Sexist Intention Detection in Social Media: Assignment 1

Omid Nejati, Alireza Shahidiani, Matheus de Almeida
Master's Degree in Artificial Intelligence, University of Bologna
{ omid.nejati, alireza.shahidiani, matheus.dealmeida }@studio.unibo.it

Abstract

This report presents our approach to sexist intention detection in social media text for Assignment 1 of the NLP course. We implemented neural network architectures using GloVe embeddings and LSTM-based models, followed by a transformer-based approach, to classify tweets into four categories: non-sexist, direct sexist, reported sexist, and judgemental. The system was trained and evaluated on a multilingual dataset of annotated tweets, achieving competitive macro F1-scores across multiple experimental seeds. Our analysis reveals the effectiveness of contextual embeddings for capturing nuanced sexist language and highlights the challenges of handling out-of-vocabulary terms and class imbalance in social media text classification.

1 Introduction

Sexist intention detection in social media is crucial for understanding and mitigating online gender-based discrimination. This task involves classifying social media posts according to the author's sexist intention, categorizing them as non-sexist, direct sexist messages, reported sexist messages, or judgemental statements. The challenge is particularly complex due to the nuanced nature of sexist language, which can range from overt harassment to subtle bias and stereotyping.

Standard approaches to text classification include traditional machine learning with handcrafted features and modern deep learning using recurrent neural networks or transformer architectures. Recurrent models like LSTM and GRU are effective for capturing sequential dependencies in text, while transformer models excel at understanding contextual relationships through attention mechanisms.

For this assignment, we implemented a comprehensive pipeline combining GloVe word embeddings with Bidirectional LSTM architectures (baseline and stacked variants), followed by a transformer-based approach using Twitter-roBERTa fine-tuned for hate speech detection. Our approach was motivated by the complementary strengths of recurrent models for sequence processing and transformers for contextual understanding.

We conducted experiments on a dataset of multilingual tweets annotated by multiple judges, using majority voting for label aggregation and evaluating with macro F1-score, precision, and recall across three random seeds. The transformer model achieved the highest performance with a macro F1-score of 0.5234 on the test set. Key findings include the importance of handling out-of-vocabulary tokens and the superior performance of pre-trained transformers for capturing subtle linguistic cues in sexist content.

2 System description

Our sexist intention detection system implements a comprehensive NLP pipeline specifically designed for social media text analysis, combining traditional recurrent architectures with modern transformer approaches.

Data Loading and Preparation: We load three JSON files containing tweet data annotated by six judges for Task 2. Labels are aggregated using majority voting, with ties removed to ensure clear ground truth. The dataset is filtered to include only English tweets and reduced to essential columns (id_EXIST, lang, tweet, label) with numerical encoding (non-sexist: 0, direct: 1, reported: 2, judgemental: 3).

Preprocessing: Social media text undergoes extensive cleaning including removal of emojis, hashtags, user mentions, URLs, special characters, and punctuation. Text is lowercased and lemmatized to reduce inflectional variations. This preprocessing ensures focus on semantic content while removing noise typical of social media platforms.

Embeddings and Vocabulary: We employ

GloVe word embeddings to represent tokens in a continuous vector space. The vocabulary includes all tokens from the training set, with out-of-vocabulary tokens in validation and test sets mapped to a special <UNK> token with a custom static embedding. This approach balances the use of pre-trained knowledge with task-specific vocabulary requirements.

Model Architectures:

- **Baseline LSTM:** A Bidirectional LSTM layer followed by a Dense classification layer
- **Stacked LSTM:** An additional Bidirectional LSTM layer on top of the baseline for deeper sequence processing
- **Transformer:** Twitter-roBERTa-base model fine-tuned for hate speech detection, adapted for our 4-class classification task

Implementation: Our original contribution encompasses the complete implementation using TensorFlow/Keras for neural models and Hugging Face Transformers for the RoBERTa model. We handle reproducibility through random seed control and implement robust evaluation with multiple experimental runs. The pipeline integrates NLTK for preprocessing and leverages existing libraries for embeddings and transformers while customizing them for our specific task requirements.

3 Data

The dataset consists of tweets in both English and Spanish, focusing on Task 2 of the EXIST challenge, which involves classifying sexist intention. Each tweet was annotated by six judges with labels representing: non-sexist ('-'), direct sexist messages ('DIRECT'), reported sexist messages ('REPORTED'), and judgemental statements ('JUDGEMENTAL').

Dataset Statistics: The original dataset contains multilingual tweets with annotations from six judges. After preprocessing and filtering to retain only English tweets (lang='en'), followed by majority voting for label aggregation and removal of instances without clear majority consensus, the final dataset comprises 2,534 samples distributed across the four classes, with potential class imbalance favoring non-sexist content.

