Explainable Classification of Internet Memes

Abhinav Kumar Thakur[1], Filip Ilievski[1,*], Hông-Ân Sandlin[2], Zhivar Sourati[1], Luca Luceri[1], Riccardo Tommasini[3] and Alain Mermoud[2]

[1]*University of Southern California, Information Sciences Institute, USA*
[2]*Cyber-Defence Campus, armasuisse Science and Technology, Switzerland*
[3]*Institut National des Sciences Appliquées, France*

Abstract

Nowadays, the integrity of online conversations is faced with a variety of threats, ranging from hateful content to manufactured media. In such a context, Internet Memes make the scalable automation of moderation interventions increasingly more challenging, given their inherently complex and multimodal nature. Existing work on Internet Meme classification has focused on black-box methods that do not explicitly consider the semantics of the memes or the context of their creation. This paper proposes a modular and explainable architecture for Internet Meme classification and understanding. We design and implement multimodal classification methods that perform example- and prototype-based reasoning over training cases, while leveraging both textual and visual SOTA models to represent the individual cases. We study the relevance of our modular and explainable models in detecting harmful memes on two existing tasks: Hate Speech Detection and Misogyny Classification. We compare the performance between example- and prototype-based methods, and between text, vision, and multimodal models, across different categories of harmfulness (e.g., stereotype and objectification). We devise a user-friendly interface that facilitates the comparative analysis of examples retrieved by all of our models for any given meme, informing the community about the strengths and limitations of these explainable methods.

Keywords

explainability, neuro-symbolic integration, case-based reasoning, internet memes

1. Introduction

In the Internet era, the real and the virtual worlds are becoming increasingly closer, and nearly every person, event, and idea have a Web counterpart. A particularly unique information medium that arises in this context is the **Internet Meme (IM)**: "a piece of culture, typically a joke, which gains influence through online transmission" [1]. An IM is based on a medium, typically an image representing a well-understood reference to a prototypical situation within a certain community. IMs have been extremely popular: according to a recent survey by Facebook, 75% of people between 13 and 36 share Internet Memes (IMs), and 30% do it daily.[1] IMs can be viral: following another study [2], 121,605 different variants of one particular meme were

NeSy 2023, 17th International Workshop on Neural-Symbolic Learning and Reasoning, Certosa di Pontignano, Siena, Italy
*Corresponding author.
akthakur@isi.edu (A. K. Thakur); ilievski@isi.edu (F. Ilievski); hongan.sandlin@ar.admin.ch (H. Sandlin); souratih@isi.edu (Z. Sourati); lluceri@isi.edu (L. Luceri); riccardo.tommasini@insa-lyon.fr (R. Tommasini); alain.mermoud@ar.admin.ch (A. Mermoud)


CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://www.facebook.com/notes/10158928003998415, accessed 17/12/2022.

posted across 1.14 million status updates.[2] As multimodal (they combine visual and language information creatively), relatable (and, thus, dependent on community and virtual context), succinct (they spread complex messages with a minimal information unit that connects the virtual circumstances to the real ones), and fluid (subject to variations and alterations), IMs provide essential data to study the flux of ideas on the Web.

Conversely, as online platforms have been already weaponized in a variety of geo-political events and social issues [4, 5, 6, 7], the scale and complexity of IMs make content moderation even more difficult. The inaccurate classification of memes can lead to inadequate moderation interventions (removal, flagging, demotion, etc.) that, combined with the lack of tracking mechanisms across platforms, has the potential to further decrease public trust in social media platforms and related moderation policies. Content moderation policies, or the lack thereof, can have serious implications on individuals, groups, and society as a whole. On the one hand, content moderators may react late, inconsistently, or unfairly, thus angering users [8], as well as contributing to reinforcing and exacerbating conspiratorial narratives [9, 6]. On the other hand, minimal content moderation may permit coordinated influence operations [10] or enable the spontaneous formation of toxic and dangerous communities, e.g., the study by Mamié et al. demonstrates how "the Manosphere", a conglomerate of men-centred online communities, may serve as a gateway to far-right movements. At this scale, moderation of IMs requires machine-augmented methods for tracking and hate speech classification of IMs.

Existing computational work on IMs has recognized the need to assist content moderators. Much of this work has focused on tracking their temporal spread over time (i.e., virality) [12, 13, 14]. The recent introduction of two tasks on the topic of hate speech, namely Hateful Memes [15] and Misogyny identification [16], has inspired methods that classify IMs based on unimodal or multimodal features, largely dominated by perceptual information. While these methods are a step in the right direction, they are typically modeled as black boxes and optimized for accuracy. Considering the complex interplay of text, vision, and background knowledge in IMs, the decisions reached by such models cannot be trusted by human stakeholders. To assist human moderators and social scientists to understand the semantics and the pragmatics of IMs at scale, these methods must be designed with explainability as a requirement.

