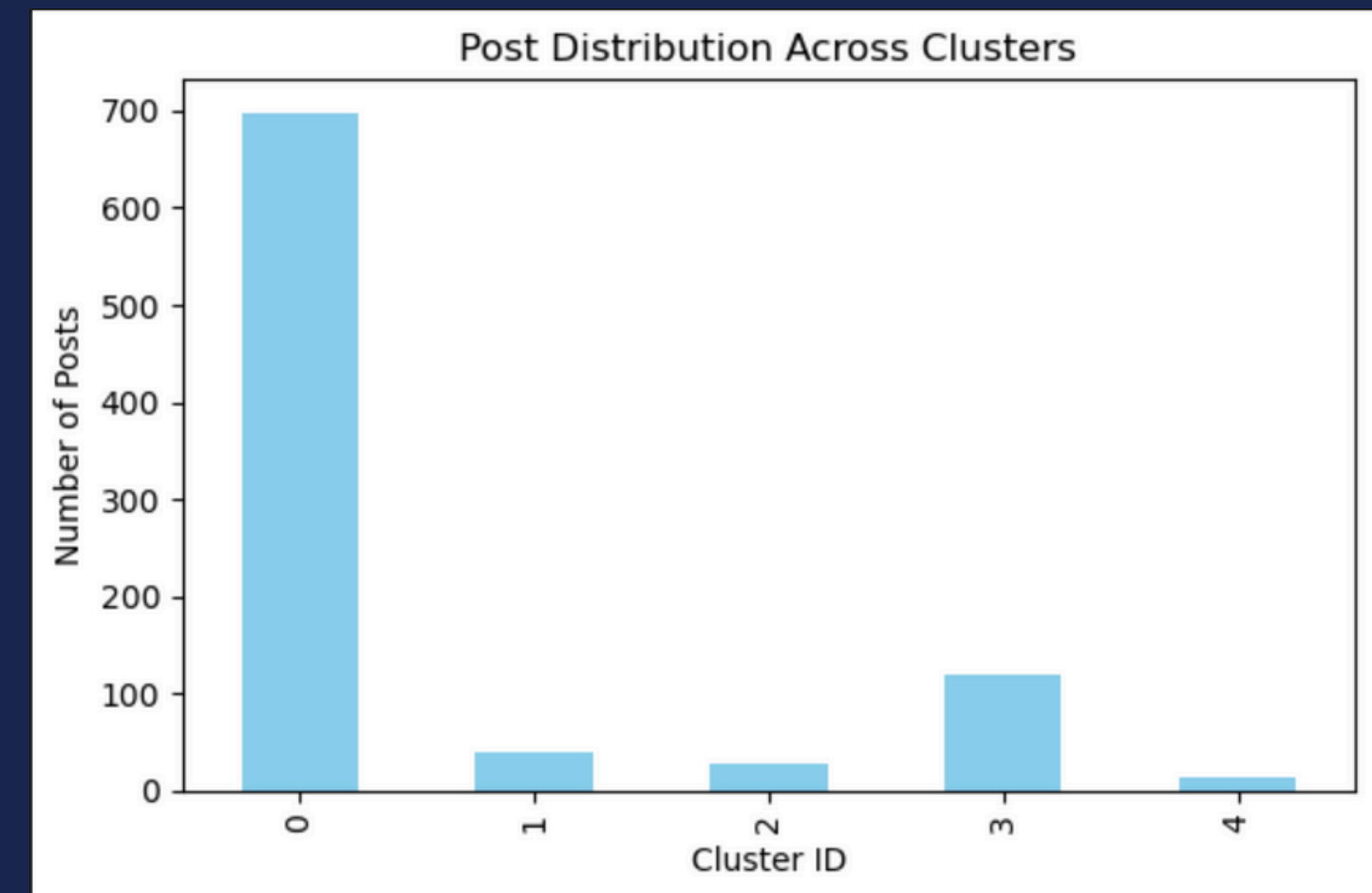


# VECTORIZATION & CLUSTERING

Goal: To group reviews by issue similarity.

