What if your excitement was flagged as negative by AI? 🤔

Lost in Translation: How AI Misinterprets Modern Online Language and Emojis

Understanding NLP Failures on Culturally Rich and Emoji-Laden Comments

Malachi Randolph

Independent Researcher | Los Angeles, CA

Background

- NLP tools often fail to interpret internet language—informal, expressive, and filled with slang and emojis.
- This language includes terms like "goat," "queen," and "fire" which express sentiment and tone beyond literal meaning.
- Misinterpretations can lead to biased outcomes, especially when language rooted in marginalized communities is treated as "noise" or "foreign."
- This project examines how sentiment and language models misread culturally influenced internet language.

💀 ☠️ 🐐 🔥 👸🏾 👑 💅

Research Questions

- How do NLP models perform on comments that use informal internet language?
- Can human annotation and active learning improve outcomes?
- What are the social implications of these model failures?

Dataset

- 27,000+ comments from Beyoncé's 2016 halftime show
- 134,000+ comments from Kendrick Lamar's 2025 halftime show
- Both performances generated high engagement with rich use of internet slang, emojis, and culturally rich language
- Collected using YouTube API, with metadata (likes, replies, timestamps

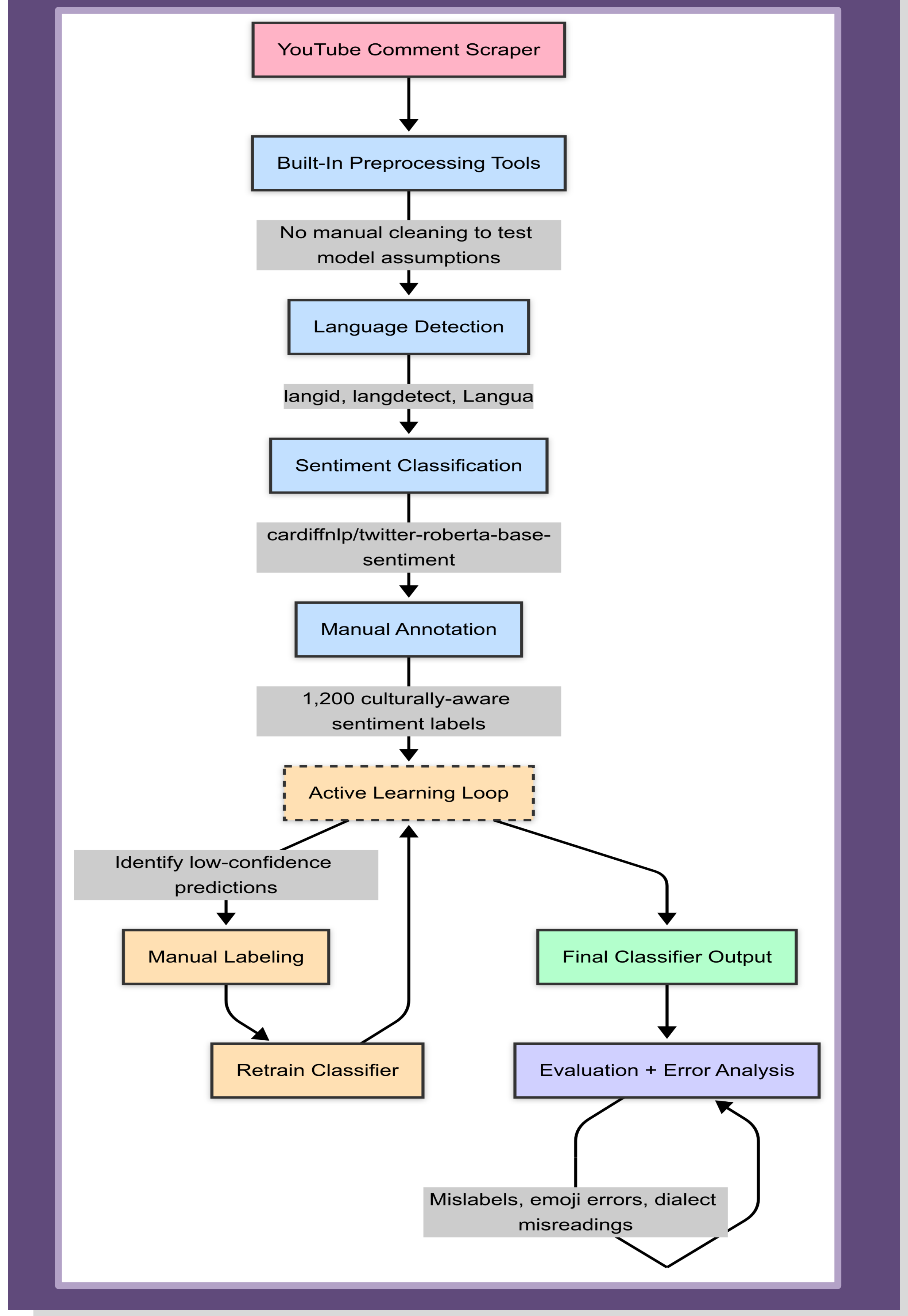
Annotation Process

- 1,200+ manually labeled comments using culturally-informed guidelines
- Sentiment categories: Positive, Neutral, Negative, Irrelevant

Modeling

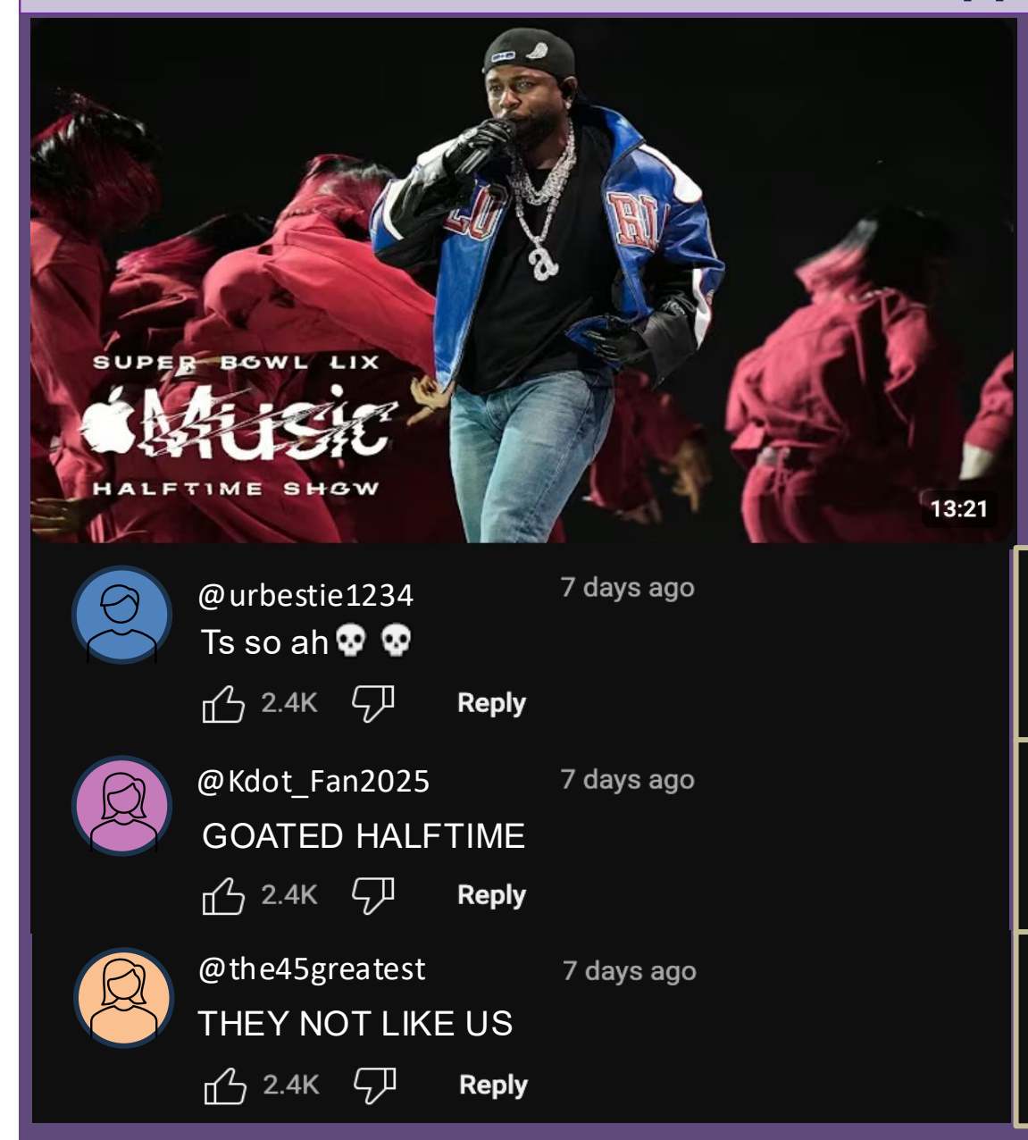
- Used `cardiffnlp/twitter-roberta-base-sentiment` for transformer classification
- Integrated `langid`, `langdetect`, and `langua` for language detection
- Two rounds of active learning (low-confidence sampling) to improve performance

