Artificial Intelligence Research and Development
K. Trejo et al. (Eds.)

doi:10.3233/FAIA250579

Hate Speech and Emojis: The Journalist Case

David LARROSA-CAMPS [a], Jaume SUAU [b] and Xavier VILASÍS-CARDONA [b]
[a] *Smart Society Research Group, La Salle-Universitat Ramon Llull*
[b] *Digilab, Blanquerna-Universitat Ramon Llull*

Abstract. Work in progress towards a simple and efficient use of emojis for hate speech detection is reported. The dataset used has been generated in the framework of studying hate speech against journalists. The results of the method are promising, yet not fully conclusive.

Keywords. Natural Language Processing, Hate Speech, Emojis

1. Introduction

Hate speech, which is strongly related with the spread of disinformation [1], has become one of the biggest challenges that democratic societies are facing today. The relation between disinformation and hate speech in journalism is twofold. First, attacks on and threats to journalists and the media constrain freedom of expression and dissemination of certain information, and generate fear and self-censorship, undermining the media's role in democratic societies [2]. Second, disinformation campaigns often exploit hate narratives to amplify polarization, delegitimize the press, and erode trust in democratic institutions.

For reasons such as the ones described above, hate speech detection has become a pivotal challenge in natural language processing (NLP). Recent advancements leverage transformer-based models, multimodal analysis, and fairness-aware techniques to improve detection accuracy while addressing biases [3,4]. Despite progress, challenges such as contextual ambiguity, cross-lingual generalization, and ethical trade-offs persist.

State-of-the-art hate speech detection systems predominantly rely on transformer architectures such as BERT [5] and RoBERTa [6], which excel at capturing implicit hate speech and sarcasm through contextual embeddings. Fine-tuning these models on domain-specific datasets (e.g., Gab, Twitter) has proven effective, though performance drops in cross-platform scenarios [7]. Spanish hate speech detection presents unique challenges due to linguistic diversity (e.g., regional dialects, Spanglish) and cultural context. Models like BETO [8] (a Spanish BERT variant) and multilingual approaches (e.g., XLM-T [9]) have shown promise, but performance varies across regions (e.g., Mexico vs. Spain).

In our analysis of data from the X platform in search for hate speech against journalists in Spain, we have remarked the heavy use of emojis. Emojis, faithful to the name, represent the emotions of the writer and are a valuable tool to identify hate speech. However, they are not part of the grammar of natural language, with which recurrent or transformer models have been trained. Because of this, these models may be confused by their presence. In this note, we investigate a strategy to use them in the detection task, combined with text analysis.

2. Data Collection

The data were collected from the social media platform X (formerly Twitter), chosen due to its high levels of toxicity and the presence of all selected journalists. The data collection process took place during the 2024 European elections and the days immediately following. A curated list of journalists to monitor was created, ensuring balance in terms of sector, gender, and other relevant criteria. Tweets and replies were automatically downloaded using Python scripts via the official X API. The endpoints used included user lookup, user tweets, and user mentions.

In total, 41,791 tweets were manually labeled, with only 4.6% (1,947 tweets) identified as containing hate speech, zresulting in a highly imbalanced dataset. The labeling was carried out by four annotators, each of whom labeled approximately 10,447 tweets. To address the class imbalance problem, the majority class was undersampled to match the minority class. This resulted in a final dataset composed of 3,894 tweets, with an equal distribution of 1,947 hateful tweets (50%) and 1,947 non-hateful tweets (50%). Among the non-hateful tweets, only 256 (6.57%) contained emojis, while 1,691 (43.43%) did not. Similarly, among the hateful tweets, 259 (6.65%) included emojis, while 1,688 (43.35%) did not. Descriptive analysis revealed a higher incidence of hate speech directed toward women and sports journalists.

3. Dealing with Emojis

When processing the data, a major challenge was the dual nature of tweets, which often contain both textual content and emojis. While most Natural Language Processing (NLP) techniques are well-suited for textual data, they are not inherently designed to interpret emojis, which convey affective and contextual nuances that can significantly alter the meaning of a message. Ignoring emojis can therefore lead to a substantial loss of sentiment-related information.

To address this issue, the preprocessing pipeline was designed to treat text and emojis as two separate input modalities. The text was cleaned by removing user mentions, retweet markers, punctuation, numbers, and any emojis. This step ensured that only clean textual content was passed to the language processing models. For the emoji data, the process involved extracting all emojis from the original tweet and assigning each one a sentiment score. To obtain these scores, we used the `emosent` Python library[1], which is based on the sentiment values published in the study ”Sentiment of Emojis” [10]. This study provides manually annotated sentiment scores—expressed as numerical continuous values represent-

[1] https://github.com/omkar-foss/emosent-py/blob/master/README.md

ing positive, neutral, or negative sentiment—for a wide range of emojis, based on human perception. However, since the original dataset is from 2015 and many new emojis have been introduced since then, we manually annotated the missing emojis to ensure completeness and consistency in the scoring process.

