Emotion Detection on hate speech and gender-based violence texts on social media

Sergio Di Donfrancesco

ISTAT | ENSAI

September 18, 2025

Outline

Outline

Context and Motivation

- Gender-based violence and online hate speech are widespread on social platforms.
- European monitoring shows that cyber violence already affects many women and girls.
- Risks are higher for women in public-facing roles.

Exposure to online harms by gender

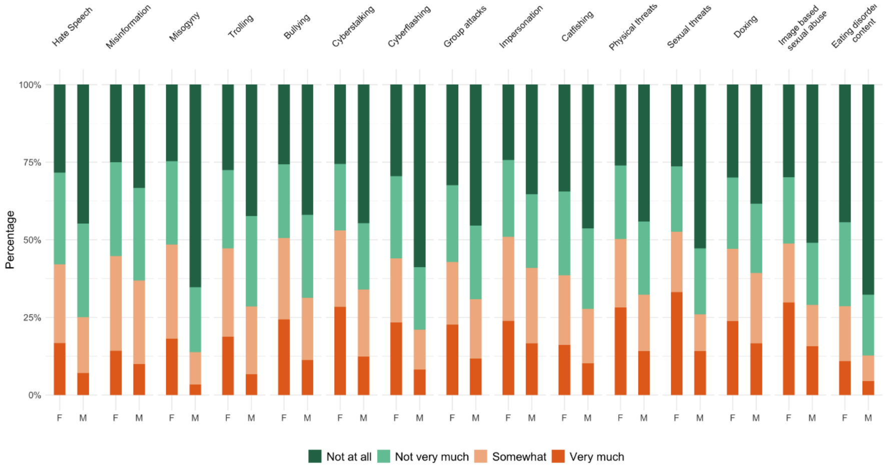


Figure: Self-reported fear of harmful online content by gender.

Research growth on online misogyny

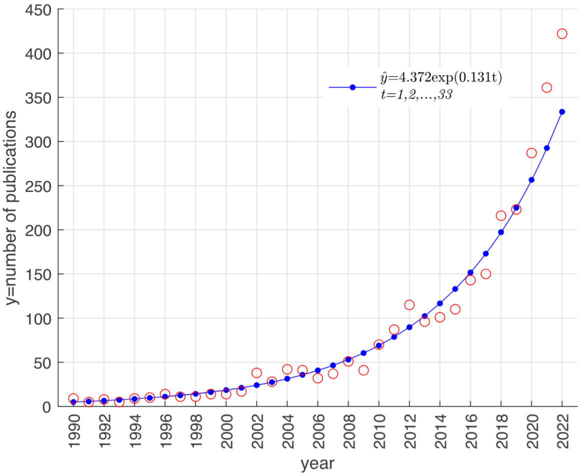


Figure: Publications per year on misogyny; observed and exponential fit.

Thesis Objectives

- Build and assess a robust Italian emotion classifier for social-media text, with a specific focus on GBV-related content.
- Study learning under data constraints: quantify the effect of class-imbalance handling and small-sample regimes, and make explicit the trade-offs between stability and capacity.
- Deliver a reproducible, extensible pipeline that can be updated when new GBV annotations become available.
- Justify design choices against these objectives through comparative results.

Outline

EMit Shared Task: Scope and Labels

- EMit is the first shared task on categorical emotion detection for Italian social media.
- Emotion detection, originally multi-label over 10 emotions.
- Our project adopts EMit's label set and framing, but trains in a single-label setting.
- Official ranking: macro-averaged F1 across classes (robust under class imbalance).

Outline

Task and Label Spaces

- GBV dataset adopts a reduced inventory because *anticipation* and *trust* are absent in ISTAT annotated sets.
- The classification head is reconfigured accordingly while keeping the same encoder and tokenizer.
- Kept 9 classes in general trainng for future-proofing and not losing data.

