FigMemes: A Dataset for Figurative Language Identification in Politically-Opinionated Memes

Chen Liu*, **Gregor Geigle***†, **Robin Krebs, Iryna Gurevych**

Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt
www.ukp.tu-darmstadt.de

Abstract

Real-world politically-opinionated memes often rely on figurative language to cloak propaganda and radical ideas to help them spread. It is not only a scientific challenge to develop machine learning models to recognize them in memes, but also sociologically beneficial to understand hidden meanings at scale and raise awareness. These memes are fast-evolving (in both topics and visuals) and it remains unclear whether current multimodal machine learning models are robust to such distribution shifts. To enable future research into this area, we first present *FigMemes*, a dataset for figurative language classification in politically-opinionated memes.[1] We evaluate the performance of state-of-the-art unimodal and multimodal models and provide comprehensive benchmark results. The key contributions of this proposed dataset include annotations of six commonly used types of figurative language in politically-opinionated memes, and a wide range of topics and visual styles. We also provide analyses on the ability of multimodal models to generalize across distribution shifts in memes. Our dataset poses unique machine learning challenges and our results show that current models have significant room for improvement in both performance and robustness to distribution shifts. The code and dataset (including splits we used for analyses) are available at: https://github.com/UKPLab/emnlp2022-figmemes.

Warning: We discuss and show memes that may be offensive to readers for illustrative purposes only. They do not represent the authors' or the affiliated institution's views in any way.

1 Introduction

Memes are transmittable units of culture that evolve fast (Dawkins, 1976), which use images and/or text to convey opinions. Figurative language[2] such as metaphors or sarcasm is often used in memes (see Figure 1) to persuade, enhance the impact of ideas, and help these ideas spread (Davison, 2012). This is especially true for internet memes, which rely on figurative language to spread propaganda (Dimitrov et al., 2021), and support violent, discriminatory or radical ideology (Tipler and Ruscher, 2019; Hakokõngäs et al., 2020). For instance, the transmission of dehumanized metaphors for women may help perpetuate negative beliefs about women's roles in modern society (Tipler and Ruscher, 2019), and the usage of rhetorical devices can attract new audiences to radicalized groups that fight against refugees (Hakokõngäs et al., 2020).

Figurative language classification has been challenging in machine learning (Wang et al., 2022; Pramanick et al., 2022; Zhang et al., 2021a) as the task requires understanding world knowledge and commonsense. Despite its difficulty, the task of developing machine learning models for classifying figurative language in politically-opinionated memes (memes that relate to politics or share views on controversial topics) can provide sociological benefits. These benefits include understanding hidden meanings, aiding humanities research, and raising awareness at scale. Yet, there is a lack of appropriate datasets for studying figurative language in these memes (see Table 1). Furthermore, prior datasets for memes classification suffer from significant limitations, such as small size or inductive bias due to keyword-based retrieval.

As politically-opinionated memes evolve quickly based on the latest news and cultural references, it also remains unclear whether current

*Equal Contributions.

†Gregor is now affiliated with WüNLP & Computer Vision Lab, CAIDAS, University of Würzburg.

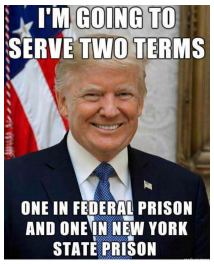
[1]**Disclaimer:** The dataset contains racial slurs and other language/images that may be offensive to the readers. This dataset should only be used for academic research or non-commercial purposes.

[2]The term "language" is overloaded to include visual non-literal expressions, e.g., visual metaphor, etc.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7069 - 7086
December 7-11, 2022

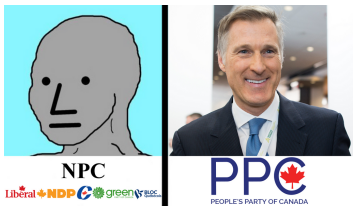
(a) *Allusion*

(b) *Exaggeration/Hyperbole*



(c) *Irony/Sarcasm*

(d) *Anthropo./Zoomorphism*



(e) *Contrast*

(f) *Metaphor/Simile*

Figure 1: Meme examples from each figurative language category, note that we tried to pick the less disturbing memes. **(a)** Alludes to World War II and Nazi Germany. The photo depicts a soldier (with a possible SS symbol partially visible on the helmet). The Luger is a type of pistol used by German soldiers during WWII. **(b)** Use of an exaggerated visual of Italy on a globe. **(c)** "Two terms" sarcastically refers to both the term limit of US presidents and prison terms. **(d)** A personified anarcho-capitalism flag. **(e)** Contrasting the leader of the PPC party (a right to a far-right political party) to leaders of other Canadian political parties, where the other leaders are also metaphorically depicted as a Non-Player Character (it's also labelled as a metaphor). **(f)** Comparing unvaccinated people to Jewish people during the Holocaust is a metaphor (it's also labelled as an allusion).

multimodal models for memes classification are robust to such distribution shifts. For example, a model trained in early 2019 may perform poorly when tested on memes referencing COVID, or one trained on photo-like memes may perform poorly when tested on hand sketches. Earlier work such as Kiela et al. (2020) has neglected this problem.

Despite recent growth in classification tasks for memes (Kiela et al., 2020; Sharma et al., 2020; Fersini et al., 2022), there has been limited work on assessing a model's ability to: 1) identify and understand figurative language in politically-opinionated memes, 2) generalize to a data distribution different from the training data, e.g., across different topics or visual categories.

To bridge these research gaps, we first present a novel and challenging multi-label memes dataset called *FigMemes (**Fig**urative language identification in **Memes**)*. Our proposed task is to identify the type of (one or more) figurative language used in a meme. Then, we conduct analyses by splitting the dataset to create distribution shifts (within each modality, e.g. text or image), aimed at assessing generalization of memes classification models.

The FigMemes dataset contains 5141 politically-opinionated memes collected from the 4chan **/pol/** (the politically incorrect) board, posted between January 2017 and December 2021. Unlike previous work, we do not build the dataset by hand-picking topics using keywords (hence less inductive biases). The dataset covers a wide range of topics and visual categories. We provide comprehensive benchmark results using several state-of-the-art unimodal and multimodal models. The best multimodal model achieves an average F1 score of 46.69% on the proposed dataset, which highlights the difficulty of this task and significant room for improvement in future research.