In this paper, we explore *explainable multimodal methods for IM classification.* We rely on the general idea of Case-Based Reasoning (CBR), where a prediction can be traced back to similar memes that the method has observed at training time. Considering the complex nature of IMs, we opt for CBR because it can provide transparent insights into the model reasoning, while still leveraging the representation learning ability of state-of-the-art (SOTA) models. Based on these premises, this paper provides three key contributions:

1. We devise an **explainable framework for IM classification**, consisting of explainable methods that perform CBR over features that represent individual memes. We adopt representative CBR methods based on example- and prototype-based reasoning over text, vision, and multimodal feature extractors.
2. We **evaluate the accuracy and the explainability** of our framework on two tasks: Hate Speech Detection and Misogyny Classification, and across different categories of

[2]While the term *meme* usually refers to biological memes [3], in this paper, we treat it as synonymous to Internet Meme (IM).

harmfulness (e.g., stereotype and objectification). We perform ablation experiments to understand the impact of different modalities and modeling choices.

3. We devise a **user-friendly interface** that facilitates the comparative analysis of examples retrieved by all of our models for any given IM. We apply the user interface to understand the ability of different explainable models to retrieve useful instances for CBR and inform future work about the strengths and limitations of these methods.

We make our code available to facilitate future research on explainable IM classification.[3]

2. Related Work

Most prior works on Internet Memes in AI have focused on understanding their virality and spread on social media over time [12, 13, 14]. Another popular direction has been detecting forms of hate speech in memes. The Hateful Memes Challenge and Dataset [15] is a competition accompanied by an open-source dataset with over 10 thousand examples, where the goal is to leverage vision and language understanding to identify memes with hateful content. Kirk et al. [17] compare memes in this challenge to memes in the 'wild', observing that extraction of captions is an open challenge, and that open-world memes are more diverse than memes in curated benchmarks. The Multimedia Automatic Misogyny Identification (MAMI) [16] challenge asks systems to identify misogynous memes, based on both text and images in the targeted input memes. Methods for these challenges typically employ Transformer-based models that incorporate vision and language, like ViLBERT [18], UNITER [19], and CLIP [20]. For a more comprehensive overview of methods for detecting hate speech in memes, we refer the reader to the recent review by Hermida and Santos [21]. Sheratt [22] aims to organize memes into a genealogy, with the goal of building a comprehensive knowledge base going forward. The combination of efforts to explain IMs with explicit knowledge and the generalization power of large visual, textual, and multimodal models holds a promise to advance the SOTA of meme understanding and classification. However, to our knowledge, no prior work has focused on such multi-faceted and explainable methods for understanding IMs. To bridge this gap, we design a modular architecture that integrates visual and textual models with prototype- and example-based reasoning methods. Our framework thus balances the goals of obtaining SOTA performance and providing transparent access to the model reasoning.

There has been a surge in using example-based explanations to enhance people's comprehension of black-box deep learning models' behavior and acquired knowledge. [23] propose and evaluate two kinds of example-based explanations in the visual domain. The extracted similar training data points help the end-users understand and recognize the capabilities of the model better. Although [24] come to the same conclusion as [23] confirming the effect of examples to boost the comprehension of the model by end-users, they do not see any evidence supporting the same effect about the trust of end-users when presented with example-based explanations. Similarly, methods for prototype-based classification have been developed for visual tasks in the past, such as xDNN [25]. However, to our knowledge, we are the first work to employ example-based and prototype-based methods for downstream tasks of IM classification.

[3] https://github.com/usc-isi-i2/meme-understanding

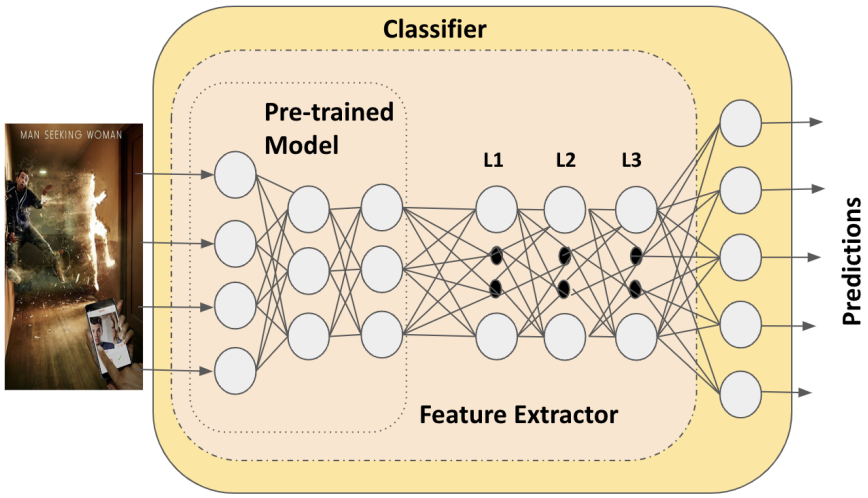


Figure 1: Classification and feature extraction model within the example-based explanation model.

