Master 2 Quantitative Finance and Financial Engineering – 272

NLP Project - Sentiment Analysis Report of FIFA World Cup 2022 Tweets

Auteurs :
Cyprien Tardivel
Flavie Bickert
Nassim Chamakh

Professeur :
Yair Lakretz

Sentiment Analysis Report of FIFA World Cup 2022 Tweets

Introduction

Subject

The objective of this project is to analyze the sentiments expressed in tweets related to the FIFA World Cup 2022. Sentiment analysis is a crucial aspect of Natural Language Processing (NLP) that aims to determine the attitude of speakers or writers towards a specific subject. In this context, we utilized a Recurrent Neural Network (LSTM) model to classify tweets into three sentiment categories: negative, neutral, and positive.

Dataset

The dataset used for this analysis is called "Tirendaz/fifa-world-cup-2022-tweets", from the website Hugging Face. This dataset contains tweets related to the FIFA World Cup 2022, along with labels indicating the sentiment associated with each tweet.

Methodology

We followed these steps to conduct our analysis:

1. **Data Preprocessing**
2. **LSTM Model Construction**
3. **Model Training**
4. **Model Evaluation**
5. **Comparison with a Logistic Regression Model**

Data Preprocessing

Loading and Creating the DataFrame

We started by loading the dataset from Hugging Face and creating a DataFrame for easier data manipulation. The data includes tweets along with their associated sentiments.

Tweet Preprocessing

Tweet preprocessing involved several steps:

- Replacing newlines and carriage returns with spaces.
- Removing URLs, mentions, and hashtags.
- Removing special characters and keeping only alphanumeric characters and spaces.
- Converting to lowercase and removing stopwords.

Tokenization and Padding

Tokenization transforms tweets into sequences of tokens, while padding standardizes sequences to a fixed length for processing by the LSTM model.

Label Conversion

Sentiment labels were converted into numerical values for use in the machine learning model. The labels "negative", "neutral", and "positive" were mapped to 0, 1, and 2 respectively.

LSTM Model Construction

We defined a class to encapsulate the construction of the LSTM model. The model includes embedding layers, bidirectional LSTM layers, and dense layers with softmax activation. These choices enable capturing contextual dependencies in tweets and effectively classifying sentiments.

Training and Evaluation

Data Preparation

The data was split into training and test sets to evaluate the model's performance on unseen data, crucial for assessing model generalization.

Model Parameters

We selected the following parameters for training the LSTM model:

- Adjusted vocabulary size: 10,000
- Embedding dimension: 50
- Maximum sequence length: 60
- Number of neurons in LSTM hidden layer: 32
- Number of classes: 3 (negative, neutral, positive)
- Number of training epochs: 4
- Batch size: 16

LSTM Model Training

The model was trained on the training data, and its performance was evaluated on the validation data at each epoch to monitor overfitting.

L STM Model Evaluation

The model was evaluated on the test set to obtain final performance metrics. Results include Test Loss and Test Accuracy, along with a detailed classification report indicating precision, recall, and F1-score for each sentiment class.

LSTM Model Results

- Test Loss: 70.56%
- Test Accuracy: 70.72%

Classification Report for LSTM Model:

- Class "negative":
  - Precision: 0.73
  - Recall: 0.70
  - F1-score: 0.71

- Class "neutral":
  - Precision: 0.65
  - Recall: 0.68
  - F1-score: 0.67

- Class "positive":
  - Precision: 0.75
  - Recall: 0.74
  - F1-score: 0.74

Overall accuracy is 71%, indicating the model's good ability to generalize to unseen data.

Comparison with Logistic Regression Model

To compare, we also trained a logistic regression model on the same data. However, this model's performance was significantly lower compared to the LSTM model.

Logistic Regression Model Results

- Test Accuracy: 38.49%

Classification Report for Logistic Regression Model:

- Class "negative":
  - Precision: 0.24
  - Recall: 0.03
  - F1-score: 0.05
- Class "neutral":
  - Precision: 0.37

    - Recall: 0.28
    - F1-score: 0.32
- Class "positive":
    - Precision: 0.40
    - Recall: 0.73
    - F1-score: 0.51

The L-BFGS optimizer used for logistic regression did not converge, suggesting either insufficient iterations or the need for data normalization.

Interpretation of Results

LSTM Model

The LSTM model showed strong performance with a test accuracy of 70.72%. The "positive" class achieved the highest precision, while the "neutral" class had the highest recall. However, slight overfitting was observed during training.

Logistic Regression Model

The logistic regression model performed significantly worse, with a test accuracy of 38.49%. Results indicate poor performance, especially for the "negative" class, with a recall of only 3%.

Conclusion

The LSTM model demonstrated strong overall performance with good precision and recall across sentiment classes. Despite slight overfitting observed during training, the model generalizes effectively to test data, achieving a test accuracy of 70.72%. The "positive" sentiment class particularly stood out with high precision, while the "neutral" class showed robust recall.

In contrast, the logistic regression model exhibited significantly lower performance compared to the LSTM model, with a test accuracy of only 38.49%. The model struggled notably with the "negative" sentiment class, highlighting challenges in capturing nuanced sentiment expressions from tweets. Convergence issues encountered with the L-BFGS optimizer suggest potential benefits from data normalization or increased iteration count during training.

The stark performance contrast between the LSTM and logistic regression models underscores the advantage of leveraging deep learning techniques, like LSTM, for sentiment analysis tasks. The LSTM's ability to capture contextual dependencies within tweets contributes significantly to its superior performance compared to more traditional machine learning approaches. An interesting extension could to explore methods that combine predictions from multiple models, including LSTM and logistic regression, to potentially achieve higher accuracy and robustness in sentiment analysis.