ARI5501 NLP Midterm Project: Sentiment Analysis

Project Overview

The goal of this project is to apply sentiment analysis techniques to determine the sentiment (positive, negative, neutral) of various text datasets. Students will train a basic model using a provided dataset, test the model on a separate English dataset, and optionally test the model using a translated Turkish dataset to evaluate the model's performance on multilingual data.

Project Objectives

1. Train a Sentiment Analysis Model:

- Obtain a dataset from Hugging Face or Kaggle or use the provided dataset.
- Train a sentiment analysis model using this dataset.

2. Evaluate Model Performance on English Data:

- Test the trained model on a separate English dataset to measure accuracy.

3. Bonus: Multilingual Sentiment Analysis:

- Translate Turkish product comments into English.
- Use the translated data to test the model trained on English data.

Dataset

English Sentiment Datasets

1. IMDB Reviews:

- **Description**: A widely-used dataset for binary sentiment classification.
- **Use with Transformers**: The dataset can be fine-tuned with transformer models such as BERT, RoBERTa, and DistilBERT.
- **Access**: Hugging Face IMDB Dataset

2. Sentiment140:

- **Description:** Contains 1.6 million tweets with sentiment labels.

- **Use with Transformers:** Suitable for fine-tuning models like BERT and RoBERTa.

- **Access:** Kaggle Sentiment140

Turkish Sentiment Dataset

1. Turkish Product Reviews:

- Description: Contains sentiment-labelled product reviews in Turkish.

- Use with Transformers: For multilingual sentiment analysis, use a model like mBERT (multilingual BERT) or translate the reviews to English and use an English transformer model.

- Access: Hugging Face Turkish Sentiment Analysis Dataset

Project Tasks

Task 1: Train a Sentiment Analysis Model

1. Data Preprocessing:

- Clean and preprocess the dataset (e.g., remove stopwords, tokenize).
- Split the data into training and validation sets.

2. Model Training:

- Choose a pre-trained model (e.g., BERT, RoBERTa, or DistilBERT).
- Fine-tune the model using the training set.
- Evaluate the model on the validation set and continue the fine-tune if necessary.

Task 2: Evaluate Model Performance on English Data

1. Testing Data:

- Obtain Sentiment140 dataset for testing.
- Ensure the testing data is not used during training.

2. Model Evaluation:

- Use the trained model to predict the sentiment of the testing dataset.
- Measure the accuracy, precision, recall, and F1-score of the model.
- Provide a detailed analysis of the model's performance.

Bonus Task: Multilingual Sentiment Analysis

1. Data Translation:

- Translate Turkish product comments into English using a translation tool or API (e.g., Google Translate API, DeepL).

2. Model Testing:

- Use the translated English data to test the trained English sentiment analysis model.
- Evaluate the model's performance on the translated data.
- Compare the performance with the English test dataset and discuss the results.

Submission Guidelines

- Jupyter Notebook:

- Submit a well-documented Jupyter notebook containing all code used for training and testing the model.

- The notebook should include:

  - Description of the datasets used.

  - Explanation of the preprocessing steps.

  - Details of the model training process.

  - Evaluation metrics and performance analysis.

  - Discussion of the multilingual sentiment analysis results (if completed).

  - Conclusion and possible improvements.

- Ensure that the notebook is executable and includes all necessary code to reproduce the results.

- Deadline:

- Submit the project by 8 June 2024

Evaluation Criteria

- Correctness and completeness of the code: 55%

- Quality of the analysis and documentation in the notebook: 20%

- Bonus task completion and analysis: 25%