3. Explainable Framework for IM Classification

An IM classification model that detects offensive or inappropriate IMs can be easily trained. However, the black-box nature of ML models makes it difficult to interpret why a meme is flagged [26], especially when misclassified. To address this limitation, we adopt a neuro-symbolic framework with explainable example- and prototype-based predictions for IM classification tasks. Both methods utilize a frozen pre-trained model to extract features from a meme in a transfer learning setup with a separate downstream classification model, which leverages the features to make a final decision. The modularity of the framework enables an easy comparison over the combination of the feature extraction model and the explanation method used.

3.1. Explainable methods

Example-Based Meme Classification [27] categorizes IMs and explains them based on similar memes found in the training data. Example-based explanation works by retrieving training examples that have a similar representation to the test example from the model's point of view to act as a proxy to understand the model's behavior. This approach provides a direct insight into the model reasoning, enabling users to analyze errors and detect latent biases in the dataset [28]. Figure 1 depicts the IM classification model, which applies a classification head (L1-L3 and Predictions layer) over the features extracted from a frozen pre-trained model for prediction. The last hidden state (output of L3) of the trained classifier is used for calculating the similarity between IMs based on cosine similarity. Then, for an unlabeled meme, we predict the labels using this classification model. The features extracted from a pre-trained model can be fed into a query engine to select similar images from a database that stores pre-computed features computed by the same pre-trained model for the training memes. To display the retrieved similar memes in a user-friendly way, we develop a visualization tool to display the model-wise

predictions and similar memes from the training dataset, thus supporting the predictions with example-based explanations.

Prototype-Based Meme Classification is based on the prototype theory [29] of categorization in psychology and cognitive linguistics. The prototype theory dictates that any given concept in any given language has a real-world example that best represents this concept, i.e., its *prototype*. For other concepts, there is a graded degree of belonging to a conceptual category, and some members are closer to the prototype than others. The prototype-based classification relies on learning label-wise prototypes from the training dataset followed by a rule-based decision algorithm for the classification, which makes these models inherently interpretable. We adopt the method of Explainable Deep Neural Networks (xDNN) [25] (Figure 5 in the appendix), which combines statistical learning and reasoning to make an explainable prediction. xDNN uses a neural backend to automate feature extraction followed by multiple statistical layers that dynamically learn from the underlying data distribution. The learning objective of xDNN is to create classwise prototypes, formalized as the local peaks for class distribution representing the data cloud in the prototype's vicinity. At inference time, given a new sample, xDNN computes similarities over the features from the feature extraction layer against all the prototypes. Based on these similarities, the new sample is classified following a rule-based decision approach. The decision-making consists of two steps: *(i) Local (per class) Decision Making:* finding the classwise prototype with the highest similarity, *(ii) Global Decision Making:* comparing the best prototypes across all the classes and choosing the one with the highest similarity to decide the final label. While the explanations of the example-based method are extracted post hoc, those in the prototype-based method are part of the decision-making model.

Both example-based and prototype-based classifiers are instances of case-based reasoning, and there has been some controversy over the superiority of one over the other. There are both claims about the superiority of prototypical examples over normal examples [30], as well as their counterparts [31] who state that a context theory of classification, which derives concepts purely from exemplars, works better than a class of theories that included prototype theory.

3.2. Pretrained Models for Feature Extraction (FE)

We chose the following pre-trained models for feature extraction to analyze the information captured by models trained over different modalities and pretraining strategies.

Textual Models. We use **BERT**$_{base}$ [32], trained on BooksCorpus (800M words) and English Wikipedia (2,500M words) using two unsupervised tasks of Masked LM and Next Sentence Prediction. We expect that BERT would help analyze explainability for general-purpose formal language. Expecting that command of slang in social media is essential for meme understanding, we use the **BERTweet** model [33] having the same architecture as BERT$_{base}$ and trained using the RoBERTa [34] pretraining procedure over 80 GB corpus of 850M English tweets. We expect BERTweet to be a better fit for encoding meme text as tweets have short text length and generally contain informal grammar with irregular slang vocabulary, similar to IMs.

Vision Models. To capture visual information, we used the **CLIP** (Contrastive Language-Image Pre-training) model [35]. CLIP is trained with Natural Language Supervision over 400 million (image, text) pairs collected from the Internet with the contrastive objective of creating similar features for an image and text pair. Because of the variety of training data and unrestricted

Table 1
MAMI dataset characteristics.

Sets	Total	Misogynous	Shaming	Stereotype	Objectification	Violence
Training	10,000	5,000	1,274	2,810	2,202	953
Test	1,000	500	146	350	348	153

text supervision, CLIP reaches SOTA-comparable zero-shot performance over various tasks like fine-grained object classification and action recognition in videos. CLIP is robust to the distribution shift between various datasets and shows better domain generalization across datasets.

Mixed Models. As IMs are based on a complex and creative interplay between textual and visual information [36, 37], we apply mixed FEs to capture both graphical and textual information simultaneously. To do so, we concatenate features from both BERTweet and CLIP together.

