

Multilingual Sentiment Analysis of Tweets Using Pre-trained NLP Models



BY: YUVAL CASPI
THE COMMUNITY TECH
DATA SCIENCES, TORONTO, CANADA
EMAIL: YUVALCASPI12@GMAIL.COM

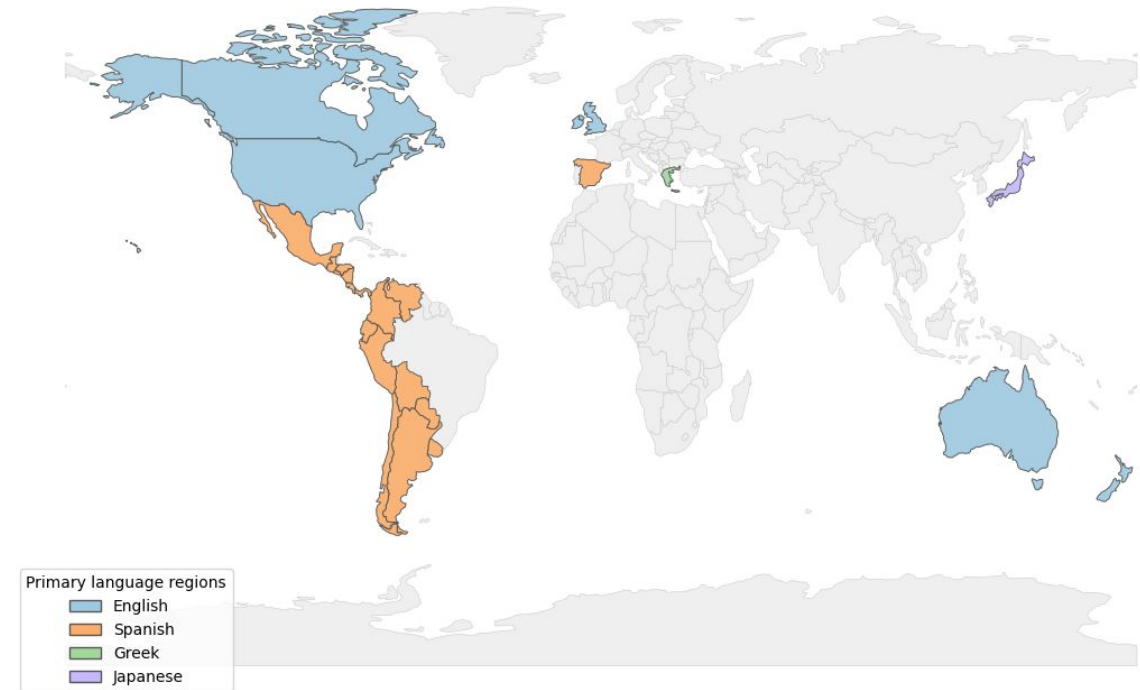
Introduction:

- Growth of online hate speech and toxic discourse highlights the importance of sentiment detection.
- Social media platforms (especially X/Twitter) are central to global communication.
- Sentiment analysis = automatic classification of text into positive, negative, or neutral tone.
- Key challenge: Most existing models are trained only on English, limiting multilingual reliability.

Automated tools can:

- Detect harmful or polarized language patterns
- Support moderation and safer digital spaces
- Provide insights for research and policy decisions

Regions Representing Dataset Languages (EN / ES / EL / JA)