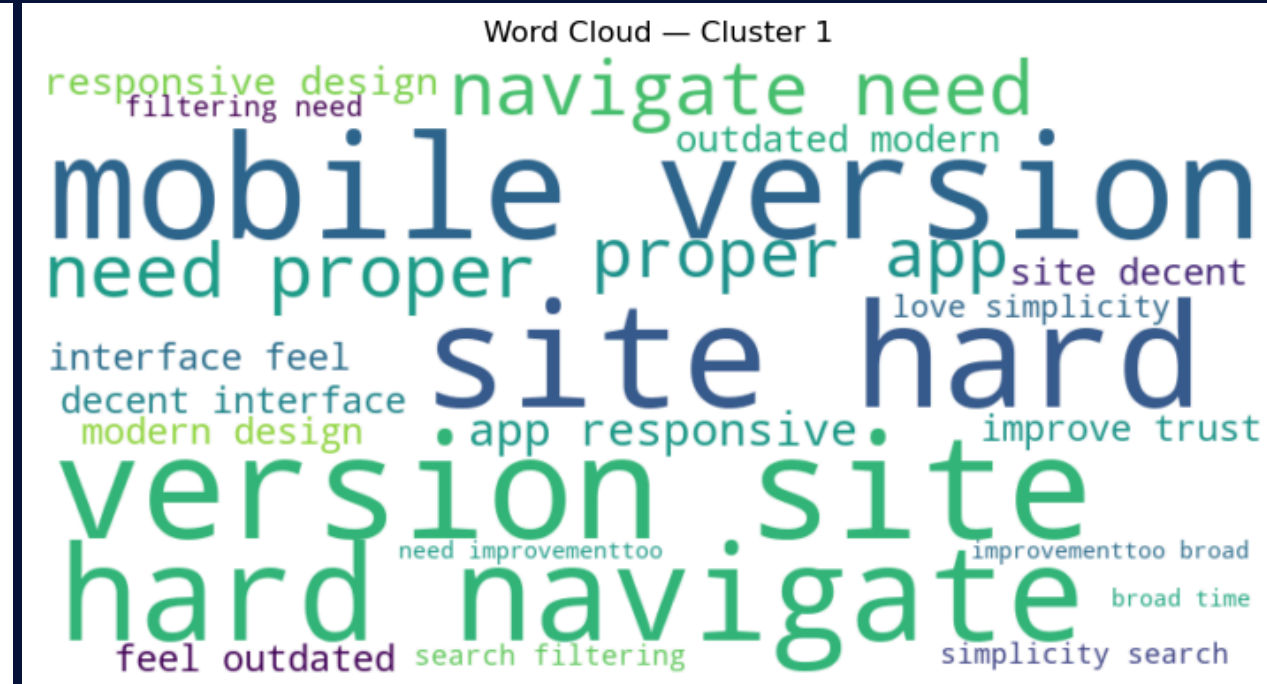
## Steps:

- Used TfidfVectorizer with n-grams (1,2) and min\_df=3
- Reduced dimensionality using TruncatedSVD (100 components, ~60% variance retained)
- Evaluated variance using cumulative explained variance plot
- Clustering via KMeans (k=5 selected)
- Evaluated using silhouette score (~0.51)
- Identified top keywords per cluster and generated cluster-specific word clouds