4. Experimental Setup

We experiment with meme classification tasks over two existing datasets: MAMI and Hateful Memes.

SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification (MAMI) [16] consists of two sub-tasks of misogyny detection and its type classification. *Sub-task A: Misogyny Detection Task* focuses on detecting whether a meme is misogynous. The inter-annotator Fleiss-k Agreement for sub-task A is 0.5767. *Sub-task B: Misogyny Type Classification Task* is a multi-class task that categorizes a meme into one or more misogynous types, namely, shaming, stereotype, objectification, and violence. A more formal description of these categories can be found in [16]. The inter-annotator Fleiss-k Agreement for sub-task B is 0.3373. The inter-annotator Fleiss-k Agreement clearly shows that sub-task B is comparatively more difficult than sub-task A. Data statistics for both sub-tasks of the MAMI dataset are presented in Table 1.

Hateful Memes [38] consists of a single task of meme hate detection. The dataset consists of 10K memes equally divided into hateful and not-hateful classes; the dev and test set consist of 5% and 10% of the dataset, respectively. The average human accuracy on this task is 84.70%.

Evaluation. We keep our evaluation of classification performance consistent with the original paper about the MAMI dataset. *Sub-task A* is evaluated using macro-average F1 measure for each class label (misogynous and not misogynous). Likewise, *Sub-task B* is evaluated using weighted-average F1 measure, weighted by the true label count for each label. For the Hateful Meme dataset, we also follow the original work and evaluate the models based on classification accuracy. In addition, we manually evaluate the example-based explanation approach using the visualization tool by analyzing the prediction and similar memes from the training dataset. We evaluate the prototype-based explanation method (xDNN) by its classification performance and manually investigating the prototypes identified from the training dataset.

Table 2
Classification results for MAMI and Hateful Memes. We report results for our four FEs and two methods. To contextualize these results, we provide statistics about the six baselines of the MAMI task and the eleven baselines of the Hateful Memes task, as well as human accuracy when available.

Method	Model	Misogyny detection	Misogyny type classification	Hateful Memes
Task Baselines	**Min**	0.481	0.467	0.520
	Mean	0.680	0.663	0.594
	Max	0.834	0.731	0.647
Prototype -based method (xDNN)	**BERT**$_{Base}$	0.537	0.524	0.496
	BERTweet	0.543	0.534	0.470
	CLIP	0.642	0.629	0.552
	CLIP + **BERTweet**	0.648	0.626	0.554
Example -based method	**BERT**$_{Base}$	0.602	0.589	0.558
	BERTweet	0.600	0.594	0.546
	CLIP	0.685	0.686	0.593
	CLIP + **BERTweet**	**0.701**	**0.688**	**0.609**
Human	-	-	-	0.847

5. Analysis

5.1. Accuracy Analysis

Table 2 shows the performance of individual models within our framework on the MAMI and Hateful memes datasets. The table compares our two methods over different feature extractor models. For context, we show the min, max, and mean scores from the tasks' papers [16, 15].

MAMI v/s Hateful Meme. Between MAMI sub-task A (misogyny detection) and Hate detection over the Hateful Meme dataset, all models perform better on the misogyny detection task. This is intuitive, as the presence of misogyny directly relates to the mention of women (or related terminology), while hate is a more open-ended and multifaceted problem. For both tasks and methods, we also observe that BERTweet, which is trained on Twitter data, performs better than the BERT-based models for the MAMI dataset, though the difference between the two models is relatively small. Thus, exposure to slang on social media data has a positive, yet limited, impact on models for meme content classification. However, for the Hateful meme task, BERT performs better than the BERTweet model due to the shift in distribution between the two datasets. Finally, we note that the result of our best model is consistently better than the average baseline result, but worse than the best result. The performance of our methods could be further improved by fine-tuning the FEs on the downstream tasks.

Prototype-Based v/s Example-Based Explanation. For both datasets and different pre-training models, the example-based method, which uses a neural classification head, performs better than the prototype-based (xDNN) on the same pre-trained model. The prototype-based models rely entirely on the pre-trained features and might lose performance on learning complex patterns, which the deep learning model can learn. However, xDNN is much faster to train than training the neural classifier head, as it needs just a single pass over the training data.