Our further analyses show that 1) making use of external knowledge, understanding commonsense, and reasoning are important for figurative language understanding in memes, 2) further development of multimodal models for memes that are robust to distribution shifts and datasets for testing such models are needed.

In summary, our contributions are:

- A novel multi-label memes dataset for figurative language classification, covering a wide range of topics and six different figurative language categories.
- A comprehensive benchmark of state-of-the-art unimodal and multimodal models on the proposed task.
- Analyses of whether multimodal memes clas-

Name	Task	w/o kwd*	Real	Multi-class	Multi-label	Size	Topics
HatefulMemes	Hateful	N/A	✗	✗	✗	10k	N/A
MultiOFF	Offensive	✗	✔	✗	✗	743	2016 U.S. Elec.
Jewtocracy	Antisemitism	✗	✔	✔	✗	3102/3509	Antisemitism
HarMemes	Harmful	✗	✔	✔	✗	3544	COVID
Memotion1,2	Emotion	✗	✔	✔	✔	8k/9871	Mixed†
TrollsWithOpinion	Trolls	✗	✔	✔	✗	8881	Mixed†
MAMI	Misogyny	N/A	✔	✔	✔	11k	Misogyny, Mixed
SemEval-2021 Task 6	Propaganda	✗	✔	✔	✔	950	Pol. Op.
FigMemes (Ours)	Fig. Lang.	✔	✔	✔	✔	5141	Pol. Op., Mixed‡

Table 1: Comparison of FigMemes to recent memes datasets. **Pol. Op.** stands for politically opinionated. *: **w/o kwd** indicates the dataset is built without pre-selecting topics based on keywords. ‡: We estimate that less than 10% of the memes are reaction memes to discussions on 4chan /pol/. †: The dataset contains the same set of memes as Memotion (Sharma et al., 2020).

sification models can generalize across distribution shifts.

2 Related Work

2.1 Memes Datasets

We provide a comprehensive overview of existing memes datasets in this section and summarize key parameters of existing memes datasets in Table 1.

HatefulMemes (Kiela et al., 2020) focuses on hateful memes that are artificially constructed from stock photos and designed layouts. Due to the construction, models trained using this dataset struggle to generalize to real internet memes (Kirk et al., 2021).

MultiOFF (Suryawanshi et al., 2020) is a dataset of 743 memes for identifying offensive memes on the topic of the 2016 U.S. election. Due to its data size, it could be limited to evaluating only sample-efficient machine learning models. In addition, the definition of offensiveness can depend on the political or cultural stance of the annotators (Hine et al., 2017; Sap et al., 2022).

Jewtocracy (Chandra et al., 2021) is a dataset for identifying antisemitism in memes. The dataset is constructed using data from social network sites such as Gab and Twitter while keeping data with defined lexicons. It focuses on a single topic and contains lexical biases.

HarMemes (Pramanick et al., 2021) focuses on identifying the harmfulness of memes on the topic of COVID-19. Similar to offensiveness, the definition of harmfulness can depend on the political or cultural stance of the annotators.

Memotion 1,2 (Sharma et al., 2020; Ramamoorthy et al., 2022) are two versions of datasets each containing three different subtasks related to the emotional effects of memes. Relevant to this work, subtask-B is a multi-label task that provides annotations of emotional responses. The dataset was built using an unknown set of keyword searches over social media sources. Based on the derivative work TrollsWithOpinion (Suryawanshi et al., 2022) (below), it is unlikely that the dataset contains a significant fraction of politically-opinionated memes.

TrollsWithOpinion (Suryawanshi et al., 2022) provides additional annotations for the Memotion datasets. The dataset focuses on identifying trolls in the topic areas of politics,[3] products, and others.

MAMI (SemEval-2022 Task 5) (Fersini et al., 2022) contains two subtasks where subtask-A is to identify misogyny in memes, and subtask-B is a multi-class classification of different types of misogyny. The dataset focuses on the topic of misogyny and likely contains a few other politically-opinionated memes.

SemEval-2021 Task 6 (Dimitrov et al., 2021) proposed three subtasks to detect fine-grained propaganda techniques in memes. This dataset was built by following an unknown set of Facebook groups based on keywords. This dataset contains 950 memes in total with more than 20 label classes. Due to the small amount of data per label, it could be limited to training and evaluating only sample-efficient machine learning models.

Features of our FigMemes dataset: To the best of our knowledge, FigMemes is the first dataset containing a wide range of figurative language in real politically-opinionated memes from the internet. Our dataset departs from existing datasets in terms of the task (see Table 1) and topic diversity. We exceed the closest memes dataset (in terms of topics, SemEval-2021 Task 6) by over 4000 memes. Our dataset contains topics such as refugees, racial minorities, U.S elections, Epstein, antisemitism,

[3]Based on an aggregation of statistics from the paper, we estimate less than 600 memes are politically-opinionated.

COVID, LGBTQ+, feminism, etc. See Appendix B for details on topics.

3 Data Collection and Annotations

3.1 Data Source and Collection Process

Images in our dataset are collected from the 4chan **/pol/** (the politically incorrect) board.[4] 4chan contains a collection of anonymous image-boards, in which **/pol/** is a sub-board known to contain a large number of memes that can be hateful, contain discriminating opinions of minorities and different gender identities, or support far-right ideologies (Chandra et al., 2021; Crawford et al., 2021). Yet, the **/pol/** board is influential in internet cultural transmission (Zannettou et al., 2017). Our data collection process used a crawler to systematically gather memes between January 2017 and December 2021 from the **/pol/** board.[5] We removed duplicated images by using the difference hash (dHash) followed by manual inspections. Additionally, we removed sexually-explicit content by hand. Unlike previous approaches, we do not use any keywords to filter the data, hence the dataset covers a wide range of topics naturally.

3.2 Annotations

We built a simple annotation interface with Python PyQt5.[6] A black and white filter was applied to the memes during annotation to protect annotators from potentially gory or obscene memes.

We followed a data-driven approach to identify a suitable set of labels. First, we randomly sampled 200 memes and labelled the most-common figurative language contained in them using free-text annotation, by prioritizing those that have been studied before (in social or computer sciences). We then grouped categories that are overlapping and do not contain enough labels individually. The following major categories emerged from the annotations: *Allusion*, *Exaggeration/Hyperbole*, *Irony/Sarcasm*, *Anthropomorphism/Zoomorphism*, *Metaphor/Simile*, and *Contrast*, which we use as labels. The task is *multi-label* classification.