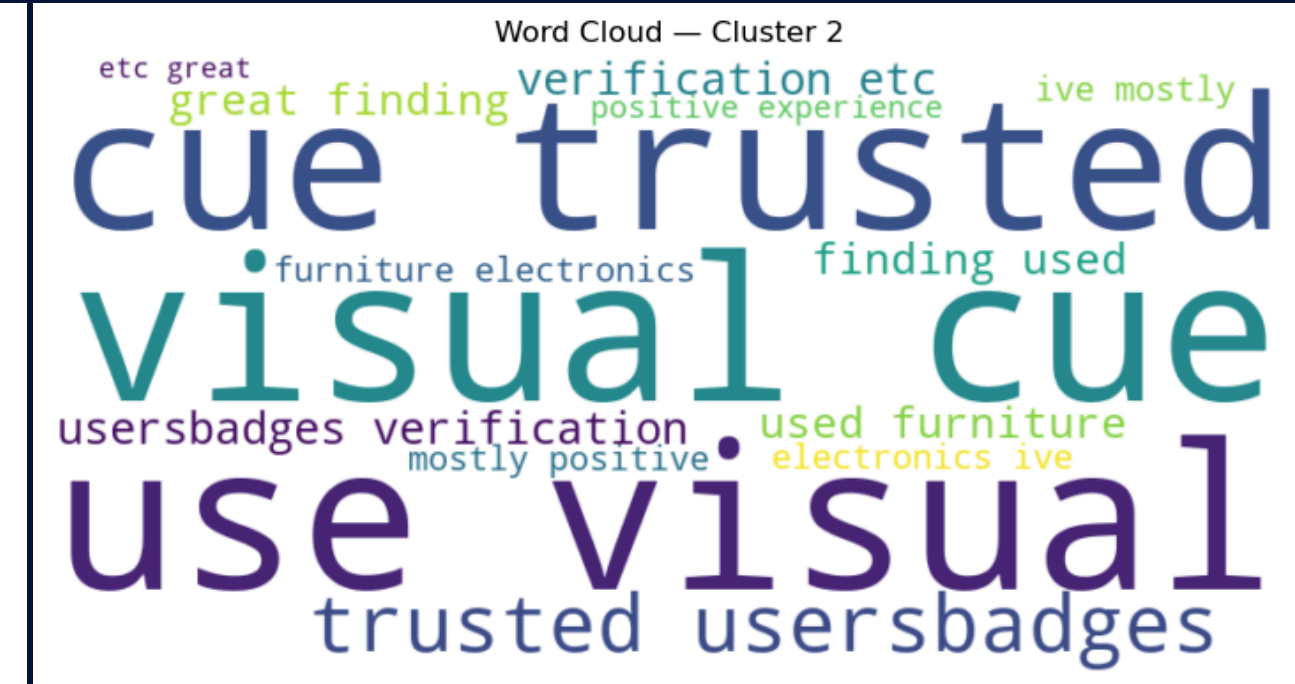




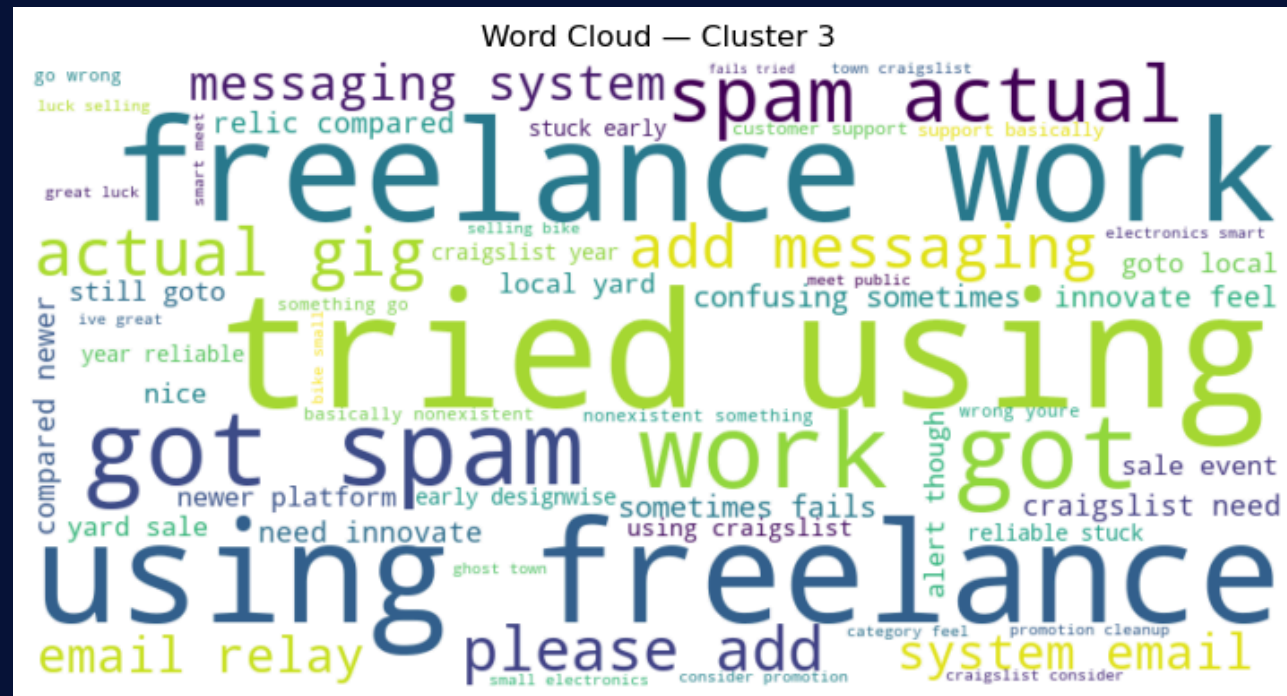
## Posting Barriers & Flagging Frustration



