LAU

الجامعة اللبْنانيّة الأميركيّة

Lebanese American University

Detection of Cyberbullying in Arabic Tweets

AAI 612: Deep Learning and its Applications

Project Report

Jad Geitani

March 15, 2025

Contents

1 Introduction

Since the introduction of social media, cyberbullying has become a serious issue. The lack of effective tools that can easily detect harmful content in Arabic texts creates a gap, and this project tries to fill this gap. This project will benefit social media platforms (like Facebook, X, Instagram…) and policymakers by providing a tool that will automatically detect cyberbullying in Arabic comments or tweets.

2 Problem Statement

The objective is to train and compare different deep learning architectures for Arabic cyberbullying detection, including:

- Fine-tuning a pretrained transformer model (AraBERT)
- Training a Convolutional Neural Network (CNN)
- Training a Recurrent Neural Network (RNN) with LSTM layers

However, the CNN and RNN models exhibited poor classification performance, with precision, recall, and F1-score for class 1 (bullying) being 0. This report investigates the possible reasons and suggests improvements.

3 Dataset And Preprocessing

The dataset was obtained from Kaggle and consists of Arabic tweets labeled as Bullying (1) or Non-Bullying (0). The preprocessing steps included:

- Standardizing labels
- Removing missing values
- Tokenizing and padding sequences for CNN and RNN
- Using AraBERT tokenizer for the transformer model
- Splitting data into training, validation, and test sets with stratification

4 Model Architectures

4.1 AraBERT (Transformer-Based Model)

- Fine-tuned using transfer learning
- Used a classification head with cross-entropy loss
- Achieved the best performance in classification

For this model, the number of epochs used were 5, but we implemented early stopping to prevent overfitting, and we loaded the best model that resulted from training. (stopped at epoch 4 and selected the model at epoch 2 which has the least validation loss: 0.104).

4.2 CNN Model

- Used an embedding layer followed by 1D convolutional layers
- Applied max pooling and dropout layers
- Ended with a dense softmax classification layer
- Poor results for class 1

For this model, the number of epochs used were 15, but with early stopping we stopped at epoch 5 because overfitting was detected.

4.3 RNN Model

- Used an embedding layer
- Two LSTM layers with dropout
- Dense layers for binary classification
- Also failed to classify class 1

For the RNN model, the number of epochs used was 10, but since overfitting was detected, we stopped at epoch 7.

5 Evaluation Results

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
AraBERT	~96.5%	0.8	0.79	0.79
CNN	Low	0	0	0
RNN	Low	0	0	0

The CNN and RNN models completely failed to classify Bullying (class 1), indicating severe bias towards non-bullying (class 0). For the AraBERT model, we realize a slight difference in the results when we froze the model. But in general, reaching an accuracy of 96.5% is incredible for this model, knowing that not a lot of such problems were solved before .

6 Possible Reasons For Failure

6.1 Data Imbalance

- If class 0 dominates the dataset, models learn to predict only 0.
- Even though compute_class_weight was used, it may not have been effective.

6.2 Feature Representation Issues

- CNNs are better suited for spatial features, while text data relies on sequential dependencies.
- LSTMs can suffer from vanishing gradients, preventing long-range dependencies from being captured effectively.

6.3 Softmax Activation Bias

- Softmax outputs probabilities for both classes, but an imbalance in learned features may push the model to favor class 0 entirely.
- Using **sigmoid activation with binary cross-entropy loss** could improve results.

6.4 Poor Tokenization

- The CNN and RNN models used a simple tokenizer, while AraBERT used a specialized transformer tokenizer.
- Arabic morphology requires subword tokenization (e.g., BPE or WordPiece) for better feature extraction.

7 Suggested Improvements

- Rebalancing the Dataset: Use data augmentation or oversampling techniques (SMOTE for text-based embeddings).
- Adjusting Class Weights: Explicitly modify loss functions to penalize incorrect class 1 predictions more heavily.
- Different Tokenization Methods: Experiment with word embeddings like Word2Vec or FastText instead of simple tokenization.
- Implement something like AraVec with CNN and RNN models to enhance the performance.
- Alternative Architectures:
  - Bidirectional LSTMs (BiLSTMs): Enhance sequence understanding.
  - Attention Mechanisms: Help the model focus on important words.
  - Hybrid CNN-RNN models: Combine CNNs for feature extraction with LSTMs for sequential learning.

8 Conclusion

This project demonstrated that transformer-based models (AraBERT) in this case outperform CNNs and RNNs in Arabic text detection in general and specifically Arabic Cyberbullying detection. As mentioned in the report, the poor performance of CNNs and RNNs is likely due to data imbalance and tokenization issues, however they were not addressed due to the limited time frame for the project and the fact that the main model used is AraBERT and the others were for comparison. To improve the results, future work should focus on data augmentation, better embeddings, and alternative deep learning architectures like Bidirectional LSTMs.
This study highlights the importance of selecting the right deep learning model based on linguistic properties and dataset characteristics.