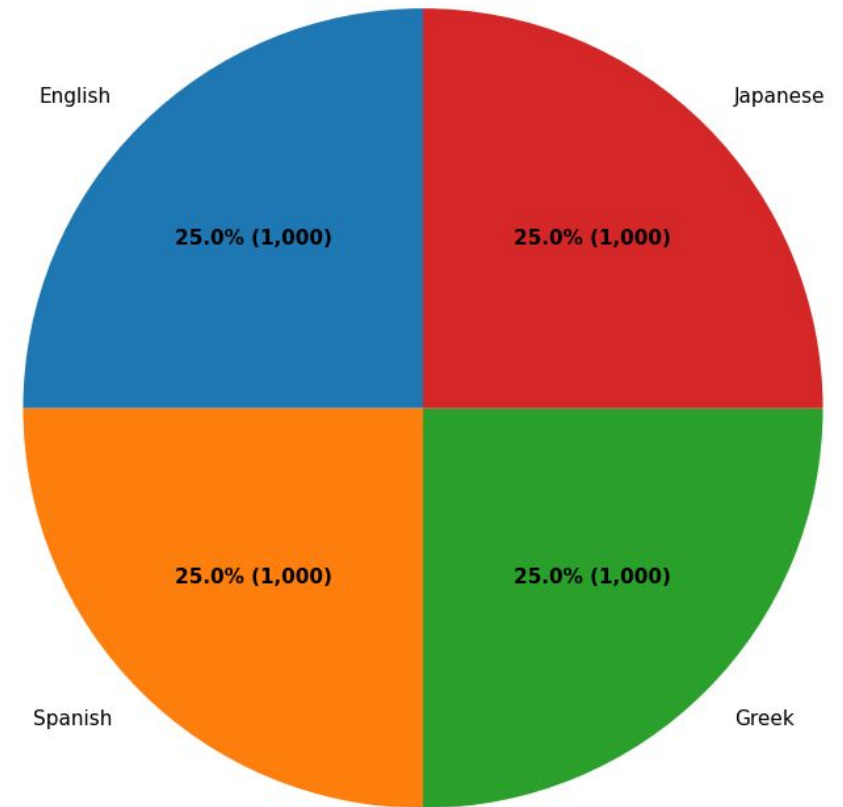


Introduction:

This study explores:

- How reliably multilingual transformer models detect sentiment across English, Spanish, Greek, and Japanese tweets.
- Comparisons between two pre-trained models:
- CardiffNLP Twitter RoBERTa (optimized for tweets)
- NLPTown Multilingual BERT (trained on reviews)
- Cross-language sentiment trends and topic distributions.
- Key challenges: cultural/linguistic bias, domain mismatch, and limited neutral prediction.

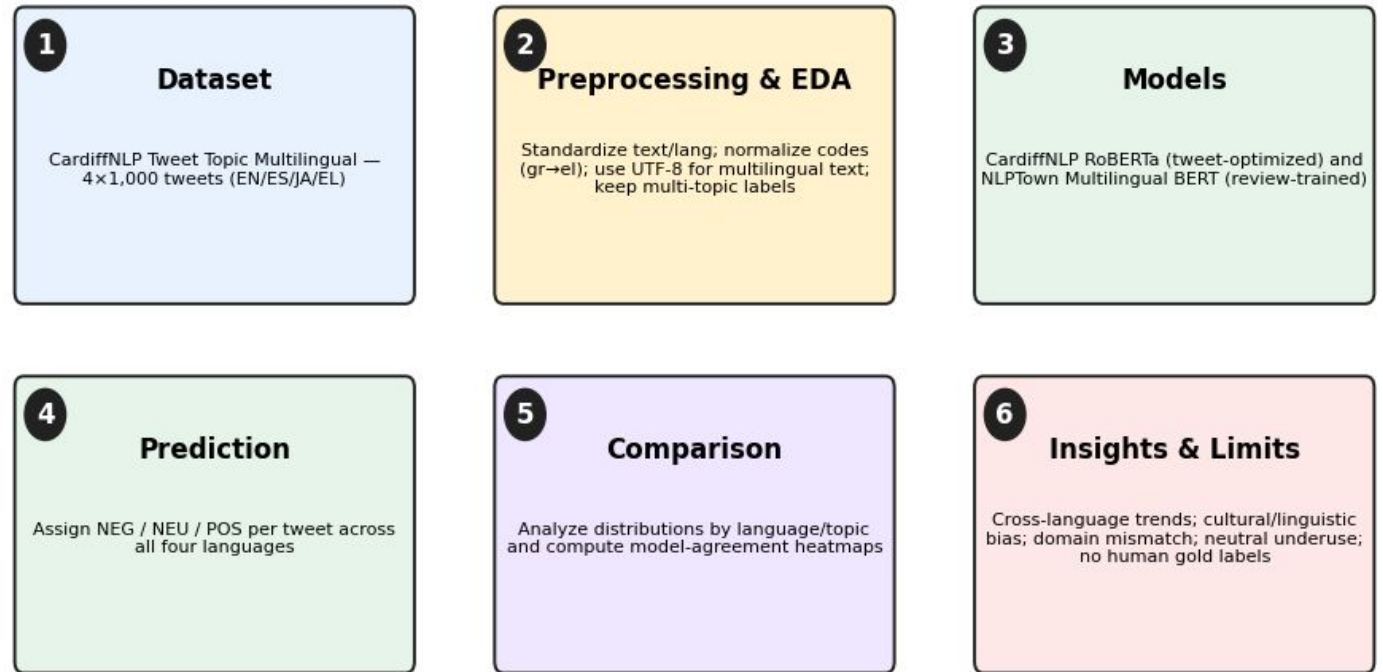
Dataset Languages — 4,000 Tweets Total



Methods:

- **Dataset:** CardiffNLP Tweet Topic Multilingual (EN, ES, JA, EL)
- **Cleaning & Prep:** Standardized text/lang columns, normalized codes (e.g., gr-el), kept multi-topic labels, and stored everything in UTF-8.
- **EDA:** Confirmed balanced languages, charted sentiment counts and topic mixes, built frequency tables to surface dominant themes.
- **Models:** Applied CardiffNLP RoBERTa (tweet-optimized) and NLPTown Multilingual BERT (review-trained) and compared predictions across languages.