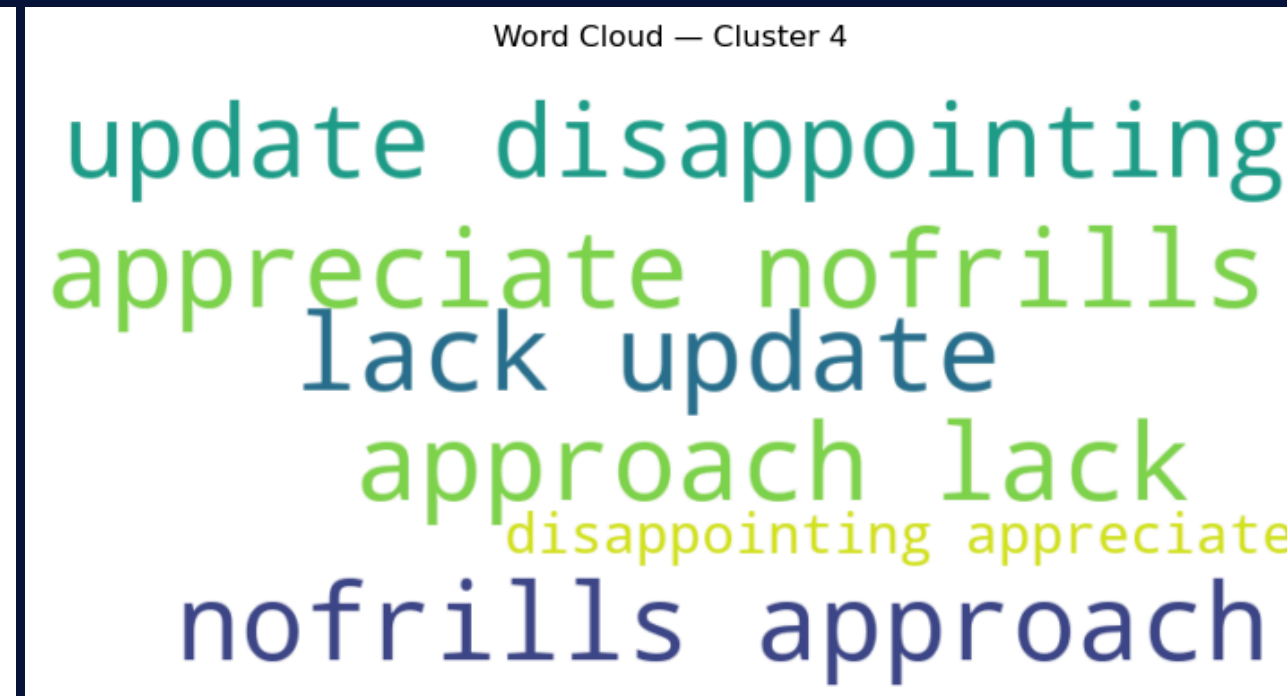
## Usability Complaints & Mobile Experience



