Evaluating AlBERTo, UmBERTo, and XLM-T in Automatic Misogyny Identification: A Comparative Study

Katya Trufanova

Abstract

This paper investigates the performance of three state-of-the-art models—AlBERTo, UmBERTo, and XLM-T—in the Automatic Misogyny Identification task when applied to two different problem modeling strategies: binary classification and ternary classification. Additionally, the impact of data augmentation on model performance is examined. The results show that AlBERTo and UmBERTo, which are specifically tailored for the Italian language, outperform the multilingual XLM-T model in both binary and ternary classification tasks. However, the data augmentation techniques used in this study led to a decrease in performance for most models and configurations, suggesting the need for more sophisticated augmentation methods in future research.

1 Introduction and Motivation

The objective of this project is to evaluate the performance of three state-of-the-art models - AlBERTo, UmBERTo, and XLM-T, in the EVALITA task of Automatic Misogyny Identification. Addressing misogyny in social media is of utmost importance, as it propagates harmful stereotypes and contributes to the normalization of discrimination against women. Through the development of effective algorithms for automatically detecting misogynistic content, timely moderation of social media platforms can be facilitated, fostering healthier online environments.

While the performance of more established and widely-used models in this task has been extensively investigated, this project aims to approach the task from two different perspectives and compare the performance of the three state-of-the-art models when applied to these two problem modeling strategies. Additionally, the variation in the performance of the models with and without data augmentation is examined. A comprehensive understanding of their capabilities is pursued by conducting twelve distinct experiments.

The task is approached from two different perspectives: first, as a binary classification problem, categorizing tweets as nonmisogynous vs. misogynous, and second, as a ternary (multi-class) classification problem, classifying tweets as nonmisogynous, misogynous (but not aggressive), or aggressive (and misogynous). The latter approach was proposed by Muti et al. in their solution to the 2020 edition of the AMI task (Muti and Barrón-Cedeño, 2020). Despite the official task being described as a sequence of two binary classification problems (misogyny identification first, and aggressiveness identification second), Muti et al. argue that modeling the two subtasks as a single multi-class problem significantly benefits the algorithm.

Considering these factors, it was determined that modeling the problem in two distinct ways and comparing the performance of different models in these two different problem settings, with and without data augmentation, would provide a deeper understanding of each model's strengths and weaknesses. This may also reveal potential improvements and adjustments that could enhance their overall performance.

It should be noted that AlBERTo has only been employed for ternary (but not binary) classification, while UmBERTo has exclusively been used for binary classification in the context of the AMI task. Consequently, this project aims to compare the two models in these two types of classification. Additionally, data augmentation has not been previously utilized with these models, presenting an opportunity to explore this aspect in the current project.

In contrast to AlBERTo and UmBERTo, XLM-T has never been applied to the Automatic Misogyny Evaluation task. Therefore, it was chosen for this project to examine its performance when used for the Automatic Misogyny Identification task, comparing its performance to that of the other state-of-the-art models in both problem definitions (binary and ternary classification), and both with and without data augmentation.

The motivation for this project stems from its potential to contribute to the existing research on automatic misogyny identification in several ways. Firstly, by comparing the performance of AlBERTo, UmBERTo, and XLM-T in both binary and ternary classification tasks, the most effective model for this specific task may be identified. This knowledge can inform future research and the development of even more accurate and efficient algorithms for misogyny identification.

Secondly, investigating data augmentation techniques is anticipated to reveal the potential benefits of these methods in enhancing model performance, particularly when available training data may be scarce. This insight can prove valuable not only for the specific task of misogyny identification but also for other classification tasks in natural language processing and machine learning.

Lastly, examining the performance of XLM-T in the Automatic Misogyny Identification task will contribute to the understanding of its capabilities and potential applications. By comparing its performance with that of other state-of-the-art models, its suitability for this task can be assessed, and areas where further research and development may be required to improve its effectiveness can be identified.

