UniBO @ AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo

Arianna Muti
Department of Modern Languages,
Literatures and Cultures - LILEC
Università di Bologna
Bologna, Italy
arianna.muti@studio.unibo.it

Alberto Barrón-Cedeño
DIT – Università di Bologna
Forlì, Italy
a.barron@unibo.it

Abstract

We describe our participation in the EVALITA 2020 (Basile et al., 2020) shared task on Automatic Misogyny Identification. We focus on task A —Misogyny and Aggressive Behaviour Identification— which aims at detecting whether a tweet in Italian is misogynous and, if so, whether it is aggressive. Rather than building two different models, one for misogyny and one for aggressiveness identification, we handle the problem as one single multi-label classification task, considering three classes: non-misogynous, non-aggressive misogynous, and aggressive misogynous. Our three-class supervised model, built on top of AlBERTo, obtains an overall F_1 score of 0.7438 on the task test set ($\mathrm{F}_1 = 0.8102$ for the misogyny and $\mathrm{F}_1 = 0.6774$ for the aggressiveness task), which outperforms the top submitted model ($\mathrm{F}_1 = 0.7406$).[1]

1 Introduction

In 2020, Twitter users in Italy amount to approximately 3.7 million and the number is expected to constantly increase by 2026.[2] Although Twitter is conceived to express personal opinions, share today's biggest news, follow people or simply communicate with friends, there has been an increasing number of users that misuse the platform by engaging in trolling, cyberbullying, or by posting aggressive and misogynous content (Samghabadi et al., 2020). Due to the sheer amount of user-generated content on social media, providers struggle to control inappropriate content. Twitter relies on the community's reports to identify and remove abusive posts from the platform, while pursuing the users' right to freedom of expression. However, it is a tricky task to determine where to draw the line between free expression and the production of harmful content, due to the subjective nature of what different users perceive as offensive. Twitter has committed to tackling this issue by releasing a policy containing a clear definition of abusive speech, according to which a user cannot promote violence against or directly attack or threaten people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.[3]

However, two main issues exist. Since Twitter mostly relies on the community subjective perception of hate speech, many posts are not subjected to report, review, and removal. Moreover, the amount of abusive posts significantly outnumbers the people that can manually control harmful content. Therefore, there is a need to improve the quality of algorithms to spot potential instances of hate speech; in particular towards women, since research shows that they are subjected to more bullying, abuse, hateful language, and threats than men on social media (Fallows, 2005).

AMI 2020 consists of two tasks (Fersini et al., 2020). Task A —Misogyny and Aggressive Behaviour Identification— aims at detecting whether a Twitter post is misogynous and, if so, whether it is aggressive (Anzovino et al., 2018). Task B —

[1] Our official submission to the task obtained $\mathrm{F}_1 = 0.6343$ ($\mathrm{F}_1 = 0.7263$ for the misogyny and $\mathrm{F}_1 = 0.5423$ for the aggressiveness task). The reason behind this poor performance was the unintended use of a mistaken transformer. See Appendix A for further details.

[2] https://www.statista.com/forecasts/1146708/twitter-users-in-italy; last visit: 6 November, 2020.

[3] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Unbiased Misogyny Identification— aims at discriminating misogynistic contents from the non-misogynist ones, while guaranteeing the fairness of the model (in terms of unintended bias) on a synthetic dataset (Nozza et al., 2019). We undertook task A and we present a system to flag misogynous and women-addressed aggressive posts on Twitter in the Italian language. Even if task A involves two sub-problems, we address it as a three-class supervised problem using AlBERTo (Polignano et al., 2019), a BERT language understanding model for the Italian language which is focused on the language used in social networks, specifically on Twitter. We built only one model to identify the three possible classes: non-misogynous, non-aggressive misogynous, and aggressive misogynous. This multi-class setting has shown to be effective. Our approach obtains an F_1 score of 0.7438, outperforming the top-ranked official submission (although our own official submission obtained $\text{F}_1 = 0.6343$ only; cf. Appendix A).

The rest of the contribution is distributed as follows. Section 2 includes some background and a brief overview of research in automatic misogyny identification. Section 3 describes the employed dataset. Section 4 describes our model. Section 5 summarizes the experiments performed and discusses the obtained results. It includes an error analysis, in order to show the error trends of the model. Section 6 draws some conclusions and discusses further possible research lines.

2 Background

Due to the subjective perception of misogyny and aggressiveness, a definition of what can be considered misogynous and aggressive is necessary:

Misogynous content expresses hating towards women, in the form of insulting, sexual harassment, male privilege, patriarchy, gender discrimination, belittling of women, violence against women, body shaming and sexual objectification (Srivastava et al., 2017). A misogynous content expresses an **aggressive attitude** when it overtly or covertly encourages or legitimizes violent actions against women.

