Automatic detection of sexist content in memes

Gaia Campisi, Silvia Corchs, Elisabetta Fersini, Francesca Gasparini and Monica Mantovani**
Department of Informatics, Systems and Communication, University of Milano-Bicocca,
Viale Sarca 336, Milan, Italy

ABSTRACT

Online social media platforms and websites have become crucial in our society as instruments to define our identity and our relationships through the content we consume. Social issues such as sexism[1] are transmitted and spread online through offensive images and texts conveying several forms of hate against women. Therefore, automatic detection of multimedia sexist content is mandatory and thus we focus on memes. To this end, we have collected and validated a dataset of 800 memes (MIME dataset) and we propose a multimodal classifier which, combining visual and textual features, is able to automatically detect sexist content. Few works are dedicated to the automatic detection of offensive content, using only one type of media. A study [4] was conducted for Youtube in order to detect violence in videos using three different types of media (audio, video and text). Dinakar & al. [2] constructed a corpus of Youtube comments on sensitive topics such as race and tried to classify them using a bag of words driven text classification. The first attempt to explore the field of automatic detection of sexist multimedia content was performed in 2018 by Gasparini et al. [3] considering advertisements. In the same year, Anzovino et al. [1] studied the detection and classification of misogynistic text collected from Twitter.

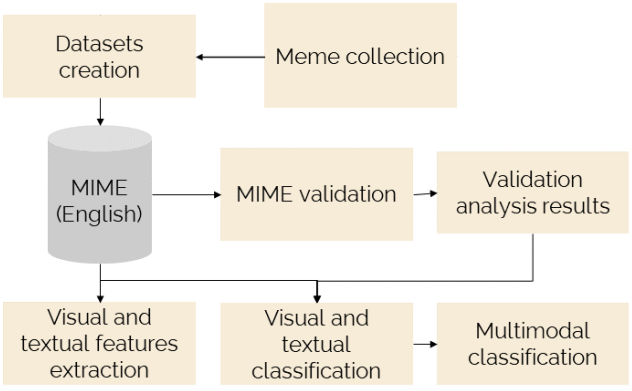

Figure 1: Pipeline for sexist automatic detection.

Our work (Figure 1) started by collecting 800 sexist and non-sexist English memes for the MIME dataset by searching through 89 different sources, including social media, websites and forums, trying to balance the two categories. In order to validate the dataset, we realized a questionnaire on the Figure Eight (https://www.figure-eight.com/) platform on the 800 collected memes. We involved 60 participants: 30 males and 30 females, distributed evenly in three ranges according to age: 20 people between 21–30, 20 people between 31–40 and 20 people between 41–50, obtaining 3 judgements for each meme. After showing a meme at a time, the questionnaire asked the participants if: i) the meme is sexist and in case of a positive answer which was the media that carries the sexist content (text, image, both); ii) the meme is aggressive; and iii) the meme is ironic. Here we focus only on the sexist validation. The results (Table 1) confirm that our starting database is balanced for the division of sexist and non-sexist memes and also, that the judgements were mainly based on text and the union of text and image contents. The image itself was rarely considered sexist by itself.

Table 1: Media involved in the sexist content identification.

Sexist			Non-sexist
46.1%			53.9%
Image	Text	Both	
6%	55.5%	38.5%	

Up to our knowledge, MIME is the first validated dataset in literature regarding multimedia sexist content and its entries contain 1) an *ID number* for univocal identification, 2) *link and source* to find it online (website or social media), 3) the image, 4) *text* of the message written on the meme; 5) the global label with respect to sexism, and 6) the media that carries the sexist content (text, image, both). On this validated dataset we have applied unimodal and multimodal classification methods. After extracting visual and textual features, we produce two different types of multimodal classifiers. In the first one we have combined two unimodal classifiers trained on visual and textual features, using the late fusion approach. In the second one we have combined textual and visual features in an early fusion approach. Comparing the results of classification of the two multimodal classifiers, the late fusion approach shows higher performance in terms of accuracy and recall: 68% early vs 75% late and 65% early vs 79% late. The most important conclusion drawn by this work is that the combination of visual and textual aspects is essential to create a sexist classifier. Future lines of research should consider better multimodal feature representation and proper classification methods.

REFERENCES

[1] 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*. Springer, 57–64.
[2] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *fifth international AAAI conference on weblogs and social media*.
[3] Francesca Gasparini, Ilaria Erba, Elisabetta Fersini, and Silvia Corchs. 2018. Multimodal Classification of Sexist Advertisements. (2018).
[4] Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. 2010. A multimodal approach to violence detection in video sharing sites. In *2010 20th International Conference on Pattern Recognition*. IEEE, 3244–3247.

*All authors contributed equally to this research.
[1]Sexism is an ideology based on discrimination on the basis of gender, often directed to women.