2 Related work

In this project, three different models were used: AlBERTo (Polignano et al., 2019), UmBERTo (Parisi et al., 2020), and XLM-T (Barbieri et al., 2022). Both AlBERTo and UmBERTo have previously been applied to automatic misogyny identification tasks in 2020 and 2018, respectively. In each instance, they surpassed the performance of the winning models from their respective task editions after the submission deadline.

More specifically, AlBERTo was utilized for ternary classification by Muti et al. in their 2020 paper (Muti and Barrón-Cedeño, 2020). The authors aimed to identify aggressive and misogynistic Italian tweets by employing a single multi-label classification model that simultaneously addressed aggressiveness and misogyny. Based on AlBERTo, this model achieved an F1 score of 0.7438 on the test set, outperforming the top submitted model. Muti et al. employed a 3-class setting, comprising non-misogynist, aggressive misogynist, and non-aggressive misogynist categories. They built their model on BERT and AlBERTo. Fine-tuning the model was done using the Pytorch instance of AlBERTo-Base and the AdamW optimizer. Optimal performance was achieved after training over 8 epochs with a batch size of 16, resulting in an F1 score 0.7438, surpassing the top submitted model and outperforming all systems submitted to the shared task.

Santini (Santini, 2021) used UmBERTo for binary classification in a study that compared three state-of-the-art models in the automatic misogyny identification task. Specifically, Santini assessed the performance of three pre-trained BERT models—AlBERTo, GilBERTo (Idb-ita, 2021), and UmBERTo—for misogyny identification in Italian tweets. After fine-tuning these models for binary classification, all three achieved state-of-the-art performances, surpassing the results of the models evaluated in the AMI Evalita 2018 campaign, with UmBERTo emerging as the best performer among the three.

In Santini's study, AlBERTo, GilBERTo, and UmBERTo were each trained in 5 separate runs, with their performances tested at each run in relation to the labeled testing set released from the AMI Evalita 2018 evaluation campaign. All models were evaluated based on their accuracies in predicting labels for the test set. The results revealed that all models, with similar training procedures, exhibited comparable performances. However, models fine-tuned on UmBERTo demonstrated the highest average accuracy rates, while those based on GilBERTo showed the lowest. This disparity may be attributed to UmBERTo being a cased language model while GilBERTo is uncased, as uppercase words can often convey different sentiments than their lowercase counterparts. This factor might also explain UmBERTo's superior performance compared to AlBERTo, the only model pre-trained on a social-media language corpus.

Notably, all models in Santini's study outperformed the systems evaluated in the AMI Evalita 2018 campaign. The most accurate system evaluated, bakarov.c.run2, was surpassed by an average increase in accuracy rate of 1.97%; compared to the baseline, the models were on average 3.36% more accurate. The study concluded that these BERT models could be considered state-of-the-art tools for solving AMI problems and, more generally, that the BERT architecture is one of the most promising solutions for text classification in deep learning.

3 Task

The task on which this project is based is Automatic Misogyny Identification (AMI) in Italian tweets, part of the Evalita 2020 campaign. This task focuses on misogyny and aggressiveness identification in Italian tweets, requiring the development of a system capable of recognizing whether a tweet is misogynous or not, and in the case of misogyny, whether it expresses an aggressive attitude. This problem is of particular importance because it addresses a significant social issue: the widespread presence of misogynous content on social media platforms, which contributes to systematic inequality and discrimination against women.

The dataset used for this task is a set of misogynous and non-misogynous tweets in the Italian language. The training set consists of tweets derived from the data collected for the 2018 edition of the AMI shared task. These tweets have been further enriched by labeling aggressive expressions according to the given definitions. The test dataset comprises approximately 1,000 tweets, collected from Twitter using a similar approach to the 2018 edition of the shared task. This choice was made to evaluate the generalization abilities of the developed models on test data collected in a different time period, characterized by higher language variability concerning the training data.

In the dataset, the tweets are annotated based on the following definitions:

1. Misogynous: a text that expresses hate towards women in particular (in the form of insulting, sexual harassment, threats of violence, stereotype, objectification, and negation of male responsibility).

2. Not Misogynous: a text that does not express any form of hate towards women.

3. Aggressive: a message is considered aggressive if it (implicitly or explicitly) presents, incites, threatens, implies, suggests, or alludes to:

   - attitudes, violent actions, hostility, or commission of offenses against women;
   - social isolation towards women for physical or psychological characteristics;
   - justify or legitimize an aggressive action against women.

4. Not Aggressive: If none of the previous conditions hold.

This task is crucial because it addresses the growing problem of online misogyny, which has significant consequences both for individuals and society. By developing models that can accurately identify misogynous and aggressive content on social media platforms, it is possible to better understand and potentially mitigate the impact of this harmful behavior. Moreover, the focus on the Italian language adds to the importance of the task, as it contributes to the development of NLP solutions that can cater to a diverse range of languages and cultures.

4 Models

As mentioned in the previous sections, in this project, three different models were used: AlBERTo, UmBERTo, and XLM-T. The following subsections provide a brief overview of each model.

4.1 AlBERTo

AlBERTo is a BERT-based language understanding model specifically pre-trained on Italian social media language. It employs a pre-trained BERT model with 110 million parameters and is trained on a corpus of Italian language collected from Twitter. AlBERTo is designed to represent the social media language, particularly Twitter, written in Italian and is suitable for text analysis tasks performed on content extracted from social media. The AlBERTo model is versatile and has been used for various NLP tasks, including hate speech detection.

AlBERTo's tokenizer carries out various data pre-processing and cleaning operations, such as normalization, annotation, HTML fixing, word segmentation, lowercasing, special character removal, whitespace normalization, repeated character reduction, and leading/trailing whitespace removal. The tokenizer supports different tokenization approaches, such as Wordpiece tokenization, character tokenization, and basic tokenization. It extends the BertTokenizer class and provides methods for tokenization, conversion between tokens and IDs, and converting tokens to a single string.

4.2 UmBERTo

UmBERTo is a RoBERTa-based language model trained on large Italian corpora. UmBERTo was trained on a variety of Italian texts, including Wikipedia articles and news articles, and has been applied to several natural language processing tasks, such as named entity recognition and part-of-speech tagging. UmBERTo was developed by a team at Musixmatch AI, including Loreto Parisi, Simone Francia, and Paolo Magnani.

4.3 XLM-T

XLM-T is a variant of the state-of-the-art XLM model, specifically designed for multilingual language models in Twitter. It is a modular framework that can be extended to include new languages and tasks. XLM-T is based on XLM-RoBERTa, a pre-trained multilingual model that outperforms multilingual BERT. XLM-T is trained on 200 million tweets for over 30 languages and includes a language model as well as a sentiment analysis model fine-tuned on a unified multilingual sentiment analysis dataset. This provides a strong multilingual baseline for sentiment analysis and other tasks.

5 Experimental setting

5.1 Experiments Overview

A total of twelve experiments were conducted, involving the three models: AlBERTo, UmBERTo, and XLM-T. The experiments were designed to compare the performance of these models in different scenarios. Each model was subjected to the following four experiments:

1. Binary classification: This experiment focused on classifying tweets as nonmisogynous or misogynous. The purpose of this experiment was to establish a baseline for the performance of each model in identifying misogynous content.

2. Ternary classification: This experiment expanded the problem definition, classifying tweets into three categories: nonmisogynous, misogynous (but not aggressive), and aggressive (and misogynous). This experiment aimed to assess the models' ability to distinguish between different levels of severity in misogynous content.

3. Binary classification with data augmentation: This experiment was similar to the first experiment but employed data augmentation techniques to address class imbalance in the dataset. The goal was to investigate the impact of data augmentation on the performance of each model in a binary classification setting.

4. Ternary classification with data augmentation: This experiment combined the ternary classification problem with data augmentation techniques. The purpose of this experiment was to determine the effectiveness of data augmentation in improving the models' performance in a more complex classification scenario.

