University of Hildesheim at SemEval-2022 task 5: Combining Deep Text and Image Models for Multimedia Misogyny Detection

Milan Kalkenings
University of Hildesheim
Hildeheim, Germany
kalkenin@uni-hildesheim.de

Thomas Mandl
University of Hildesheim
Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract

This paper describes the participation of the University of Hildesheim at the SemEval task 5. The task deals with Multimedia Automatic Misogyny Identification (MAMI). Hateful memes need to be detected within a data collection. For this task, we implemented six models for text and image analysis and tested the effectiveness of their combinations. A fusion system implements a multi-modal transformer to integrate the embeddings of these models. The best performing models included BERT for the text of the meme, manually derived associations for words in the memes and a Faster R-CNN network for the image. We evaluated the performance of our approach also with the data of the Facebook Hateful Memes challenge in order to analyze the generalisation capabilities of the approach.

1 Introduction

Hate in Social Media continues to be a societal problem. The identification of problematic content based on text has made progress, but the performance is still not satisfying (MacAvaney et al., 2019; Modha et al., 2020b). Visual content and multi-modal construction on semantics is a reality in social media today (Dancygier and Vandelanotte, 2017). Systems for realistic scenarios in social media platforms (e.g. (Modha et al., 2020a) require image processing (Sai et al., 2022).

The Multimedia Automatic Misogyny Identification (MAMI) Challenge (SemEval-2022 task 5) is addressing this problem (Fersini et al., 2022). MAMI provides a testbed for algorithms which are capable of processing text and image of memes in one system. The experiments described in this paper measure the effectiveness of different models and their combination into a fusion system. We implemented a basic text classifier based on BERT and an image processing system based on the Faster R-CNN network. In addition, the generalization capabilities between collections are tested. We conducted the experiments with the MAMI dataset as well as with the dataset provided by the Facebook Memes Challenge (Kiela et al., 2021).

2 Previous Work

The detection of Hate Speech can be considered part of Natural Language Processing. Current research is driven by benchmark data and deep learning algorithms have shown to provide best performance.

Data sets such as Offensive Language Detection in Spanish Variants (MeOffendEs@IberLEF 2021) (Plaza-del Arco et al., 2021) and DEtection of TOXicity in comments In Spanish (DETOXIS) (Gonzalo et al., 2021) focus on general concepts of offensive content while other data sets are dedicated to more specific topics than general offensive content. The SemEval 2019 Task-5 (Basile et al., 2019) focused on the detection of hate speech against immigrants and women in Spanish and English messages extracted from Twitter. Besides the main binary task to detect hate speech, there was a fine grained task to further classify into aggressive attitude and the target harassed, to distinguish whether a message contains incitement against an individual rather than a group. The best performing system (Indurthi et al., 2019) trained a SVM model with a RBF kernel using Google's Universal Sentence Encoder (Cer et al., 2018) as features.

The shared task HASOC (Modha et al., 2019) created a large multilingual dataset for hate Speech identification. The first HASOC track focused on the identification of Hate Speech in Indo-European languages (Hindi, English and German). HASOC introduced a binary classification into problematic content and other content.

While most data sets include English data, several recent shared tasks have created new collections for other languages such as Greek (Pitenis et al., 2020) and Turkish (Çöltekin, 2020). Spe-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 718 - 723
July 14-15, 2022 ©2022 Association for Computational Linguistics

cific forms of Hate Speech based on the targeted groups have also been analyzed automatically. For this work, the detection of misogyny is especially relevant. In the Automatic Misogyny Identification (AMI) task at Evalita, a Twitter collection of misogynous messages was assembled. Overall, 10.000 tweets were available and classified into misogynous and non-misogynous tweets. They were also further analyzed into more fine-grained classes(Fersini et al., 2018). The second edition of the Automatic Misogyny Identification (AMI) task in 2020 followed up on binary classification. It also included the prediction of aggressiveness as a binary concept for the misogynous tweets and provided a subtask for the analysis of bias in the models (Fersini et al., 2020). Multi-modal processing of text and image simultaneously has made great progress recently. There are approaches for late fusion which first analyze image and text and combine the representations. Early fusion systems process both text and image in parallel in order to benefit from the dependencies. Systems like ImageBERT (Qi et al., 2020) and Uniter (Chen et al., 2020) have achieved promising results. Uniter relates text and image parts to one another and tries to capture their interaction.