Pipeline

- Cleaning and normalization of datasets.
- Deduplication and conflict resolution, label harmonization, stratified splits.
- Two-stage training: Stage 1 on EMit+ISTAT general set and Stage 2 on GBV set that continues from the best checkpoint.
- Evaluation and logging: accuracy and macro-F1 on test sets. Experiment tracking and reproducibility.
- Inference on a large unlabeled Q4-2023 stream (Facebook/Instagram/Twitter) and comparison with IRIDE tags.

Outline

Initial Datasets and Splits

- EMit: Multi-annotator scheme; release with regenerated IDs and GDPR compliance.
- ISTAT datasets have predefined train/valid/test splits and three folders: all (general), mix (alternative general sampling), gbv (GBV-related). Overlaps across folders handled explicitly during integration.

Pre-processing and Integration

- Filter EMit train/test to single-label instances.
- Join ISTAT all and mix folders per split; normalize, deduplicate exact matches, and drop conflicting texts.
- Label harmonization: collapse *Disgust* into *Anger* in EMit; map ISTAT Italian labels to the EMit English inventory for a shared label space.
- Stratified split on EMit train to derive a validation set preserving class proportions.
- Merge EMit with ISTAT general for each split; keep GBV (ISTAT only) separate.
- Note: GBV lacks *Anticipation* and *Trust*.

Datasets sizes

- Final post-integration datasets sizes:

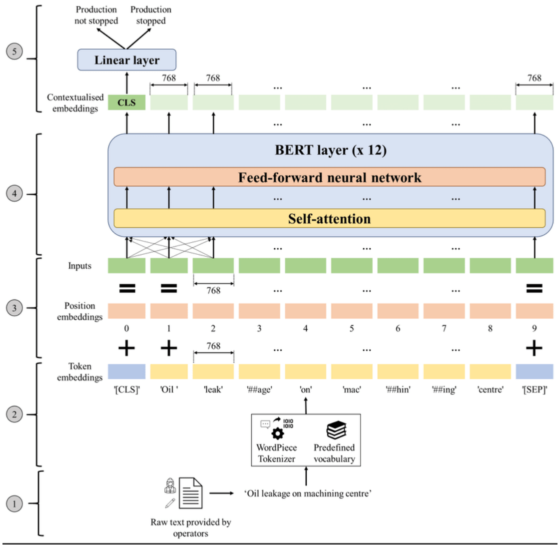
Split	General (ISTAT+EMit)	GBV (ISTAT only)
Train	5 461	366
Valid	1 347	64
Test	781	108

Outline

LLMs, Transformer, and BERT

- Large Language Models learn general-purpose linguistic representations by predicting tokens over very large corpora.
- Transformer encoders replace recurrence with multi-head self-attention + feed-forward layers, with residuals and layer norm for stable training.
- Encoder-only models read full bidirectional context, which suits short, noisy social posts.
- BERT for sequence classification: take the pooled `[CLS]` representation $\rightarrow$ linear head $\rightarrow$ softmax; train with cross-entropy.
- Why using BERT: emotion detection needs robust encoders, not text generation; BERT fine-tunes reliably with modest compute and has strong Italian checkpoints.

BERT architecture example



Italian Backbones: AlBERTo vs UmBERTo

- Both encoder-only Transformers pre-trained with masked language modeling; they differ mainly in pre-training data, tokenization, and resulting inductive biases.
- Pre-training domain: AlBERTo is trained primarily on Italian social media; UmBERTo is trained on broader, general-domain Italian.
- Tokenization: both use subwords, but vocabularies differ; this affects how emojis, hashtags, user mentions and creative spellings are segmented.
- Trade-offs and choice: AlBERTo aligns better with noisy social language, capture cues helpful for minority emotions in social contexts; weaker coverage on formal registers.
- UmBERTo offers stronger coverage on standard Italian, may underfit highly informal slang.
- We prefer AlBERTo for this task while using UmBERTo as a meaningful comparator.