5.2 Implementation Details

All models were implemented using PyTorch Transformers and trained on Google Colaboratory, leveraging the free GPU provided by the platform. The pre-trained models were obtained from the Transformers API.

The hyperparameters for the experiments were determined based on prior research and the limitations of the available GPU resources. In the case of AlBERTo and UmBERTo, the models were trained for 8 epochs with a batch size of 16, following the findings of Muti et al. For XLM-t, due to GPU availability constraints, the model was trained for 5 epochs with a batch size of 8. The AdamW optimizer was employed for all models in all experiments.

5.3 Data Cleaning

Data cleaning is an essential pre-processing step in Natural Language Processing tasks, as it can help improve model performance by removing irrelevant or distracting elements from the data and ensuring that the model focuses on the critical aspects of the task. Additionally, data cleaning can reduce the noise and inconsistencies in the input data, leading to more accurate and reliable predictions.

In this project, the dataset was cleaned during the pre-processing phase to remove mentions and links ('<MENTION_N>' and '<URL>') that were anonymized in the original dataset to protect the identity of the tweet authors and their interlocutors.

Additionally, the AlBERTo tokenizer provided various data pre-processing and cleaning operations, specifically aimed at twitter posts, such as normalization, annotation, HTML fixing, word segmentation, lowercasing, special character removal

5.4 Data Augmentation

Data augmentation is an essential technique in Natural Language Processing tasks that can help improve model performance, particularly when the available training data is scarce or imbalanced. By generating new training instances through various transformations, data augmentation can enhance the diversity and representativeness of the training set, leading to more robust and accurate models.

In this project, data augmentation was applied to both the binary and ternary classification experiments to address class imbalance in the dataset. The degree of augmentation performed depended on the imbalance in the dataset for each classification problem. For binary classification, the dataset was only slightly imbalanced, and some augmentation was performed. In contrast, for ternary classification, the dataset was significantly imbalanced, necessitating more extensive data augmentation.

The distribution of data before and after augmentation for both binary and ternary classification is as follows:

	0	1
Before augmentation	2362	2047
After augmentation	2362	2362

Table 1: Data distribution before and after augmentation for binary classification.

	0	1	2
Before augmentation	2362	478	1569
After augmentation	2362	2362	2362

Table 2: Data distribution before and after augmentation for ternary classification.

Several different data augmentation techniques were considered, with the intention of choosing the techniques most appropriate for the problem domain (Feng et al., 2021). In particular, three different data augmentation techniques were applied: random character swap, random character insert, and random character deletion. These methods were chosen because they closely resemble common typos found in social media text-based posts, which involve somewhat random character manipulations. The data augmentation was performed using NLPAug, a Python library for data augmentation in Natural Language Processing tasks.

6 Evaluation

6.1 Metrics

The performance of the models was evaluated using two distinct metrics in accordance with the problem settings:

For the binary classification task, the accuracy metric, as proposed by the 2018 AMI campaign, was employed to evaluate the models and enable a comparison with the models officially ranked. For the ternary classification task, a weighted F1-score, as proposed by Muti et al., was utilized. It should be noted that the results obtained when reproducing Muti et al.'s work showed a significant difference from the results presented in their paper. This discrepancy occurred despite an attempt to reproduce their code, which is available on GitHub, as closely as possible. One possible explanation for this difference is that Muti et al. used a non-anonymized version of the dataset, where Twitter mentions (and possibly URLs) were not censored. This is in contrast to the dataset provided for the task, where Twitter mentions are replaced with <MENTION_N> and URLs are replaced with <URL>. Another hypothesis is that the code available on GitHub was further modified before the publication of the paper, and these changes were not reflected in the repository.

