CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.21956

Emotion detection for online gender-based stereotypes: a comparative study

Elena Catanese[1]
[1]National Institute of Statistics ISTAT, Italy.

How to cite: Catanese, E. (2026). Emotion detection for online gender-based stereotypes: a comparative study. In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 1-3 July 2026. https://doi.org/10.4995/CARMA2026.2026.21956

Abstract

Monitoring online discourse related to violence against women is increasingly relevant for official statistics, as social media provide timely signals on emerging forms of gender-based violence (GBV), stereotypes, and public attitudes. Istat currently relies on IRIDE, a proprietary natural language processing (NLP) platform, to perform emotion detection on social media posts. This study evaluates whether a transparent and reproducible open-source alternative can provide comparable performance in this socially sensitive domain. We develop an Italian transformer-based emotion classifier using AlBERTo and adopt a two-stage fine-tuning strategy combining a general social media emotion dataset and a GBV-specific corpus. Rather than proposing a novel architecture, the study contributes an operational benchmarking framework for comparing open-source and proprietary systems in official statistics settings. Results show that the open-source model achieves competitive benchmark performance, reaching higher accuracy and comparable macro-F1 scores relative to IRIDE on held-out datasets. However, when applied to a large corpus of approximately 150,000 social media posts related to GBV, the two systems produce substantially different emotion distributions. In particular, the open-source model exhibits a tendency to assign emotionally ambiguous content to positive categories such as joy, whereas IRIDE predicts higher shares of anger and sadness, which appear more aligned with prior expectations for GBV-related discourse. These findings suggest that standard supervised metrics alone may be insufficient for evaluating deployment validity in socially sensitive monitoring tasks and highlight the need for richer validation strategies when integrating NLP systems into official statistics

Keywords: *emotion detection; social stereotypes against women; BERT; open-source .*

Article in Press

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

1. Introduction

Violence against women remains a persistent social problem with substantial consequences for public health, equality, and social cohesion. Official statistical systems increasingly complement traditional survey instruments with digital traces and large-scale online data to monitor emerging social phenomena. In this context, social media represent an important source of timely information for identifying narratives, stereotypes, and reactions associated with gender-based violence (GBV). However, the use of natural language processing systems in official statistics raises methodological concerns related to transparency, reproducibility, and auditability. Istat currently uses IRIDE, a proprietary NLP platform, to support emotion detection on social media posts related to violence against women. Although proprietary systems may benefit from large-scale pretrained models and industrial optimization pipelines, their closed nature limits methodological transparency, prevents retraining on domain-specific corpora, and reduces reproducibility. These constraints are particularly relevant in official statistics, where methodological transparency and replicability are essential.

The objective of this study is therefore not to introduce a novel NLP architecture, but to evaluate whether an open-source transformer-based solution can provide a transparent and reproducible alternative while maintaining acceptable performance in a socially sensitive monitoring task. To this end, we develop an emotion classifier based on AlBERTo and compare it with IRIDE using both benchmark evaluation metrics and large-scale prediction behavior.

The study makes three contributions. First, it presents a reproducible open-source pipeline for Italian emotion detection in GBV-related discourse. Second, it provides an operational comparison between an institutional proprietary system and an open-source alternative. Third, it highlights a potentially important methodological issue: strong benchmark performance does not necessarily translate into plausible large-scale prediction behavior in domain-sensitive applications.

The paper is structured as follows: in section 2 annotated data are described; section 3 presents the methodology while results are shown in section 4. Conclusions are drawn in section 5.

2. Data

2.1. EMit EVALITA 2023 campaign

EVALITA provides a shared framework for the evaluation of different systems and approaches on separate tasks, all for Italian. For the 2023 edition Emit (Emotions in Italian, Araque et al. 2023) was the first shared task on categorical emotion detection in Italian social media. It comprises two subtasks: emotion detection and target detection. Emotion detection is formulated as a multi-label classification problem over ten labels corresponding to Plutchik's