Outline

Stage 1: General Training

- Optimization: AdamW with cosine scheduler and warmup ratio 0.06; label smoothing $= 0.03$; gradient clipping.
- Imbalance handling: tried weighted CE $\rightarrow$ switched to light *dynamic* oversampling (early boost to minority classes, then anneal) to stabilize learning without overfitting.
- Capacity control: partial freezing at start (lower blocks + embeddings) for stability, then unfreeze to recover capacity.

Stage 2: GBV Training

- Initialize from best Stage-1 checkpoint and same tokenizer; adapt head.
- Conservative learning rate; keep cosine + warmup 0.06; lighter regularization; selective freezing in the first epochs.
- Dynamic oversampling with small factors to mitigate scarcity while containing memorization risk.

Imbalance and Data Regime

- Weighted CE increases penalty on rare classes but can destabilize optimization and degrade calibration when weights are large.
- Simple oversampling boosts minority gradients but shifts train vs validation distribution and can speed up overfitting.
- Final choice: light dynamic oversampling (early boost, annealed later) $\Rightarrow$ steadier curves and better minority recall without inflating validation metrics.

Regularization and Stability

- Label smoothing $= 0.03$: discourages overconfidence, improved calibration without hurting separability.
- Early stopping on macro-F1: aligns the stopping rule to the target metric under class imbalance.
- Gradient clipping: avoids large, unstable steps in early epochs; leaves small updates unchanged.

Trainable Capacity: Freezing vs LoRA

- Full fine-tuning of $\sim$ 110M params risks overfitting and adds compute on modest datasets.
- Considered LoRA (low-rank adapters) to reduce trainable params;
- In the end, preferred partial freezing warm-up for simplicity and fewer hyperparameters.
- Effect: reduced variance at start, faster stabilisation of validation F1, then recovery of capacity when unfreezing deeper layers.

Trainable Capacity: Freezing vs LoRA

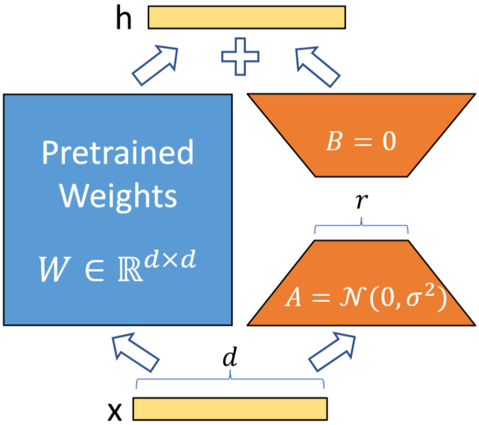


Figure: Schematization of the LoRA decomposition.

Optimizer and LR Policy

- AdamW: decoupled weight decay improves generalization and predictable tuning on BERT-like encoders.
- Cosine schedule with warmup 0.06: smoother late updates than linear, helpful for minority classes; warmup stabilizes head/upper layers before larger steps.
- Base LR: tuned per stage (general slightly higher; GBV more conservative) to limit drift from a well-initialized checkpoint.

Optimizer and LR Policy

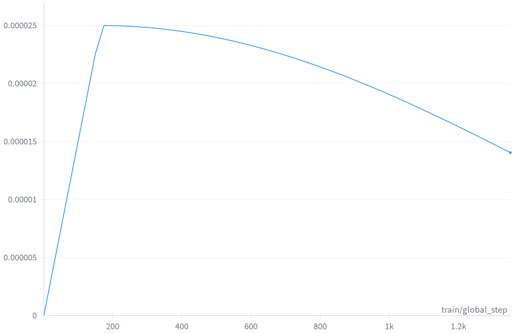


Figure: Learning rate evolution during training.

Outline

Evaluation setup and metrics

- Models are evaluated on their respective test splits; we report per-class precision, recall, F1, and macro-F1.
- GBV test is very small and imbalanced: per-class metrics are volatile; comparisons are indicative rather than definitive.
- Interpretation focuses on class-wise behaviour.

