Arabic Offensive Language Detection

A Machine Learning Approach for Social Media Text Classification

Team member:

Ahella Mohamed 21100824

Hanen Ehab 21100792

Ali Yasser 21100801

Table of Contents

1. Abstract

This report details the development of an Arabic offensive language detection system using BERT-based Transformers. The project involves data preprocessing, model training, and evaluation using the OSACT2022 dataset. Key findings include the model's ability to effectively classify social media text into offensive and non-offensive categories.

2. Introduction

- Background Information

Offensive language detection is essential to maintaining a healthy online environment. This project addresses this need by developing a BERT-based model to classify Arabic social media text.

- Purpose of the Report

This report aims to document the development process, methodologies, and results of building a machine learning model for offensive language detection.

- Scope and Objectives

The report covers data acquisition, preprocessing, model development, evaluation, and future work recommendations.

- Overview of the Document

The document provides a comprehensive overview of the project's lifecycle, from data collection to model deployment, and evaluates the model's performance on real-world data.

3. Project Overview

Project Description:

The project involves building an NLP model using BERT to detect offensive language in Arabic social media text.

Goals and Objectives:

- Develop a robust machine learning pipeline for text classification.
- Train and evaluate the model on Arabic social media datasets.
- Deploy the model for practical use in monitoring social media content.

Technologies and Tools Used:

- Programming Languages: Python
- Libraries: PyTorch, Transformers, Pandas, NLTK
- Frameworks: Hugging Face Transformers, Gradio
- Hardware: GPU for model training and evaluation

4. Data Acquisition

Data Sources:

The OSACT2022 dataset was the primary source, containing labeled Arabic text data for offensive language detection.

Data Acquisition Process:

The dataset was downloaded from online repositories and loaded into the Jupyter Notebook environment for processing.

Challenges Faced:

- Handling Arabic text, including normalization and stopword removal, presented significant preprocessing challenges.
- Ensuring data was balanced and representative for training.

5. Data Analysis

Data Analysis Results:

Extensive text cleaning and preprocessing steps were applied to the dataset, resulting in a well-structured input for the model.

Insights and Patterns:

- The dataset showed an imbalance between offensive and non-offensive content.
- Common offensive terms were identified, aiding in feature engineering.

6. Model Development

Model Details

The project focuses on developing a BERT-based model for detecting offensive language in Arabic social media text. The model is implemented using the "qarib/bert-base-qarib" pre-trained model from the Hugging Face Transformers library.

Development Approach

To create a user-friendly interface for the model, Gradio, a Python library for creating web-based interfaces for machine learning models, was used. The interface allows users to input text and receive predictions on whether the content is offensive.

This includes:

- Installing Gradio and Importing Necessary Libraries
- **Loading the Model and Tokenizer**: The "qarib/bert-base-qarib" model and tokenizer were loaded to handle the classification task.
- Preprocessing and Prediction Functions
- **File Download Functionality**: A function to save the classification results to a temporary file, allowing users to download the results as a text file.
- **Gradio Interface Setup:** The Gradio interface was designed with a simple and clean theme, providing a text box for input, buttons for

classifying the text and downloading the results, and an output area to display the classification results.

- **Tools and Libraries Used:**
  Gradio: For creating the user interface.
  PyTorch: For loading and using the BERT model.
  Transformers: For accessing the pre-trained BERT model and tokenizer.
  Tempfile, os: For handling temporary files and managing file downloads.

This integration of Gradio into the model development process makes the offensive language detection model accessible to users through a simple web interface, allowing for easy interaction and practical application

7. Results and Evaluation

Evaluation Metrics:

- Accuracy
- Precision
- Recall
- F1-Score

Model Performance:

The BERT-based model achieved a high accuracy rate, demonstrating its effectiveness in classifying offensive language in Arabic text.

1269

	precision	recall	f1-score	support
0	0.91	0.88	0.90	865
1	0.76	0.81	0.78	404
accuracy			0.86	1269
macro avg	0.84	0.85	0.84	1269
weighted avg	0.86	0.86	0.86	1269

8. Conclusion

Key Findings:

The BERT-based model successfully classified Arabic social media text with high accuracy, confirming the utility of Transformers in NLP tasks.

Contributions of the Project:

This project contributes to the development of tools for monitoring and moderating offensive content in Arabic social media.

Limitations and Future Work:

- Further hyperparameter tuning could improve model performance.
- Future work could involve exploring different Transformer architectures and deploying the model in real-time environments.

9. References

- training data
- development data
- Test data
- subtask A OFF (or NOT_OFF)

10.Output: