HateItOff at MultiPRIDE: Linguistic and Sentiment Cues in Reclaimed LGBTQ+ Slur Detection*

Greta Damo[1,*,†], Nicolás Benjamín Ocampo[2,*,†]

[1]Université Côte d'Azur, CNRS, Inria, I3S
Route des Colles 930, 06903 Sophia Antipolis, France
[2]Centrum Wiskunde & Informatica
Science Park 123, 1098 XG Amsterdam, The Netherlands

Abstract

Reclaimed slurs challenge automatic hate speech detection, as terms historically used against the LGBTQ+ community may be reappropriated with positive or self-identifying intent. Traditional lexicon-based approaches often fail to capture these nuances, leading to misclassification errors. To address this challenge, this paper presents our participation in the MultiPRIDE shared task at EVALITA 2026. We addressed both textual (Task A) and contextual (Task B) reclaimed language detection for Italian and Spanish. We explored transformer-based models enriched with sentiment analysis information and linguistic features. Our results demonstrate that these features improve performance over the baselines. In Task A, our system ranked 8th out of 24 submissions for Italian and 6th out of 19 submissions for Spanish. In Task B, we achieved 11th place out of 18 submissions for Italian and 5th place out of 12 submissions for Spanish. An error analysis highlights persistent challenges related to pragmatic ambiguity, in-group criticism, and meta-linguistic uses of slurs.

Keywords

Hate speech, Reclaimed slurs, Social Media

Content Warning: This paper contains examples of language which may be offensive to some readers.

1. Introduction

Social media has become a primary medium for communication and self-expression, enabling users to share opinions, identities, and experiences at an unprecedented scale. For the LGBTQ+ community, online platforms can represent spaces of visibility and empowerment, but also environments where language is contested and meaning is highly context-dependent. In particular, terms historically used as slurs may be reappropriated and reclaimed by community members, acquiring positive or self-identifying meanings that differ substantially from their derogatory use [1, 2, 3, 4]. This linguistic phenomenon poses a significant challenge for automatic language processing systems [5]. Standard approaches for hate speech and abusive language detection often rely on offensive terms or expressions that are commonly used in hateful messages [6]. As a consequence, systems may incorrectly classify reclaimed language as harmful, leading to biased moderation outcomes and the silencing of marginalized voices. Addressing this issue requires models capable of capturing subtle semantic and contextual cues beyond the presence of potentially offensive terms [7, 8, 9]. The MultiPRIDE shared task [10] was proposed to foster research on reclaimed language within the LGBTQ+ community. The task formulates a binary classification problem aimed at determining whether a term related to the LGBTQ+ context in a message is used with a reclamatory intent or not. MultiPRIDE is articulated into two main tasks: Task A, which focuses on the textual content of the message, and Task B, which additionally allows the use of contextual information derived from the author's profile. Both tasks are further divided into language-specific subtasks covering Italian, Spanish, and English.

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT
*Corresponding author.
†These authors contributed equally.
greta.damo@univ-cotedazur.fr (G. Damo); n.b.ocampo@cwi.nl (N. B. Ocampo)
https://grexit-d.github.io/damo.greta.github.io// (G. Damo); https://www.nicolasbenjaminocampo.com/ (N. B. Ocampo)
0009-0009-8204-9513 (G. Damo); 0009-0001-0077-4626 (N. B. Ocampo)

In this paper, we describe our participation in the MultiPRIDE shared task, addressing both Task A (textual content) and Task B (contextual content) for Italian and Spanish. We extend transformer-based models by incorporating two additional feature sets. First, we hypothesize that reclaimed slurs are frequently used with non-negative sentiment. To capture this aspect, we integrate an external sentiment analysis tool and include both the predicted sentiment label and its confidence score as additional model features. Second, we hypothesize that linguistic patterns such as the use of first-person pronouns and specific verb types may signal reclaiming intent. We therefore automatically detect these linguistic cues, calculate their frequency, and embed them as model features. Our results show consistent improvements over the baseline in Task A for both Italian and Spanish. In Task B, improvements are observed for Spanish, though not for Italian. We rank 8th out of 24 submissions for Italian and 6th out of 19 submissions for Spanish in Task A. In Task B, we place 11th out of 18 submissions for Italian and 5th out of 12 submissions for Spanish. These results suggest that the proposed features contribute positively to reclaimed slur detection and can enhance model performance beyond standard transformer-based approaches.[1]

2. Methodology

In this section we describe the data used for the task, that were provided by the organizers, and how we splitted them into train, dev, test, and the systems we utilized.