After assigning sentiment values to each emoji, we computed an aggregated sentiment score for the emoji set in each tweet using a frequency-weighted formula, $Emoji\ score = \sum_{n=0}^{N} f(n) \cdot s(n)$, where $f(n)$ is the frequency of the n-th emoji and $s(n)$ is its sentiment score. This formula takes into account all the emojis present in the tweet, giving more weight to the most frequently used ones, thus emphasizing the predominant sentiment. Finally, the resulting score was normalized to fall within a predefined range, ensuring consistency across tweets regardless of the total number of emojis used.

4. Experiments

Experiments were conducted by splitting the dataset into three blocks: one for training, one for testing, and one for final evaluation, with proportions of 2726, 584, and 584 samples respectively.

We present four different approaches to emoji treatment, all based on a Recurrent Neural Network (RNN). To validate the hypothesis that emojis can improve hate speech classification, the following models were designed: (1) an RNN with BATO embeddings, without emoji information; (2) an RNN with BATO fine-tuned, also without emoji information; (3) an RNN with BATO fine-tuned, incorporating emojis as sentiment scores; and (4) an RNN with BATO fine-tuned, incorporating emojis as one-hot encoded vectors. The models were designed to assess whether incorporating emojis contributes to the classification task. Each model incrementally improves the representation of textual and emoji content by exploring increasingly complex ways of encoding emoji semantics.

	RNN (embeddings)	**RNN (fine-tuning BETO)**
F1 score	$0.73 \pm 0.17\times10^{-3}$	$0.76 \pm 0.12\times10^{-3}$
Accuracy	$0.73 \pm 0.17\times10^{-3}$	$0.76 \pm 0.12\times10^{-3}$
Precision	$0.74 \pm 0.19\times10^{-3}$	$0.77 \pm 0.21\times10^{-3}$
Recall	$0.74 \pm 0.26\times10^{-3}$	$0.75 \pm 0.44\times10^{-3}$
ROC-AUC	$0.73 \pm 0.17\times10^{-3}$	$0.76 \pm 0.12\times10^{-3}$

Table 1. Performance comparison: RNN with embeddings vs. fine-tuned BETO.

	RNN + MLP (score)	**RNN + MLP (one-hot)**
F1 score	$0.76 \pm 0.40\times10^{-3}$	$0.76 \pm 0.22\times10^{-3}$
Accuracy	$0.76 \pm 0.39\times10^{-3}$	$0.76 \pm 0.22\times10^{-3}$
Precision	$0.77 \pm 1.06\times10^{-3}$	$0.77 \pm 0.77\times10^{-3}$
Recall	$0.78 \pm 5.06\times10^{-3}$	$0.77 \pm 1.42\times10^{-3}$
ROC-AUC	$0.76 \pm 0.37\times10^{-3}$	$0.77 \pm 0.22\times10^{-3}$

Table 2. Performance comparison: RNN + MLP with emoji score vs. one-hot encoding.

5. Discussion and Conclusion

Treating separately the emojis seems to reinforce the performance of the detection system. However, we cannot consider these results conclusive. The amount of data is scarce, due to the imbalance and may limit the capability of the RNNs. Further data collection and labeling is under way to confirm the analysis. ANOVA and Tukey HSD tests confirmed statistically significant differences between all models ($p < 0.001$), with RNN + MLP (emoji score) outperforming the rest. This model also showed the highest recall, though without introducing bias, as its F1 score remained balanced and differences were not statistically significant.

6. Acknowledgments

Authors want to thank Alvaro García-Piquer and Luis Servera for the data collection under the EU project Media Councils in the Digital Age. This work is supported by the project *Disargue*, reference TED2021-130810B-C22, from the Spanish MICINN.

References

[1] Anat Ben-David and Ariadna Matamoros Fernández. Hate speech and covert discrimination on social media: Monitoring the facebook pages of extreme-right political parties in spain. *International Journal of Communication*, 10(0), 2016.

[2] M. Clark and A. Grech. *Journalists Under Pressure: Unwarranted Interference, Fear and Self-censorship in Europe.* Council of Europe, 2017.

[3] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. The risk of racial bias in hate speech detection. *Proc. ACL*, pages 1668–1678, 2019.

[4] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proc. AAAI*, 35:14867–14875, 2021.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proc. NAACL*, 1:4171–4186, 2019.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[7] C. J. Kennedy, G. Bacon, A. Sahn, and C. von Vacano. Constructing interval variables via faceted rasch measurement and multitask deep learning: A hate speech application. *PLoS ONE*, 15(12):e0243700, 2020.

[8] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. Spanish pre-trained bert model and evaluation data. *Proc. PML4DC at ICLR*, 2020.

[9] F. Barbieri, L. Espinosa-Anke, and J. Camacho-Collados. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. *Proc. LREC*, pages 258–266, 2022.

[10] Borut Sluban Igor Mozetič Petra Kralj Novak, Jasmina Smailović. Sentiment of emojis. *PLoS ONE 10(12): e0144296*, 2015.