From a computational point of view, misogyny detection is a text classification task. Text classification in Natural Language Processing has been widely explored and it is typically addressed by using supervised models (Mirończuk and Protasiewicz, 2018). Past research shows the effectiveness of diverse neural-network architectures to learn text representations, such as convolutional models, recurrent networks and attention mechanisms (Sun et al., 2019). Recent work shows that pre-trained models such as BERT achieve state-of-the-art results in text classification tasks and spare time, since they prevent you from training models from scratch (Sun et al., 2019).

For what concerns misogyny identification, a shared task took place at IberEval 2018, focusing on English and Spanish tweets (Fersini et al., 2018b). Whereas task A concerned misogyny identification, task B proposed a multi-class problem to classify misogynous sentences into seven categories: discredit, stereotype, objectification, sexual harassment, threats of violence, dominance, and derailing. The most used supervised models were support vector machines, ensembles of classifiers and deep-learning models. Participants mostly used n-grams and word embeddings to represent the tweets.

As for misogyny identification in Italian, the first edition of the AMI shared task took place in 2018 (Anzovino et al., 2018). The task A was again misogyny identification, while the task B aimed at recognizing whether a misogynous content is person-specific or generally addressed towards a group of women, and at classifying the positive instances in the aforementioned categories. The best-performing approach obtained an F_1 score of 0.844, using TF-IDF weighting combined with singular value decomposition for language representation and an ensemble of supervised models (Fersini et al., 2018a).

3 Dataset

As mentioned above, the aim of our model is to flag misogynous contents and aggressive attitudes towards women in Italian tweets. To address this task, a dataset was provided by the task organizers: $5,000$ tweets, manually labelled according to two classes, misogyny and aggressiveness. The first one defines whether a tweet has been flagged as misogynous (positive class) or not (negative class). If a tweet has been flagged as misogynous, it is further determined whether it is considered as aggressive (positive class) or not (negative class).

The training dataset is fairly balanced in terms of misogyny. It contains $2,337$ misogynous and $2,663$ non-misogynous instances. A total of

	batch size	
epochs	16	32
8	**0.8491**	0.8392
10	0.8485	0.8298
15	0.8283	0.8351
20	0.8342	0.8087

Table 1: F$_1$ performance of the 3-class model with different batch sizes after diverse numbers of epochs using AlBERTo

$1,783$ of the former are also considered as aggressive, whereas only 554 are not. The test set was composed of $1,000$ tweets.

Since we opted for a constrained approach, we only used the data provided by the organizers. We randomly split the supervised data into training and validation sets: $4,700$ instances for the former and 300 for the latter.

4 Description of the System

Since the identification of aggressiveness is related to the identification of misogynous tweets, we opt for a 3-class setting, based on one single model. The three classes are hence non-misogynist, aggressive misogynist, and non-aggressive misogynist. The idea is to determine how well a multilabel classifier can perform when addressing these two related problems; handling aggressiveness as a consequential class of the misogyny one.

We decided to base our model on BERT (Bidirectional Encoder Representations from Transformers), a task-independent language representation model based on the transformers architecture (Devlin et al., 2019). BERT uses a masking approach that randomly masks some input tokens within a sentence and then predicts the removed tokens based on the context. It is bidirectional because it makes use of Transformers that consider both the left and right context at once with respect to the hidden word to make the prediction upon. We decided to use AlBERTo, a variation of BERT in Italian, trained on Twitter posts (Polignano et al., 2019), which includes emojis, links, hashtags, and mentions. AlBERTo was trained on 200M tweets randomly sampled from the TWITA corpus (Basile et al., 2018).

As for the pre-processing, we used the pretrained AlBERTo tokenizer for text tokenization, and then we encoded the data. We set the maximum length to 256 characters, since that was the length of the longest instance in the training material (even if Twitter allows up to 280 characters).

team	run	constrained	score
UniBO[a]	2	yes	**0.7438**
jigsaw	2	no	0.7406
jigsaw	1	no	0.7380
fabsam	1	yes	0.7343
YNU_OXZ	1	no	0.7314
fabsam	2	yes	0.7309
NoPlaceForHateSpeech	2	yes	0.7167
YNU_OXZ	2	no	0.7015
fabsam	3	yes	0.6948
NoPlaceForHateSpeech	1	yes	0.6934
AMI_the_winner	2	yes	0.6869
MDD	3	no	0.6844
PoliTeam	3	yes	0.6835
MDD	1	yes	0.6820
PoliTeam	1	yes	0.6810
MDD	2	no	0.6679
AMI_the_winner	1	yes	0.6653
PoliTeam	2	yes	0.6473
UniBO[b]	1	yes	0.6343
AMI_the_winner	3	yes	0.6259
NoPlaceForHateSpeech	3	yes	0.4902

[a] Run submitted after the deadline.