2.1. Dataset

The datasets used in this work were provided by the organizers of the MultiPRIDE shared task. The multilingual corpus covers three languages—Italian, Spanish, and English—and was constructed by combining existing resources collected between 2020 and 2022 from social networks, blog posts, and TV series. For Italian, data were sourced from the TWITA collection [11], while Spanish data were drawn from a dataset of LGBTQ+-related messages [12]. The English portion aggregates content from multiple sources, including Twitter, Reddit, and TV series dialogues [13]. To build the final datasets, the organizers applied a consistent filtering methodology across languages. First, keyword-based filtering was performed using terms related to homosexuality extracted from the Hurtlex lexicon [14]. Subsequently, the selection was refined by identifying messages more likely to contain reappropriative language, based on the presence of positive terms associated with pride and community belonging (e.g., pride, queer, LGBT, rainbow). The final datasets consist of 1,811 instances for Italian, 1,461 for Spanish, and 1,711 for English. For our experiments, we focused on Italian and Spanish, in line with our participation in both Task A and Task B for these two languages. Starting from the training data released by the organizers, we created *train*, *development*, and *test* splits combining the instances of the 3 languages provided. We use a stratified sampling strategy to preserve label distribution. Specifically, we allocated 70% of the data for training, and split the remaining 30% evenly into development and test sets. All splits were generated using a fixed random seed to ensure reproducibility. Table 1 shows the label distribution of each split. Throughout the paper, we use the terms training, development, and test sets to refer to the data splits derived from the portion of the dataset released by the organizers during the "Data Release" phase of the MultiPRIDE task. We refer to the split retained by the organizers and used during the "Evaluation Window" as the *submission* set. It is important to note that this set is not accessible to participants, and that only two systems per task and per language could be submitted for evaluation. Moreover, although we officially participated in the Italian and Spanish tracks, we also included English-language instances in our data splits in order to leverage them during training. This decision was motivated by the very limited number of instances containing reclaimed slurs: by combining data from all three languages, we obtained 300 such instances to be used as training material. However, we decided to participate for only Spanish and Italian.

[1]The code, trained models, and instructions for reproducing our experiments are publicly available at: https://github.com/grexit-d/HateItOff_MultiPRIDE.

Lang	Label	Train		Dev		Test		Submission
		#	%	#	%	#	%	#
EN	recl	62	0,0864	13	0,0844	13	0,0844	685
	non-recl	656	0,9136	141	0,9156	141	0,9156	
ES	recl	93	0,1517	20	0,1527	20	0,1515	585
	non-recl	520	0,8483	111	0,8473	112	0,8485	
IT	recl	145	0,1908	31	0,1902	31	0,1902	725
	non-recl	615	0,8092	132	0,8098	132	0,8098	

Table 1
Data splits and label distribution for text messages labeled as reclaimed (recl) and non reclaimed (non-recl) of the MultiPRIDE Shared Task dataset.

2.2. Description of the System

For our participation in the MultiPRIDE shared task, we submitted two systems per task and per language. We adopted an ablation-based approach to identify the most suitable systems for submission. Our ablation study starts from a vanilla transformer-based model and incrementally incorporates additional features. These features fall into two main categories: linguistic features and sentiment features. Each resulting system configuration is described below.

Baseline. All the proposed systems rely on pre-trained transformer encoders fine-tuned for binary sequence classification. Given an input message, the model predicts whether a term related to the LGBTQ+ context is used with a reclamatory intent.

Sentiment-enriched models. We hypothesize that the use of reclaimed slurs tends to rely on non-negative sentiment. In this direction, we trained transformer models with injected `sentiment label` and `sentiment score` from the sentiment analysis model `tabularisai/multilingual-sentiment-analysis`. The predicted sentiment label and/or the sentiment score are appended to the original input text in the form of a natural-language prompt. This enables the classifier to use sentiment cues without modifying the transformer encoder architecture.

Linguistically-enriched models. We hypothesize that reclamation is often associated with the use of first-person personal pronouns, such as *"yo/nosotros"* in Spanish and *"io/noi"* in Italian, as well as with specific call-to-action expressions, for example *"Speak up, queer voices matter."*. Therefore, we experimented with linguistic features automatically extracted using spaCy, including counts and normalized ratios of personal pronouns (first, second, and third person), the number of verbs, and the presence of imperative forms. The extracted linguistic features are concatenated with the transformer's [CLS] representation and fed into a feed-forward classification head.

Implementation details and ablation study. We used BERT, ModernBERT, and XLM as our baseline models. These models serve as backbones, later enriched with the three strategies described: `sentiment label`, `sentiment score`, and `linguistic cues`. We selected these models to maintain simple, adaptable strategies that could later incorporate additional features. To measure the impact of each strategy while minimizing noise, we preferred these transformer models over LLMs. The ablation study considers the set of strategies $S = \{\texttt{sentiment label}, \texttt{sentiment score}, \texttt{linguistic cues}\}$ and all possible subsets of S, i.e., the power set $\mathcal{P}(S)$. Using the set of backbone models $M = \{\texttt{BERT}, \texttt{ModernBERT}, \texttt{XLM}\}$, we compute the Cartesian product of M and $\mathcal{P}(S)$. Considering two tasks, A and B, this results in 3 models × 8 strategy subsets × 2 tasks = 48 runs.

All models were implemented using the Hugging Face `Transformers` library. For BERT, ModernBERT, and XLM models, we used the `bert-base-uncased`, `answerdotai/ModernBERT-base`, and `cardiffnlp/twitter-xlm-roberta-base` versions available on Hugging Face, respectively. Training and evaluation were performed with a batch size of 8. Models were evaluated and checkpointed at the end of each epoch, retaining up to two checkpoints, and the best-performing model was selected based on the macro-averaged F1-score on the development set. Training logs were recorded every 50 steps. All experiments were conducted with a fixed random seed on a single NVIDIA A100 GPU. Experiments

Task	Submitted Systems	Spanish			Italian		
		no-recl	recl	macro	no-recl	recl	macro
A ES	S1 BERT + ling. cues	0,9335	0,5405	0,7370	-	-	-
	S2 ModernBERT + sent. label. + sent. scores	0,9234	0,3651	0,6442	-	-	-
A IT	S1 BERT + ling. cues	-	-	-	0,9571	0,8046	0,8809
	S2 UmBERTo + ling. cues	-	-	-	0,9549	0,7619	0,8584
B ES	S1 BERT + ling. cues + sent. label	0,9247	0,4762	0,7005	-	-	-
	S2 XLMT + sent. label + sent. scores	0,9280	0,4186	0,6733	-	-	-
B IT	S1 ModernBERT + sent. label	-	-	-	0,9416	0,7222	0,8319
	S2 ModernBERT + sent. label. + sent. scores	-	-	-	0,9497	0,7426	0,8462

Table 2
Results of System 1 and 2 on tasks A and B for Spanish and Italian on the submission test set.

were run for both 3-epoch and 6-epoch training schedules, resulting in $48 \times 2 = 96$ runs. In addition to the ablation study, we also evaluated for Task A four additional configurations using UmBERTo, an extension of the RoBERTa model for Italian trained with Whole Word Masking. The evaluated configurations include UmBERTo with `sentiment label` only, `linguistic cues` only, and the combination of `sentiment label` and `linguistic cues`. The motivation behind these experiments was to assess whether an Italian-specific language model could further improve performance on the task. In total, we conducted 100 runs for all the systems proposed in this shared task.

3. Results

The results of the ablation study across all experimental runs are reported in Table 3 for the 3-epoch setting and Table 4 for the 6-epoch setting (both tables are provided in Appendix A). The results indicate that certain features, such as `linguistic cues` and `sentiment labels`, have a positive impact on the prediction of slur reclamation, particularly for BERT and ModernBERT models. Regarding the differences between Task A and Task B, performance for Spanish decreases when contextual information is included, whereas for Italian it remains unchanged, yielding comparable results with and without context. In our experiments, incorporating users' *BIO* information did not contribute to improved detection performance.

We selected models for submission based on both performance over the baseline and system diversity. Although we could have chosen the two best-performing configurations for each task and language, we aimed to reduce the risk of selecting models that performed well on our internal test split but might not generalize to the submission set.

Table 2 shows the selected systems with their results on the *submission* set. We can see that models perform particularly well on reclaim slur detection for Italian, achieving a macro score of 0.8809, with the S1 BERT + linguistic cues system being the best. For Spanish, the best system achieves a macro score of 0.7370. We also observe that our models generally perform better when no context is used as seen in the ablation study.

4. Error Analysis

4.1. Italian Language

Manual inspection of the misclassified Italian messages reveals that errors can be grouped into a limited number of recurring categories. These categories highlight both annotation ambiguities and modeling limitations when dealing with reclaimed language.

Borderline or Ambiguous Reclaimed Usage. Some messages contain slur-related terms in contexts that are neither clearly positive nor explicitly negative. In these cases, reclamation is ambiguous and highly dependent on pragmatic interpretation, making both annotation and automatic classification difficult. Examples:

1. *”sono sempre loro che ci adescano mannaggia a loro, con le loro frociate...”* (it_887);
2. *“Che carini tutti questi che sono spuntati adesso a difendere i diritti dei froci.”* (it_502).

These messages adopt an ironic or critical tone, and the reclaimed intent is not explicit, suggesting potential inconsistencies in the gold labels.

Reclaimed Slurs Combined with Negative or Aggressive Language Several instances show reclaimed self-references co-occurring with insults, profanity, or aggressive discourse directed at others. While the slur may be reclaimed, the overall negative tone can mislead sentiment- or toxicity-aware models. Examples:

3. *“Orgoglioso di essere ‘frocio’! Omofoba di merda!”* (it_1750);
4. *“Ripetiamolo insieme non potete dire frocio se non siete lgbtq+... Are u that stupid?”* (it_474);
5. *“Ci sono ‘finocchi’ che lo fanno orgogliosamente col culo proprio.”* (it_695).

These cases highlight the difficulty of disentangling reclaimed identity markers from general offensive language.

In-Group Criticism and Community-Internal Discourse Some misclassified messages contain reclaimed slurs used within critical reflections on the LGBTQ+ community itself. Although reclamation is present, the critical stance complicates interpretation. Examples:

6. *“Omofobia interiorizzata da parte di froci che si credono etero...”* (it_1057);
7. *“alle frocie queer non le fanno né partecipare...”* (it_1242).

Here, reclaimed terms are used for internal critique rather than affirmation, a phenomenon that current models struggle to capture.

Quotation, Meta-Discussion, and Explanatory Uses In these cases, slurs appear inside quotations, explanations, or meta-linguistic discussions about their meaning or usage, rather than being directly employed as insults or self-labels. Examples:

8. *“il termine frocio sia attribuito a qualcuno negativamente...”* (it_1179);
9. *“‘Frocio’ me l’hanno fatto uscire dalle orecchie quando andavo a scuola”* (it_1401);
10. *“Nuovo epiteto ricevuto degno di nota...”* (it_979).

These uses are typically reclaimed or reflective, but surface-level lexical cues can trigger misclassification.

Pragmatic and Contextual Cues Beyond Textual Content Some messages rely heavily on emojis, hashtags, or shared cultural context (e.g., Pride events), which are not always fully exploited by text-based models. Examples:

11. *“Buongiorno, specialmente ai miei amici froci...* 🏳️‍🌈 🏳️‍⚧️ ❤️ *”* (it_631);
12. *“*🏳️‍🌈🏳️‍🌈🏳️‍🌈 *Fatelo per lui...”* (it_1401).

Although clearly supportive or reclaimed to a human reader, these signals may be underweighted by the model.

Community Membership Without Explicit Reclaimed Slurs Finally, some false positives occur in messages where the author self-identifies as part of the LGBTQ+ community, but no reclaimable slur is actually used. Examples:

13. *“un’attrice lella che seguo è incinta...”* (it_1222);
14. *“Le persone trans sono a volte guidate dall’omofobia...”* (it_269).

These cases suggest that models may over-rely on topical cues or community-related vocabulary.

Overall, the error analysis shows that misclassifications are often due to pragmatic ambiguity, mixed sentiment, and meta-linguistic usage of slurs. These findings confirm that reclaimed language detection cannot be reduced to lexical or sentiment-based signals alone and requires deeper modeling of speaker intent, discourse function, and community context.

4.2. Spanish Language

We also inspected the misclassified Spanish messages. They can be grouped into the following recurring categories:

Criticism and Non-Textual Content In Spanish, messages containing reclaimed slurs often rely on emojis and hashtags commonly used to label LGBTQ+ topics or to signal in-group membership. However, these same markers may also appear in messages that criticize or oppose the community, even from individuals who explicitly claim membership in it. Examples:

15. *"No se quien sea el organizador del PRIDE 🏳️‍🌈 2022 pero en que cabeza cabe invitar a* ***#RominaMarcos*** *que ella y su familia son homofóbicos misoginos y machistas no conforme con eso invitan al señor* ***#Alfredoadame*** *que usa palabras como joto y marica para ofender* ***#LGTBI"****.* (es_535).

16. *""@USER Esto me da a mí que roza la ilegalidad, uds sois una entidad oficial de un estado con bandera 🇪🇸,* ***no sois circo mariconchi, vergüenza****. Marlaska no puede dar más asco. #LGTBI #PRIDE2020 PD.* ***soy gay****."* (es_1872).

Exaggeration Some other instances use intensifiers or exaggeration to deliver hate to the community.

17. *"****Primero un dia y ahora un mes****. Al final el dia de los maricas sera dia no laborable #OrgulloLGTBI".* (es_465).

Sarcasm There are slurs that might look as if they were used by members of the community but actually were used with sarcasm.

18. *"Feliz día de la mariconeria #LGTBI #Pride2022"* (es_372).

19. *"****YO SOY EL COLECTIVO!!!!!*** *#OrgulloLGTBI* ***pues si****, hasta que llegasteis las maricas* ***rojas-podemitas-fascistoides****, la GS era una reivindicación del colectivo, y a punto estuvimos de conseguirlo como los israelíes. (es_275). F-I-N URL"*

"mariconeria" is a slight variation of the reclaimed slur *"maricón"* that is used with negative connotations while"YO SOY EL COLECTIVO" is a way of mocking which is later revealed with by calling the *"rojas-podemitas-fascistoides"*

5. Conclusion

In this study, we focus on the detection of messages containing reclaimed LGBTQ+ slurs within the MultiPRIDE shared task. We participated in Task A, which relies solely on the single instance of the message, and Task B, which incorporates both the message and its context to assess whether contextual information improves detection. Our team, *HateItOff*, participated in both tasks for Spanish and Italian. Our proposal and underlying hypothesis is that reclamation typically depends on a speaker who belongs to the targeted community using a reclaimed slur to express belongingness or identity. A common way this is expressed in natural language is through the use of pronouns and verbs, usually in a non-negative manner. Based on this intuition, we experimented with transformer models that incorporate linguistic features and sentiment features, providing cues related to pronoun and verb usage as well as the overall sentiment of the message. We conducted an Ablation Study and observed that adding these features resulted in a performance boost over the baseline for both languages. For Task A, our best-performing system surpassed the baselines provided by the organizers, ranking 8th out of 24 submissions for Italian and 6th out of 19 submissions for Spanish. For Task B, we achieved 11th place out of 18 submissions for Italian and 5th out of 12 submissions for Spanish. In this task, our system outperformed the organizers' baseline for Spanish only.

Acknowledgments

This work has been supported by the French government, through the 3IA Cote d'Azur investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001. This work was also carried out during the tenure of an ERCIM 'Alain Bensoussan' Fellowship Programme.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] D. L. de Sa, B. Cepollaro, The successes of reclamation, Synthese 202 (2023) 1–19. URL: https://philpapers.org/rec/DESTSO-18. doi:10.1007/s11229-023-04393-y.

[2] C. Bianchi, Slurs and appropriation: An echoic account, Journal of Pragmatics 66 (2014) 35–44. URL: https://www.unisr.it/attachments/Bianchi-C.-(2014)-%E2%80%9CSlurs-and-appropriation--an-echoic-account%E2%80%9D,-Journal-of-Pragmatics-66,-pp.-35-44,-DOI-10.1016-j.pragma.2014.02.009./be0c1563-7e7d-47b4-87ee-996c68fcf241/70483ac4-c1d0-4514-ae7b-be89664df395.pdf. doi:10.1016/j.pragma.2014.02.009, published in *Journal of Pragmatics* 66:35–44. Copyright Elsevier B.V.

[3] E. Nossem, Queer, frocia, femminiellə, ricchione et al. – localizing "queer" in the italian context, gender/sexuality/italy 6 (2019). URL: https://www.gendersexualityitaly.com/1-queer-frocia-femminiell%C9%99-ricchione-et-al-localizing-queer-in-the-italian-context/. doi:10.15781/31yc-ys20.

[4] B. Cepollaro, Linguaggio d'odio, in: In: *Linguaggio d'odio*, pp. 145–156, 2022. URL: https://hdl.handle.net/20.500.11768/144456.

[5] D. O. Thiago, A. D. Marcelo, A. Gomes, Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online, Sexuality & Culture 25 (2021) 700–732.

[6] N. B. Ocampo, E. Sviridova, E. Cabrio, S. Villata, An in-depth analysis of implicit and subtle hate speech messages, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1997–2013. URL: https://aclanthology.org/2023.eacl-main.147/. doi:10.18653/v1/2023.eacl-main.147.

[7] E. W. Pamungkas, V. Basile, V. Patti, Do you really want to hurt me? predicting abusive swearing in social media, in: Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020), European Language Resources Association, Marseille, France, 2020, pp. 6237–6246. URL: https://aclanthology.org/2020.lrec-1.765.

[8] E. W. Pamungkas, V. Basile, V. Patti, Investigating the role of swear words in abusive language detection tasks, Language Resources and Evaluation 57 (2023) 155–188. URL: https://doi.org/10.1007/s10579-022-09582-8. doi:10.1007/s10579-022-09582-8.

[9] L. Draetta, C. Ferrando, M. Cuccarini, L. James, V. Patti, Reclaim project: Exploring italian slurs reappropriation with large language models, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), 2024, pp. 335–342.

[10] C. Ferrando, L. Draetta, M. Madeddu, M. Sosto, V. Patti, P. Rosso, C. Bosco, J. Mata, E. Gualda, Multipride at evalita 2026: Overview of the multilingual automatic detection of slur reclamation in the lgbtq+ context task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[11] V. Basile, M. Lai, M. Sanguinetti, et al., Long-term social media data collection at the university of

turin, in: Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), CEUR-WS, 2018, pp. 1–6.
[12] J. Mata, E. Gualda, A dataset of spanish tweets on people and communities lgbtqi+ during the covid-19 pandemic 2020-2022 [lgbtqi+ dataset 2020-2022_es], 2025. Dataset.
[13] E. Zsisku, A. Zubiaga, H. Dubossarsky, Hate speech detection and reclaimed language: Mitigating false positives and compounded discrimination, in: Proceedings of the 16th ACM Web Science Conference, ACM, 2024, pp. 241–249.
[14] E. Bassignana, V. Basile, V. Patti, et al., Hurtlex: A multilingual lexicon of words to hurt, in: CEUR Workshop Proceedings, volume 2253, CEUR-WS, 2018, pp. 1–6.

A. Ablation Study for System Selection

In this section we show the results of the ablation study carried out between the transformer models used and the feature strategies. Table 3 and 4 show the results on the test set that we used for our experiments with transformers trained on 3 and 6 epochs. Table 5 shows our additional experiments carried out with UmBERTo for Task A.