Our task is defined by the question: "What are the types of figurative language used in this meme?" The definitions for the categories of figurative language are as follows:

- **Allusion:** Referencing historical events, figures, symbols, art, literature or pop culture.[7]
- **Exaggeration/Hyperbole:** Use of exaggerated terms for emphasis, including exaggerated visuals (including unrealistic features portraying minorities).
- **Irony/Sarcasm:** Use of words that convey a meaning that is the opposite of its usual meaning/mock someone or something with caustic or bitter use of words.
- **Anthropomorphism/Zoomorphism:** Attributing human qualities to animals, objects, natural phenomena or abstract concepts or applying animal characteristics to humans in a way that conveys additional meaning.[8]
- **Metaphor/Simile:** Implicit or explicit comparisons between two items or groups, attributing the properties of one thing to another. This category includes dehumanizing metaphors.
- **Contrast:** Comparison between two positions/people/objects (usually side-by-side).

We provide examples of each figurative language category in Figure 1, along with explanations. The final dataset contains 5141 memes in total. We take the majority vote of 3 annotators to determine the labels. The annotations are done by the authors of the paper with an agreement of 0.42 Fleiss Kappa (which we consider moderate). 70% of the memes are annotated with at least one figurative language, and 8.5% of the memes (438) are without text. We used the Google Vision API[9] to extract the texts within memes.

The dataset is randomly partitioned into train, validation, and test sets with proportions of 60%, 10%, and 30%. We refer to this split as ***standard evaluation*** to distinguish between this split and those we used for distribution-shift analysis. Detailed dataset statistics are in Table 2.

4 Distribution Shifts in Memes and Analysis Settings

To evaluate a machine learning model's ability to generalize across distribution shifts, we simulate

[4] http://boards.4chan.org/pol/

[5] We use the API from https://4plebs.org/

[6] https://pypi.org/project/PyQt5/

[7] Memes using photos from a movie are not automatically included.

[8] e.g. In our context, memes with animal/object-based characters are not automatically considered Anthropomorphism/Zoomorphism. For example, SpongeBob or Pepe the Frog etc.

[9] https://cloud.google.com/vision

distribution shifts in individual modalities (i.e., image or text) separately. The images and extracted text are grouped into different categories (visual categories and topic clusters) and used to create different distribution-shifted evaluation settings.

Text Topic Clusters: We create text distribution shifts by using topic clusters based on the text extracted from memes. First, we used Sentence-BERT (Reimers and Gurevych, 2019)[10] to compute feature vectors from the extracted text, as SentenceBERT provides good sentence embeddings for semantic search and paraphrase mining. We then used the hierarchical clustering algorithm[11] to identify clusters. The clusters generally make sense, forming topics such as *exercise* or *school*, see Table 9 for examples.

We define in-distribution (ID) and out-of-distribution (OOD) evaluation scenarios. In the ID scenario, train/validation/test data are all from the same text topic cluster. In the OOD scenario, train/validation data are from a different cluster than test data. This results in one ID evaluation set and two OOD evaluation sets.

Visual Categories: Based on the approach of DomainNet (Peng et al., 2019), a domain adaptation benchmark across different visual categories, we manually labelled and grouped the memes into different visual categories. Our dataset consists of the following visual categories: **1) Artistic** (including clips-arts, illustrations, anime, hand drawings, etc.), **2) Real** (including manipulated photos, realistic CGIs etc.), **3) Infographic**,[12] **4) Mixed**. Please see Figure 5 in the Appendix for examples.

We propose the following two evaluation schemes, and use the Infographic and Mixed categories only as test sets, due to their small data size:

- Training on memes in the Real category. Evaluating on Real memes (ID) and Artistic, Infographic, Mixed memes (3 OOD sets).
- Training on memes in the Artistic category. Evaluating on Artistic memes (ID) and Real, Infographic, Mixed memes (3 OOD sets).

[10]We used the *all-distilroberta-v1* model from https://www.sbert.net/index.html.

[11]https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

[12]The visual elements in most of the infographic-style memes are hand drawings.

4.1 Analysis Settings

Here, we aim to study how distribution shifts affect classification models for memes. We conduct analyses on the proposed FigMemes dataset, as well as two additional datasets (that are multi-label and large in size) to understand whether our findings also exist in memes classification datasets at large.

The datasets included in our analyses are FigMemes (our proposed dataset), Memotion 2 (subtask B) and MAMI (subtask B). Note that our goal is not to achieve state-of-the-art, but rather to understand the effects of distribution shifts. Hence, we created new train/validation/test sets for all datasets.

FigMemes (proposed): We created visual categories using the method described in §4. We used topic clustering to identify ID train/validation/test sets and a single text cluster as the OOD evaluation set. Since there are 438 image-only memes in our dataset, they naturally serve as the second OOD (text) evaluation set.

Memotion 2 (subtask B): Ramamoorthy et al. (2022) is a multi-class multi-label classification task for Funny, Sarcasm, Offensive and Motivational emotional responses. We created topic and visual categories using the methods described in §4 (labelled using the same categories) and combined the training and validation memes to create the distribution-shifted splits. We did not include the test spits from Memotion 2 as the labels are not publicly released.

MAMI (subtask B): Fersini et al. (2022) is a multi-class multi-label classification task on classifying types of misogyny memes, namely Stereotype, Shaming, Objectification and Violence. We created topic and visual categories using the methods described in §4 (labelled using the same categories) and combined all of the training/validation/test memes in this dataset to create the distribution-shifted splits.

Detailed statistics about the splits used for analyses are in Table 7 and Table 8 in the Appendix D.

5 Baseline Models

We provide results of baseline models for the proposed multimodal multi-label classification task. Our selection covers state-of-the-art unimodal text- and image-only models, and multimodal models with and without multimodal pre-training.

Text-only: We use **BERT** (Devlin et al., 2019) (bert-uncased-base) and **DeBERTa** (He et al.,

Split	Allusion	Exag.	Irony	Anthropo.	Metaphor	Contrast	None	Total	1-label	2-label	2^{+}-label	Labels per Meme
Train	515	650	629	282	685	273	853	3084	1558	556	117	0.98
Validation	84	67	92	48	79	55	193	515	236	71	15	0.83
Test	265	265	320	131	286	171	495	1542	715	277	55	0.93
Total	864	982	1041	461	1050	499	1541	5141	2509	904	187	0.95

Table 2: Statistics of the labels. The first word from the category is used if a category consists of more than one type of figurative language (E.g. exaggeration/hyperbole).