Methods Overview

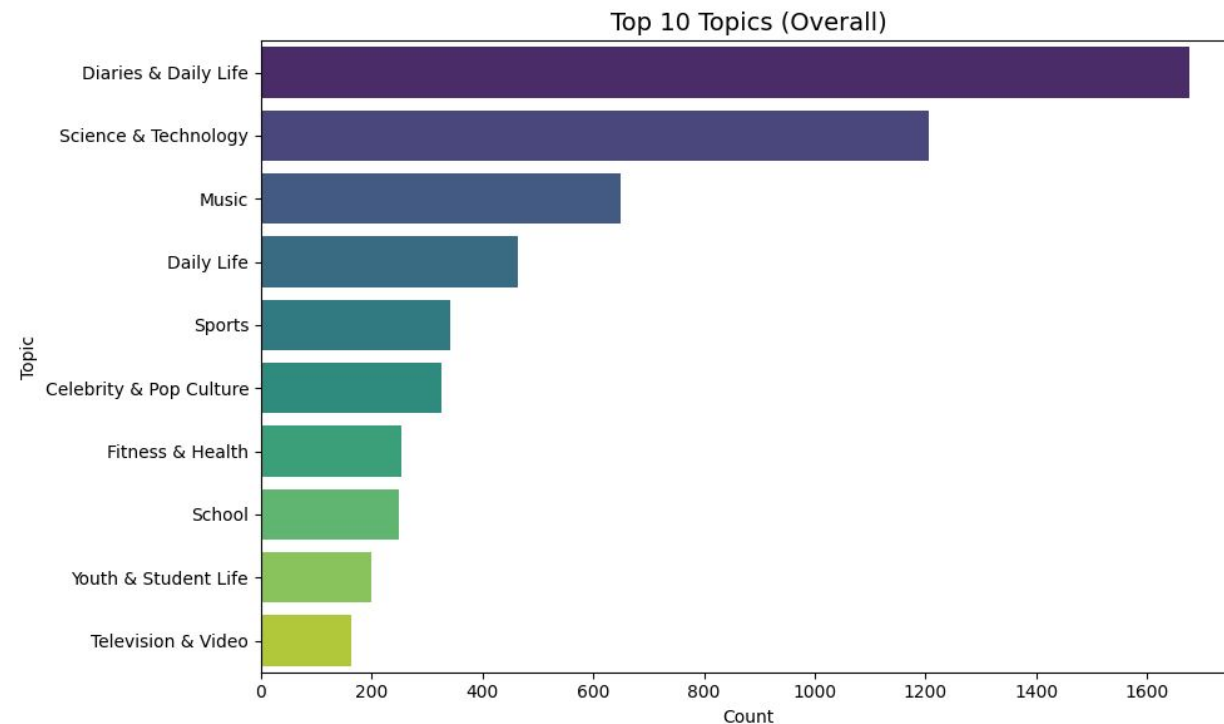


Results (Dataset Overview):