Methodology Flow Chart


YouTube Comment Scraper
Built-In Preprocessing Tools
No manual cleaning to test model assumptions
Language Detection
langid, langdetect, Langua
Sentiment Classification
cardiffnlp/twitter-roberta-base-sentiment
Manual Annotation
1,200 culturally-aware sentiment labels
Active Learning Loop
Identify low-confidence predictions
Manual Labeling
Retrain Classifier
Final Classifier Output
Evaluation + Error Analysis
Mislabels, emoji errors, dialect misreadings

Misclassification Examples

YouTube: Kendrick Lamar's Apple Music Super Bowl Halftime Show



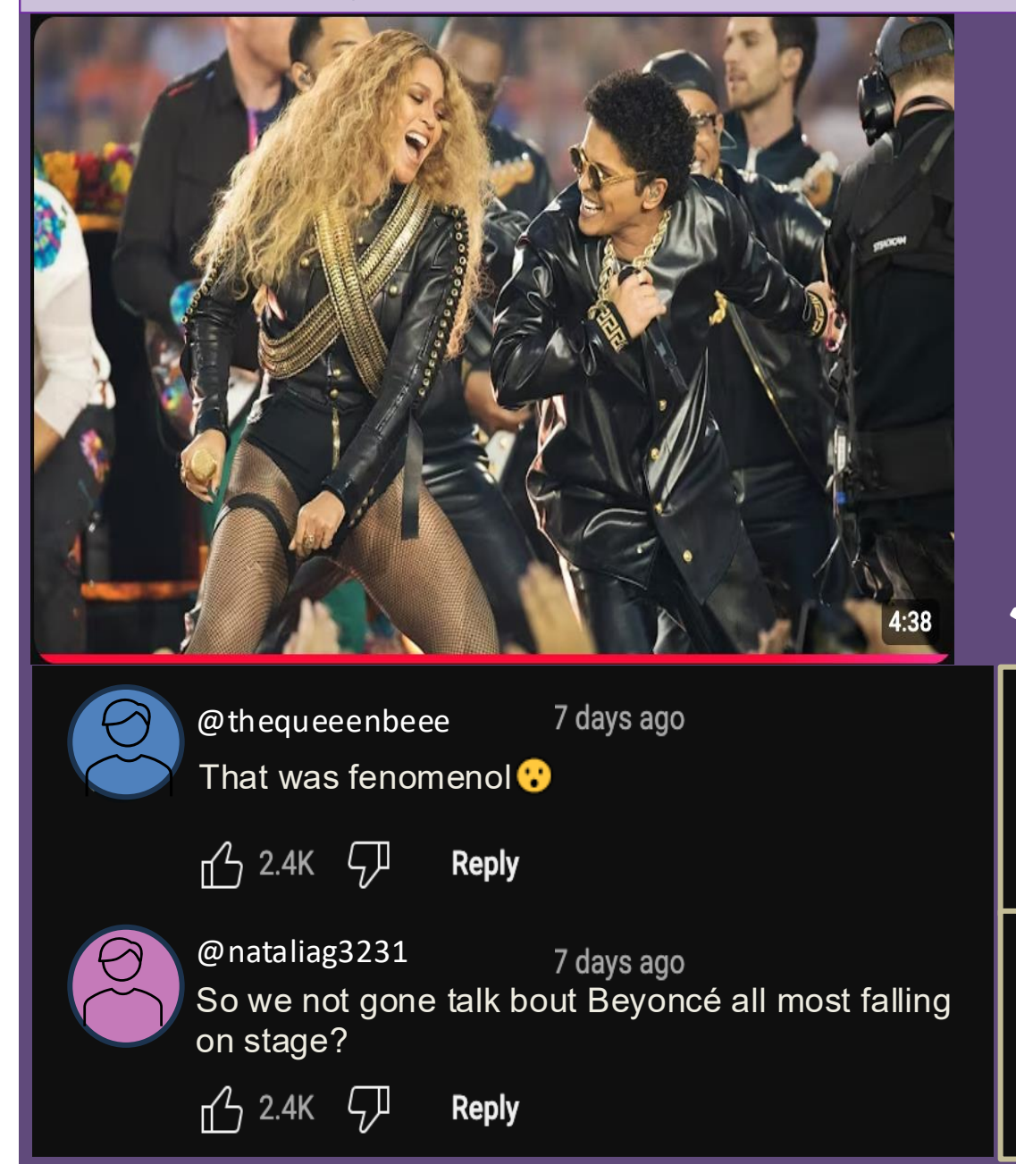
@urbestie1234 · 7 days ago
Ts so ah 💀 💀
2.4K · Reply