Article in Press

eight basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust), plus love and neutral. EMit dataset consists of 6966 Italian tweets discussing RAI TV shows and series; 5966 are used for training and 1000 for test. If only single-label, after removing duplicates, are considered, records for train and test become respectively 4532 and 818. These are sampled during 2020 by using the main hashtags of Italian tv shows. Although it may seem that the domain is out of target, some programs such as #amorecriminale, #chilhavisto specifically focus on gender violence. In addition to that participant to the Emit campaign were evaluated in (Emotions in Italian, Araque et al. 2023). Most systems participating in EMit relied on supervised approaches based on fine-tuning large language models. Several BERT-based architectures were tested, together with sequence-to-sequence models such as ChatGPT, Flan-T5, mT5, IT5, and LLaMA. A battery of experiments was carried out with classic BERT models and the best system turned out to be based on AlBERTo (Calò et al 2023). While semi-supervised approaches achieved the best results (Hromei et al., 2023), in this study the AlBERTo model was adopted to ensure comparability with the IRIDE architecture. By collapsing disgust into anger, and disregarding classes such as anticipation and trust we get, in Table1

Table 1 Training dataset and test set EMit

Class	Count Training	Relative share	Count Test	Relative share
Love	203	6.4%	39	6.9%
Joy	382	12.1%	66	11.6%
Neutral	1210	38.3%	210	37.0%
Fear	37	1.2%	6	1.1%
Anger	675	21.4%	131	23.1%
Sadness	305	9.6%	60	10.6%
Surprise	349	11.0%	56	9.9%
Total	3161		568	

2.2. Istat datasets

Unlike EMit dataset this dataset (IRIDE ISTAT) has a single-emotion label, and 7 emotion classes such as: love joy, neutral, fear, anger, sadness and surprise. It consists of 2,727 records, split into 70% training, 20% test and 10% validation, stratified by emotion to preserve relative shares. Tweets were sampled among Italian posts published between 1 June 2020 and 30 September 2020 that contained at least one gender-stereotype keyword (example (e.g. "*violenza"* violence AND "*donna"* woman). Some tweets appeared multiple times in the original dataset and were not excluded in order to preserve the naturally skewed distribution observed in the social media stream. In addition, not all annotations were identical, as inter-annotator agreement measured by Cohen's κ was 0.6, but in the final dataset each record even

Article in Press

if can repeated has a unique annotation (for instance out of 545 records in the test set only three records are repeated for a total of 6 records)

Table 2. IRIDE ISTAT dataset (left) Gender-based violence (right). Source: Istat (2020).

Emotion	Relative share	Relative share
Love	3.9 %	24.2 %
Joy	10.8 %	19.3 %
Neutral	27.5 %	9.1 %
Fear	14.4 %	2.2 %
Anger	19.0 %	25.6 %
Sadness	15.9 %	4.4 %
Surprise	8.5 %	15.2 %

A second, smaller dataset (GBV) of 363 tweets was annotated with a stronger focus on gender-based violence (e.g. posts containing keywords such as *rape*, *killing*, and *woman*). Relative shares of these dataset are reported in Table 2.

3. Methodology

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model introduced by Devlin et al. (2019). It is based on the Transformer encoder architecture and is trained using a masked language modeling objective, which allows the model to learn deep bidirectional contextual representations of text. Unlike earlier sequential models, BERT processes the entire sentence simultaneously, enabling it to capture contextual dependencies between words more effectively. After pre-training on large-scale corpora, BERT can be fine-tuned on downstream tasks such as text classification, named entity recognition, and question answering with relatively limited labeled data.

AlBERTo (Polignano et al., 2019) is an Italian adaptation of BERT specifically designed for social media language. It is based on the RoBERTa architecture and was pretrained on a large corpus of Italian tweets (≈200 million). This domain-specific pretraining makes AlBERTo particularly suitable for handling informal language, hashtags, abbreviations, and other platform-specific linguistic phenomena. It is available on github[1] and was selected following the EVALITA campaign results (Araque et al., 2023; Polignano et al., 2019). The training conditions of the two systems are not fully comparable. IRIDE relies on proprietary pretrained embeddings and an existing sentiment analysis architecture that was already optimized before its specialization for the Istat project. In contrast, the open-source model was trained starting

