NLP Assignment 1 report

Tara Sabooni, Yasaman Samadzadeh, Bahar Hamzehei, Hamidreza Bahmanyar
Master's Degree in Artificial Intelligence, University of Bologna
{ tara.sabooni, yasaman.samadzadeh, bahar.hamzehei, hamidreza.bahmanyar }@studio.unibo.it

Abstract

This report examines sexism detection on social media, finding that while LSTM models performed similarly, the Transformer-based RoBERTa model outperformed them with an F1 score of 0.8515, effectively handling nuanced and context-dependent language.

1 Introduction

The growing presence of harmful content on social media has underscored the importance of automated systems for identifying sexist language. This report analyzes a tweet dataset to determine whether posts are sexist or non-sexist.

2 Background

Sexism detection on social media requires models that capture nuanced, context-dependent language. This project compared LSTM models, which learn sequential dependencies but train slowly, with the pre-trained Transformer model "twitter-roberta-base-hate." Fine-tuning the Transformer model on labeled English tweets highlighted its effectiveness for detecting harmful online content.

3 System description

3.1 Building the vocabulary

This section constructs a vocabulary and initializes embeddings using 50-dimensional GloVe vectors. The vocabulary includes all tokens from the training set and GloVe, along with special tokens `[PAD]` and `[UNK]`. An embedding matrix is created with GloVe embeddings for known words, random initialization for unknown words, a zero vector for `[PAD]`, and the mean of known embeddings for `[UNK]`.

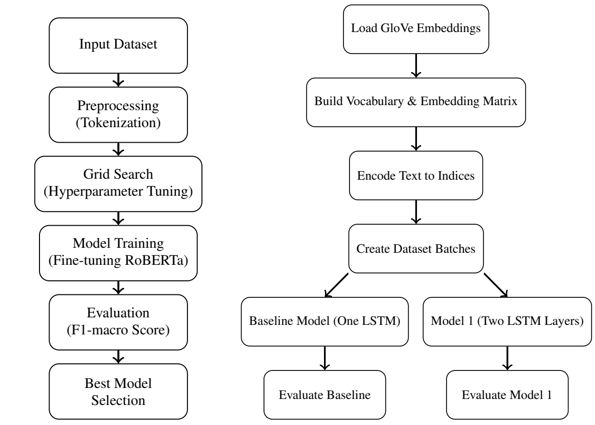
3.2 LSTM model

Two LSTM-based models were implemented: the Baseline Model with a single bidirectional LSTM layer and Model 1 with two stacked bidirectional LSTM layers. Both use an embedding layer initialized with a prebuilt embedding matrix. A grid search was performed over hyperparameters, including LSTM units, batch size, and embedding trainability, to find the optimal configurations. Each model was trained with three random seeds, and the mean and standard deviation of macro-F1 scores were calculated to ensure robustness.

3.3 Transformer model

The architecture used in this project is based on the `AutoModelForSequenceClassification` from the HuggingFace Transformers library. The model architecture itself is not an original contribution; it is a pre-trained RoBERTa model specifically fine-tuned for hate speech detection. Our original contribution lies in the implementation of a grid search process to identify the optimal hyperparameter configuration for this task.

Figure 1 shows the pipeline architecture for both the Transformer and LSTM models. The left side of the figure (a) illustrates the Transformer model pipeline, while the right side (b) presents the pipeline for the LSTM model.


Input Dataset
Preprocessing (Tokenization)
Grid Search (Hyperparameter Tuning)
Model Training (Fine-tuning RoBERTa)
Evaluation (F1-macro Score)
Best Model Selection
Load GloVe Embeddings
Build Vocabulary & Embedding Matrix
Encode Text to Indices
Create Dataset Batches
Baseline Model (One LSTM)
Model 1 (Two LSTM Layers)
Evaluate Baseline
Evaluate Model 1

(a) Transformer model pipeline

(b) LSTM model pipeline

Figure 1: Comparison of Transformer and LSTM model.

4 Data

4.1 Dataset

The dataset used in this assignment is a curated collection of tweets aimed at identifying sexist content in social media posts. The classification task involves categorizing text into two distinct labels: yes for sexist and no for Non-sexist. The dataset is divided into three separate `.json` files: `training.json`, `validation.json`, `test.json`.

4.2 Preprocessing Work

The preprocessing of tweets involved cleaning and transforming text to prepare it for machine learning models. Key steps included:

- **Text Cleaning:** Removed emojis, hashtags, mentions, URLs, special characters, and standalone numbers. Text was also converted to lowercase, and URLs were eliminated.

- **Lemmatization:** Applied POS tagging and NLTK's `WordNetLemmatizer` to reduce words to their base forms.

5 Experimental setup and results

The baseline LSTM model included an embedding layer, a bidirectional LSTM with 64 units, and a sigmoid-activated dense output layer. Model 1 enhanced this by adding another bidirectional LSTM layer. The Transformer model fine-tuned Twitter-RoBERTa with a classification head. Models were assessed using accuracy during training and F1 score on validation. Hyperparameter tuning was performed via grid search. Key hyperparameters included LSTM units (64, 256), learning rate (1e-3 for LSTM, 5e-5/3e-5 for Transformer), batch size (32, 64 for LSTM; 8, 16 for Transformer), and dropout (0.3). Robustness was ensured by evaluating each configuration across three random seeds.

The results of our experiments are summarized in Table 1.

Model	Validation F1 (mean ± std)	Test F1 (mean ± std)
Baseline Model	0.7829 ± 0.0159	0.7383 ± 0.0089
Model 1	0.7727 ± 0.0164	0.7421 ± 0.0097
Transformer	0.8812	0.8515

Table 1: Model performance metrics for validation and test sets.

6 Discussion

6.1 Discussion of Error Patterns in Misclassified Tweets

The analysis of misclassified tweets from both the **LSTM models** and the **Transformer model** reveals several common types of tweets that both models struggled to classify correctly. These include:

- **Tweets Discussing Gender-Related Topics and Harassment:** Both models faced challenges in accurately classifying tweets about gender roles, harassment, or women's rights due to subtle language and cultural context.

- **Tweets with Complex Sentences and Contextual Language:** Both models had difficulty classifying tweets containing long, complex sentences or academic language.

- **Tweets with Sarcasm or Ambiguity:** Both models struggled with tweets containing sarcasm, irony, or ambiguous language. These tweets require an understanding of tone and implicit meaning, which surface-level text analysis often fails to capture.

- **Tweets Discussing Violence or Harsh Topics:** Tweets discussing violent actions or harsh realities were often misclassified by both models. The models might interpret such tweets as offensive or inappropriate, even when they are factual statements.

- **Tweets with Informal Language or Slang:** Informal language, slang, or colloquial expressions also posed a challenge for both models. These variations of informal speech may not be fully represented in the training data.

7 Conclusion

This report evaluated sexism detection in social media posts, finding that while LSTM models performed moderately, the Transformer model (Twitter-RoBERTa) achieved superior performance (F1: 0.8515). Challenges include sarcasm, informal language, and complex context. Future work should explore richer datasets, multi-task learning, and enhanced context handling. Transformer models demonstrate strong effectiveness for nuanced text classification.

References

[1] Cardiff NLP. twitter-roberta-base-hate: A RoBERTa-based model fine-tuned for hate speech detection. Available at: https://huggingface.co/cardiffnlp/twitter-roberta-base-hate.