@Kdot_Fan2025 · 7 days ago
GOATED HALFTIME
2.4K · Reply

@the45greatest · 7 days ago
THEY NOT LIKE US
2.4K · Reply

Model Sentiment Label	Human Sentiment Label	langid language label	langdetect language label	langua language label
negative	positive	Khmer	Somali	Sotho
neutral	positive	English	German	English
negative	neutral	English	Vietnamese	English

YouTube: Beyoncé & Bruno Mars Crash the Pepsi Super Bowl 50 Halftime Show | NFL



@thequeeenbeee · 7 days ago
That was fenomenol 😮
2.4K · Reply

@nataliag3231 · 7 days ago
So we not gone talk bout Beyoncé all most falling on stage?
2.4K · Reply

Model Sentiment Label	Human Sentiment Label	langid - Language Label	langdetect - Language Label	langua language label
negative	positive	English	English	Malay
negative	neutral	French	English	English

Results – Language Detection

- The `langua` model had highest accuracy (~80%)
- The `langdetect` model mislabeled 27% of internet-language-heavy English comments as foreign
- Most errors occurred with short or emoji-filled phrases

Results – Sentiment Classification

- Transformer model consistently misunderstood culturally coded internet expressions
- Emojis like the skull were read literally, not contextually
- Active learning improved prediction accuracy and confidence

Internet Language & Cultural Influence

- Much of what is considered Gen Z or internet slang is rooted in African American Vernacular English (AAVE)
- Examples:
  - G.O.A.T. — stands for Greatest of All Time / admiration
  - Fire — approval / excitement
  - Queen — empowerment / admiration
- These phrases carry strong emotional tones, but are often misread by NLP tools not trained on culturally diverse data
- Treating such language as abnormal introduces risk of digital exclusion

Responsible AI Insight

- Annotations were guided by lived experience and supported by clear labeling guidelines
- Even with limited resources, the project prioritized transparency and fairness
- Highlights need for diverse annotator teams, culturally aware training data, and inclusive evaluation

Implications & Future Work

- Mislabeling internet language can marginalize voices and distort meaning in social data
- Recommendations:
  - Dialect-aware open-source datasets
  - Community-involved annotation
  - Responsible model design with the Bender Rule in mind
  - Incorporate AAVE with Standard American English when training models
- Broader goal: ensure NLP systems don't erase or distort culturally rich communication

Scan here to see my LinkedIn

Scan here to see the GitHub repo