A Metric Learning Approach to Misogyny Categorization

Juan M. Coria and **Sahar Ghannay** and **Sophie Rosset** and **Hervé Bredin**
Université Paris-Saclay, CNRS, LIMSI
{coria, ghannay, rosset, bredin}@limsi.fr

Abstract

The task of automatic misogyny identification and categorization has not received as much attention as other natural language tasks have, even though it is crucial for identifying hate speech in social Internet interactions. In this work, we address this sentence classification task from a representation learning perspective, using both a bidirectional LSTM and BERT optimized with the following metric learning loss functions: contrastive loss, triplet loss, center loss, congenerous cosine loss and additive angular margin loss. We set new state-of-the-art for the task with our fine-tuned BERT, whose sentence embeddings can be compared with a simple cosine distance, and we release all our code as open source for easy reproducibility. Moreover, we find that almost every loss function performs equally well in this setting, matching the regular cross entropy loss.

1 Introduction

Whether it is at the word or at the sentence level, learning robust representations allows neural networks to consolidate knowledge that can later be transferred to other tasks and domains. Many approaches have dealt with this problem in different ways, for instance with CBOW or skip-gram from word2vec (Mikolov et al., 2013) for context-independent word embeddings, or more recently with BERT's (Devlin et al., 2019) sentence embeddings and contextual word embeddings.

In order to learn sentence representations, a neural encoder enc needs to learn a mapping from an initial representation x_i to a target vector space. In a metric learning approach, the distances between each pair of sentence embeddings $(\mathrm{enc}(x_i), \mathrm{enc}(x_j))$ should be low if classes $y_i = y_j$ (intra-class compactness) and high if $y_i \neq y_j$ (inter-class separability). To achieve this objective, the angle θ_{ij} separating a pair of embeddings (as depicted in Figure 1) can be used to redefine the model's loss function.

In the domain of face recognition, many loss functions (Schroff et al., 2015; Wen et al., 2016; Liu et al., 2017; Wang et al., 2018; Deng et al., 2019) have been proposed to learn better face representations, motivated by high intra-class variability due to lighting, position or background. Other studies have experimented with these methods in different domains with similar characteristics, like speaker verification (Bredin, 2017; Chung et al., 2018; Yadav and Rai, 2018), and even as an enhancement of BERT's sentence representations (Reimers and Gurevych, 2019) for semantic textual similarity. A recent study (Srivastava et al., 2019) has also focused on comparing these methods on face verification, showing that angular margin losses achieve superior performance.

On the other hand, the automatic misogyny identification (AMI) evaluation campaign (Fersini et al., 2018a) was proposed to address misogyny on tweets. Included tasks were identification (i.e. misogynous or not), categorization over five different misogyny types, and target identification (to an individual or a group). However, no participant has proposed a metric learning model. The best system (Ahluwalia et al., 2018) uses a bidirectional LSTM with word embeddings of size 100 for the identification task, and ensemble methods with feature engineering for category and target classification. They achieve a macro F1 score of 36.1 on the misogyny categorization part of sub-task B, which is the one we address as well. A different architecture (Caselli et al., 2018) uses a multi-layer character bidirectional LSTM for categorization, obtaining a macro F1 score of 14.1.

In this paper, we focus on five metric learning losses for the task of misogyny categorization, using the AMI (Fersini et al., 2018a) dataset. Our hypothesis was that metric learning might reduce

Proceedings of the 5th Workshop on Representation Learning for NLP (RepL4NLP-2020), pages 89–94

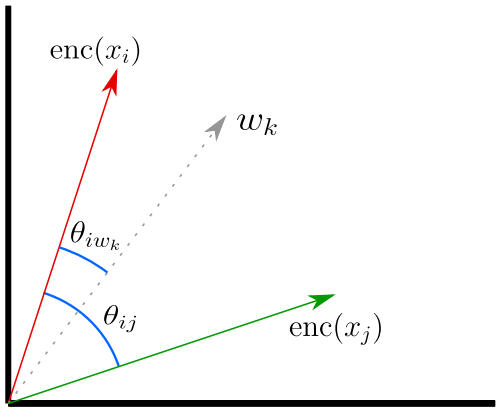

Figure 1: Depiction of embeddings in two dimensions. The dotted vector w_k represents a centroid for some class k, while the other vectors are sentence embeddings. θ values are angles separating two vectors.