2021) (deberta-v3-base) as the text-only models. DeBERTa is a recently proposed state-of-the-art text Transformer model.

Image-only: **ConvNeXt** (Liu et al., 2022) (convnext-base-224), is a state-of-the-art CNN model. **CLIP-CNN**, uses the CNN component of CLIP (RN50x4, see below). We experiment with full fine-tuning (**-FT**) and linear probing (**-LP**) (i.e. freezing all weights except for the classification head), for both models.

Multimodal: CLIP (Radford et al., 2021) is a contrastively trained state-of-the-art multimodal model. We concatenated the text and image representations from CLIP for classification (**CLIP-MM**). We experimented with both full fine-tuning (**-FT**) and linear probing (**-LP**) with the RN50x4 model.

VinVL (Zhang et al., 2021b) is a multimodal Transformer that utilizes image features from a Faster R-CNN-based (Ren et al., 2015) object-detection model.[13]

BERT+CLIP is a multimodal Transformer initialized with BERT weights. For the image input, we used the penultimate pre-pooling feature maps from CLIP (RN50x4) and applied an adaptive max-pooling to reduce the feature size to 6×6. These features were flattened and concatenated with input text tokens as input to the Transformer, with the CNN component frozen during training.

CLIP-MM-OOD (distribution shifts analysis only). LP-FT (Kumar et al., 2022) is a recently developed training strategy for improved out-of-distribution generalization in image classification tasks. This method consists of a 2-step tuning process that performs linear probing followed by fine-tuning. We applied this training method to CLIP and refer to it as **CLIP-MM-OOD**.

We train all models with binary cross-entropy loss (each label was treated as a separate binary classification task) and re-weighted all positive instances to address class imbalances. The detailed hyperparameters and discussions of hyperparameter search methods are given in Appendix C.

[13] We used the top 36 regions as input for the model.

To evaluate the performance of the models, we report averaged (macro) F1 scores (averaged over 4 runs) in our experiments. The averaged F1 is computed by taking the classification task as 6 binary classification tasks.

6 Baseline Results and Discussions

Table 3 shows our benchmark results on FigMemes. Overall, text-only models performed poorly on this task, with the worst F1 scores in the Anthropomorphism/Zoomorphism category.

Vision is more important than expected. Notably, image-only models performed significantly better than text-only models in all categories, especially in Allusion, Exaggerations, Anthropomorphism/Zoomorphism and Contrast. For example, both ConvNeXt-FT and CLIP-CNN-FT vision-only models achieved significantly higher F1 scores in Exaggerations (7.96 points and 12.92 points higher) and Anthropomorphism/Zoomorphism (12.14 points and 26.62 points higher) than the best text-only model. We hypothesize that both of these categories may be easy to identify with a vision module, as in our data the Exaggerations category contains a sizable number of memes with visual reference to minority stereotypes. Similarly, the Anthropomorphism/Zoomorphism category often contains memes that use animal characters in an image. Consequently, better image representations would be critical to perform well on these memes.

In addition, a better language model (DeBERTa) does not seem to improve results significantly.

CLIP-based models achieve better results. In general, CLIP-based models achieved better results than other models for both unimodal (image only) and multimodal inputs, with CLIP-CNN-LP (unimodal) achieving the best F1 score. The CLIP model being contrastively trained over a large corpus from the internet may explain the superior results. It is also interesting to note that the BERT+CLIP model outperformed the multimodal pre-trained VinVL model (which uses BERT

Model	Allusion	Exag.	Irony	Anthrop.	Metaphor	Contrast	Avg. F1
BERT	32.83±0.44	28.31±1.39	43.44±2.39	16.72±0.59	33.05±0.65	41.34±2.20	32.62
DeBERTa	33.83±2.62	29.75±1.93	44.90±1.68	14.69±1.95	35.59±0.48	45.58±1.81	34.06
ConvNeXt-LP	40.71±0.23	32.34±0.38	37.98±0.56	27.43±0.79	34.26±0.45	41.72±0.97	35.74
ConvNeXt-FT	36.90±2.26	37.71±0.91	36.04±0.84	26.83±1.68	35.77±1.22	51.53±2.21	37.46
CLIP-CNN-LP	**52.32**±0.21	**44.00**±0.35	**49.77**±0.14	41.10±0.56	**44.87**±0.37	55.71±0.48	**47.96**
CLIP-CNN-FT	50.66±2.27	42.67±2.18	41.40±3.67	41.31±1.42	41.54±2.01	54.32±1.88	45.32
VinVL	45.16±1.21	40.57±1.36	40.81±3.67	33.09±2.32	37.97±2.00	47.37±3.56	40.83
BERT + CLIP	50.57±1.94	39.86±1.44	45.07±2.12	39.38±2.15	40.01±1.05	54.41±2.27	44.88
CLIP-MM-LP	49.19±0.51	43.53±0.38	46.38±1.12	36.53±1.22	40.21±0.47	54.02±0.20	44.98
CLIP-MM-FT	51.88±3.40	42.43±1.28	44.57±4.38	**41.76**±2.20	42.59±3.46	**56.91**±1.88	46.69

Table 3: Benchmark results (F1 scores) on the standard evaluation setting. The top value in each column is highlighted in **bold**. Results are averaged over 4 runs.

weights). This could further indicate that CLIP provides a better image representation than object-detector-based features for meme-like tasks.

6.1 Error Analysis

Even though a pure vision model achieved better overall F1 in our current experiment, a good multimodal model should perform well on multimodal data, as well as when a modality is missing or non-informative. To understand what is required to achieve good results on the proposed task, we performed error analysis on the best-performing multimodal model CLIP-MM-FT by randomly sampling 60 prediction errors (10 per class) and analyzing them manually.

Memes classification requires knowledge of external references or contexts. In our analysis, nearly 60% of the prediction errors are either due to the model's lack of understanding of external references (for example, the use of the yellow badge during the Holocaust) when making predictions or failure to jointly reason over text and image with context. Some typical wrongly-predicted examples are shown in Figure 2.

(a) A false negative in Metaphor.

(b) A false negative in Anthropomorphism/Zoomorphism.

