Assignment 1

Tommaso Bernardini, Samuele Gasbarro, Carmine Santella,
Master's Degree in Artificial Intelligence, University of Bologna
{ tommaso.bernardini6, samuele.gasbarro, carmine.santella }@studio.unibo.it

Abstract

This report details our approach to Task 2 of the EXIST 2023 challenge: categorizing tweets into four classes based on sexist intention (*Non-sexist, Direct, Reported, Judgmental*). Addressing the high variability of linguistic forms in social media, we explored two complementary architectures: Recurrent Neural Networks (BiLSTMs) and Transformer-based models. We designed a robust evaluation pipeline to compare Baseline and Stacked BiLSTMs initialized with three pre-trained embeddings: GloVe-Wiki, GloVe-Twitter, and FastText. We experimented both with frozen and trainable embeddings, giving comparable results. Ultimately, the Transformer-based approach, with **RoBERTa** (Antypas and Camacho-Collados, 2023), demonstrated superior performance over the best RNN configuration. We further validated the cross-lingual generalization of Transformers using **pysentimiento** (Pérez et al., 2024) for the Spanish subset.

1 Introduction

Sexism detection involves navigating complex linguistic forms and contextual subtleties. The **EXIST 2023 Task 2** challenge addresses this by categorizing the author's *intention* into *Direct* (harming), *Reported* (sharing), or *Judgmental* (condemning) sexism—a distinction made difficult by the aggressive vocabulary shared across these intent classes. In this report, we explore two approaches: statistical RNNs and modern Transformers.

In this report, we explore two complementary approaches to tackle this problem: statistical Recurrent Neural Networks (RNNs) and modern Transformer architectures. For the recurrent approach, we implemented a shared preprocessing pipeline to clean text and normalize tokens (lemmatization). We constructed a unified vocabulary to initialize three pre-trained embedding matrices, which are the following. **GloVe Wiki-Gigaword (Pennington et al., 2014) :** This is a general-purpose baseline. **GloVe Twitter:** Selected to better handle the informal syntax and slang typical of the dataset. **FastText (Grave et al., 2018) :** Chosen for its ability to handle rare words and sub-word information, which is crucial for noisy social media text. We evaluated these embeddings in both **frozen** and **fine-tuned** settings to assess the impact of updating pre-trained representations during training. Using these encodings, we designed two RNN architectures: a **Baseline BiLSTM** and a deeper **Stacked BiLSTM**, training each across five random seeds to ensure statistical robustness. In parallel, we investigated Transformer-based architectures to leverage their dynamic attention mechanism. Additionally, to address the multilingual aspect of the dataset, we experimented with **pysentimiento**, a Transformer-based toolkit optimized for Spanish sentiment analysis, to assess cross-lingual performance capabilities.

2 System description

Data Processing: For the English subset, we applied a cleaning pipeline (removing emojis, URLs, mentions, hashtags) and performed **Lemmatization** (using NLTK with POS tagging) to normalize the input for RNNs. For RoBERTa, we used the raw text as contextual models benefit from full sentence structures. For the Spanish tweets analysis we used Stanza (Qi et al., 2020), a collection of accurate and efficient tools for the linguistic analysis of many languages.

Embeddings & Strategies: We experimented with three pre-trained static embeddings: GloVe Wiki (100d), GloVe Twitter (100d), and FastText (300d). We also investigated the impact of fine-tuning by comparing **Trainable** and **Frozen** embedding layers. **RNN Architectures:** We implemented two core RNN structures using TensorFlow/Keras:

- **Baseline BiLSTM:** Embedding layer → BiLSTM (64) → Dense (64, ReLU) → Dropout (0.5) → Softmax.

- **Stacked BiLSTM:** Adds a second BiLSTM layer (32) to capture deeper dependencies.

Transformers: We implemented a fine-tuning pipeline using RoBERTa along with the HuggingFace Trainer API. We also experimented a different version of RoBERTa, presented in (He et al., 2021), as we wanted to try an improved version of BERT which used disentagled attention. **Multilingual Analysis:** For the Spanish subset, which presents distinct linguistic challenges, we utilized pysentimiento. It is based on *RoBERTuito* (Pérez et al., 2022), a *RoBERTa* model trained in Spanish tweets. Moreover, we decided to make another experiment, i.e. training RoBERTa with english tweets and testing it on the English test set.