the natural intra-class variability within misogyny categories, making representations robust to writing styles, irony, insults, etc. The loss functions we experiment with are contrastive loss (Hadsell et al., 2006), triplet loss (Schroff et al., 2015), center loss (Wen et al., 2016), congenerous cosine loss (Liu et al., 2017) and additive angular margin loss (Deng et al., 2019), as well as cross entropy loss. We optimize these loss functions with two different architectures: a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) and BERT (Devlin et al., 2019), and we evaluate their performance using a simple K-nearest neighbors (KNN) classifier to better measure representation quality.

Our main contributions consist of new state-of-the-art performance for the misogyny categorization task, as well as empirical evidence that these methods do not perform better than cross entropy loss on closed-set sentence classification. Moreover, our code is released as open source for easy reproducibility.

2 Loss Functions

In this section, we present the loss functions chosen for our study, which can be separated into contrast-based and classification-based, according to how they are computed.

2.1 Contrast-based losses

The contrastive loss (Hadsell et al., 2006) uses pairs annotated as similar/dissimilar (also called positive/negative). It brings representations from similar examples closer together, while separating dissimilar ones explicitly:

$$\mathcal{L} = \sum_{i=1}^{P_+} (D_i)^2 + \sum_{i=1}^{P_-} \max(m - D_i, 0)^2 \quad (1)$$

where P_+ is the number of similar pairs, P_- the number of dissimilar pairs, $D_i = 1 - \cos\theta_i$ the distance between embeddings of the ith pair, and m a margin.

The triplet loss (Schroff et al., 2015) is calculated over triplets composed of a reference example known as the anchor, a positive and a negative, both the latter with respect to the anchor. Following the idea introduced by Gelly and Gauvain (2017), we define this loss using the sigmoid function:

$$\mathcal{L} = \sum_{i=0}^{T} \text{sigmoid}(\alpha\,(\cos\theta_i^n - \cos\theta_i^p))) \quad (2)$$

where T is the number of triplets, α a scaling hyper-parameter, θ_i^p the angle separating the anchor and the positive embeddings, and θ_i^n the angle separating the anchor and the negative ones.

Taking Figure 1 as an example, contrast-based losses encourage the cosine distance between embeddings i and j to be larger if $y_i \neq y_j$, and smaller if $y_i = y_j$. This is achieved a single pair at a time with contrastive loss, while triplet loss does it jointly using both the positive and negative inside the triplet.

2.2 Classification-based losses

These loss functions derive from the cross entropy loss, either by modifying how the classification layer output is calculated or working as a penalization term. The cross entropy loss is defined as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \log\,\text{softmax}(\sigma_i, y_i) \quad (3)$$

where N is the number of training examples, σ_i the output of the classification layer, and y_i the class of the ith example.

The congenerous cosine (CoCo) loss (Liu et al., 2017) interprets the weights w_k of the classification layer as class centroids, learning to maximize the cosine similarity between a representation and its centroid. The classification layer output σ_i is redefined as:

$$\forall k \;\; \sigma_{ik} = \alpha \cdot \cos\theta_{iw_k} \quad (4)$$

where θ_{iw_k} is the angle separating the ith representation and w_k, and α a scaling hyper-parameter.

The additive angular margin (AAM) loss (Deng et al., 2019) goes one step further adding a margin in angular space to penalize the distance between a representation and its centroid:

$$\forall k \quad \sigma_{ik} = \alpha \cdot \cos(\theta_{iw_k} + \delta_{ik}\, m) \qquad (5)$$

where m is a margin, and $\delta_{ik} = 1$ if $k = y_i$ and 0 otherwise.