Modality Performance Analysis (Text v/s Image v/s Mixed). For each meme dataset

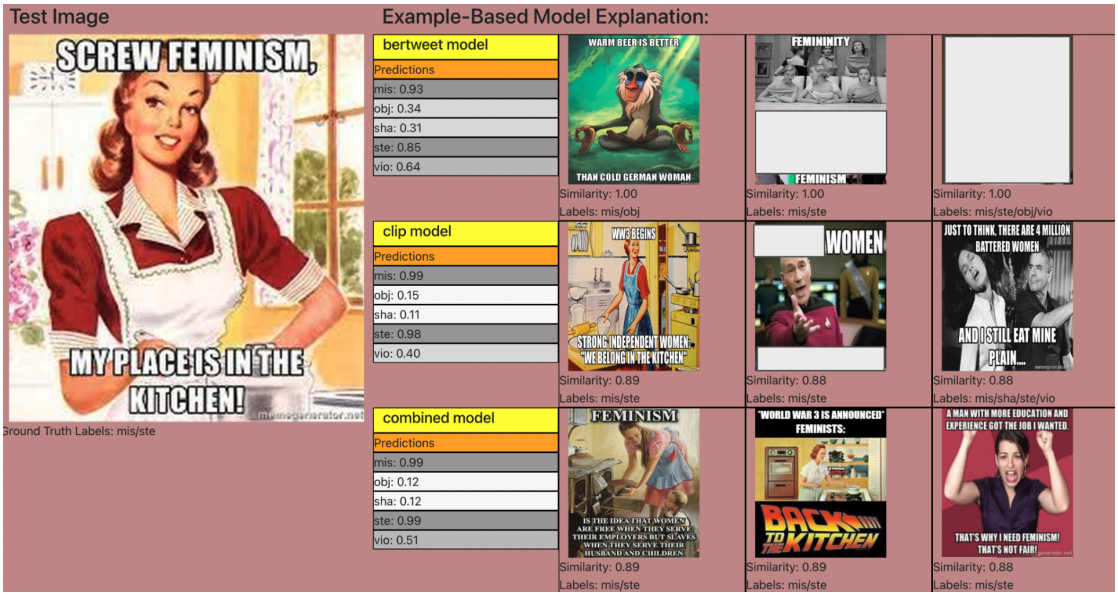
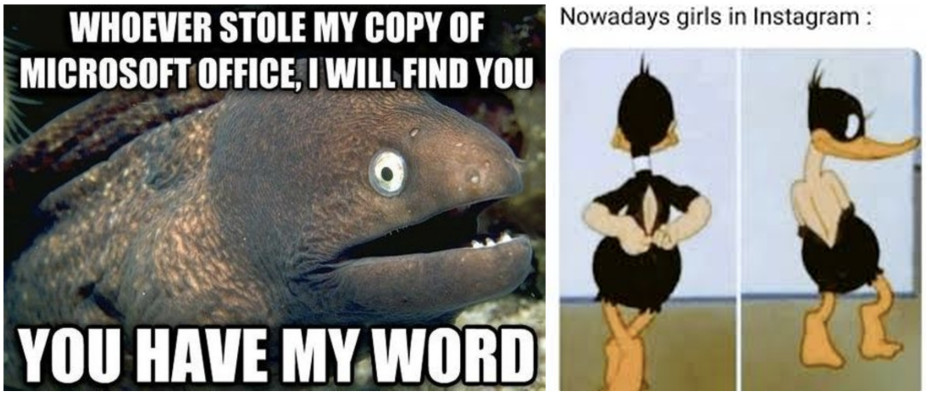


Figure 2: Explanatory interface for our example-based classification method. We cover highly explicit or offensive content with white boxes.

and method combination, the CLIP-based image model performs comparatively better than BERT-based text models. The combined model using CLIP and BERTweet features outperforms all models, including those using CLIP alone. However, the improvement of the joint model over CLIP is relatively low (0.5-1.5 points) compared to the improvement over BERT models (6-10 points) for both explanation strategies. This means that either visual information is more important than text, the CLIP model can also capture the textual information in the IM, or both. For misogyny classification, the combination of CLIP and BERTweet also performs the best, with the CLIP-only model performing very closely as the second best.

5.2. Explanability Analysis

Example-Based Classification Figure 2 shows the most similar IMs retrieved by the example-based classification for one test image, visualized by our custom-made user interface. The interface displays the model-wise predictions for BERTTweet, CLIP, and their combination, together with similar memes from the training dataset for explainability. The test image is misogynous, portraying a stereotype about women. The predictions from each model are correct with high confidence about the misogyny detection and stereotype classification, which is explainable to some degree by looking at similar examples from the training dataset. Focusing on the most similar images per model, we observe that the combined model retrieves three images that also depict misogyny (red background) and stereotyping. The interplay between the text and the vision components is consistent across these IMs, two of which refer to the relationship between the kitchen and women within the context of feminist discourse. This example shows that the instances retrieved by the combined multimodal model are the most reliable, which also correlates with its best performance. The examples by CLIP and BERTTweet



Prototype according to BERTTweet (not misogynous)

Prototype according to CLIP and CLIP+BERTTweet (misogynous: objectification, stereotyping)

Figure 3: Prototypes for the running example IM according to the xDNN method with BERTTweet extractor (left) and CLIP/CLIP+BERTTweet (right).

are partially useful, with CLIP retrieving more relevant images than BERTTweet in most cases. In both the image-only and text-only settings, the retrieved memes relate to the same overall topic e.g., feminism, but their intended meaning is somewhat different than the input IM. This confirms that, although CLIP is retrieving images that are related to the test image and can encode text to some extent, it is still beneficial to combine the features of CLIP with a dedicated language model that understands slang, like BERTTweet. Focusing on the text-setting only, although the model is performing well, we can observe that none of the memes are accurately about the subject of discussion in the test image (women in the kitchen). This confirms our expectation that IM classification should be evaluated both in terms of accuracy and explainability simultaneously, as models might predict correctly for the wrong reasons.