[b] Buggy run submitted on the deadline (cf. Appendix A).

Table 2: Full shared task leaderboard plus our unofficial top-performing submission. The score is the average of the F$_1$ measures for the misogyny and the aggressiveness tasks.

We used the Pytorch instance of AlBERTo-Base, Italian Twitter lower cased[4] and fine-tuned it to the downstream task. We used a softmax output layer with three neurons to produce the classification.

In order to tune the network, we used the AdamW optimizer, which decouples weight decay from gradient computation, with a learning rate of 1e-5 (Loshchilov and Hutter, 2017).[5]

5 Results

We explored different batch sizes over an increasing number of learning epochs. Table 1 shows the performance evolution on the validation set. The best combination was to train the model over 8 epochs with a batch size of 16. This combination leads to an F$_1$ score of 0.8491 on the three-class problem. It is worth noting that these scores are not comparable against those for the actual task, which consists of two independent binary decisions: whether a tweet is considered misogynist and, if the answer is yes, whether it is aggressive.[6]

[4] `https://github.com/marcopoli/AlBERTo-it,`

[5] The implementation is available at `https://github.com/TinfFoil/unibo_ami2020/`.

[6] Indeed, the official task score is computed as the average of the F$_1$ measures for the misogyny and the aggressiveness classifications.

	tweet	misogynist		aggressive	
		actual	pred	actual	pred
Type-1 errors					
1.2	La cosa del filo del telefono attorno alla gola è presa da una lite di Stefanie e Brooke in Beautiful #chilhavisto	no	yes	no	yes
1.2	"Quanti schiaffi, #Malinconico. Li meriteresti tutti, perché anche questa volta sei riuscito a lasciare il segno.E no, non era facile…"	no	yes	no	yes
1.3	Tutti quelli che ho conosciuto mi hanno sempre detto che a primo impatto sembro stronza e psicopatica ma in verità sono una cucciola	yes	no	no	no
Type-2 errors					
2.1	@Fedoc5 in ginocchio con te che comandi il ritmo del pompino e decidi quando sborrarmi in gola, perchè un pompino deve sempre finire con l'ingoio	no	yes	no	yes
2.2	@Nigagalsen @matteosalvinimi un follower dal 2016…e ti permetti di aprire quella lurida fogna di bocca che hai.	no	yes	no	no
2.3	Antonio ti prenderei a schiaffi, come fa a dire mi manchi, quando con Ilaria fai tutto. Coglione	no	no	no	no
2.4	Posso volerlo vedere cagarsi in mano e prendersi a schiaffi finché non diventano dispari o sono nazista?	no	no	no	no
Type-3 errors					
3.1	La cena è pronta. Lo squalo balena affonda nei banchi di pesce per saziare la sua immensa mole. Non è un abitudine frequente visto che filtra dall'acqua i microorganismi come le balene.	no	yes	no	no
3.2	Comunque le pringles più buone sono quelle alla panna acida e cipolla	no	yes	no	no

Table 3: Instances from the test partition including their actual class and the one predicted by our model for both misogyny and aggressiveness.

Given these results, we trained a new model on the full trained and development sets during 8 epochs, using a batch size of 16, and predicted on the test set. Such model obtains $F_1 = 0.7438$, resulting from 0.8102 on the misogyny task and 0.6774 from the aggressiveness one.

Table 2 shows the AMI shared task leaderboard. It highlights both our official submission `UniBO run 1` (cf. Appendix A) and our post-deadline submission `UniBO run 2`. Run 2 tops all the systems submitted to the shared task. Indeed, modelling the two tasks as one single multi-class problem (and using transformers for the right language) helps the algorithm significantly.

Error Analysis After the release of the gold labels, we performed an analysis of the classification errors. We analyzed 300 instances, taken randomly from the test set (100 at the beginning, 100 in the middle and 100 at the end). As observed from the reported performance, our model struggled mostly with the identification of aggressive instances. As a result, there are relatively few cases in which our model correctly labels non-aggressive misogynous instances. We noticed that most of the time, when our model labels an instance as misogynist, it also labels it as aggressive. On the contrary, the system performs very well in identifying non-misogynous instances and aggressive-misogynous instances. The most common mistakes are grouped into three categories:

1. The system identifies as aggressive the instances that contain verbs expressing an aggressive attitude.[7]

2. The system identifies as misogynous (and most of the time also aggressive) instances

[7] One potential reason behind this confusion is that we suspect that there are aggressive tweets in the dataset which, not having been identified as misogynist in the first place, are mislabeled as non-aggressive. This hypothesis should be further explored.

that are neither misogynous nor aggressive, but contain typical misogynous sentences.