Finally, the center loss (Wen et al., 2016) penalizes the cross entropy loss with the distance to jointly learned centroids c_k external to the classification layer:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \frac{\lambda}{2} \sum_{i=1}^{N} (1 - \cos\theta_{ic_{y_i}})^2 \qquad (6)$$

where λ is a hyper-parameter controlling the effect of penalization.

To see the effect of classification-based losses more intuitively, consider embeddings and centers in Figure 1. If $y_i = k$, then both congenerous cosine loss and center loss will penalize the loss value with the distance from embedding i to w_k (or c_k in the case of center loss), hence bringing all vectors from class k close to the centroid k. The additive angular margin loss follows the same principle, but penalizing further by artificially augmenting the distance of embedding i to w_k with the angular margin.

3 Task

The term misogyny is defined as hatred towards women. Hate speech of this nature is unfortunately common in social Internet interactions, and current language models are generally unable to accurately detect and classify it. The AMI task and corpus were proposed in the context of the IberEval 2018 (Fersini et al., 2018b) and Evalita 2018 (Fersini et al., 2018a) evaluation campaigns, allowing researchers to train models focused specifically on misogyny. The corpus consists of an ensemble of tweets with three different types of annotations: misogyny (binary), misogyny category and target (active or passive).

We use the same dataset as in Fersini et al. (2018a) and we focus exclusively on misogyny categorization, using an additional class for non misogynous tweets. Our results are thus compared to the categorization part of sub-task B. An explanation of misogyny categories according to the definitions given in Fersini et al. (2018a) can be found in Table 2.

Class	Train	Dev	Test
derailing	74	18	11
discredit	811	203	141
dominance	118	30	124
sexual harassment	282	70	44
stereotype	143	36	140
non misogynous	1,772	443	540
total	3,200	800	1,000

Table 1: Number of sentences per class for each partition of the AMI dataset. Note that classes are greatly imbalanced.

As the corpus does not provide a development set, one was constructed from the training set following the same class distribution. The final Train set is composed of 3200 tweets, and the Dev and Test sets of 800 and 1000 tweets respectively. Class distribution is described in detail in Table 1. The task is evaluated using the macro F1 score.

4 Experiments

4.1 Experimental protocol

As different losses rely on different hyper-parameters, we perform a hyper-parameter search including learning rates, margins m, scalings α, and λ. The values we have experimented with are shown in Table 3. Each configuration is trained on Train for 60 epochs and validated using a KNN classifier on Dev. As we deal with a rather small dataset, the best configuration for each loss and each architecture is then trained and validated from scratch 10 times to reduce the effect of randomness. Reported results are the mean macro F1 score and standard deviation on Test over these 10 runs.

In all experiments we use the cosine distance to compare embeddings, as congenerous cosine loss and additive angular margin loss can only be optimized in this way. Additionally, a linear classification layer is jointly trained with the sentence encoder when optimizing classification-based loss functions.

4.2 Architecture

We experiment with two different encoder architectures. The first one is a one-layer bidirectional LSTM (Hochreiter and Schmidhuber, 1997) with output size 768 (to match BERT) and word embeddings of size 300 obtained from a word2vec CBOW model (Mikolov et al., 2013) trained on 2-billion-word Wikipedia dumps. The second one is

Category	Description	Example
derailing	"to justify women abuse, rejecting male responsibility"	"if rape is real why aren't more people reporting it? just another feminist lie"
discredit	"slurring over women with no other larger intention"	"this b*** is a s***"
dominance	"to assert the superiority of men over women to highlight gender inequality"	"#didyouknow the male brain is 3.4 times larger than the female brain? #maledominance"
sexual harassment	"sexual advances, harassment of a sexual nature, etc."	"come on box I show you my c*** darling"
stereotype	"a widely held but fixed and oversimplified image or idea of a woman"	"these people are hysterical. it's like a commercial for why men should never marry [...]"

Table 2: Misogyny categories as described by the corpus authors (Fersini et al., 2018a) along with examples found in the training set.