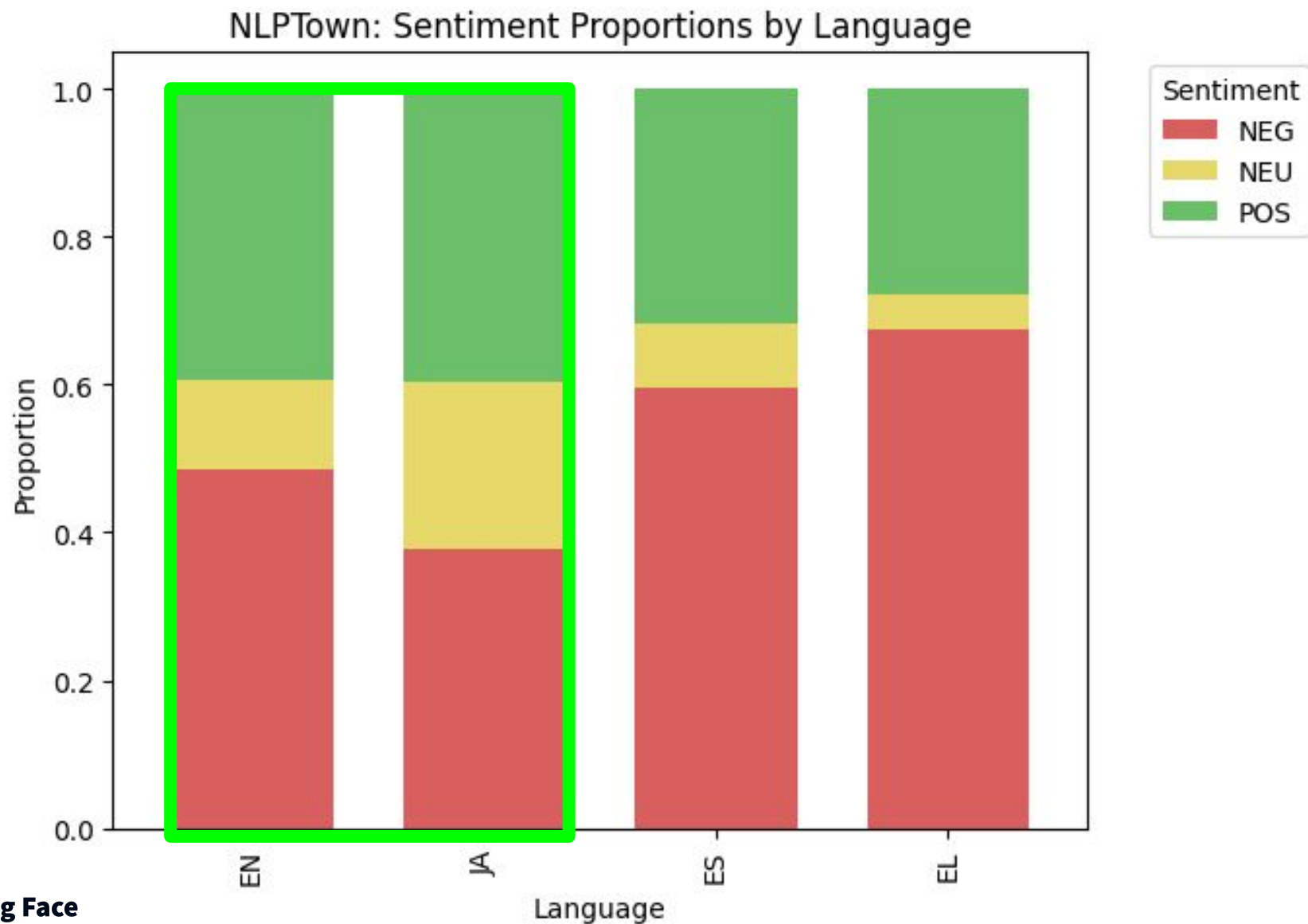
- Balanced dataset: 4,000 tweets total
- 1,000 each: English, Spanish, Japanese, Greek
- Provides baseline for sentiment variation analysis

Common categories:

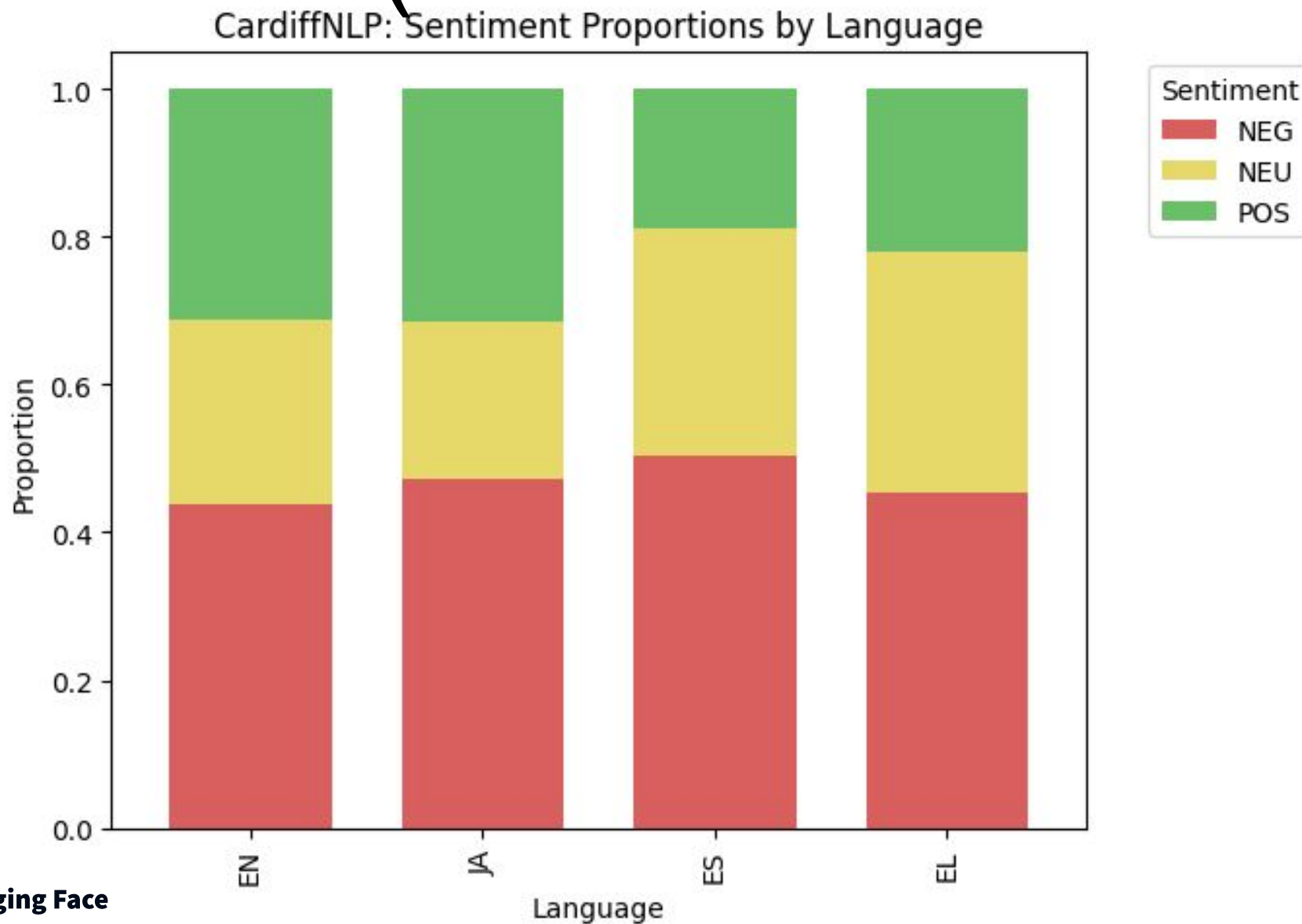
- Diaries & Daily Life
- Science & Technology
- Music