## Trust Signals & Verified Listings



## Spam & Messaging System Limitations



## Mixed Sentiment Toward Simplicity & Progress

# MODELS APPLIED

Goal: To build predictive models to classify new forum posts.


## Algorithms Tested:

- Logistic Regression
- Naive Bayes
- Decision Tree
- Support Vector Machine (SVM)
- Random Forest

**Each model was trained on TF-IDF-transformed versions of the cleaned forum posts (cleaned\_review) with validated against the VADER labeling.**

## Preprocessing:

- Text converted to numeric features using TF-IDF vectorization
- Balanced class distribution using oversampling techniques
- 80/20 train-test split used for evaluation

 Model Accuracy Comparison

	Model	Accuracy
0	SVM	73.89
1	Random Forest	73.89
2	Logistic Regression	72.78
3	Decision Tree	63.33
4	Naive Bayes	62.22

# VALIDATION

## Evaluation Metrics:

- Accuracy, F1-Macro, Confusion Matrix
- Target threshold: Accuracy > **70%**, F1-Macro > **0.70**

## Best Models:

- **Random Forest** and **SVM** showed top individual performance
- Voting Ensemble (Hard Voting) combined top 3 models: **Logit, SVM, RF**
- **Final Accuracy: 74%, Macro F1: 0.73**