3 Multimedia Datasets

The data for the SemEval-2022 task 5 (Multimedia Automatic Misogyny Identification, MAMI) is described in the overview paper (Fersini et al., 2022). The system presented here is aiming at a binary classification. It did not use the fine-grained classification on kinds of misogyny. In addition and in order to observe how well the multimodal system generalizes over similar datasets from similar tasks, we also processed the Facebook Hateful Meme Challenge (HM) dataset (Kiela et al., 2021). This dataset provides examples for hateful memes in general and includes also other kinds of problematic content than misogyny. However, because the tasks are related a system might also work across these two datasets. Table 1 gives an overview over the two sets. Another multimodal dataset is available for English. Some 700 memes related to the presidential election in the USA in 2016 were collected and annotated (Suryawanshi et al., 2020).

4 System Description

Our system includes six single models. They were all tested as classifiers and we explored several combinations in the experiments.

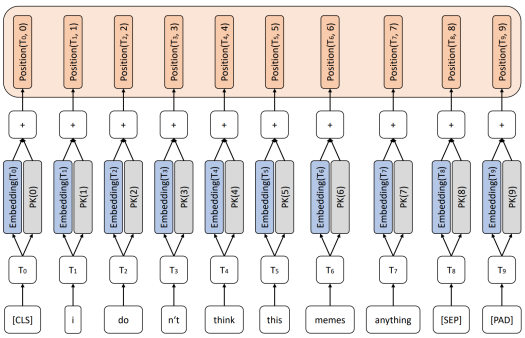

Figure 1: Transformer encoding of meme texts

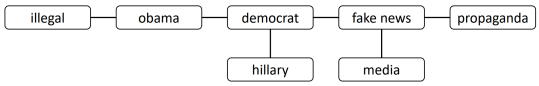

Figure 2: Examples for associations from the knowledge graph

4.1 Text classification with BERT

The first system is processing the text sequence associated with the meme (Devlin et al., 2019). We used the the model *bert-base-uncased* [1]. It creates a transformer based presentation of the text. The principle of BERT is illustrated in Figure 1.

4.2 Associations from Text

The association system includes semantic knowledge to enrich the representation. The assumption is that memes might often use words in another meaning than the obvious one. Looking for associations could help to enrich the representation. The associations might also be misleading. The associations were manually assembled into a knowledge graph. They were extracted after looking at many of the examples from the dataset and observing the intended meaning of many tweets.

For all words in the text sequence, the system looks for associated words in the knowledge graph. For example, the token *Obama* is related to *democrat* and to *illegal* (see figure 2). These relations represent world knowledge and prejudice which can be helpful for understanding the memes. A similar approach to incorporate knowledge graphs in classification systems has been taken by Liu et. al. (Liu et al., 2020). The approach is illustrated in Figure 3.

[1] https://huggingface.co/bert-base-uncased

Feature	HM	MAMI
Number of memes	9.000	10.000
Number of hateful memes	3.300	5.000
Characters per text meme	62	101
Recognized objects per meme	2	2
Memes with recognized objects	7856	7777
Recognized associations per meme	2	2
Memes with recognized associations	3968	5419

Table 1: Dataset Statistics

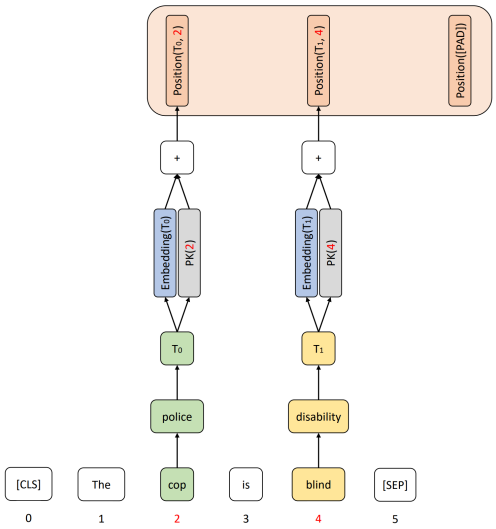

Figure 3: Graph encoding

Figure 4: Object Detection for one meme

The tokens for the associations are extracted from the BERT model and given the position encoding of the word from the original text. All associations found are used. If more associations are found than the sequence length, then the last ones are cut off.