Prototype-Based Classification To our surprise, for both datasets, xDNN creates prototypes equal to the training supports for the class, i.e., the number of prototypes coincides with the number of IMs in the training data. The memes, even though belonging to the same category, can have very different textual/visual information content and representation. To understand better the predictions of the prototype-based classifier, we investigate the most similar prototypes for the same input meme presented in Figure 2. We show the nearest prototype according to BERTTweet on the left of Figure 3, and the prototypical image according to CLIP and the combined model on the right of this figure. The prototype according to BERTTweet is an image that is not misogynous and seemingly unrelated to the input meme. However, the prototypical IM according to CLIP and the mixed feature extractor is more relevant, as it is misogynous and expresses stereotype and objectification, similarly to the input meme. As this IM depicts a cartoon character, its surface similarity to the input is low, but CLIP is seemingly able to connect it to the input meme based on abstract similarity relating to objectification and sexuality. Yet, we note that the retrieved prototypes with CLIP are less relevant compared to the example-based retrieval.

6. Conclusions and Outlook

In this work, we implemented and analyzed example- and prototype-based approaches for explainable Misogyny Identification and Hate Speech Detection in IMs. Our experiments revealed that methods for IMs classification can balance the goals of explainability and good accuracy. While the example-based method is simpler, it achieved higher performance and its explanatory power was more intuitive, as demonstrated through our tool-supported analysis. Among the feature extractors, we observed that vision models were more effective than language models, and the combination achieved the best performance.

A key future direction is integrating background knowledge and figures of speech. The example in Figure 2 includes a test IM which depicts a woman in a kitchen. The stereotype and misogyny, in this case, are most likely linked to assumed background knowledge, such as the women's social status in the 1960s, the second wave of feminism, and the more expansive link between housewives and the kitchen. While the text and the image separately already hint at misogyny, it is their combination that exhibits non-ambiguous misogyny. This is also apparent from the examples extracted by models. Particularly, the most similar meme in terms of content and references, the central image in the bottom row has explicit references to World War 3 and the discussion revolving around two opposing opinions: (1) many Gen Z and Millennial women worried about being drafted, and (2) women wanting equality only until they have to be a part of the draft and joining the military. Moreover, by focusing on the other retrieved IMs, we observe references to various sources. For instance, in the center IM in the last row, we observe the title of Back to the Future movie that substitutes "future" with "kitchen", implying the relation between these two terms. In the center IM in the middle row, we see the Annoyed Picard, the Star Trek character that has long been associated with implying irritability or disappointment that is also extended here towards women. While these cues are captured to some degree by the combined model, still, we see a gap that should be filled by background commonsense and factual knowledge, as well as internet folklore to build robust and explainable meme classification methods. This lack of knowledge is also apparent in the remaining IMs, which seem relevant on the surface but are in fact dissimilar to the input meme.

We intend to explore methods for injecting background knowledge and figures of speech, for instance, by integrating the framework with our Internet Meme Knowledge Graph to tap into rich background knowledge [39]. We also plan to integrate a wider set of models, e.g., VilBERT [18], explore the role of large language model prompting [40], and improve our explanations to be more informative for users, e.g., by using features generated by CLIP in combination with the misogynous type classification predictions. As test data for internet memes is limited, we intend to work towards creating larger and more diverse data for the identification and explanation of hate speech in memes. We make our code available in the hope that the community will help us in pursuing these challenges together.

7. Acknowledgments

The first two authors have been supported by armasuisse Science and Technology, Switzerland under contract No. 8003532866.

References

[1] P. Davison, 9. the language of internet memes, in: The social media reader, New York University Press, 2012, pp. 120–134.

[2] E. A. L. Adamic, T. Lento, P. Ng, The evolution of memes on facebook, Facebook Data Science (2014).

[3] R. Dawkins, The Selfish Gene, Oxford University Press, 1976.

[4] F. Pierri, L. Luceri, E. Ferrara, How does twitter account moderation work? dynamics of account creation and suspension during major geopolitical events, arXiv preprint arXiv:2209.07614 (2022).

[5] G. Nogara, P. S. Vishnuprasad, F. Cardoso, O. Ayoub, S. Giordano, L. Luceri, The disinformation dozen: An exploratory analysis of covid-19 disinformation proliferation on twitter, in: 14th ACM Web Science Conference 2022, 2022, pp. 348–358.

[6] E. Chen, J. Jiang, H.-C. H. Chang, G. Muric, E. Ferrara, et al., Charting the information and misinformation landscape to characterize misinfodemics on social media: Covid-19 infodemiology study at a planetary scale, Jmir Infodemiology 2 (2022) e32378.