General test: classification report

Class	Precision	Recall	F1	Support
Anger	0.7547	0.8451	0.7973	142
Anticipation	0.5714	0.4000	0.4706	10
Fear	0.9634	0.8587	0.9080	92
Joy	0.8312	0.6737	0.7442	95
Love	0.5000	0.3103	0.3830	29
Neutral	0.8419	0.8458	0.8438	214
Sadness	0.7222	0.8053	0.7615	113
Surprise	0.8462	0.7213	0.7788	61
Trust	0.3556	0.6400	0.4571	25
Macro avg	0.7096	0.6778	0.6827	781
Weighted avg	0.7903	0.7785	0.7796	781

GBV test: classification report

Class	Precision	Recall	F1	Support
Anger	0.9200	0.8214	0.8679	28
Fear	1.0000	0.5000	0.6667	2
Joy	0.6818	0.7500	0.7143	20
Love	0.7097	0.8148	0.7586	27
Neutral	0.6667	0.6000	0.6316	10
Sadness	0.9333	0.8750	0.9032	16
Surprise	0.6000	0.6000	0.6000	5
Macro avg	0.7874	0.7087	0.7346	108
Weighted avg	0.7885	0.7778	0.7794	108

Synthesis across splits: strengths and differences

- Consistent strengths on high-represented emotions: Anger and Sadness are well captured in both settings; Fear is strong in the general set with adequate support.
- Neutral remains stable when support is large; this anchors macro-level performance and reduces variance.
- GBV adaptation improves focus on GBV-salient emotions (Anger, Sadness) while preserving the general encoder's robustness.
- Macro-F1 levels are comparable across settings, indicating effective transfer despite domain shift and smaller GBV data.

Synthesis across splits: weaknesses, similarities, conclusions

- Persistent weaknesses for Anticipation and Trust (general only); Love can be confused with Joy; Neutral boundaries stay fuzzy in GBV news/reporting.
- Rare classes show volatile per-class metrics; imbalance remains the main driver of variance even with dynamic oversampling.
- Similar error patterns across splits suggest a need for richer supervision and clearer operational definitions.

Outline

Testing the model on a new dataset

- Italian social media posts related to GBV from the last trimester of 2023.
- Data collected from Twitter, Instagram and Facebook.
- Emotions classified by IRIDE®, workflow and details are mostly unknown.

Comparison between models: Confusion Matrix

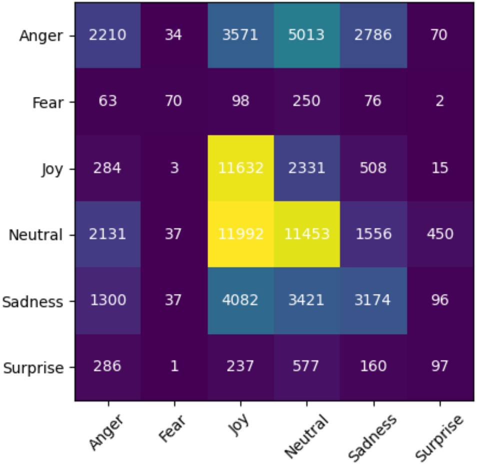


Figure: Confusion matrix between IRIDE (rows) and our model (columns) predictions.

Comparison between models: Word Cloud

JOY vs ANGER

Figure: Word Cloud for texts predicted "Joy" by our model and "Anger" by IRIDE's model.

Outline

Criticalities and Improvements

- Data scarcity and imbalance (especially GBV): rare classes unstable and high variance in per-class metrics.
- Need for more annotated and recent data, with a focus on GBV: include under-represented emotionsand clarify boundaries between positive and neutral classes.
- Move from single-label to multi-label classification when texts clearly convey mixed affect.
- Extend from emotion recognition to the automatic detection of GBV-specific signals to increase operational usefulness.

Operational Takeaways and Outlook

- A first reliable prototype for Italian emotion detection with a GBV focus, based on a reproducible pipeline and motivated design choices.
- Operational perspective: human-in-the-loop pipeline for periodic updates, drift monitoring, and transparent reporting to support monitoring use cases.

Thank you for the attention!