4.3 Sentiment Analysis in Text

In the approach presented here, sentiment analysis is used on the text of the meme. The rationale is that highly emotional texts could indicate a tendency toward hatefulness. Overall, six values are collected. The system VADER suggested by Hutto [2] is used to obtain the first two values for the overall sentiment and intensity (Hutto and Gilbert, 2014). The third and fourth value record the maximum and minimum values for sentiment for all tokens. The method of Loria [3] was used to obtain a measure of subjectivity and of sentiment.

[2]https://github.com/cjhutto/vaderSentiment
[3]https://textblob.readthedocs.io/

4.4 Faster R-CNN-Network

The first visual feature classifier is built with an object recognition system. It uses a Faster R-CNN-Network (Ren et al., 2017) as available online [4]. This system identifies interesting regions which contain much information. The visual features of these regions and their location embedding are fed into a ResNet system. Only the N most likely objects are used. Layer normalization is applied to obtain a final embedding. An example for a result of the object recognition for the dataset is given in figure 4.

4.5 Tile Approach for Image Analysis

A tile approach is splitting the image into 196 rectangular tiles of equal size. These are tiles are analyzed by CNNs and processed as suggested by (Dosovitskiy et al., 2021) and (Lin et al., 2021). The resulting feature vector is associated with a

[4]https://pytorch.org/vision/stable/models.html,

number indicating the position of the tile.

The ResNet architecture is used to obtain an embedding of the entire image. The output embedding is split in two parts of the same size which are used for further processing (Huang et al., 2020).

5 Experiments

For combining the single classifiers, they are fed into a fusion system. For that processing step, a transformer is used. After a layer normalization (Ba et al., 2016), all embedding values are concatenated and fed into a transformer. A sigmoid function is used for the final prediction. First, the models were tested individually. Then, combinations were tested. For finding the optimal fusion of the classifiers above, we applied Sequential Forward Selection (SFS) and Sequential Backward Elimination (SBE). For all experiments, the models were fully trained. Learning rates were adapted for each underlying model so that models converging faster did not overfit. Models with larger embeddings were assigned a higher dropout in the fusion system.

For SFS, all systems were tested individually first and the best system is used as the first component of the combined system. Afterwards, SFS iteratively adds further components. In each iteration, the current version of the combined system is extended by each individual system that is not part of the combined system yet. The combination that leads to the best improvement is taken as the new best combination. SFS converges, when no further improvement can be accomplished. SBE works similar to SFS but starts with a combination of all systems and iteratively detaches individual systems from the combined system. The submitted result was assigned the team name milan_kalkenings.

6 Results and Discussion

The main results refer to the training on the two datasets. Furthermore, we used the two datasets for training and testing respectively. These cross-dataset experiments are reported in the subsequent section.

6.1 Experiments within Datasets

First, the classification by each system individually was tested. As Table 2 shows, the best performance was given by the text classification system. It was followed by the associations system which is another system based on text analysis.

System	AUC-ROC
Text	0.6617
Sentiment	0.5706
Associations	0.6588
Image	0.5958
Tiles	0.5633
Object detection	0.5607

Table 2: Results for each system

The SFS selection method for the SemEval task 5 (MAMI) led to the following optimal combination: text, objects and associations (0.8509 AUC-ROC score). The SFS selection method for HM led to the following optimal combination: text, sentiment, associations and tiles (0.7136 AUC-ROC score). SBE led to the best performance for MAMI with the following combination: text, sentiment, associations and tiles (0.7136 AUC-ROC score). For HM, SBE gave best performance for this set of featues: text, associations and objects (0.8556 AUC-ROC score). The fusion led to improved scores as compared to processing one single modality. It is obvious, that text based metrics are more often in the optimal set.

Selection method	Systems	AUC-ROC
SFS	text, object detection and associations	0.8509
SBE	text, sentiment, associations and tiles	0.7136

Table 3: Results for fusion systems on the MAMI dataset

After the optimal fusion of single systems was determined, we obtained the performance on the test data. Table 4 reports the experiments for the optimal combination as well as for late fusion.

The HM dataset includes many benign confounders which either modify the text of a meme to

Experiment	AUC-ROC	Recall	Precision
HM	0.7146	0.6134	0.5334
HMlate	0.7069	0.3131	0.6712
MAMI	0.8421	0.8217	0.7261
MAMIlate	0.846	0.8529	0.7321

Table 4: Results of experiments with the two datasets

change its class or use the text of a hateful meme in another image. These were introduced to make the task more challenging (Kiela et al., 2021). Leaving out these duplicates changes the performance drastically. Leaving out the memes with identical text increases the performance by some 10%. On the contrary, leaving out the memes with identical images decreases the performance by some 20%. This shows again the impact of the text for this task.

