Analyzing Emotional Intensity of Tweets: A Statistical Machine Learning Approach

By
Mohammed Farooq basha S
Codalab Username : Farooqasha008

Abstract:

This technical paper presents a comprehensive analysis of emotional intensity in tweets using a combination of text processing, statistical modeling, and machine learning techniques. The study aims to provide insights into the intensity of emotions expressed in social media conversations. The research includes data preprocessing, model comparison using PyCaret, and the development of a Ridge regression model tuned for optimal performance. The processed dataset is exported and used for training and evaluation. The findings are based on evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R2), Root Mean Squared Logarithmic Error (RMSLE), and Mean Absolute Percentage Error (MAPE).

1. Introduction

1.1 Background

The widespread use of social media platforms has provided an unprecedented opportunity to analyze the emotions and sentiments expressed by users in real-time. Twitter, as one of the largest social media platforms, serves as a rich source of user-generated content that can offer valuable insights into various aspects of human emotions.

1.2 Objective

The objective of this research is to analyze the emotional intensity of tweets and develop a statistical machine learning model to predict and quantify emotional intensity accurately. By understanding the emotional dynamics of tweets, we can gain insights into the emotional responses of users and their sentiment towards different topics.

1.3 Contribution

This study contributes by providing a comprehensive approach to analyze emotional intensity in tweets. It combines data preprocessing techniques, model comparison using PyCaret, and the development of a Ridge regression model to accurately predict emotional intensity. The research findings offer valuable insights into the intensity of emotions expressed in tweets, facilitating a better understanding of user sentiments in social media conversations.

2. Data Collection and Preprocessing

2.1 Data Source

The data used in this study is collected from Twitter, specifically focusing on tweets related to a specific topic or event. The dataset consists of text content, emotion labels, and corresponding scores representing emotional intensity.

2.2 Text-to-CSV Conversion

To facilitate further analysis and modeling, the raw text data is converted into a structured format. The provided code snippet demonstrates the process of converting the text data into a CSV file, where each row represents a tweet with associated emotion labels and scores.

2.3 Data Preprocessing using PyCaret

The preprocessing step involves transforming the raw text data into a suitable format for modeling. PyCaret, a powerful Python library, is utilized for automating the preprocessing tasks, including text cleaning, tokenization, feature engineering, and data transformation. The preprocessed dataset is then exported for subsequent modeling steps.

3. Model Comparison using PyCaret

3.1 PyCaret Overview

PyCaret provides a streamlined workflow for model comparison and selection. It offers a range of pre-processing techniques, model training, hyperparameter tuning, and performance evaluation measures.

3.2 Model Comparison Workflow

In this research, PyCaret is employed to compare various machine learning models. The code snippet demonstrates the use of PyCaret's `compare_models()` function to evaluate multiple models using default configurations and performance metrics.

3.3 Selection of Best Model

Based on the model comparison results, the best-performing model is selected for further analysis and evaluation. PyCaret helps in identifying the model with the highest performance based on the chosen evaluation metric.

4. Statistical Machine Learning with Ridge Regression

4.1 Data Preparation

The preprocessed dataset is split into training and testing sets, and the necessary input features and target variables are extracted. The code snippet showcases the process of preparing the data for model training and evaluation.

4.2 Model Development

Ridge regression, a popular linear regression technique, is employed to build a statistical machine learning model. The code snippet illustrates the development of the Ridge regression model using the training dataset.

4.3 RidgeCV for Alpha Selection

RidgeCV, a variant of Ridge regression, is utilized to automatically select the optimal value of the regularization parameter (alpha). The code snippet demonstrates the usage of RidgeCV with cross-validation to determine the best alpha value.

4.4 Performance Evaluation Metrics

Various evaluation metrics are employed to assess the performance of the Ridge regression model. The research utilizes metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R2), Root Mean Squared Logarithmic Error (RMSLE), and Mean Absolute Percentage Error (MAPE).

4.5 Results and Discussion

The evaluation metrics provide insights into the predictive performance of the Ridge regression model in quantifying emotional intensity in tweets. The results are discussed, highlighting the model's accuracy and its ability to capture the intensity of emotions expressed in social media conversations.

The evaluation metric scores are as follows:

.

1. Mean Absolute Error (MAE): 0.0510570074666143
2. Mean Squared Error (MSE): 0.011451054954837744
3. Root Mean Squared Error (RMSE): 0.10700960216185156
4. R-squared (R2): 0.7095968416530064
5. Root Mean Squared Logarithmic Error (RMSLE): 0.07229241877765256
6. Mean Absolute Percentage Error (MAPE): inf

The R-squared score is 0.70, and the rest of the scores are also in the better ranges depicting that the model’s performance is good.

5. Conclusion

5.1 Summary of Findings

This research successfully analyzes the emotional intensity of tweets using a statistical machine learning approach. The developed Ridge regression model demonstrates promising performance in predicting emotional intensity. The study provides valuable insights into user sentiments and emotional dynamics in social media conversations.

5.2 Future Work

The present study sets the foundation for further research in the field of sentiment analysis and emotion detection in tweets. Future work could focus on exploring advanced deep learning models, incorporating contextual information, and considering additional linguistic features for enhanced emotion intensity prediction.

References:

- Using Hashtags to Capture Fine Emotion Categories from Tweets. Saif M. Mohammad, Svetlana Kiritchenko, Computational Intelligence, Volume 31, Issue 2, Pages 301-326, May 2015.
- Crowdsourcing a Word-Emotion Association Lexicon, Saif Mohammad and Peter Turney, Computational Intelligence, 29 (3), 436-465, 2013.
- Ekman, P. (1992). An argument for basic emotions. Cognition and Emotion, 6 (3), 169-200.
- #Emotional Tweets, Saif Mohammad, In Proceedings of the First Joint Conference on Lexical and Computational Semantics (*Sem), June 2012, Montreal, Canada.
- Portable Features for Classifying Emotional Text, Saif Mohammad, In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 2012, Montreal, Canada.
- Strapparava, C., & Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In Proceedings of SemEval-2007, pp. 70-74, Prague, Czech Republic.
- From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales, Saif Mohammad, In Proceedings of the ACL 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), June 2011, Portland, OR.
- Plutchik, R. (1980). A general psychoevolutionary theory of emotion. Emotion: Theory, research, and experience, 1(3), 3-33.
- Stance and Sentiment in Tweets. Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media, In Press.