[1] https://github.com/marcopoli/AlBERTo-it

Article in Press

from a publicly available pretrained language model without prior task-specific adaptation. For this reason, additional training data were required: the EMit dataset was used to increase the number of emotion-labelled examples, while the GBV dataset was used in a second fine-tuning stage to enrich the model with domain-specific signals. AlBERTo was fine-tuned through a two-stage training procedure to obtain an emotion classifier. Training follows a two-stage fine-tuning strategy. In the first stage, the model is fine-tuned on a mixed social media corpus (EMit and ISTAT data) using the IRIDE-class emotion inventory. The task is formulated as single-label, multi-class emotion classification of Italian social media texts using a Transformer-based encoder architecture. The datasets were harmonized to ensure consistency across sources. Specifically, the category *disgust* was merged with *anger*, while *anticipation* and *trust* were excluded from the analysis. Any remaining multi-label records were also removed so that only single-label observations were retained. In addition, to mitigate overfitting observed during preliminary experiments, only unique texts were considered in the training data. This preprocessing resulted in training dataset of 4326 records following the distribution reported in Table 3, (for EMIT validation set was set to 20% of the original training set reported in Table1, while for IRIDE we used the validation set described in 2.2)

Table 3. Training (including validation) dataset and test set (EMit+IRIDE)

Class	Count Training	Relative share	Count Tesr	Relative share
Love	289	5.5%	60	5.4%
Joy	613	11.7%	125	11.3%
Neutral	1804	34.3%	359	32.3%
Fear	347	6.6%	85	7.7%
Anger	1034	19.7%	233	21.0%
Sadness	481	9.1%	106	9.5%
Surprise	692	13.2%	143	12.9%
Total	5260		1111	

In the second stage, the model is further fine-tuned on a Gender based violence (GBV)-specific dataset to specialize the model for online violence, although very reduced, with conservative hyper-parameters.

Optimization is performed using the AdamW optimizer with a cosine learning rate schedule and linear warmup. Gradient clipping is applied to control the norm of the gradients, while early stopping based on validation loss is used to prevent overfitting. Training efficiency was improved through mixed-precision computation (FP16) and gradient accumulation to obtain larger effective batch sizes while reducing GPU memory usage. To reduce the number of trainable parameters — given limited hardware resources and training data — Low-Rank Adaptation (LoRA) was employed (Hu et al., 2021): pretrained weights were frozen and small

Article in Press

low-rank trainable matrices were injected into selected attention projection layers; only these additional parameters were optimized

Model performance is evaluated on held-out test sets using accuracy and macro-F1 score.

4. Results

If we compare the accuracy, we have that the first-stage open-source (O-S) model and IRIDE reach respectively and accuracy of 77.9 % vs 71.0 % while macro F1 score is 68.3 vs 70.0. However, if we split results by class, we get very different results as shown in Table 4.

Table 4. Model performance evaluation

Class	Precision		Recall		F1	
	O-S	IRIDE	O-S	IRIDE	O-S	IRIDE
Anger	75.50	82.1	84.5	83.7	79.7	**83.0**
Fear	96.30	92.5	85.9	78.5	**90.1**	85.0
Joy	83.10	68.8	67.4	74.6	**74.4**	72.0
Love	50.00	69.6	31.0	76.2	**38.3**	**73.0**
Neutral	84.20	94.4	84.6	90.0	84.4	**92.0**
Sadness	72.20	71.7	80.1	87.4	76.1	**78.8**
Surprise	84.60	78.4	72.1	63.0	**77.9**	69.9

Even if the test set differ by several reasons (EMit included, duplicates removed, inconsistent labels removed) except that for *love* which provides poor results in our simulation (small support and probably training on sub-samples with low resources is not efficient) we have that IRIDE outperforms anger, love, neutral and sadness, while our model has a better F1 score for fear, joy and surprise The O-S model achieves in the second fine-tuning stage an accuracy of 77.8% and macro-F1 of 73.5 on the GBV dataset.

We then analyzed predictions of the two models for 2023Q4 (≈150k posts) gathered from FB, Instagram and X. The resulting emotion distributions predicted by the two models are reported in Table 5 together with their main cross-model correspondences.