Figure 2: Examples of prediction mistakes. **(a)** This is a Metaphor, the meme is comparing unvaccinated people to Jewish people during the Holocaust. **(b)** This is Anthropomorphism as it is a personified Arkansas state.

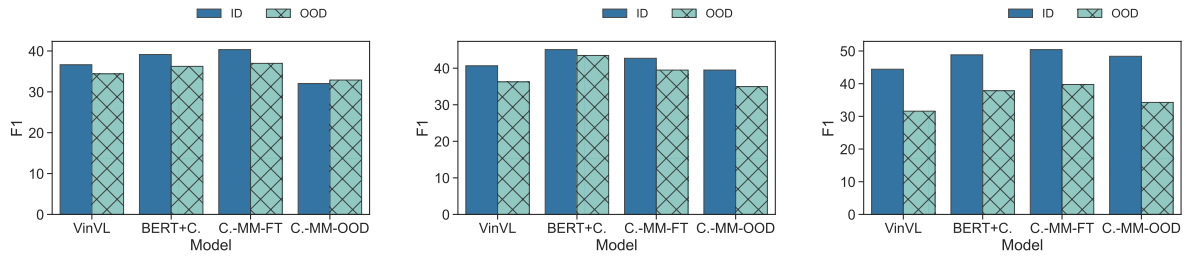
Exploitation of spurious correlation in vision. Most of the positive training examples in the *Contrast* category follow a left-to-right two-panel layout. From our error analysis, we found that 9 out of 10 prediction mistakes in *Contrast* are likely due to the model learning spurious correlations (i.e. by paying attention to the layout rather than the content). This could also explain why all multimodal models perform the best in *Contrast* compared to their results in other categories (as there is an easy spurious correlation). Other categories may contain less-exploitable visual patterns and require the machine learning model to make use of a diverse set of "references" to world knowledge or commonsense (e.g., the US President cannot be under the age of 18, hence not a child). This further suggests that models that can make judicious use of external knowledge will likely perform better on the FigMemes task.

7 Distribution Shift Analysis

As memes evolve fast based on the latest news and cultural references, in this section we aim to analyze models trained on different memes classification tasks to understand 1) how models perform under data distribution shifts on FigMemes (proposed); and 2) do the findings also translate to other memes datasets?

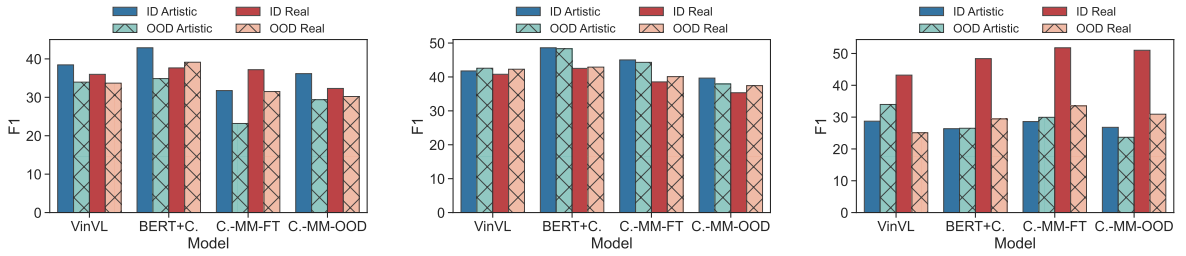
We use three previously mentioned datasets (§4.1) in this analysis with the following multimodal models: VinVL, BERT+CLIP, CLIP-MM-FT, and CLIP-MM-OOD.

In general, we found the models perform better when trained and evaluated on data from the same distribution (ID) than evaluated on data that are OOD, see Figure 3 and Figure 4. However, there are anomalies between datasets. We further observed that CLIP-MM-OOD does not close the gaps between ID and OOD evaluation results as it did in image-only tasks (Kumar et al., 2022), which



(a) FigMemes (b) Memotion 2 (c) MAMI

Figure 3: Text-based in-distribution versus out-of-distribution evaluation results. C. is CLIP.



(a) FigMemes (b) Memotion 2 (c) MAMI

Figure 4: Image-based in-distribution versus out-of-distribution evaluation results. ID Artistic means train and test on Artistic visual style data. OOD Artistic means train on Artistic and test on all *except* Artistic style data. Similar for ID/OOD Real. C. is CLIP.

could be due to multiple factors (choice of backbone models, hyperparameters, etc.) and we plan to investigate this in future work.

Models trained with FigMemes are less sensitive to text-based distribution shifts. We do not observe any significant drop in averaged F1 scores for FigMemes between ID and OOD evaluation (Figure 3a) across all evaluated multimodal models. In fact, poor OOD test results mainly came from the OOD test set without any text (see Appendix E.2 for results per OOD test set). This is important since an ideal multimodal model should not *only* be able to utilize features from each input modality, *but also* maintain performance on a task when an input mode is missing (and the nature of the task is unchanged).

Figure 4a shows that models trained on FigMemes are more sensitive to visual category shifts, which aligns with our previous findings. We also found that training with memes that are Artistic in style creates a harder transfer scenario.

ID/OOD performance gaps are dataset/task-dependent. Contrary to findings on FigMemes, models evaluated on the Memotion 2 dataset do not show much difference between ID/OOD results in the visual category shifts. We suspect that this is due to the extreme imbalance of the dataset.[14]

On the other hand, MAMI (misogyny classification) is quite sensitive to distribution shifts in both text and images (with the largest drop of 14.11 and 20.09 in F1, respectively).[15] The most prominent drop is when training on memes in the Real style, but testing on memes from other styles (especially when tested on memes in the Artistic style, see Table 12). Similar trends were also observed in other splits (see Appendix E.3 for details).

Based on these results, we recommend that future work on multimodal datasets perform analyses on distribution shifts and identify potential failure modes, as well as provide splits for evaluating these failures. Datasets should help facilitate the development of machine learning models that are robust to distribution shifts, which also have implications for their real-world applicability.

[14]Memotion 2 (subtask B) is a task on classifying humour, sarcasm, offensive and motivational memes. In the dataset, nearly 86.5% of the data are labelled as humour, and only 4.2% of the data are labelled as motivational.

[15]Both MAMI and Memotion 2 are more sensitive to the text shifts in OOD test set 1, see Table 11 for details. We recommend future works use the test set 1 to study OOD generalization for these two tasks.