[7] F. Pierri, B. L. Perry, M. R. DeVerna, K.-C. Yang, A. Flammini, F. Menczer, J. Bryden, Online misinformation is linked to early covid-19 vaccination hesitancy and refusal, Scientific reports 12 (2022) 1–7.

[8] H. Habib, R. Nithyanand, Exploring the magnitude and effects of media influence on reddit moderation, Proceedings of the International AAAI Conference on Web and Social Media 16 (2022) 275–286. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/19291. doi:10.1609/icwsm.v16i1.19291.

[9] L. Luceri, S. Cresci, S. Giordano, Social media against society, The Internet and the 2020 Campaign (2021) 1.

[10] R. DiResta, S. Grossman, Potemkin pages & personas: Assessing gru online operations, 2014-2019, White Paper https://fsi-live. s3. us-west-1. amazonaws. com/s3fs-public/potemkin-pagespersonas-sio-wp. pdf (2019).

[11] R. Mamié, M. Horta Ribeiro, R. West, Are anti-feminist communities gateways to the far right? evidence from reddit and youtube, in: 13th ACM Web Science Conference 2021, 2021, pp. 139–147.

[12] G. Marino, Semiotics of spreadability: A systematic approach to internet memes and virality (2015).

[13] V. Taecharungroj, P. Nueangjamnong, The effect of humour on virality: The study of internet memes on social media, in: 7th International Forum on Public Relations and Advertising Media Impacts on Culture and Social Communication. Bangkok, August, 2014.

[14] C. Ling, I. AbuHilal, J. Blackburn, E. De Cristofaro, S. Zannettou, G. Stringhini, Dissecting the meme magic: Understanding indicators of virality in image memes, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–24.

[15] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, 2021. arXiv:2005.04790.

[16] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, Semeval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp.

533–549.
[17] H. R. Kirk, Y. Jun, P. Rauba, G. Wachtel, R. Li, X. Bai, N. Broestl, M. Doff-Sotta, A. Shtedritski, Y. M. Asano, Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset, arXiv preprint arXiv:2107.04313 (2021).
[18] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Advances in neural information processing systems 32 (2019).
[19] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Universal image-text representation learning, in: European conference on computer vision, Springer, 2020, pp. 104–120.
[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[21] P. C. d. Q. Hermida, E. M. d. Santos, Detecting hate speech in memes: a review, Artificial Intelligence Review (2023) 1–19.
[22] V. Sherratt, Towards contextually sensitive analysis of memes: Meme genealogy and knowledge base (2022).
[23] C. J. Cai, J. Jongejan, J. Holbrook, The effects of example-based explanations in a machine learning interface, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 258–262. URL: https://doi.org/10.1145/3301275.3302289. doi:10.1145/3301275.3302289.
[24] C. Ford, E. M. Kenny, M. T. Keane, Play mnist for me! user studies on the effects of post-hoc, example-based explanations &; error rates on debugging a deep learning, black-box classifier (2020). URL: https://arxiv.org/abs/2009.06349. doi:10.48550/ARXIV.2009.06349.
[25] P. Angelov, E. Soares, Towards explainable deep neural networks (xdnn), 2019. URL: https://arxiv.org/abs/1912.02523. doi:10.48550/ARXIV.1912.02523.
[26] R. Andrews, J. Diederich, A. B. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge-Based Systems 8 (1995) 373–389. URL: https://www.sciencedirect.com/science/article/pii/0950705196819204. doi:https://doi.org/10.1016/0950-7051(96)81920-4, knowledge-based neural networks.
[27] A. Renkl, Toward an instructionally oriented theory of example-based learning, Cognitive Science 38 (2014) 1–37. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12086. doi:https://doi.org/10.1111/cogs.12086. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12086.
[28] I. Sigler, Example-based explanations to build better ai/ml models, 2022. URL: https://cloud.google.com/blog/products/ai-machine-learning/example-based-explanations-to-build-better-aiml-models.
[29] E. H. Rosch, Natural categories, Cognitive Psychology 4 (1973) 328–350. URL: https://www.sciencedirect.com/science/article/pii/0010028573900170. doi:https://doi.org/10.1016/0010-0285(73)90017-0.
[30] M. K. Johansen, J. K. Kruschke, Category representation for classification and feature inference, J. Exp. Psychol. Learn. Mem. Cogn. 31 (2005) 1433–1458.
[31] D. L. Medin, M. M. Schaffer, Context theory of classification learning, Psychol. Rev. 85 (1978) 207–238.

[32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[33] D. Q. Nguyen, T. Vu, A. Tuan Nguyen, BERTweet: A pre-trained language model for English tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 9–14. URL: https://aclanthology.org/2020.emnlp-demos.2. doi:10.18653/v1/2020.emnlp-demos.2.

[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692. doi:10.48550/ARXIV.1907.11692.

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020. doi:10.48550/ARXIV.2103.00020.

