Automated Sentiment Analysis for Brand Monitoring

A MACHINE LEARNING APPROACH TO TWITTER DATA

FINAL GROUP 1 PROJECT PRESENTATION

NLP & MACHINE LEARNING ANALYSIS

Business Understanding

Business Overview

- Social media platforms generate millions of customer opinions daily
- Manual analysis of tweets is time-consuming, costly, and not scalable
- Organizations need automated methods to understand customer sentiment in real time

Business Problem

- Twitter data is unstructured but contains valuable insights about brand perception.
- Companies struggle to track public sentiment toward products like Apple and Google
- There is a need to automatically classify tweets as positive, negative, or neutral

Project Objective

- Apply Natural Language Processing (NLP) and Machine Learning techniques
- Accurately classify sentiment toward Apple and Google products
- Support brand monitoring, reputation management, and data-driven decision making

Data Understanding

Dataset Overview

- Source: Judge Emotion About Brands and Products dataset
- Size: 9,093 tweet entries
- Labels: Human-annotated sentiment toward brands or products
- Columns: tweet_text, emotion_in_tweet_is_directed_at, is_there_an_emotion_directed_at_a_brand_or_product

Key Data Challenges

- Missing values in brand/product column
- Ambiguous sentiment expressions Common in social media
- Class imbalance, with neutral sentiment dominating the dataset

Data Cleaning and Preparation

Cleaned data: Removed rows with missing tweet text
Filtered samples: Excluded ambiguous entries and standardized labels
Retained columns: tweet_text and sentiment
Text normalization:

- Lowercasing
- Removed URLs, mentions, and hashtag symbols
- Removed non-alphabetic characters
- Removed stopwords (preserving negations)
- Lemmatization using WordNet

Exploratory Data Analysis

Key Findings:

- **Class distribution:** Highly imbalanced; neutral dominates, negative is rare but important
- **Tweet length:** Average ~10 words → suitable for bag-of-words
- **Vocabulary:** Large relative to dataset → high sparsity; justifies feature limits in vectorization
- **Brand distribution:** Apple-related tweets dominate → potential model bias toward Apple sentiment

Modeling Strategy

Binary Classification Approach:

- Focus: **Positive vs. Negative sentiment**
- Excluded: Neutral & ambiguous tweets
- Rationale: Strong sentiment provides the **most actionable insights**

Feature Representation:

- **TF-IDF Vectorization** (Unigrams only)
- Max features: 5000

Modeling Strategy

Models Evaluated:

- Logistic Regression (with class balancing)
- Multinomial Naive Bayes
- Linear Support Vector Classification (LinearSVC)

Evaluation Strategy:

- **Train-test split:** 80/20 (stratified)
- **Metrics:** Recall, F1-score, Confusion Matrix
- **Business priority:** Minimize missed negative sentiment

Model Performance and Comparison

Model	Accuracy	Negative Recall	Positive Recall	Weighted F1-score	Limitation
Logistic Regression	85%	0.69	0.88	–	Misses ~31% of negative tweets
Multinomial Naive Bayes	85%	0.04	1.00	–	Fails to detect negative sentiment
LinearSVC (Selected)	**88%**	0.61	0.93	0.88	Best overall balance; reliable for brand monitoring

Final Model Selection: LinearSVC

Why LinearSVC?

- Highest overall accuracy (88%)
- Best balance between positive and negative recall
- Most reliable for brand monitoring use case
- Fewer extreme misclassifications

Performance Summary

- Strong at identifying explicit sentiment
- Struggles with ambiguous/neutral language
- Acceptable error rate for proof-of-concept
- Provides actionable insights for brand teams

Error Analysis

False Positives (Predicted Positive, Actually Negative)

- Often contain neutral or factual language(product updates, rumors, comparisons)
- Model over-weights brand mentions
- **Business Impact**: May overestimate positive sentiment

False Negatives (Predicted Negative, Actually Positive)

- Subtle or implicit positivity
- Casual endorsements without strong language
- Short or ambiguous phrasing
- **Business Impact**: Misses genuine positive feedback

N/B Errors reflect common NLP challenges in social media data

Model Interpretability: Feature Importance

Positive Sentiment Indicators:

- Words: "cool", "free", "great", "awesome", "fun", "winning", "smart"
- Context terms: "party", "week", "everywhere“

Negative Sentiment Indicators:

- Words: "fail", "hate", "headache", "crash", "faulty", "disappointment"
- Negation terms: "not", "won't", "didn’t”

Validation:

- Model captures **meaningful sentiment patterns** rather than random correlations

Limitations

Model & Feature Limitations:

- Bag-of-words ignores **word order and context**
- Struggles with **sarcasm, irony, and nuanced expressions**
- **Class imbalance** biases toward positive sentiment
- Missing metadata: no emoji or hashtag semantics
- Binary classification excludes **neutral sentiment**

Dataset Limitations:

- **Apple-dominated content** → potential bias
- Limited negative examples
- Tech-specific terminology

Recommendations for Improvement

Enhanced Features

- Add n-grams and emoji handling
- Incorporate sentiment lexicons
- Include metadata features

Advanced Modeling

- Explore transformer-based embeddings (BERT, etc.)
- Implement cost-sensitive learning
- Add neutral class for 3-way classification

Deployment Considerations

- Human-in-the-loop review system
- Real-time monitoring dashboard
- Regular model retraining with new data

Business Value and Conclusion

Key Business Applications:

- **Brand Health Monitoring:** Track sentiment trends over time
- **Crisis Detection:** Early identification of negative spikes
- **Campaign Analysis:** Measure impact of marketing efforts
- **Competitive Intelligence:** Compare brand sentiment vs. competitors

Project Success:

- Developed **working sentiment classifier** (88% accuracy)
- Extracted **actionable insights** from noisy social media data
- Built **interpretable, business-aligned model**
- Created foundation for **production deployment**

Future Vision:

- Scalable, **real-time brand sentiment monitoring platform**

Thank you!

https://github.com/sheillamacharia/twitter_brand_sentiment_nlp.git

Presented by:

Sheilla Macharia

Margaret Mondia

Benard Kariuki

Bravian Ashono

Isaac Ndung'u