Multilingual Sentiment Analysis of Tweets Using Pre-trained NLP Models

BY: YUVAL CASPI

THE COMMUNITY TECH
DATA SCIENCES, TORONTO, CANADA

EMAIL: YUVALCASPI12@GMAIL.COM

THE Community TECH

Hugging Face

Background:

- Online hate speech is rising across various social media platforms, including X (formerly known as Twitter).
- Harmful content spreads quickly across multiple languages, not just English
- Most existing tools focus on English; low-resource languages are overlooked.
- There's a growing need for scalable, language-agnostic sentiment tools to track emotional tone and toxic trends globally.

THE Community TECH

Hugging Face

HOW SENTIMENT ANALYSIS WORKS?

1. Text Preprocessing

Cleaning and preparing raw text for analysis.

2. Feature Extraction

Converts text into a structured format for analysis.

3. Sentiment Classification

Uses machine learning models to analyze and categorize text sentiment.

4. Sentiment Scoring

Assigns numerical values to represent sentiment intensity.

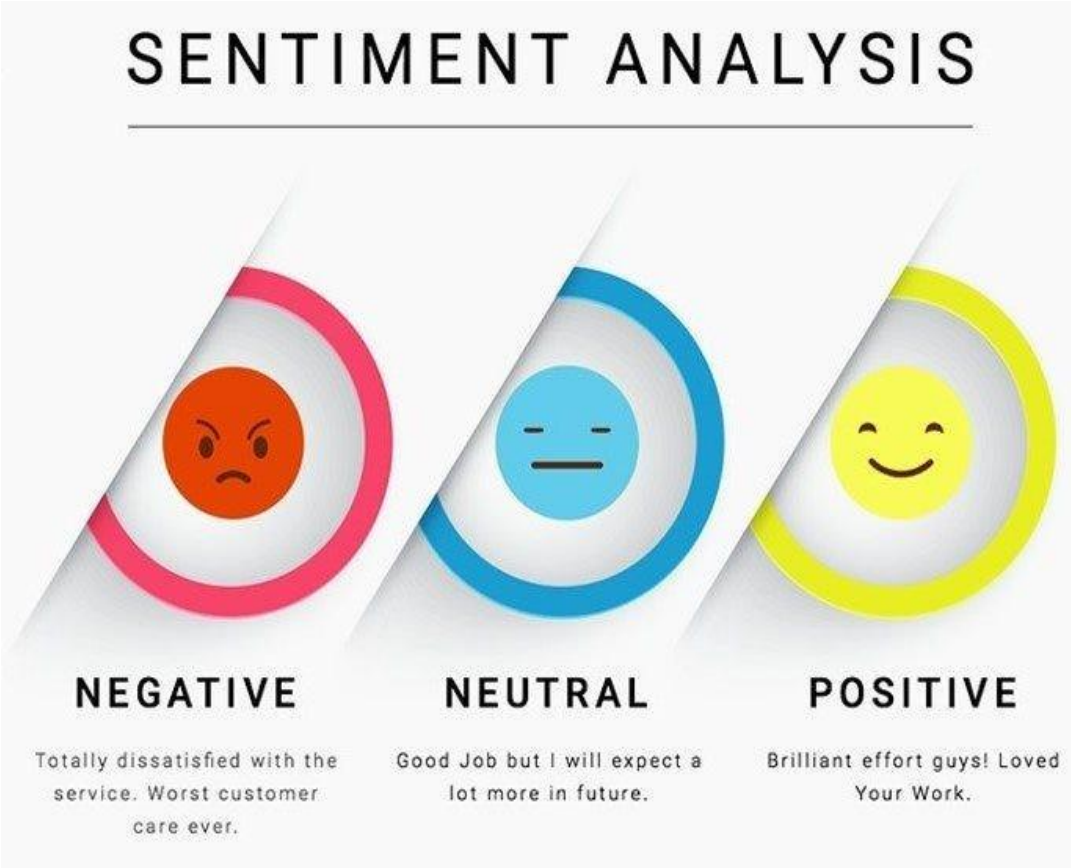
5. Post-processing and Visualization

Refining analysis results and displaying them for clear interpretation.

