Spring 5-1-2024

Computational Linguistics and Multilingualism: A Comparative Analysis with Spanish and English Data

Evelyn Lawrie
Chapman University, lawrie@chapman.edu

Computational Linguistics and Multilingualism: A Comparative Analysis with Spanish and English Data

Evelyn Lawrie

Department of World Languages and Cultures

Introduction

- Computational linguistics is the field of linguistics in which techniques of computer science are applied to the analysis and synthesis of language
- There is a significant disparity between the accuracy and prevalence of models for English and Spanish language analysis
- The lack of Hispanic perspectives in technology is exacerbated by the scarcity of resources for analyses in Spanish and Latin American indigenous languages
- Sentiment analysis is the process of analyzing text to obtain the tone/polarity of the message (positive, negative, or neutral)
- Conducting computational linguistic analyses on datasets containing multiple languages is shown to increase the model's accuracy

Methods

Data Preprocessing

- URLs, mentions, and hashtags were removed from all tweets
- The tweets were tokenized to separate every word of each tweet for the model to more easily analyze its polarity

Spanish Sentiment Analysis

- Sentiment analysis was performed on a dataset containing 30,000 tweets in Spanish from the social media platform Twitter
- The pysentimiento model (a pre-trained language model for Spanish content) was utilized for analysis of tweet polarity (Figure 1)

English Sentiment Analysis

- Sentiment analysis was implemented on a dataset of 1,600,000 English tweets
- A Support Vector Machines (SVM) classification algorithm was used to predict the tweets' polarities (Figure 2)

Multilingual Sentiment Analysis

- Datasets were combined to analyze the impact of multilingual data on sentiment analysis accuracy
- The Spanish tweets dataset was added to the English tweets' testing set to increase variation in test data

Tweet Text	Sentiment Score
Lo mejor del discurso de es cuando le sacude la mano a parece que la estuviera regañando jajaja	Output = POS Probas = {POS: 0.532, NEU: 0.351, NEG: 0.117})

Figure 1. Example Output of Spanish Tweet Polarity Analysis

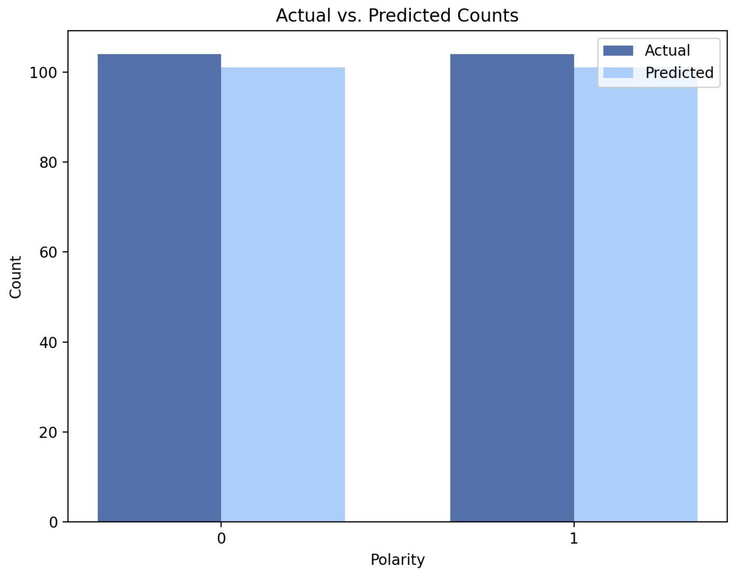

Figure 2. Actual and Predicted English Tweet Polarity (Negative and Positive)

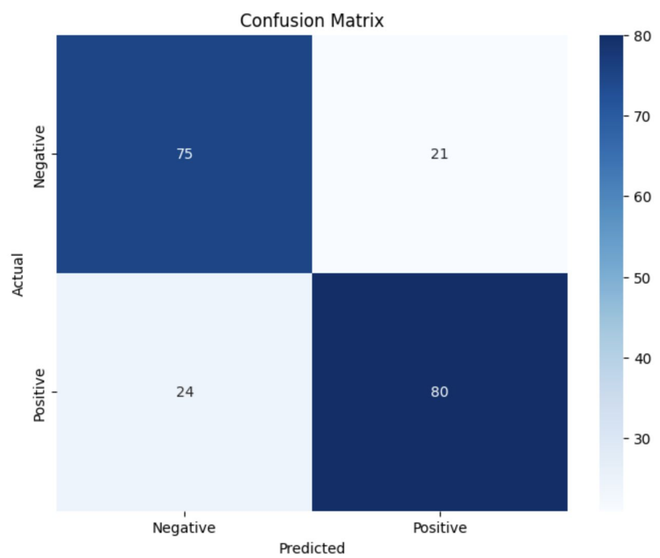

Figure 3. Visual Representation of English Model's Correct and Incorrect Predictions with Confusion Matrix

Results

- Sentiment analysis toolkits are more thoroughly developed and accurate for English than Spanish
- The robustness of English dataset compared to Spanish contributes to the increased accuracy of the English model (Figure 4)
- Given that the English dataset contained polarity labels, it was possible to perform supervised machine learning
  - Supervised methodologies allow for improvement of accuracy measurement techniques (Figure 3)

Sentiment Analysis Model Type	Accuracy Score
Spanish (pysentimiento)	0.70
English (SVM)	0.83

Figure 4. Accuracy of Spanish and English Models

Multilingual Sentiment Analysis

- Given that Spanish tweets are unlabeled, they could not be used in the training dataset of the English model
- Their incorporation into the testing dataset decreased the overall prediction accuracy

Acknowledgements

Thank you to Dr. Loustau for advising this project, and to Dr. Rivera for providing insights on the shortcomings of resources for indigenous language translation through a personal interview.

References