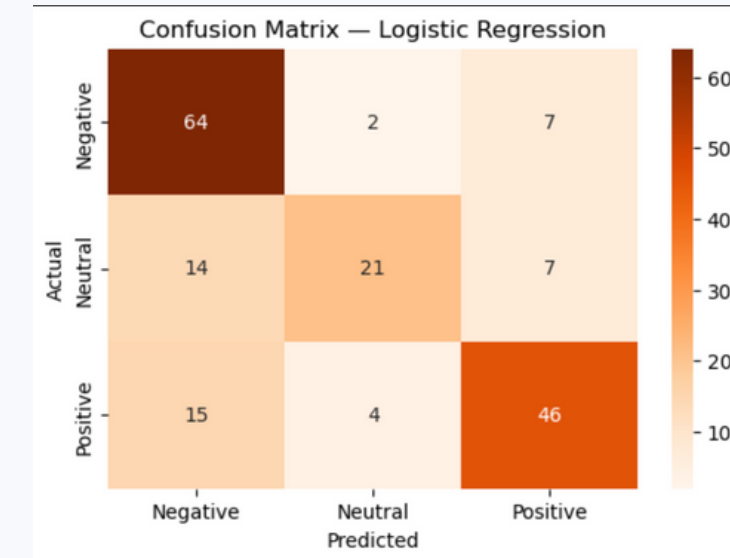
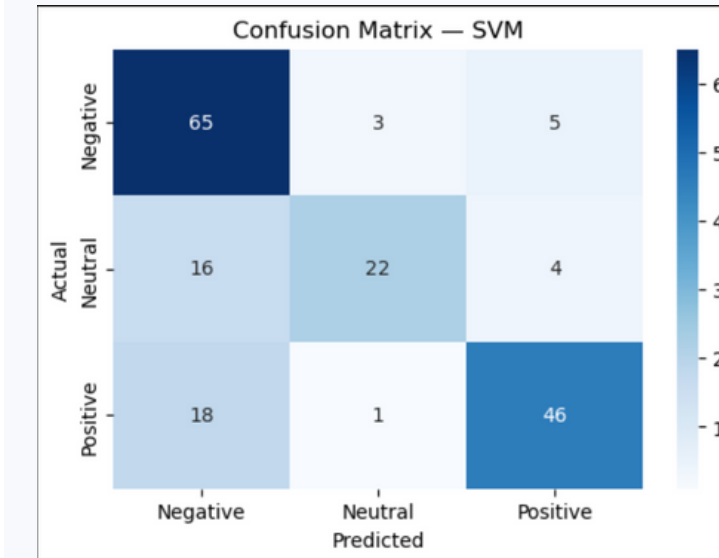
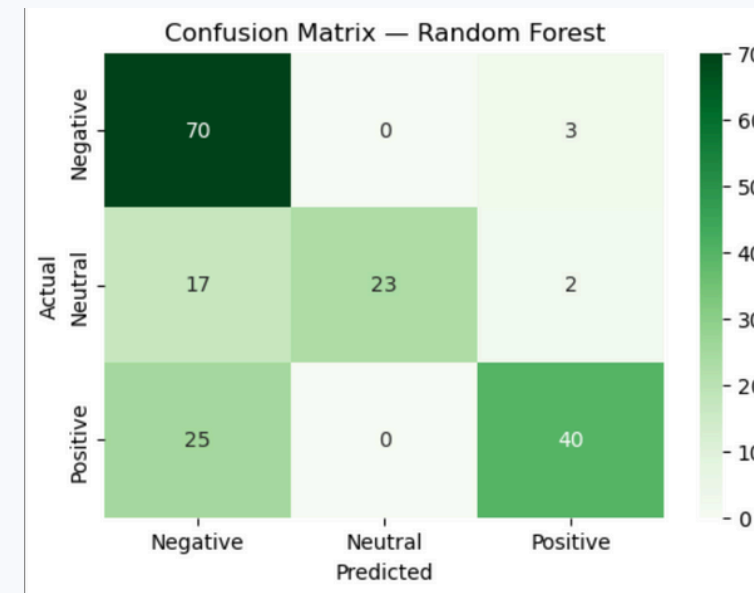
🔍 Classification Report – Voting Ensemble				
	precision	recall	f1-score	support
Negative	0.66	0.92	0.77	73
Neutral	0.88	0.55	0.68	42
Positive	0.85	0.68	0.75	65
accuracy			0.74	180
macro avg	0.80	0.71	0.73	180
weighted avg	0.78	0.74	0.74	180



# ENSEMBLE & INSIGHTS

## 💡 Why Voting Ensemble?

- Combines predictions of top 3 classifiers
- Reduces model variance
- Improves generalization across classes
- Smooths out weaknesses of individual models