8 Conclusions

In this paper, we introduce a novel multi-label memes dataset – *FigMemes*, for classifying figurative language in politically-opinionated memes. The dataset contains annotated memes covering a wide range of figurative language and topics over a period of 5 years. We provide comprehensive benchmarks using state-of-the-art unimodal and multimodal machine learning models and show that the dataset poses unique challenges to existing models. Our error analysis and distribution shifts analysis reveal the limitations of state-of-the-art models and the need to develop robust models that understand world knowledge and commonsense. Our dataset presents unique challenges and opportunities for future research in robust multimodal models that can benefit computer science research and humanities research in political sciences and sociology.

9 Limitations

One limitation of our work is the single data source, 4chan **/pol/** board. In the future, we would like to extend the work to other data sources. The set of figurative language used in the proposed dataset is also limited to six, which we aim to expand to more categories in the future.

Our distribution shifts analysis is based on distribution shifts in each modality, independently. In the future, we would like to extend the work to include joint distribution shifts, in both image and text. We would also like to extend the work and create an adversarial test set and facilitate the development of models that are robust to pseudo-correlation in the future.

10 Ethics Considerations

We aim to study figurative language in politically-opinionated memes, where they can be potentially misused, such as to develop models to automatically generate persuasive or even harmful politically-opinionated memes. The dataset may also serve as unintentional advertisement of certain online communities. However, since all of the memes in our dataset are freely available (and archived) on the internet, we do not provide any additional advantage to such actors. We argue that by focusing the research community on models that can better understand figurative language in these memes, we can help to develop countermeasures against such actors, bring awareness, and potentially mitigate spreading of the harmful content on the internet. To further mitigate risk, we do not allow our dataset to be indexed by search engines, and we will require all users to provide their academic affiliation as a condition to access the data.

The research study was examined with the help of the self-assessment checklist of the Ethics Committee of Technical University of Darmstadt and found to be ethically unobjectionable, which is why, in accordance with local regulations, a formal vote of the Ethics Committee was dispensed with.

Acknowledgements

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 13N15897 (MISRIK), and the LOEWE initiative (Hesse, Germany) within the emergenCITY center.

We thank Jan-Christoph Klie and Luke Bates for their valuable feedback and suggestions on a draft of this paper. We thank the anonymous reviewers for their detailed and insightful comments.

References

Mohit Chandra, Dheeraj Reddy Pailla, Himanshu Bhatia, AadilMehdi J. Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "Subverting the Jewtocracy": Online antisemitism detection using multimodal deep learning. In *WebSci '21: 13th ACM Web Science Conference 2021, Online, United Kingdom, June 21-25, 2021*, pages 148–157. ACM.

Blyth Crawford, Florence Keen, and Guillermo Suarez-Tangil. 2021. Memes, radicalisation, and the promotion of violence on chan sites. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, Online, June 7-10, 2021*, pages 982–991. AAAI Press.

Patrick Davison. 2012. *9. The Language of Internet Memes*, pages 120–134. New York University Press.

Ricahrd Dawkins. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

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.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 Task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Bangkok, Thailand, August 5-6, 2021*, pages 70–98. Association for Computational Linguistics.

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 on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Eemeli Hakoköngäs, Otto Halmesvaara, and Inari Sakki. 2020. Persuasion through bitter humor: Multimodal discourse analysis of rhetoric in internet memes of two far-right groups in finland. *Social Media + Society*, 6(2).

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Online, Austria, May 3-7, 2021*.

Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 92–101. AAAI Press.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Online, December 6-12, 2020*.

Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M Asano. 2021. Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 26–35, Online. Association for Computational Linguistics.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986. IEEE.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1406–1415. IEEE.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2783–2796. Association for Computational Linguistics.

Shraman Pramanick, Aniket Roy, and Vishal M. Patel. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 546–556. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, Online, 18-24 July 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, Suryavardan S, Aishwarya N. Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit P. Sheth, Asif Ekbal, and Chaitanya Ahuja. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *DE-FACTIFY@AAAI*, volume 3199 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (Online), December 12-13, 2020*, pages 759–773. International Committee for Computational Linguistics.

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*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, Suzanne Little, and Paul Buitelaar. 2022. TrollsWithOpinion: A dataset for predicting domain-specific opinion manipulation in troll memes. *Multimedia Tools and Applications*.

Caroline N. Tipler and Janet B. Ruscher. 2019. Dehumanizing representations of women: the shaping of hostile sexist attitudes through animalistic metaphors. *Journal of Gender Studies*, 28(1):109–118.

Jiquan Wang, Lin Sun, Yi Liu, Meizhi Shao, and Zengwei Zheng. 2022. Multimodal sarcasm target identification in tweets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8164–8175, Dublin, Ireland. Association for Computational Linguistics.

Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 Internet Measurement Conference*, IMC '17, page 405–417, New York, NY, USA. Association for Computing Machinery.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021a. MultiMET: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214–3225, Online. Association for Computational Linguistics.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. VinVL: Revisiting visual representations in vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Online, June 19-25, 2021*, pages 5579–5588. IEEE.

A Examples of Different Visual Categories

As described in our main paper, we defined four different visual-style categories: Artistic/Real/Mixed/Info-graph. Figure 5 shows examples from different visual categories.

B Additional Dataset Statistics

Since our data covers topics spanning 5 years, to understand which words are important per year of our dataset, we compute 10 top-weighted tokens per year (tokens attributed to the most occurrences in the entire corpus, in a year). We construct the vocabulary by selecting tokens that occur more than 5 times in different memes, and only count unique tokens per meme, with the removal of stopwords[16], URLs, and numerical values. We construct per year token count f_{year} and total token count f_{total}. The weights of the tokens per year are calculated as $f_{\text{year}}/f_{\text{total}}$. These tokens highlight the significant events that occurred during that time period, ranging from Syrian refugees and Epstein to COVID-19.

Figure 6 shows the normalized histogram of the number of tokens in our dataset. We truncate the token count to 100 tokens since the token distribution is long-tailed.

C Hyperparameters

We report the hyperparameters used in our benchmark experiments in Table 5. We used the AdamW (Loshchilov and Hutter, 2019) optimizer, and all learning rates used are listed in Table 6. The maximum sequence length for BERT and DeBERTa is 96. We use a maximum sequence length of 96 for text, and 36 for image features for both VinVL and BERT+CLIP. The maximum sequence length for multimodal CLIP experiments is 77 (original).