QuestionPro

Objectives:

- Apply pre-trained multilingual sentiment classifiers to tweets in English, Spanish, Japanese and Greek.
- Evaluate cross-language sentiment trends using descriptive statistics and visualizations.
- Assess model consistency across languages and topics.
- Provide insight into cultural or linguistic variation in emotional tone on social media.

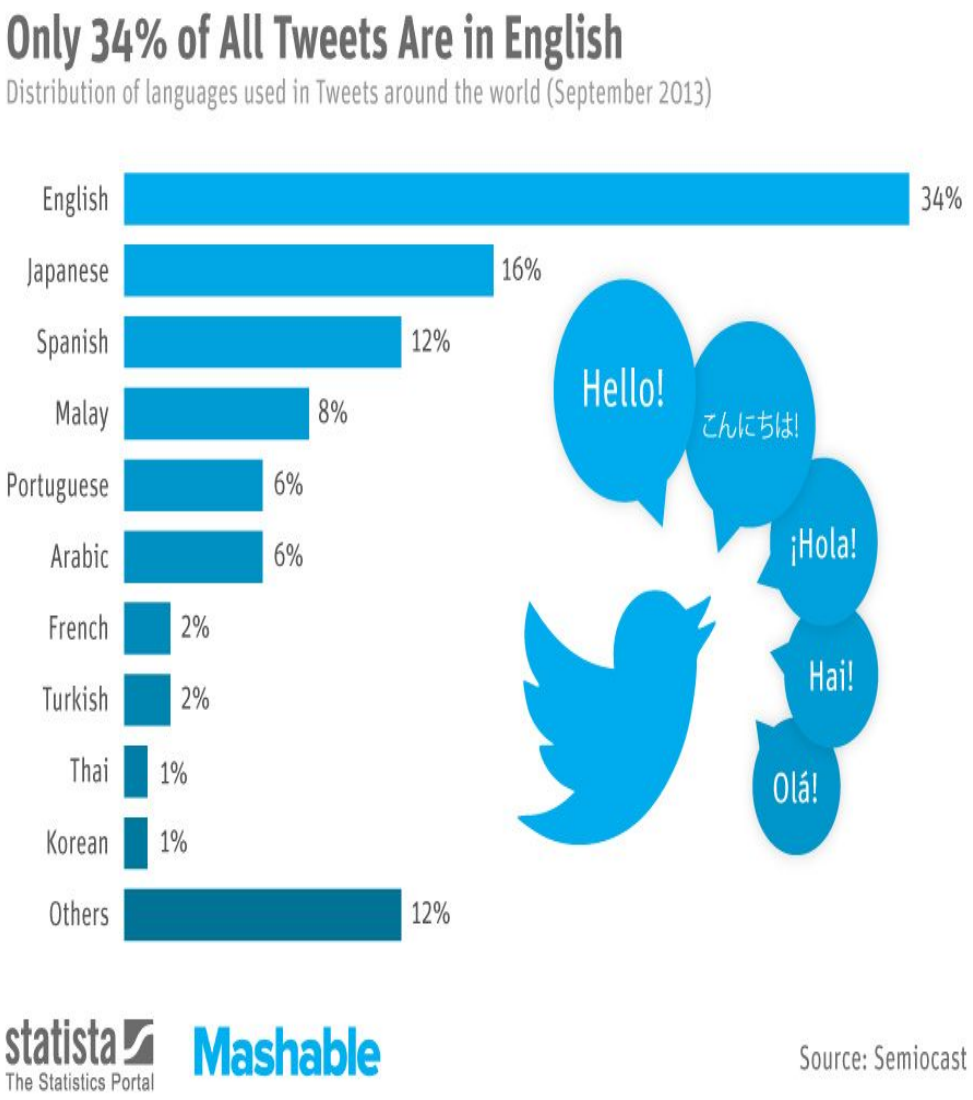


THE Community TECH

Hugging Face

Problem Statement:

- Most sentiment analysis tools are designed for English, despite the multilingual nature of social media.
- It's unclear whether pre-trained multilingual models can accurately detect sentiment in less represented languages.
- **Key questions:**
  - Can these models correctly classify positive, negative or neutral sentiment in English, Spanish, Japanese, and Greek?
  - Do they perform consistently across all languages?