Results (Model Performance):



Results (Model Performance):



Results (Model Performance):



NLPTown Multilingual BERT:

- Higher Positive predictions in EN & JA, but very negative in EL & ES.
- Broader cross-lingual coverage, but domain mismatch with tweets creates inconsistencies.
- Delivers smoother sentiment mixes in English and Japanese, but struggles in Greek and Spanish.

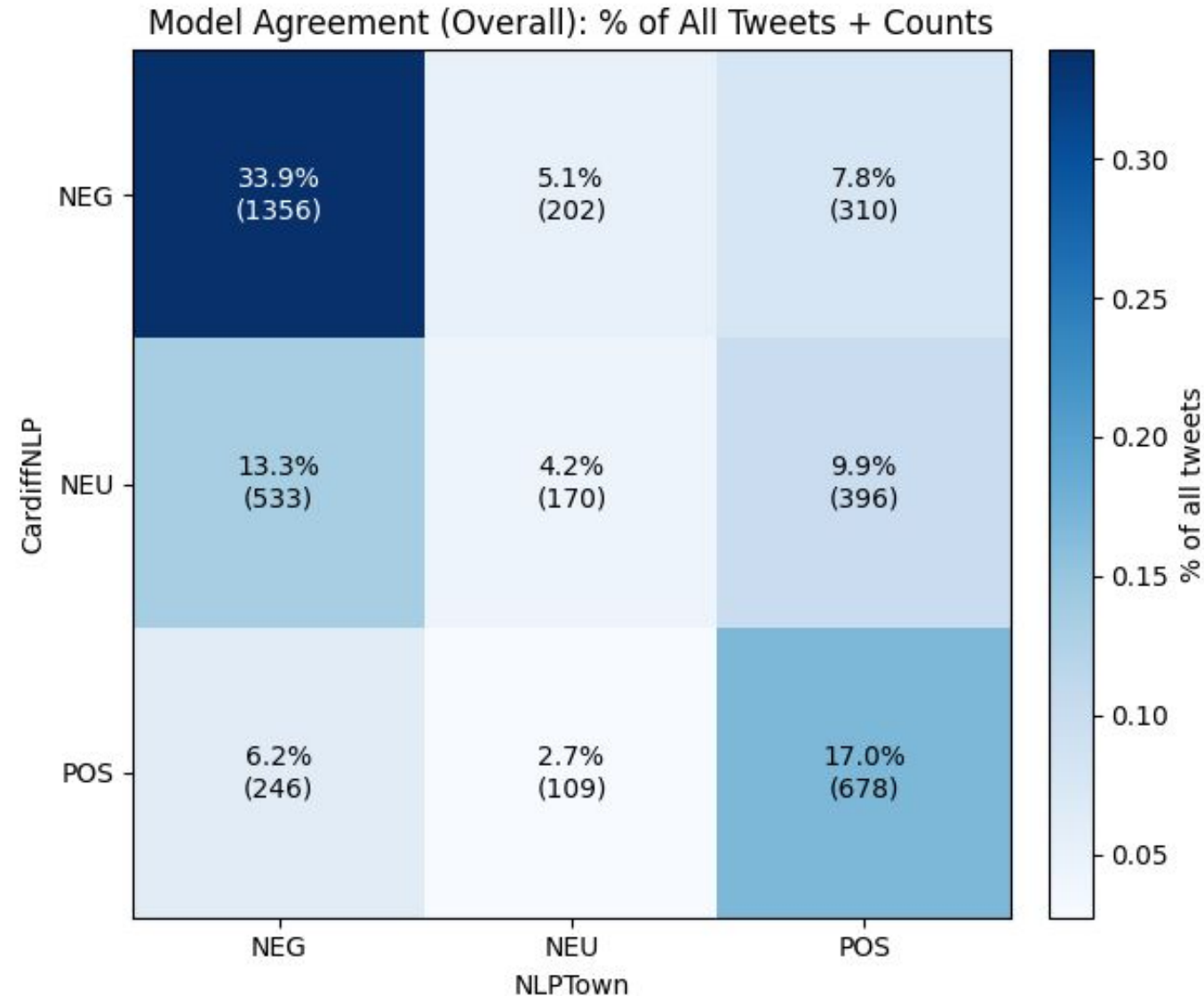
CardiffNLP RoBERTa:

- Consistently negative-leaning, especially in ES, JA, and EL; more balanced in EN.
- Reliable for short, informal content.
- Consistently skews negative across all four languages, particularly Spanish, Greek, and Japanese, but remains more balanced in English.

 [cardiffnlp/twitter-roberta-base-sentiment](#) 

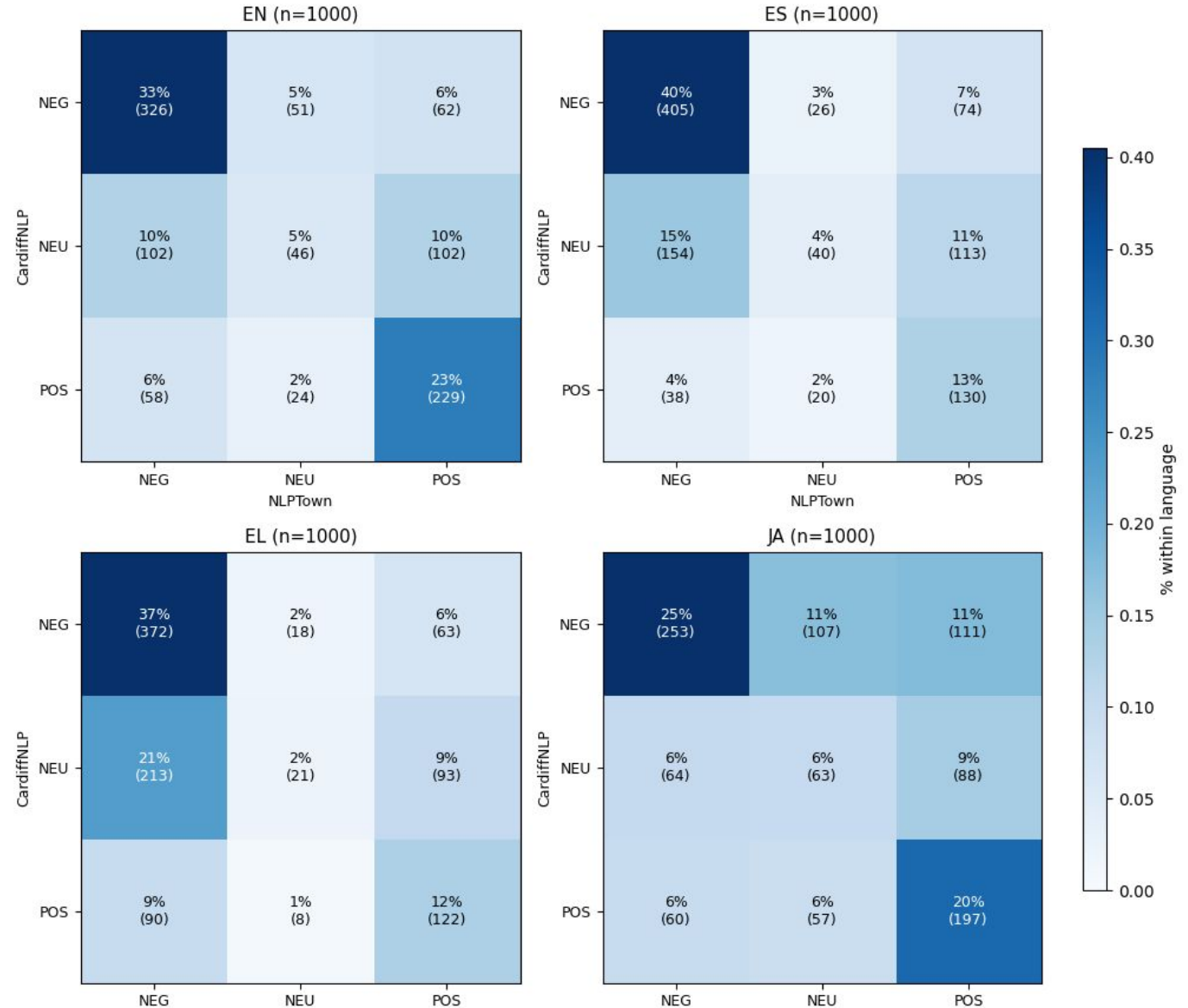
 [nlptown/bert-base-multilingual-uncased-sentiment](#) 

Results (Model Agreement):



Results (Model Agreement):

- Agreement varies widely across languages.
- Greek and Japanese show the largest inconsistencies, reflecting potential cultural and linguistic bias.
- Spanish and English results are more stable, but still skew differently depending on sentiment polarity.



