Multilingual Sentiment Analysis of Tweets Using Pre-trained NLP Models

By: Yuval caspi
Community Tech - Data Sciences, Toronto, Canada
Email: yuvalcaspi12@gmail.com

Background: As online discourse continues to grow across platforms like X (formerly known as Twitter), the need for scalable language-agnostic sentiment analysis becomes increasingly important. While sentiment analysis tools exist for English, there is limited application of such tools to low-resource or less commonly represented languages. Pre-trained transformer models offer an accessible and efficient solution to this gap without the need for training data or complex pipelines. *Objectives:* This project aims to explore sentiment trends in a multilingual tweet dataset using pre-trained models. Specifically, I will apply existing multilingual sentiment classifiers to Twitter data in four languages - English, Spanish, Japanese and Greek, to automatically assign sentiment labels (positive, negative or neutral). The project seeks to:
(1) Evaluate cross-language sentiment trends using descriptive statistics and visualizations
(2) Assess model consistency across languages and topics
(3) Provide insight into cultural or linguistic variation in emotional tone on social media.

*Problem Statement**:*** Can existing multilingual sentiment analysis models accurately understand how people feel in different languages? Most tools are designed for English, but people use many languages on social media. Can pre-trained models correctly detect positive, negative, or neutral tones in tweets written in English, Spanish, Japanese, and Greek, and whether they work equally well across all languages. *Methods:* The CardiffNLP Tweet Topic Multilingual dataset will serve as the primary source, containing 1,000 tweets per language. Pre-processing will include minimal cleaning to preserve the natural format of tweets. Sentiment classification will be performed using pre-trained models available via the Hugging Face transformers library, including the Multilingual BERT Sentiment Classifier by NLPTown and the Twitter RoBERTa-base Sentiment Model developed by the CardiffNLP team. Sentiment outputs will be aggregated and visualized using Python-based tools, such as pandas, matplotlib and seaborn. *Dashboarding:* An interactive Power BI dashboard will be provided for users to view sentiment distribution by language and topic, compare relationships and identify potential inconsistencies across languages.

Links:

1. Dataset: LINK
2. Multilingual BERT Sentiment Classifier (NLPTown): LINK
3. Twitter RoBERTa-base Sentiment Model (CardiffNLP): LINK
4. Project: LINK
5. PPT: LINK