The learning rate and weight decay parameters are determined by using a grid search when training on the standard evaluation set. The weight decay grid values were $\{0.05, 10^{-8}\}$. The learning rate grid values were $\{10^{-3}, 5\times10^{-3}, 10^{-4}, 5\times10^{-4}, 10^{-5}\}$ for linear probing and $\{2\times10^{-5}, 3\times10^{-5}, 5\times10^{-5}\}$ for all others.

Note that for the CLIP-MM-OOD model, we first train CLIP-MM-LP for 20 epochs, then reload the weights for full fine-tuning for another 20 epochs.

We evaluated each model after every epoch and loaded the checkpoint with the best average F1 score at the end of training for testing.

All experiments were performed on NVIDIA P100 or A100 GPUs.

[16]NLTK:https://www.nltk.org/book/ch02.html

D Data Statistics for Analysis

D.1 Splits

We report the statistics of the splits used in our distribution shifts analysis in Table 7 and Table 8. Note that Infographic is excluded from visual-based evaluation for Memotion 2 and MAMI due to its data size. The training data size of 900 memes was picked to keep the same number of training examples across all distribution-shift scenarios and ensure there are enough memes with labels in the test set.

Since Memotion 2 and MAMI are larger than the proposed FigMemes dataset, a majority of the memes from these two datasets are used for evaluation. We further provide additional results on Memotion 2 and MAMI on different splits with much larger training sets in Appendix E § E.3. The trends we observed in our analysis persist when using a larger training set.

D.2 Token Statistics

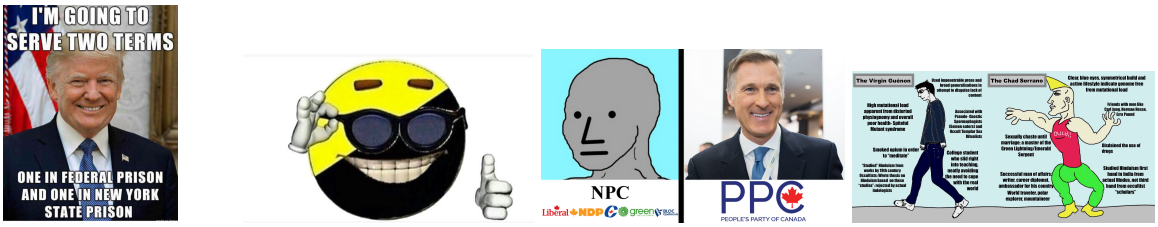
We also included top tokens (by TFIDF, min. count=5, max document frequency=0.8) for all of the test sets used in our distribution shifts analysis in Table 9.

E Additional Results

E.1 Standard Evaluation

We include benchmark results on the Exact Match (EM) score and Hamming Loss (HL) in this section. EM is the total number of predictions that match the targets in all 6 categories, divided by the total number of predictions. HL is defined as follows: given a multi-label classification task with N classes and a dataset of size M, let the true target of a data point be y_{nm} and the predicted target of a data point be $\hat{y}_{nm}$. The HL is the average Hamming distance between y and $\hat{y}$ over the dataset given by $\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\text{XOR}(y_{nm},\hat{y}_{nm})$. A lower HL is better.

Table 10 shows the Hamming loss of benchmark models under the standard evaluation setting. CLIP-

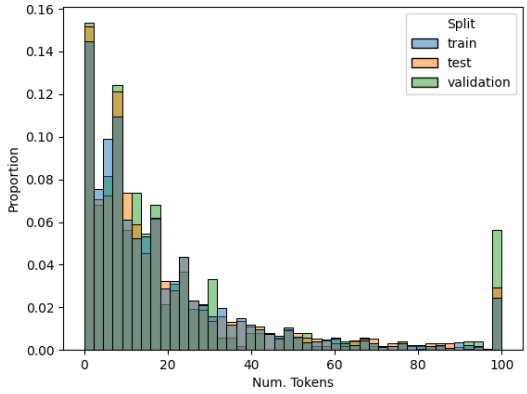


(a) *Real* (b) *Artistic* (c) *Mixed* (d) *Infographic*

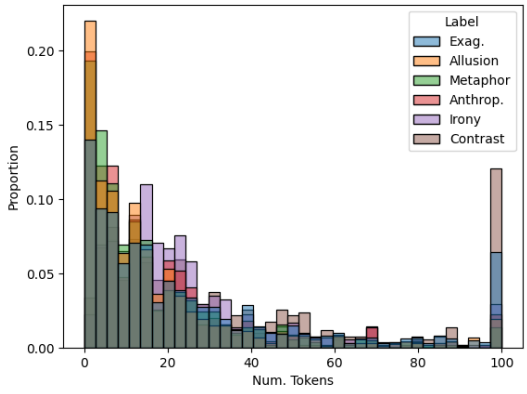
Figure 5: Examples of memes from different visual categories.

Year	Tokens
2017	bitches, saving, syrian, catholic, excellent, occupy, dd, grand, saved, refugees
2018	gives, fashion, purple, dollars, levels, immigrants, isis, ga, po, cucks
2019	cake, record, epstein, immigration, yang, murder, despite, island, month, nordic
2020	unlimited, virus, corona, color, leaders, coronavirus, hide, tall, banks, voting
2021	vaccinated, ancestry, spanish, laws, consequences, chud, mossad, terrorism, fb, medical

Table 4: Top weighted tokens per year.



(a) Histogram (normalized) of the number of tokens in different splits.



(b) Histogram (normalized) of the number of tokens in different labels.

Figure 6: Distribution of number of tokens in different data split (a) or label (b).

Name	Value
Optimizer	AdamW
Warmup steps	0
Learning rate schedule	linear
Weight decay	0.05
Batch size	64
Epochs	$20^{\alpha\beta}$, 40^{γ}

Table 5: **Training Setup**: $^{\alpha}$: Text-based Distribution Shifts. $^{\beta}$: Image-based Distribution Shifts. $^{\gamma}$: CLIP-MM-OOD with 20 epochs trained for linear probing and 20 epochs trained for fine-tuning.

Name	Learning rate
BERT	0.00003
DeBERTa	0.00003
CLIP-CNN-LP	0.005
CLIP-CNN-FT	0.00002
ConvNeXt-LP	0.005
ConvNeXt-FT	0.00005
VinVL	0.00003
BERT+CLIP	0.00003
CLIP-MM-LP	0.005
CLIP-MM-FT	0.00002

Table 6: Learning rates used in our experiments. Note that CLIP-MM-OOD uses the same learning rate as CLIP-MM-LP during the linear probing training and uses the same learning rate as CLIP-MM-FT during fine-tuning.

