A Metric Learning Approach to Misogyny Categorization

Juan M. Coria, Sahar Ghannay, Sophie Rosset, Hervé Bredin

5th Workshop on Representation Learning for NLP (RepL4NLP)
at ACL 2020

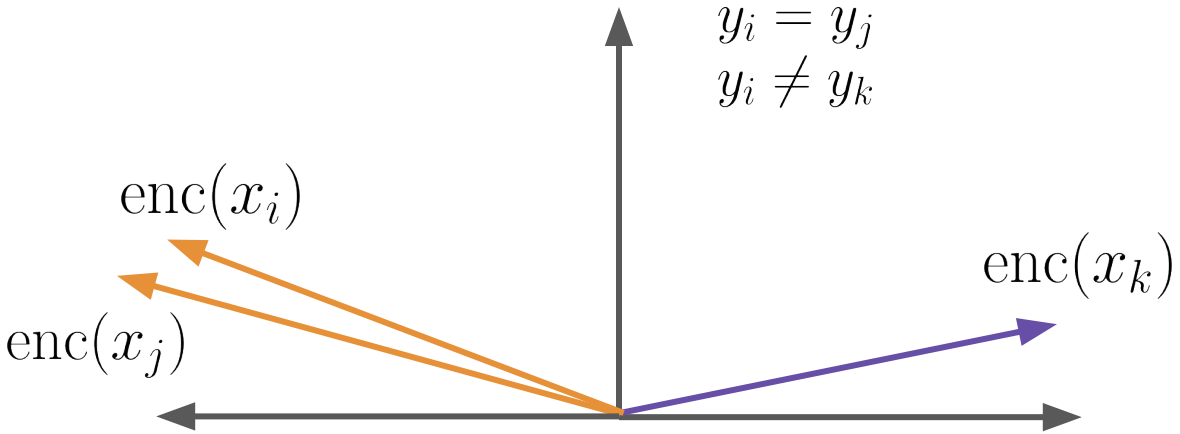
Motivation

Misogyny identification and categorization do not receive as much attention as other NLP tasks.

The Automatic Misogyny Identification (AMI) task of the Evalita 2018 evaluation campaign targeted misogyny identification, categorization, and target detection (Ahluwalia et al., 2018).

Motivation: Metric Learning

Metric learning aims at optimizing a representation function enc so that the distance between representations $\mathrm{enc}(x_i)$ and $\mathrm{enc}(x_j)$ is low if $y_i = y_j$, and high if $y_i \neq y_j$



This is achieved by modifying the model's loss function

Motivation: Metric Learning

Several loss functions have proven useful in face recognition tasks to reduce intra-class variability.

Can we improve sentence representations for misogyny categorization by reducing intra-category variability? (e.g. writing styles, irony, insults, etc.)

We experiment with 5 popular loss functions and 2 different architectures to find an answer to this question.

Loss Functions

We chose to work with the following loss functions:

- contrastive loss
- triplet loss

→ **contrast-based**

- congenerous cosine loss
- additive angular margin loss
- center loss

→ **classification-based**

- cross entropy

→ **baseline**

Corpus

The corpus of the AMI task includes Italian and English versions.

It consists of tweets with three types of annotation:

- Is the tweet misogynist?
- **What type of misogyny is it? (5 categories)**
- Is it targeted to an individual or to a group of people?

We focus on **misogyny categorization in English with an additional class for non-misogynous tweets**.