[36] E. Zenner, D. Geeraerts, One does not simply process memes: Image macros as multimodal constructions, Cultures and traditions of wordplay and wordplay research 6 (2018) 9783110586374–008.

[37] B. Dancygier, L. Vandelanotte, Internet memes as multimodal constructions, Cognitive Linguistics 28 (2017) 565–598.

[38] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, 2020. URL: https://arxiv.org/abs/2005.04790. doi:10.48550/ARXIV.2005.04790.

[39] R. Tommasini, F. Ilievski, T. Wijesiriwardene, IMKG: The Internet Meme Knowledge Graph, in: Extended Semantic Web Conference, 2023.

[40] R. Cao, R. K.-W. Lee, W.-H. Chong, J. Jiang, Prompting for multimodal hateful meme classification, arXiv preprint arXiv:2302.04156 (2023).

8. Appendices

8.1. Example-based Classification Process

The example-based classification method process at training and inference time is illustrated in Figure 4.

8.2. Architecture of xDNN

Our architecture for prototype-based explainable classification, called Explainable Deep Neural Networks (xDNN) is shown in Figure 5.

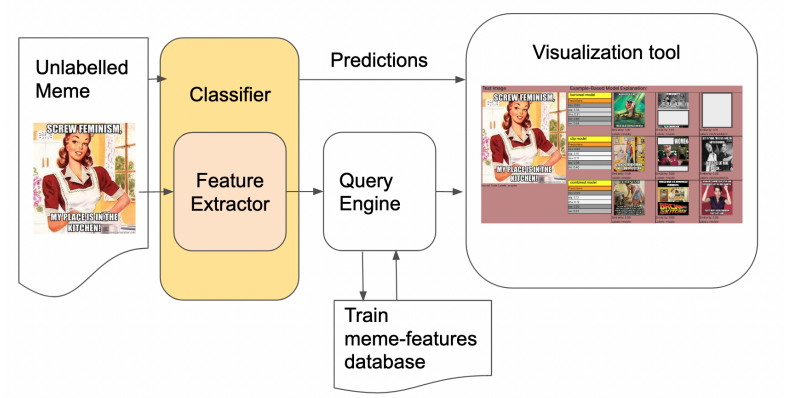


Figure 4: Example-based explanation based on similarity-based meme search. The Train meme-features database contains pre-computed features using the Feature Extractor module.

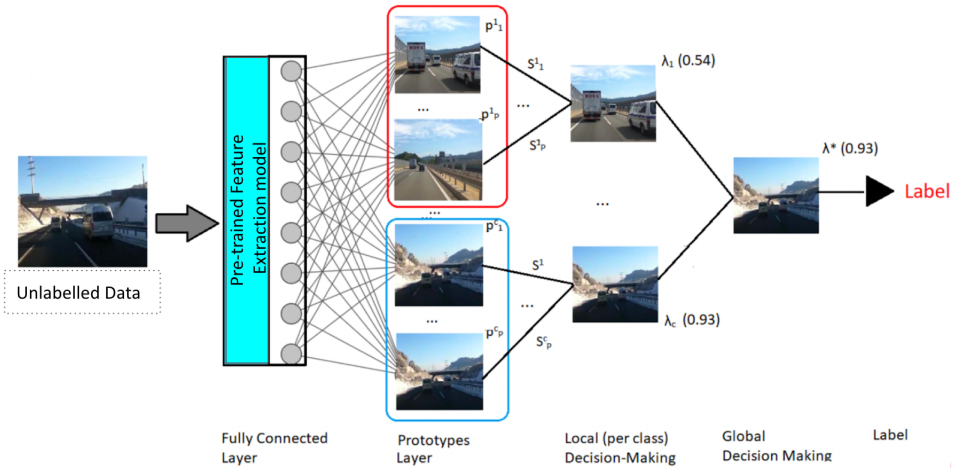


Figure 5: Our architecture for prototype-based explainable classification, called Explainable Deep Neural Networks (xDNN). Figure reused from the original paper [25].

8.3. Model Training Details

The classification model (Figure 1) used in the example-based explanation setup applies a trainable neural head over frozen pre-trained models, which is trained with the Binary Cross Entropy Loss using the Adam optimizer with a learning rate of 10^{-4}. Table 3 describes each layer of the classification head, and the hidden state of **L3** is used for feature extraction for similar example searches over the training dataset.

xDNN [25] is a generative model, i.e., it learns prototypes and respective distributions automatically from the training data with no user/problem-specific parameters. We reuse the publicly available xDNN implementation and experiment with different pre-trained models

described in the subsection on Pretrained Models.[4] Table 3 describes each layer of the classification head, and the hidden state of **L3** is used for feature extraction for similar example searches over the training dataset.

Table 3
Classification Head parameters for the Example-based method.

Layers	Dimension	Activation
L1	Feature length * 512	ReLU
L2	512 * 256	ReLU
L3	256 * 128	ReLU
Prediction	128 * Label count	Sigmoid

[4]https://github.com/Plamen-Eduardo/xDNN---Python