Results (Key Insights):

Category trends:

- **Diaries & Daily Life**

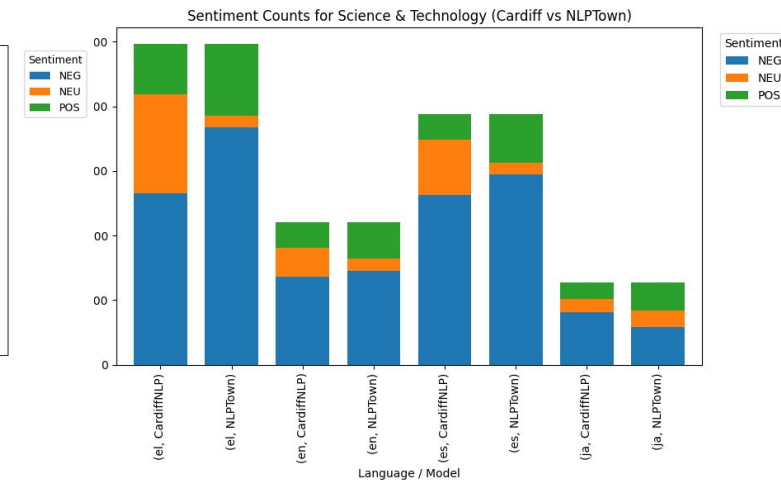
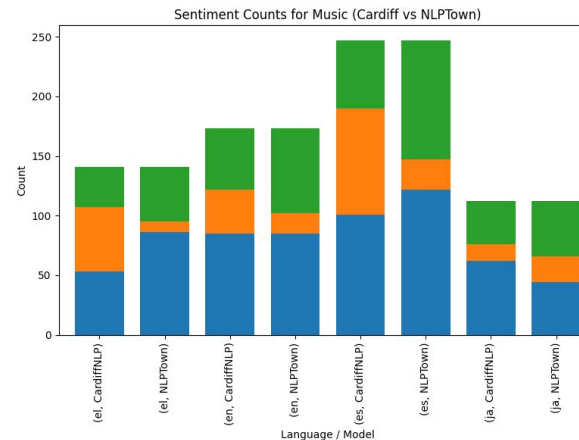
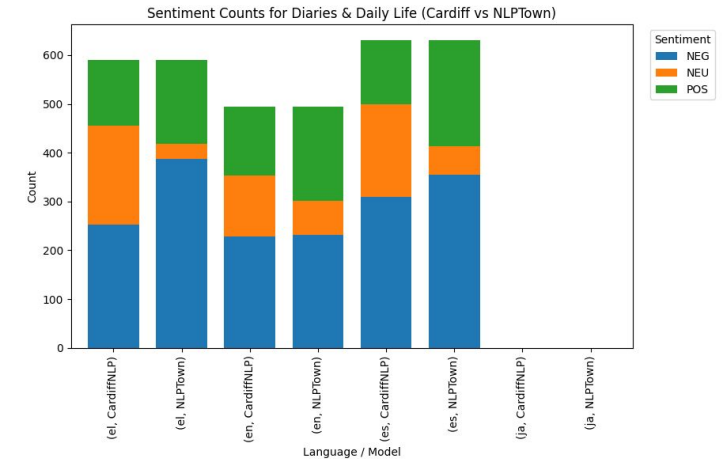
- Less polarizing, showing how everyday discourse tends to balance sentiment across languages.

- **Science & Technology**

- Both models generally agree that Science & Technology is less polarizing, aligning with its focus on factual or optimistic content rather than emotional or critical discourse.

- **Music**

- The divergence here highlights Cardiff's tendency to read cultural/language differences as neutral or negative, while NLP Town reflects its training bias toward positive sentiment in entertainment/review contexts.



Results (Key Insights):

Limitations:

- Since the dataset only provided topics, sentiment predictions cannot be benchmarked against human annotation. Results represent model agreement/disagreement rather than true accuracy.
- CardiffNLP was trained on Twitter data, while NLPTown was trained on product reviews. This mismatch explains divergences in Greek and Japanese.
- Sentiment predictions may reflect biases in the training corpora rather than actual sentiment expressed in the tweets (e.g., Cardiff leaning negative in JA/EL, NLPTown overly negative in EL/ES).
- Across languages, NLPTown tends to underpredict Neutral, compressing tweets into Negative or Positive categories.
- While 4,000 tweets ensures balance by language, some individual topics are still relatively small, limiting the strength of per-topic generalizations.