Corpus: Misogyny Categories

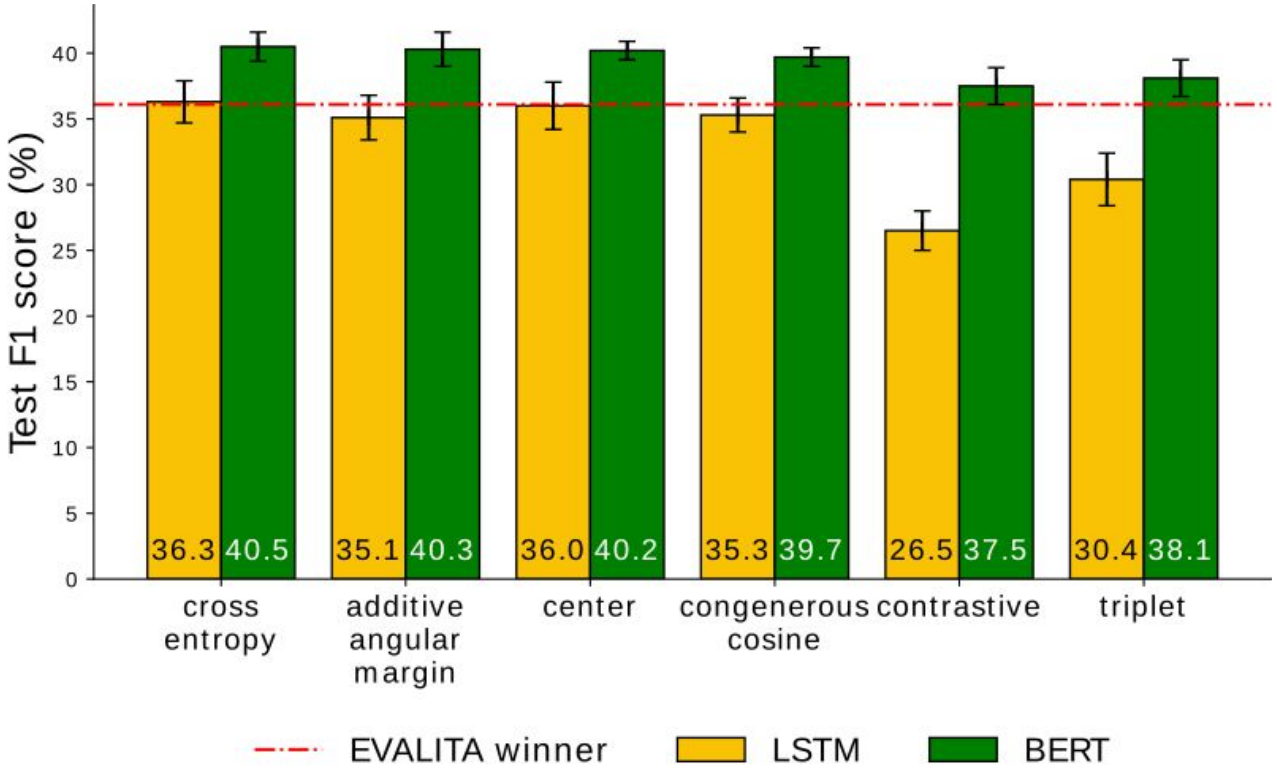
Category	Description	Example
derailing	"to justify women abuse, rejecting male responsibility"	"if rape is real why aren't more people reporting it? just another feminist lie"
discredit	"slurring over women with no other larger intention"	"this b*** is a s***"
dominance	"to assert the superiority of 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 [. . .]"

Experiments

- **Architectures:**
  - Single-layer BiLSTM with word embeddings of size 300 from a CBOW model
  - BERT base uncased
- **Hyper-parameter Search:**
  - 486 different configurations for learning rate and loss parameters
  - Best configurations trained and evaluated 10 times
- **Evaluation**
  - Weighted K-nearest neighbors (K=10) to better measure representation quality
  - Macro F1 score

Code available at github.com/juanmc2005/MetricAMI

Results



1. Contrast-based losses perform poorly and might need larger architectures to perform competitively
2. No loss function outperforms the regular cross entropy loss
3. Our fine-tuned BERT outperforms the best Evalita 2018 model

Discussion

Reduction of intra-class variability does not seem to improve sentence representations for this task.

We think the advantage of metric learning may lie in open-set tasks (like face verification), rather than closed-set tasks (like sentence classification).

Thank you

juanmc2005.github.io

juan.coria@limsi.fr

github.com/juanmc2005/MetricAMI