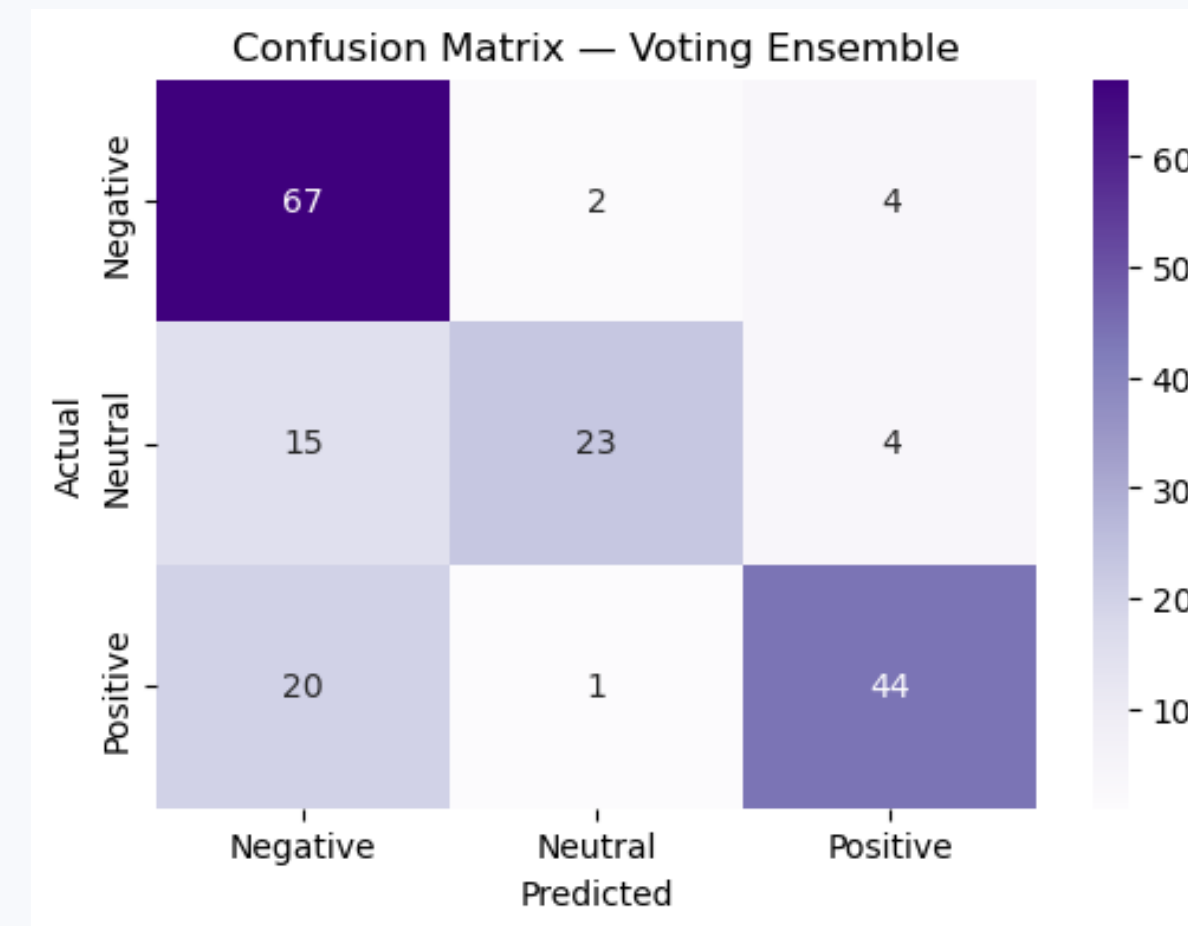


## 🔍 Performance Highlights:

- Negative recall: 92%
- Neutral precision: 88%

## 🧠 Key Insight:

- Misclassifications dropped significantly after applying the frustration keyword override during labeling. This ensured more reliable training signals and better separation of true negatives.





# SOLUTION & IMPACT

## Solution

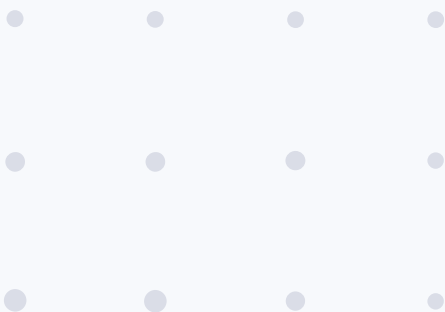
A full-stack NLP pipeline that automates sentiment detection, topic extraction, and complaint classification.

## Value

- Real-time feedback surfacing at scale
- Identification of root-cause frustrations (flagging, hidden posts)
- Enables faster decision-making by product & moderation teams
- Reduces reliance on manual community monitoring

## Deployability

Ready to integrate as a Forum Mining Analytics System into the existing tech stack at Craigslist.



# CONCLUSION

- **Most frequent complaints stem from post visibility, flagging, and confusing moderation.**
- **Spam and duplicate replies are common in freelance categories.**
- **Many users are satisfied when issues are resolved — showing potential for positive sentiment recovery.**
- **Our final model can be used for live monitoring or future triaging of user issues.**

**THANK YOU**

**For Your Time & Attention**