Disagree

CardiffNLP = NEG (0.93) | NLPTown = POS (0.42)

Agree

CardiffNLP = NEG (0.95) | NLPTown = NEG (0.94)

Ambiguous

CardiffNLP = NEU (0.36) | NLPTown = NEG (0.35)



Name

@user_es_5831

...

Váyanse plp con sus novios que les dan cariñitos cuando se sienten mal, los odio 👍👍👍

Go away, people, with your boyfriends who give you cuddles when you're feeling bad, I hate you 👍👍👍

12:00 PM · Jun 1, 2021



Name

@Username

...

Quarantine? Isolation? Help me God! A public official has been forcing me into isolation for over 35 years in order to cover-up a crime. What did I do wrong? Absolutely nothing. I have never committed a crime. I don't owe any taxes.

12:00 PM · Jun 1, 2021



Name

@user_el_4682

...

Η οικονομία. Έριξε τον (άσχετο) Τσίπρα που φορολόγησε άγρια τη μεσαία τάξη για να συσσωρεύει πλεονάσματα Σηκώνει τον Μητσοτάκη που μειώνει φορολογία, παρά τις παγκόσμιες δυσκολίες κ με τις ιδιωτικοποιήσεις αναπτρώνει την οικονομία δίνοντας αέρα στη μεσαία τάξη. Χαώδης διαφορά..

The economy. It brought down the (incompetent) Tsipras who heavily taxed the middle class in order to accumulate surpluses. It lifts up Mitsotakis who lowers taxes, despite global difficulties, and with privatizations revitalizes the economy, giving breathing space to the middle class. A vast difference..

12:00 PM · Jun 1, 2021



Discussion (findings):

main takeaways:

- Pre-trained models show strong multilingual coverage, but reveal cultural & linguistic biases
- CardiffNLP: excels with Twitter-style texts, NLPTown: better at cross-lingual sentiment
- Sentiment varies by language & topic - model choice is critical

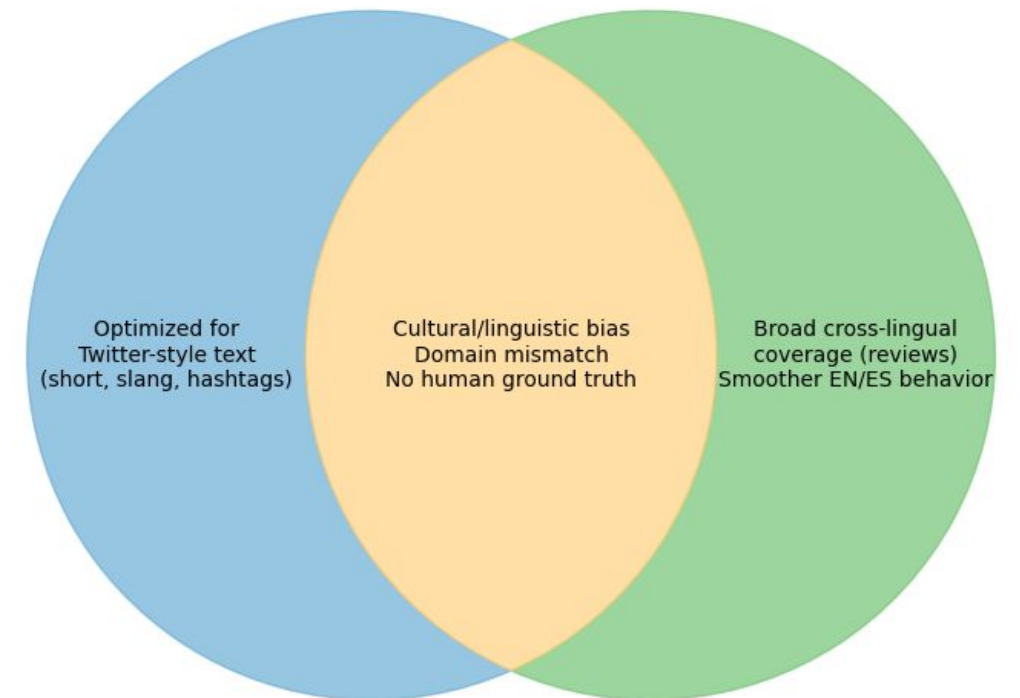
Limitations:

- No ground truth labels
- Domain mismatch
- Cultural and linguistic bias

Future directions:

- Human-labeled benchmarks
- Improved cross-lingual robustness

Discussion: Strengths & Shared Challenges



CardiffNLP (Twitter RoBERTa)

NLPTown (Multilingual BERT)