3 Experimental setup and results

All models were trained using the AdamW optimizer and Sparse Categorical Cross-entropy loss. To ensure statistical significance, we trained every configuration on 5 random seeds. We used Early Stopping (patience=3) and ReduceLROnPlateau to prevent overfitting.

Model	Config	Prec.	Rec.	F1
Recurrent Neural Networks (English)				
Baseline	GloVe Wiki (100d)	0.467	0.413	0.422 ±0.04
Stacked	GloVe Wiki (100d)	0.378	0.357	0.355 ±0.04
Baseline	GloVe Twit (100d)	0.442	0.366	0.361 ±0.01
Stacked	GloVe Twit (100d)	0.464	0.415	0.420 ±0.02
Baseline	FastText (300d)	0.311	0.365	0.334 ±0.02
Stacked	FastText (300d)	0.372	0.368	0.348 ±0.03
Transformer Models				
RoBERTa	**Pre-trained (En)**	**0.581**	**0.524**	**0.505**
DeBERTa	Pre-trained (En)	0.422	0.430	0.403
pysentimiento	Pre-trained (Es)	0.490	0.485	0.487

Table 1: Test set performance (Macro-Avg). Best model in bold.

RNN Results: The *Baseline BiLSTM with GloVe Wiki* was the best RNN (F1: 0.422). Interestingly, increasing complexity (Stacked) hurt performance with Wiki embeddings but helped with Twitter embeddings (F1: 0.420), suggesting domain-specific embeddings require deeper networks to be fully utilized. FastText consistently underperformed, likely due to high dimensionality (300d) causing overfitting on this small dataset. Experiments with Frozen embeddings showed comparable performances (F1 ~0.40). **Transformer Results:** RoBERTa significantly outperformed all RNNs (F1: 0.505), demonstrating the superiority of contextualized embeddings and attention mechanisms in detecting subtle intent. The second BERT model didn't achieve the expected results, as we obtained an F1 of about 0.40. **Spanish Analysis:** The pysentimiento model achieved a competitive F1 of 0.487 on the Spanish subset, validating the effectiveness of language-specific pre-training.

4 Discussion

Architectural Comparison: The leap in performance from RNNs (0.422) to RoBERTa (0.505) highlights the limitations of static embeddings. Static models struggle with polysemy and the pragmatic context required to distinguish between a user *reporting* sexism and *being* sexist.

Error Analysis: We analyzed the confusion matrices and misclassified examples:

1. **Direct vs. Judgmental:** This was the most frequent error. Both categories contain "sexist keywords." RNNs often flagged *Judgmental* tweets (which quote sexism to condemn it) as *Direct* sexism because they reacted to the aggressive vocabulary without grasping the corrective intent. RoBERTa reduced this error but still struggled with sarcastic condemnations.

2. **Reported vs. Non-Sexist:** Neutral descriptions of sexist events (*Reported*) were often misclassified as *Non-sexist* because they lacked overt aggressive sentiment.

3. **Class Imbalance:** Despite majority voting, the "Non-sexist" class dominance in training led to a bias towards the negative class in the frozen embedding experiments.

5 Conclusion

We developed a comprehensive sexism detection system. Our experiments confirm that while simple BiLSTMs with GloVe embeddings provide a decent baseline, Transformer-based architectures (RoBERTa) are essential for capturing the intent behind the text. We also highlighted the viability of off-the-shelf tools like pysentimiento for multilingual support. Future work should focus on data augmentation to address class imbalance and contrastive learning to better separate the embedding space of *Direct* and *Judgmental* sexism.

References

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv e-prints*, page arXiv:2111.09543.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Juan Manuel Pérez, Damián A. Furman, Laura Alonso Alemany, and Franco Luque. 2022. Robertuito: a pre-trained language model for social media text in spanish.

Juan Manuel Pérez, Mariela Rajngewerc, Juan Carlos Giudici, Damián A. Furman, Franco Luque, Laura Alonso Alemany, and María Vanina Martínez. 2024. pysentimiento: A python toolkit for opinion mining and social nlp tasks.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.