3 Epochs Training with Data in Spanish, Italian, and English							
		Spanish			Italian		
Task	Model	no-recl	recl	macro	no-recl	recl	macro
Task A	bert_multil	0,9231	0,4000	0,6615	0,9627	0,8276	0,8951
	+ ling. cues	0,9224	0,4375	0,6800	0,9734	0,8889	0,9311
	+ sent. score	0,9099	0,3226	0,6162	0,9434	0,7541	0,8487
	+ sent. label	0,9231	0,4000	0,6615	0,9704	0,8571	0,9138
	+ sent. score	0,9145	0,3333	0,6239	0,9368	0,7018	0,8193
	+ sent. label	0,9091	0,3636	0,6364	0,9635	0,8077	0,8856
	+ sent. score	0,9099	0,3226	0,6162	0,9697	0,8710	0,9203
	+ sent. score	0,9185	0,3871	0,6528	0,9742	0,8727	0,9234
	modernbert_multil	0,8987	0,3784	0,6385	0,9509	0,7869	0,8689
	+ ling. cues	0,9153	0,2857	0,6005	0,9552	0,7931	0,8742
	+ sent. score	0,9185	0,3871	0,6528	0,9670	0,8302	0,8986
	+ sent. label	0,8996	0,3429	0,6212	0,9630	0,8214	0,8922
	+ sent. score	0,9185	0,3871	0,6528	0,9632	0,8148	0,8890
	+ sent. label	0,9099	0,3226	0,6162	0,9627	0,8276	0,8951
	+ sent. score.	0,9217	0,4706	0,6962	0,9663	0,8475	0,9069
	+ sent. score	0,9130	0,4118	0,6624	0,9478	0,7586	0,8532
	xlmt_multil	0,9177	0,4242	0,6710	0,9699	0,8667	0,9183
	+ ling. cues	0,9198	0,2963	0,6081	0,9429	0,6522	0,7975
	+ sent. score	0,9198	0,2963	0,6081	0,9734	0,8889	0,9311
	+ sent. label	0,9018	0,4500	0,6759	0,9655	0,8615	0,9135
	+ sent. score	0,9180	0,0000	0,4590	0,8949	0,0000	0,4475
	+ sent. label	0,9180	0,0000	0,4590	0,8949	0,0000	0,4475
	+ sent. score.	0,9180	0,0000	0,4590	0,8949	0,0000	0,4475
	+ sent. score	0,9205	0,2400	0,5803	0,9736	0,8852	0,9294
Task B	bert_multil	0,9004	0,3030	0,6017	0,9704	0,8571	0,9138
	+ ling. cues	0,9043	0,3529	0,6286	0,9701	0,8621	0,9161
	+ sent. score	0,9277	0,4138	0,6707	0,9738	0,8814	0,9276
	+ sent. label	0,9185	0,3871	0,6528	0,9699	0,8667	0,9183
	+ sent. score	0,9362	0,4828	0,7095	0,9478	0,7586	0,8532
	+ sent. label	0,9205	0,2400	0,5803	0,9630	0,8214	0,8922
	+ sent. score	0,9205	0,2400	0,5803	0,9738	0,8814	0,9276
	+ sent. score	0,9198	0,2963	0,6081	0,9627	0,8276	0,8951
	modernbert_multil	0,9130	0,4118	0,6624	0,9776	0,8966	0,9371
	+ ling. cues	0,9145	0,3333	0,6239	0,9738	0,8814	0,9276
	+ sent. score	0,9237	0,3571	0,6404	0,9776	0,8966	0,9371
	+ sent. label	0,9060	0,2667	0,5863	0,9742	0,8727	0,9234
	+ sent. score	0,9198	0,2963	0,6081	0,9888	0,9492	0,9690
	+ sent. label	0,8987	0,3784	0,6385	0,9630	0,8214	0,8922
	+ sent. score	0,8938	0,3684	0,6311	0,9848	0,9355	0,9602
	+ sent. score	0,8996	0,3429	0,6212	0,9697	0,8710	0,9203
	xlmt_multil	0,9145	0,3333	0,6239	0,9701	0,8621	0,9161
	+ ling. cues	0,9129	0,0870	0,4999	0,9701	0,8621	0,9161
	+ sent. score	0,9167	0,1667	0,5417	0,9263	0,4878	0,7071
	+ sent. label	0,9145	0,3333	0,6239	0,9588	0,8136	0,8862
	+ sent. score	0,9198	0,2963	0,6081	0,9776	0,8966	0,9371
	+ sent. label	0,9205	0,2400	0,5803	0,9559	0,7778	0,8668
	+ sent. score	0,9231	0,4000	0,6615	0,9520	0,7636	0,8578
	+ sent. score	0,9121	0,1600	0,5361	0,9559	0,7778	0,8668

Table 3
Ablation Study on Tasks A and B for the Spanish and Italian languages using models trained with 3 epochs.