3. The system identifies as misogynous instances that are neither misogynous nor aggressive, but they contain *double-entendre* words typically used to insult women.

Table 3 shows some examples for all three kinds of errors. Regarding the errors of type 1, in instance 1.1 the action of winding up a telephone cable around the neck was perceived as aggressive, despite the speaker did not express a misogynous or aggressive attitude towards a woman, and indeed she is just commenting on something watched on TV. In instance 1.2, the sentence *meritare gli schiaffi* (deserving slaps) denotes violence, but it is not addressed towards a woman. This kind of mistake might be overcome by implementing a model trained on the misogynist partition of the data only. Finally, instance 1.3 represents the bias related to the subjectivity nature of what is perceived to be misogynous. According to the annotation guidelines, a tweet should be flagged as misogynous if it expresses hating towards women. In this case, the poster of the tweet is not expressing any misogynous attitude, but she is reporting what she is been told by males. Therefore, our system flagged the instance as non-misogynous and we could agree.

As for the errors of type 2, if we look at the text only, the instances could seem misogynous sentences. However, in the instances 2.1 and 2.2 the hashtag tells us that it is referred to a man and the system fails to understand that. On the contrary, the system performs well when a masculine name or a masculine pronoun is specified, instead of an hashtag, as we can observe in the instances 2.3 and 2.4. In these cases our system could understand that the aggressive actions, that usually tend to be classified as aggressive-misogynous, are not referred to a woman.

For the type 3 errors, in instance 3.1 *balena* (whale/fat woman) and in 3.2 *acida* (acid/peevish) could confuse the model causing it to flag such instances as misogynous.

6 Conclusions and Further Work

In this paper we described our approach to the EVALITA 2020 task on misogyny and aggressiveness identification in Italian tweets —AMI. The purpose of our participation was to determine whether a multi-label classifier is a good way to address this two-step task. Although the task seems to be conceived to be addressed with two different models, one for the identification of misogyny and the other for aggressiveness, we decided to try a different approach and build a single model that could identify three cases: non-misogynous, non-aggressive misogynous and aggressive misogynous tweets.

We built our model on top of AlBERTo, an Italian version of BERT, and we trained the model using only the dataset provided by the task organizers. We experimented by setting different batch sizes over an increasing number of epochs. The highest F_1 score on the validation set was reached by a batch size of 16 during 8 epochs. When evaluated on the test set, our model obtained an overall F_1 score of 0.7438; 0.8102 for the misogyny and 0.6744 for the aggressiveness task. We hypothesize that the model struggles to identify misogynist aggressive instances partly because it gets confused by non-misogynist aggressive tweets which are labeled simply as non-misogynous. The implementation is publicly available for research purposes.

For what concerns further experiments, we plan to build two separate models: one to detect misogyny and the other trained only on already-flagged misogynous tweets to identify instances of aggressiveness. Another step to undertake would be to use an unconstrained approach and increase the number of instances for the training set, so that the model will have more data to learn from.

References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Valerio Basile, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term social media data collection at the university of turin. In *Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Turin, Italy.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and*

Speech Tools for Italian. Final Workshop (EVALITA 2020), Online. CEUR.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, June. ACL.

Deborah Fallows. 2005. How women and men use the internet. Technical report, Pew Internet & American Life Project, December.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 12-13 December 2018, Naples*, pages 59–66. Torino: Accademia University Press.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the task on automatic misogyny identification at ibereval 2018. In *Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, Sevilla, Spain.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. Ami @ evalita2020: Automatic misogyny identification. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Marcin M. Mirończuk and Jarosław Protasiewicz. 2018. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, 106:36–54.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 149–155, Thessaloniki, Greece.

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, Bari, Italy. CEUR.

Niloofar S. Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*.

Kalpana Srivastava, Suprakash Chaudhury, P.S. Bhat, and Samiksha. Sahu. 2017. Misogyny, feminism, and sexual harassment. *Industrial psychiatry journal*, 26(2):111–113.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

A Official English-BERT-based Submission

Our official submission used a pre-trained BERT model trained only on the English language. The experimentation and tuning were identical to the one applied when using AlBERTo (cf. Section 5). Table 4 shows the tuning evolution. The best configuration of this model, derived from the English BERT, obtains an F_1 score of 0.8222 on the validation set when dealing with our three-class problem. Nevertheless, the performance dropped to $F_1 = 0.6343$ on the test set.

	batch size		
epochs	8	16	32
5	0.8126	0.8042	0.7955
8	0.8067	**0.8222**	0.8004
10	0.8042	0.8069	0.8141
15	0.8095	0.8037	0.8121
20	0.7895	0.8178	0.8153

Table 4: F_1 performance of the 3-class model with different batch sizes after diverse numbers of epochs using BERT for English.