Parameter	Values
LR	$\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}^{\bullet}$ $\{10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}^{\circ}$
m	$\{0.02, 0.05, 0.25, 0.5, 0.75\}$
α and λ	$\{0.01, 0.1, 1, 10, 100, 1000\}$

Table 3: Values tested during initial hyper-parameter search, totaling 486 configurations. LR stands for learning rate, and m, α and λ are loss parameters (see Section 2). Values with $\bullet$ are LSTM only and values with $\circ$ are BERT only.

the standard monolingual uncased BERT (Devlin et al., 2019) from the huggingface library (Wolf et al., 2019) pretrained on Wikipedia.

To obtain a sentence embedding from an encoder, we perform a max pooling over the hidden states of the last layer, leaving us with sentence embeddings of size 768 on both models.

4.3 Implementation details

All sentences are pre-tokenized using the `TweetTokenizer` from the NLTK toolkit (Bird et al., 2009) in order to correctly deal with Twitter-specific tokens like hashtags, mentions, and even emojis. During this process we remove handles and URLs. When training BERT, we do a second pass of tokenization with BERT's pretrained tokenizer. We use a batch size of 32 sentences and RMSprop as optimizer, reducing the learning rate by half every 5 epochs of no improvement. The best configurations found during hyper-parameter search for each architecture and loss function are shown in Table 4.

Our code is released as open source, available at `github.com/juanmc2005/MetricAMI`.

4.4 Evaluation

We evaluate each model with the macro F1 score of a KNN classifier with $K = 10$ fit with all sentence embeddings from Train. However, given the high class imbalance, the a priori probability of a random embedding being closer to a *non-misogynous* embedding is higher than for a *discredit* one (see Table 1). To circumvent this issue, we penalize the vote for class k by the number of examples from k in Train. We believe this simple classifier to be a better measure for representation quality, as it relates to the separability and compactness properties that we expect from a metric learning model.

5 Results

The results are summarized in Figure 2. With a fixed architecture, it is clear that all loss functions perform equally, with the exception of LSTM with contrastive and triplet loss. As the LSTM encoder is rather shallow (4.4M parameters) in comparison to BERT (110M parameters), it is possible that contrast-based losses need bigger models to perform competitively.

The fact that almost all losses perform equally well shows that, contrary to what we thought, metric learning models perform no better than cross entropy, in contrast to other findings (Srivastava et al., 2019) on face verification. One possible explanation is that the AMI dataset may not contain enough examples or classes for these models to exploit. However, another factor might be responsible for this behavior. One of the key differences of AMI with respect to face verification is the closed-set nature of the problem. An open-set task is evaluated with unseen *classes*, while a closed-set task is evaluated with unseen *instances* of the train-

Loss	Hyper-parameters
Cross entropy	LR $= 10^{-3}$ •
	LR $= 10^{-5}$ ◦
AAM	LR $= 10^{-3}$, $m = 0.05$, $\alpha = 100$ •
	LR $= 10^{-5}$, $m = 0.05$, $\alpha = 100$ ◦
Center	LR $= 10^{-4}$, $\lambda = 1000$ •
	LR $= 10^{-5}$, $\lambda = 0.1$ ◦
Congenerous cosine	LR $= 10^{-3}$, $\alpha = 10$ •
	LR $= 10^{-5}$, $\alpha = 100$ ◦
Contrastive	LR $= 10^{-4}$, $m = 0.25$ •
	LR $= 10^{-6}$, $m = 0.25$ ◦
Triplet	LR $= 10^{-4}$, $\alpha = 1000$ •
	LR $= 10^{-6}$, $\alpha = 1000$ ◦

Table 4: Best hyper-parameter configurations found per loss function. LR stands for learning rate, and m, α and λ are loss parameters (see Section 2). Rows with • correspond to LSTM and rows with ◦ to BERT.