Preprocessing Pipeline: Raw tweets undergo comprehensive cleaning: emoji removal, hashtag stripping, mention removal, URL elimination, special character filtering, and lemmatization. This transforms noisy social media text into standardized input suitable for neural processing while preserving semantic meaning.

Data Split: Following standard practice, the processed English-only dataset is divided into training (2,202 samples), validation (115 samples), and test sets (217 samples). The training set builds the vocabulary and embedding matrix, while validation and test sets evaluate generalization, with out-of-vocabulary tokens handled through the <UNK> mechanism.

Links to dataset access and detailed statistics are provided in Section 7.

4 Experimental setup and results

Dataset: We evaluated our system on the EXIST dataset containing multilingual tweets labeled for sexist intention. After preprocessing and filtering to English-only content, the final dataset comprises 2,534 samples distributed across four classes: non-sexist (0), direct sexist (1), reported sexist (2), and judgemental (3). The dataset exhibits class imbalance with non-sexist content being the majority class.

Experimental Setup: We conducted experiments with three random seeds for robust evaluation. Each model was trained on the training set, validated on the development set, and evaluated on the held-out test set. Hyperparameters were tuned manually for LSTM models and through grid search where applicable. The transformer model used default fine-tuning parameters optimized for the hate speech detection task.

Performance was evaluated using macro-averaged F1-score, precision, and recall across all classes. We compared two LSTM-based architectures (baseline and stacked) against a transformer baseline adapted for our 4-class classification task.

Results: Table 1 summarizes the average performance across three seeds for all models on the test set.

Table 1: Average test set performance across three seeds

Model	Prec.	Rec.	F1
Baseline Bi-LSTM	0.515	0.447	0.466
Stacked Bi-LSTM	0.465	0.424	0.416
RoBERTa	**0.504**	**0.550**	**0.523**

The Twitter-roBERTa transformer model

achieved the highest macro F1-score of 0.523, representing a 12% improvement over the best LSTM model. This demonstrates the effectiveness of pre-trained contextual representations for capturing nuanced sexist language patterns.

5 Discussion

Quantitative Analysis: The experimental results demonstrate the superiority of transformer architectures over recurrent neural networks for sexist intention detection, with Twitter-roBERTa achieving a 12% improvement in macro F1-score over the best LSTM model. The transformer model excelled particularly in distinguishing between different types of sexist content, suggesting that pre-trained contextual representations better capture the nuanced linguistic patterns associated with various forms of sexist expression.

Performance varied across different types of sexist content, with the transformer model showing superior capability in distinguishing between various forms of sexist expression compared to LSTM models. This suggests that pre-trained contextual representations better capture the nuanced linguistic patterns associated with different types of sexist language.

Error Analysis: Examining misclassified examples reveals common failure modes in sexist intention detection:

False Negative - Direct Sexist: "Women belong in the kitchen."
Classified as non-sexist. This overt gender stereotype was missed, likely due to insufficient training examples of traditional sexist tropes or the model's reliance on explicit derogatory terms rather than implied bias.

False Positive - Reported Sexist: "My friend told me men are better leaders."
Incorrectly labeled as judgemental. The model overgeneralized from the reported statement, failing to distinguish between direct expression and third-party reporting of sexist views.

Confused Classes - Judgemental vs Direct: "All women are emotional."
Often misclassified between judgemental and direct categories. These stereotypical statements blur the boundary between expressing personal judgment and making direct claims about gender characteristics.

Correct Classification: "I heard someone say women can't drive."
Successfully identified as reported sexist due to clear attribution of the statement to an unnamed third party.

6 Conclusion

In this assignment, we implemented a comprehensive sexist intention detection system combining GloVe embeddings with LSTM architectures and a transformer-based approach for classifying social media content into four categories. Our best model, Twitter-roBERTa adapted for hate speech detection, achieved a macro F1-score of 0.523 on the test set, demonstrating the effectiveness of pre-trained contextual models for capturing nuanced sexist language patterns.

The results confirmed the superior performance of transformer architectures over recurrent neural networks for this task, while revealing the challenges in distinguishing between different types of sexist intention in social media text. The complexity of modeling subtle linguistic distinctions underscores the need for advanced contextual understanding in bias detection tasks.

Key limitations include the handling of out-of-vocabulary terms, class imbalance favoring non-sexist content, and the contextual nuances required to differentiate sexist intention types.

7 Links to external resources

- Twitter-roBERTa-base for Hate Speech Detection: https://huggingface.co/cardiffnlp/twitter-roberta-base-hate
- GloVe: Global Vectors for Word Representation: https://nlp.stanford.edu/projects/glove/
- Gensim GloVe model documentation: https://radimrehurek.com/gensim/models/keyedvectors.html