Table 5 Predicted emotion distribution on 2023Q4 social media corpus (% Relative share)

Class	Total predicted		Anger IRIDE	Joy O-S	Sadness IRIDE	Neutral IRIDE	Neutral O-S
	IRIDE	O-S	O-S Pred	O-S Pred	O-S Pred	O-S Pred	IRIDE Pred
Love	3.7%	5%	6.2%	3.2%	1.7%	3.6%	1.9%
Joy	12.9%	**40.5**	**21.4%**	36.9%	**32.1%**	39.7%	9.2%
Neutral	33.1%	32.9	**38.6%**	34.7%	**28.2%**	41.5%	54.4%
Fear	0.7%	0.3	0.3%	0.3%	0.3%	0.1%	1.2%

Article in Press

Anger	**22.0%**	8.9%	17.0%	11.3%	10.7%	7.7%	**23.8%**
Surprise	4.5%	1.0%	0.5%	0.8%	0.8%	1.6%	2.1%
Sadness	**23.0%**	11.8	16.0%	12.9%	26.2%	5.6%	**7.4%**

The two models' predicted emotion distributions diverge substantially — the O-S model overestimates joy, whereas IRIDE yields higher shares of anger and sadness, which align more closely with expectations for GBV-related discourse. By analyzing how joy O-S is predicted by IRIDE we observe that 74.7 records are in "adjacent" classes such as love, neutral, joy, the rest being probably wrong. This is somehow in contrast with the higher F1 score attained. If we analyze the world clouds of the posts' predicted as joy by O-S and neutral by IRIDE (see Figure 1) we get that most of them are about the demos' for the international day against violence, which can be indeed positive (we are against violence, we want to eliminate violence), while the ensemble joy-O-S and anger IRIDE seems more negative (see figure 1). We then most probably deduce the algorithm often predicts joy instead of neutral. Conversely 60% and 60.3% of the records classified respectively as anger and sadness by IRIDE are either neutral or joy for O-S. Indeed 31.4% of records classified as Neutral by O-S are predicted as anger or sadness by IRIDE. It is worth to finally notice that even if the O-S model achieves a poor F1 score on *love* both models recognize moreover the same amount of records

The two models therefore produce substantially different prediction distributions despite having similar overall accuracy and macro-F1 scores. It seems that the two models apply a different decision thresholds. A random human-annotated sample would be needed to assess which distribution better reflects the true labels. Predictions of the O-S model led to an unrealistic emotion distribution because in GBV-related discourse, negative emotions are naturally more prevalent. This may be due to the presence of Emit dataset that introduces noise for the model, or differences among annotator guidelines across datasets. Both aspects likely affect model thresholds and class bias.



Figure 1. word clouds for posts predicted as joy by O-S and anger (left) neutral (right) by IRIDE.

Article in Press

5. Conclusions

This study evaluated whether an open-source transformer-based model could provide a transparent alternative to a proprietary emotion detection system currently used in the Istat monitoring framework for gender-based violence discourse. Results indicate that competitive benchmark performance can be achieved using open architectures and limited computational resources. However, the comparison also reveals an important limitation: models with similar benchmark performance may produce substantially different prediction distributions when deployed at scale.

The divergence observed between the open-source model and IRIDE suggests that conventional supervised metrics such as accuracy and macro-F1 may not be sufficient to assess deployment validity in socially sensitive domains. Although the open-source model achieved strong benchmark results, its tendency to assign emotionally ambiguous content to positive categories raises questions regarding domain adaptation and calibration.

These findings highlight both the opportunities and challenges of integrating transparent NLP systems into official statistics. Open-source approaches improve reproducibility, auditability, and methodological flexibility, but additional validation strategies—including human annotation, calibration analysis, and qualitative error assessment—remain necessary before deployment in operational contexts. Future work should focus on expanding domain-specific annotations and strengthening validation procedures to support fully transparent and reproducible NLP pipelines for monitoring online discourse related to violence against women.

References

Araque, O., S. Frenda, R. Sprugnoli, D. Nozza, and V. Patti (2023). “EMit at EVALITA 2023: Overview of the Categorical Emotion Detection in Italian Social Media Task”; G. Calò, F. Massafra, B. De Carolis, C. Loglisci, TASK A at EVALITA 2023: Overview of the Emotion Hunters Approach to the Categorical Emotion Detection in Italian Social Media;Hromei, C.d., Croce, D., Basile, V., Basili, R. (2023). ExtremITA at EVALITA 2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme

. *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2023) – CEUR Workshop Proceedings. Vol. 3473. CEUR Workshop Proceedings, 1–8.* url: https://ceur-ws.org/Vol-3473/paper1.pdf.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). *“*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 4171–4186.* url: https://arxiv.org/abs/1810.04805

Hu, E. J. et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. url: https://arxiv.org/abs/2106.09685

Article in Press

Polignano P, Basile.P, De Gemmis M., & Basile.V AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, pages 312–317

Article in Press