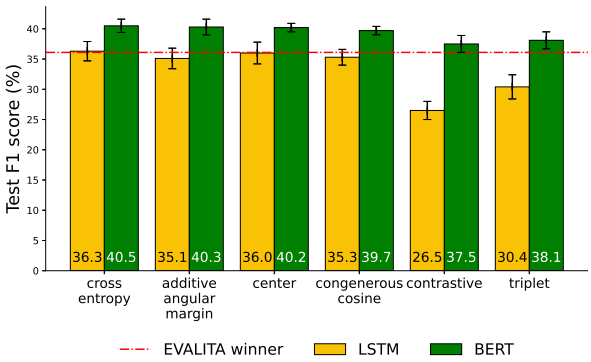

Figure 2: F1 scores on Test for each architecture and loss function. Scores are calculated as the mean of 10 runs and standard deviation is shown as error bars. The baseline of the Evalita 2018 winner (Ahluwalia et al., 2018) is shown for reference.

ing classes. It is possible that open-set verification tasks are more suitable for metric learning than closed-set tasks, meaning that the power of metric learning might in fact lie in generalizing to unseen classes rather than unseen class instances. The fact that verification tasks more closely resemble the training objective than exact class prediction could provide an explanation for this.

On the other hand, our fine-tuned BERT outperforms the Evalita winner baseline (Ahluwalia et al., 2018), setting new state-of-the-art for misogyny categorization, with the added benefit of having comparable embeddings with a simple cosine distance.

As a final note, results in Table 4 suggest that congenerous cosine loss and center loss hyper-parameters could be more sensitive to architecture changes than other losses, as they are the only ones whose best configurations differ from one architecture to the other. Perhaps not surprisingly, we also observe that additive angular margin loss works better with lower margins. This is consistent with the margin's role, serving as an upper bound for the distance between an embedding and its centroid, while the margin in contrastive loss serves as a lower bound for the distance between two negatives.

6 Conclusion

In this work we have addressed the problem of misogyny categorization from a metric learning perspective, comparing the performance of several loss functions. We hypothesized that reducing intra-class variability in this way would be beneficial. However, we have shown that none of the considered losses can outperform the regular cross entropy on the task. Our results suggest that metric learning approaches might not be suited to closed-set sentence classification tasks.

Finally, our fine-tuned BERT sets new state-of-the-art performance, with a macro F1 score of 40.5.

Acknowledgements

This work has been partially funded by the LIHLITH project (ANR-17-CHR2-0001-03), and supported by ERA-Net CHIST-ERA, and the "Agence Nationale pour la Recherche" (ANR, France). It has also been made possible thanks to the Saclay-IA computing platform.

Finally, we would like to thank the reviewers for their useful comments.

References

Resham Ahluwalia, Himani Soni, Edward Callow, Anderson Nascimento, and Martine De Cock. 2018. Detecting Hate Speech Against Women in English Tweets. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian*, pages 194–199. Accademia University Press.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Herve Bredin. 2017. TristouNet: Triplet loss for speaker turn embedding. In *2017 IEEE International Conference on Acoustics, Speech and Signal*

Processing (ICASSP), pages 5430–5434, New Orleans, LA. IEEE.

Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. Tweetaneuse@ AMI EVALITA2018: Character-based Models for the Automatic Misogyny Identification Task. In *Proceedings of the Final Workshop*, volume 12, page 13.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Interspeech*, pages 1086–1090. ISCA.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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, Minnesota. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian*, pages 59–66. Accademia University Press.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *IberEval@ SEPLN*, pages 214–228.

G. Gelly and J.L. Gauvain. 2017. Spoken Language Identification Using LSTM-Based Angular Proximity. In *Interspeech*, pages 2566–2570. ISCA.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1735–1742, New York, NY, USA. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Yu Liu, Hongyang Li, and Xiaogang Wang. 2017. Rethinking Feature Discrimination and Polymerization for Large-scale Recognition. *ArXiv*, abs/1710.00870.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *ArXiv*, abs/1301.3781.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Yash Srivastava, Vaishnav Murali, and Shiv Ram Dubey. 2019. A Performance Comparison of Loss Functions for Deep Face Recognition. *ArXiv*, abs/1901.05903.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A Discriminative Feature Learning Approach for Deep Face Recognition. In *Computer Vision – ECCV 2016*, volume 9911, pages 499–515, Cham. Springer International Publishing.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Sarthak Yadav and Atul Rai. 2018. Learning Discriminative Features for Speaker Identification and Verification. In *Interspeech*, pages 2237–2241.