6 Epochs Training with Data in Spanish, Italian, and English							
		Spanish			Italian		
Task	Model	no-recl	recl	macro	no-recl	recl	macro
Task A	BERT	0,9244	0,5641	0,7443	0,9430	0,7619	0,8524
	+ ling. cues	0,9264	0,4848	0,7056	0,9582	0,8254	0,8918
	+ sent. score	0,8957	0,2941	0,5949	0,9627	0,8276	0,8951
	+ sent. label	0,9292	0,5789	0,7541	0,9621	0,8387	0,9004
	+ sent. score	0,9130	0,4118	0,6624	0,9509	0,7869	0,8689
	+ sent. label	0,9091	0,3636	0,6364	0,9585	0,8197	0,8891
	+ sent. score	0,9310	0,5000	0,7155	0,9398	0,7333	0,8366
	+ sent. score	0,9153	0,2857	0,6005	0,9774	0,9000	0,9387
	ModernBERT	0,8966	0,2500	0,5733	0,9489	0,7308	0,8398
	+ ling. cues	0,9099	0,3226	0,6162	0,9430	0,7619	0,8524
	+ sent. score	0,8978	0,4103	0,6540	0,9552	0,7931	0,8742
	+ sent. label	0,9170	0,4571	0,6871	0,9506	0,7937	0,8721
	+ sent. score	0,9185	0,3871	0,6528	0,9738	0,8814	0,9276
	+ sent. label	0,9030	0,1481	0,5256	0,9630	0,8214	0,8922
	+ sent. score	0,9123	0,4444	0,6784	0,9627	0,8276	0,8951
	+ sent. score	0,9224	0,4375	0,6800	0,9742	0,8727	0,9234
	XLM	0,9224	0,4375	0,6800	0,9697	0,8710	0,9203
	+ ling. cues	0,9180	0,0000	0,4590	0,8949	0,0000	0,4475
	+ sent. score	0,9180	0,0000	0,4590	0,8949	0,0000	0,4475
	+ sent. label	0,9185	0,3871	0,6528	0,9615	0,8485	0,9050
	+ sent. score	0,9180	0,0000	0,4590	0,8949	0,0000	0,4475
	+ sent. label	0,9180	0,0000	0,4590	0,8949	0,0000	0,4475
	+ sent. score	0,9115	0,4737	0,6926	0,9494	0,8116	0,8805
	+ sent. score	0,9180	0,0000	0,4590	0,8949	0,0000	0,4475
Task B	BERT	0,9270	0,4516	0,6893	0,9588	0,8136	0,8862
	+ ling. cues	0,8918	0,2424	0,5671	0,9660	0,8525	0,9092
	+ sent. score	0,9304	0,5294	0,7299	0,9549	0,8000	0,8774
	+ sent. label	0,9217	0,4706	0,6962	0,9736	0,8852	0,9294
	+ sent. score	0,9106	0,2759	0,5933	0,9704	0,8571	0,9138
	+ sent. label	0,9174	0,0909	0,5041	0,9531	0,7347	0,8439
	+ sent. score	0,9205	0,2400	0,5803	0,9565	0,7600	0,8583
	+ sent. score	0,9283	0,3704	0,6493	0,9813	0,9153	0,9483
	ModernBERT	0,9091	0,3636	0,6364	0,9701	0,8621	0,9161
	+ ling. cues	0,9091	0,3636	0,6364	0,9699	0,8667	0,9183
	+ sent. score	0,9138	0,3750	0,6444	0,9740	0,8772	0,9256
	+ sent. label	0,9099	0,3226	0,6162	0,9740	0,8772	0,9256
	+ sent. score	0,9138	0,3750	0,6444	0,9738	0,8814	0,9276
	+ sent. label	0,9091	0,3636	0,6364	0,9850	0,9333	0,9591
	+ sent. score	0,8860	0,2778	0,5819	0,9701	0,8621	0,9161
	+ sent. score	0,8850	0,3158	0,6004	0,9773	0,9032	0,9402
	XLM	0,9180	0,0000	0,4590	0,8949	0,0000	0,4475
	+ ling. cues	0,9180	0,0000	0,4590	0,8949	0,0000	0,4475
	+ sent. score	0,9114	0,2222	0,5668	0,9630	0,8214	0,8922
	+ sent. label	0,9180	0,0000	0,4590	0,8949	0,0000	0,4475
	+ sent. score	0,9160	0,2308	0,5734	0,9774	0,9000	0,9387
	+ sent. label	0,9180	0,0000	0,4590	0,8949	0,0000	0,4475
	+ sent. score	0,9060	0,2667	0,5863	0,9740	0,8772	0,9256
	+ sent. score	0,9180	0,0000	0,4590	0,8949	0,0000	0,4475

Table 4
Ablation Study on Tasks A and B for the Spanish and Italian languages using models trained with 6 epochs.

Model	no-recl	recl	macro
UmBERTo	0,9542	0,8125	0,8833
UmBERTo + sent. label	0,9582	0,8254	0,8918
UmBERTo + ling. cues	0,9704	0,8571	0,9138
UmBERTo + ling. cues + sent. label	0,9663	0,8475	0,9069

Table 5
Additional UmBERTo runs on Task A