6.2 Experiments Across Datasets

Across datasets, we first trained a classifier for distinguishing between the two datasets. That turned out to be fairly easy for the system (0.94 AUC-ROC). It seems that there are inherent features inserted during data creation which make that distinction easy for systems. Pretraining with the other dataset does not lead to a better performance overall. Only for the MAMI dataset, a performance close to the best overall performance was achieved.

7 Conclusion

The experiments have shown that the identification of hateful memes is still a challenging problem. In our experiments, text features are the most beneficial ones for the system. The influence of associations in particular needs to be further analyzed. First analysis seems to suggest that the number of found associations has a correlation with the performance for the problematic class. The performance across the two datasets is not optimal. Further datasets are needed to analyze the generalization across different collections. In future work, we intend to analyze the impact of each system for the fusion in more detail. We also plan to experiment whether training with a misogyny text dataset can be beneficial for a multimodal system.

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proc.13th Intl. Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for english. In *Proc. Conf. on Empirical Methods in Natural Language Processing, EMNLP: Brussels, Oct. 31 - Nov. 4*, pages 169–174.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Çagri Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proc. 12th Language Resources and Evaluation Conference, LREC Marseille, France, May 11-16*, pages 6174–6184. European Language Resources Association.

Barbara Dancygier and Lieven Vandelanotte. 2017. Viewpoint phenomena in multimodal communication. *Cognitive Linguistics*, 28(3):371–380.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop onSemantic Evaluation (SemEval)*. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 task on automatic misogyny identification (AMI). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. AMI @ EVALITA2020: automatic misogyny identification. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Julio Gonzalo, Manuel Montes-y-Gómez, and Paolo Rosso. 2021. Iberlef 2021 overview: Natural language processing for Iberian languages. In *Proc. Iberian Languages Evaluation Forum (IberLEF) co-located with Conference of the Spanish Society for Natural Language Processing (SEPLN), XXXVII*, volume 2943 of *CEUR Workshop Proceedings*, pages 1–15. CEUR-WS.org.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *CoRR*, abs/2004.00849.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. The AAAI Press.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proc. 13th Intl. Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA. ACL.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, et al. 2021. The hateful memes challenge: competition report. In *NeurIPS 2020 Competition and Demonstration Track*, pages 344–360. PMLR.

Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, Jie Zhang, Jianwei Zhang, Xu Zou, Zhikang Li, Xiaodong Deng, Jie Liu, Jinbao Xue, Huiling Zhou, Jianxin Ma, Jin Yu, Yong Li, Wei Lin, Jingren Zhou, Jie Tang, and Hongxia Yang. 2021. M6: A Chinese multimodal pretrainer. *CoRR*, abs/2103.00823.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: enabling language representation with knowledge graph. In *Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI, Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI, New York, NY, USA, Febr. 7-12*, pages 2901–2908. AAAI Press.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8):1–16.

Sandip Modha, Prasenjit Majumder, Thomas Mandl, and Chintak Mandalia. 2020a. Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance. *Expert Syst. Appl.*, 161:113725.

Sandip Modha, Thomas Mandl, Prasenjit Majumder, and Daksh Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European Languages. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, Dec. 12-15*, volume 2517, pages 167–190. CEUR-WS.org.

Sandip Modha, Thomas Mandl, Prasenjit Majumder, and Daksh Patel. 2020b. Tracking hate in social media: Evaluation, challenges and approaches. *SN Comput. Sci.*, 1(2):105.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proc. 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Flor Miriam Plaza-del Arco, Marco Casavantes, Hugo Jair Escalante, M Teresa Martín-Valdivia, Arturo Montejo-Ráez, Manuel Montes, Horacio Jarquín-Vásquez, Luis Villaseñor-Pineda, et al. 2021. Overview of MeOffendEs at IberLEF 2021: Offensive language detection in Spanish variants. *Procesamiento del Lenguaje Natural*, 67:183–194.

Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. ImageBERT: Cross-modal pre-training with large-scale weak-supervised image-text data. *CoRR*, abs/2001.07966.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.

Siva Sai, Naman Deep Srivastava, and Yashvardhan Sharma. 2022. Explorative application of fusion techniques for multimodal hate speech detection. *SN Comput. Sci.*, 3(2):122.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, May 2020*, pages 32–41. European Language Resources Association (ELRA).