6.2 Results

The outcomes of the experiments are summarized in the tables below:

Table 3: Binary Classification (Accuracy)

	Accuracy	
Model	**Without DA**	**With DA**
AlBERTo	0.822	0.808
UmBERTo	0.821	0.796
XLM-T	0.698	0.732

Table 4: Ternary Classification (Weighed F1-score)

	Weighed F1-score	
Model	**Without DA**	**With DA**
AlBERTo	0.642	0.580
UmBERTo	0.610	0.622
XLM-T	0.568	0.492

6.3 Result Analysis

In the binary classification task, AlBERTo and UmBERTo exhibited very similar performance, with accuracy scores of 0.822 and 0.821, respectively. In contrast, XLM-T lagged behind, achieving an accuracy of 0.698. When data augmentation was applied, the performance of AlBERTo and UmBERTo decreased, with AlBERTo scoring 0.808, UmBERTo 0.796, while XLM-T had a better result, with 0.732 in accuracy.

In the ternary classification task, AlBERTo outperformed the other models with a weighted F1-score of 0.642, followed by UmBERTo with 0.610 and XLM-T with 0.568. When data augmentation was employed, the performance of AlBERTo and XLM-T decreased to 0.580 and 0.492, respectively, while UmBERTo's performance improved slightly, reaching a weighted F1-score of 0.622.

It is worth noting that data augmentation led to a decrease in performance for most of the models and configurations. This outcome may be attributed to the specific types of augmentation chosen. More sophisticated data augmentation techniques could potentially result in performance improvements rather than decreases, presenting an opportunity for future research. Alternative explanations for the observed performance decline include the possibility that the augmentation methods introduced noise or inadvertently altered the meaning of the original text.

The results also indicate that AlBERTo and UmBERTo, which are models specifically tailored for the Italian language, perform better than the multilingual XLM-T model in both binary and ternary classification tasks. This may be attributed to the fact that these models have been pretrained on large Italian corpora, allowing them to better capture the nuances and idiomatic expressions of the language. In contrast, the XLM-T model, being a multilingual model, may not possess the same level of linguistic understanding for each individual language, including Italian. This observation emphasizes the importance of domain- and language-specific models for NLP tasks, especially when dealing with informal and colloquial text like tweets.

The evaluation results provide valuable insights into the strengths and weaknesses of the three state-of-the-art models—AlBERTo, UmBERTo, and XLM-T—when applied to the Automatic Misogyny Identification task. In both binary and ternary classification tasks, AlBERTo and UmBERTo demonstrated superior performance compared to XLM-T. However, performance was in most cases negatively impacted by the data augmentation techniques used in this study. Future work should explore more sophisticated data augmentation methods that may lead to performance improvements.

7 Conclusion

Overall, this project provided an in-depth analysis of the performance of three state-of-the-art models in the context of Automatic Misogyny Identification in Italian tweets. The results highlight the importance of domain- and language-specific models, as evidenced by the superior performance of AlBERTo and UmBERTo compared to the multilingual XLM-T model. This finding suggests that investing in dedicated models for specific languages and tasks may yield better performance than relying on general-purpose multilingual models.

Another key finding of this study is the negative impact of the employed data augmentation techniques on model performance, indicating that the choice of augmentation methods is critical in achieving the desired outcome. To this end, future work should focus on developing and evaluating more sophisticated data augmentation techniques that can better preserve the original meaning and linguistic cues of the text while effectively addressing class imbalance issues.

In conclusion, this project contributes to the ongoing effort to develop effective and reliable NLP tools for Automatic Misogyny Identification, which is an important step towards combating online gender-based violence and promoting a more inclusive and respectful digital environment.

References

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Steven Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp.

Idb-ita. 2021. Gilberto. `https://github.com/idb-ita/GilBERTo`. Accessed on June 28, 2023.

Arianna Muti and Alberto Barrón-Cedeño. 2020. Unibo @ ami: A multi-class approach to misogyny and aggressiveness identification on twitter posts using alberto.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. `https://github.com/musixmatchresearch/umberto`.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. 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)*, volume 2481. CEUR.

Cristian Santini. 2021. Comparison of three BERT language models for Automatic Misogyny Identification on Italian tweets.