THE Community TECH

Hugging Face

Methods:

- **Dataset:** CardiffNLP Tweet Topic Multilingual dataset with 1,000 tweets per language.
- **Processing:** minimal cleaning to keep the tweet format natural.
- **Models**:
  - Multilingual BERT (NLPTown) - a pre-trained transformer; fine-tuned on multilingual product review sentiment. Supports six languages.
  - Twitter RoBERTa-base (CardiffNLP) - trained on 58 million tweets and fine-tune for English sentiment classification.
- **Tools:** sentiment scores aggregated and visualized.

nlptown / **bert-base-multilingual-uncased-sentiment**

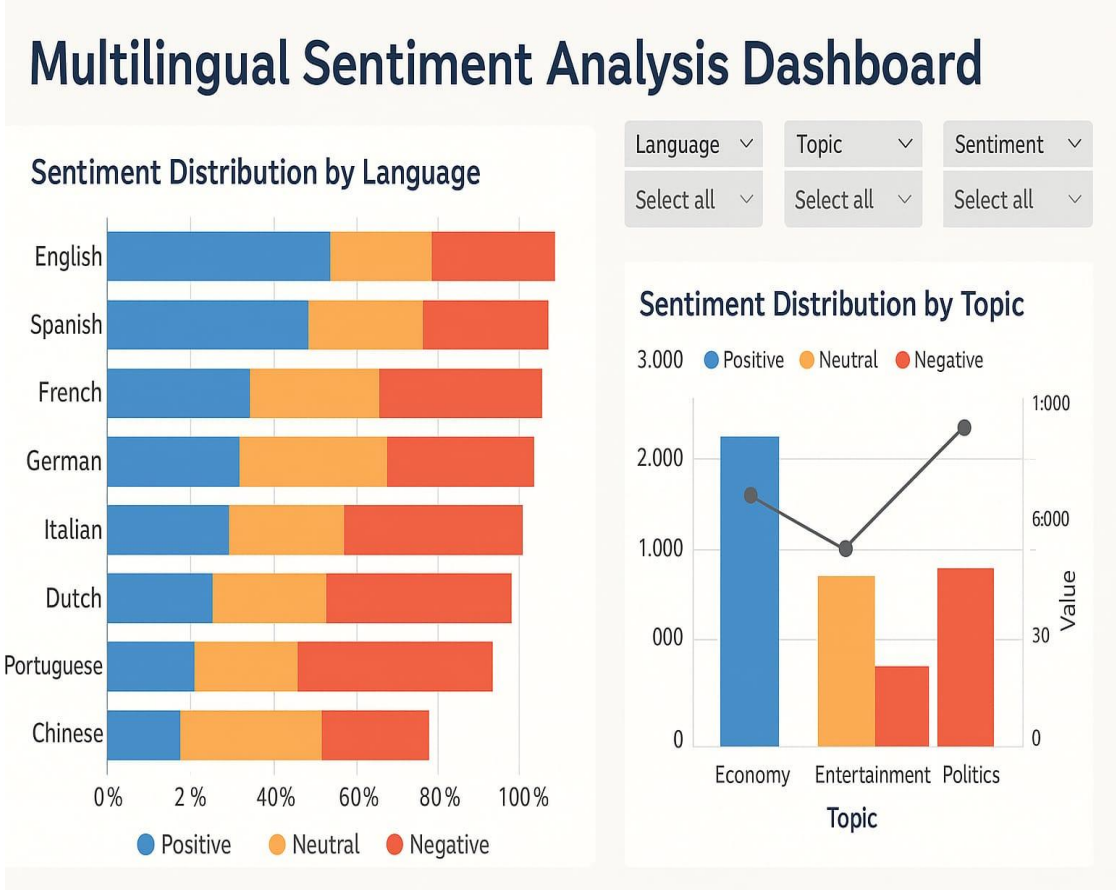
cardiffnlp / **twitter-roberta-base-sentiment**

THE Community TECH

Hugging Face

Dashboard:

- Create an interactive Power BI dashboard to explore sentiment analysis results.
- Visualize sentiment distribution by language and topic.
- Enable users to compare patterns and spot inconsistencies across languages.



THE Community TECH

Hugging Face