Dataset	Train	Val.	ID	OOD 1	OOD 2
FigMemes	900	150	2102	1551	438
Memotion 2	900	150	5997	1043	410
MAMI	900	150	5969	3284	697

Table 7: Statistics on text-based distribution-shift splits.

Dataset	Style	Train	Val.	Test
FigMemes	Artistic	900	150	1152
	Real	900	150	1090
	Infographic	-	-	152
	Mixed	-	-	647
Memotion 2	Artistic	900	150	696
	Real	900	150	5430
	Infographic†	-	-	1
	Mixed	-	-	273
MAMI	Artistic	900	150	913
	Real	900	150	7605
	Infographic†	-	-	4
	Mixed	-	-	378

Table 8: Statistics on visual-based distribution-shift splits. †: Infographic is excluded from evaluation for Memotion 2 and MAMI due to its data size.

MM-FT is the best-performing model based on the Hamming loss.

E.2 Distribution Shifts

The complete results from our topic distribution-shift analysis are shown in Table 11. The complete benchmark results for visual distribution-shift are shown in Table 12.

E.3 Distribution-Shift Comparisons Between 2 Splits.

We observed no clear changes in the Memotion 2 dataset or generalization issues with the MAMI dataset when experimenting with distribution splits described in our paper and § D.1. To verify that these patterns are not due to model under-fitting (as the training size for the previous splits was 950 for all categories), we used alternative splits with up to 5000 memes for training in both datasets. The detailed statistics are in Table 13 and Table 14.

The distribution-shift evaluation results for Memotion 2 and MAMI are shown in Figure 7 and Figure 8, respectively. Regardless of the training set size, we still observed large generalization gaps for the MAMI dataset in both modalities. A very small gap was observed under text-distribution shifts for the Memotion 2 dataset.

F Dataset Statement

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? The dataset contains internet memes collected from 4chan. The dataset contains photos, drawings and other user-generated content.

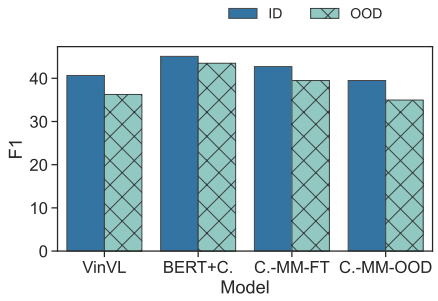
What data does each instance consist of? Each instance will consist of an image (i.e. the meme), text extracted by OCR, and labels.

Are there recommended data splits (e.g., training, development/validation, testing)? Yes, we describe the splits in the main section of the paper. We will release the dataset splits.

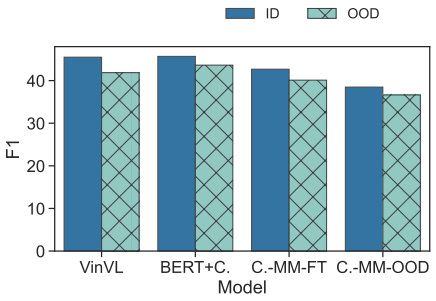
Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? We aim to study figurative language in politically-opinionated memes. As our dataset source is the 4chan **/pol/** board, it contains content that uses racial slurs, stereotypes, Nazi references, etc. Our dataset is intended to be used for building robust multimodal models that can classify such memes, aid humanities research, especially in political science and sociology, and build automatic systems that can raise awareness at scale. This dataset should only be used for academic research or non-commercial purposes.

FigMemes	
Split/Style	**Tokens**
ID	autism, butt, ce, coronavirus, disease, horse, mask, painting, roll, sit
OOD 1	amerimutt, ancestor, ancestors, ancestry, animals, australia, authoritarian, bank, banks, bible
OOD 2†	-
Artistic	nope, gotta, reddit, important, pls, question, rare, bit, favorite, hide
Real	memri, bert, supporters, ay, daughter, feminism, realise, taste, tits, trans
Mixed	mar, syrian, blacked, greek, tha, wonder, sea, ancient, empire, gee
Infographic	bell, curve, gf, libertarian, authoritarian, chad, actual, anon, fun, virgin
Memotion 2	
Split/Style	**Tokens**
ID	adolf, ads, alex, alive, alright, app, april, army, art, basically
OOD 1	workouts, bench, exercising, itches, reps, treadmill, cardio, intense, lifting, squat
OOD 2	essays, project, assignment, essay, exam, assignments, studied, teacher, semester, homework
Artistic	pathetic, feet, intellectual, scale, lower, sneeze, spongebob, fix, throw, art
Real	ac, acting, adam, adapt, agents, ah, alabama, alarm, aliens, alive
Mixed	kalm, panik, aunt, bikini, council, dentist, greatest, military, pr, slaps
Infographic†	-
MAMI	
Split/Style	**Tokens**
ID	absolutely, actor, africa, african, airbags, americans, amy, arab, aunt, balcony
OOD 1	interview, isekai, lover, manga, mmonopoly, smarter, waitress, breakup, condoms, farts
OOD 2	responsibilities, wwiii, oppressed, feminists, equal, equality, wave, patriarchy, rights, adultswim
Artistic	derp, dining, ffffff, fffuu, opinions, pokemon, considered, fall, hentai, japanese
Real	absolute, ac, action, actor, admit, ads, adultswim, advantage, advertising, advice
Mixed	kalm, panik, shef, ii, miles, medium, nd, programmers, rd, film
Infographic†	-

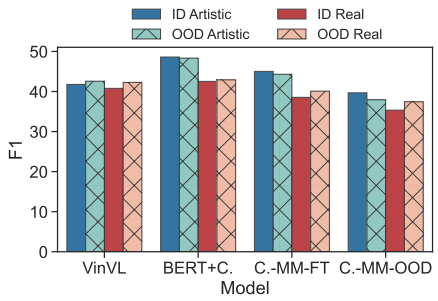
Table 9: Top tokens by TFIDF for each test split for all datasets used in our analysis. †: For FigMemes, the second OOD test set consists of memes without text, as it is a natural text-based topic shift. Since both Memotion 2 and MAMI contain less than 5 memes in the style of Infographic, we did not compute the top tokens.



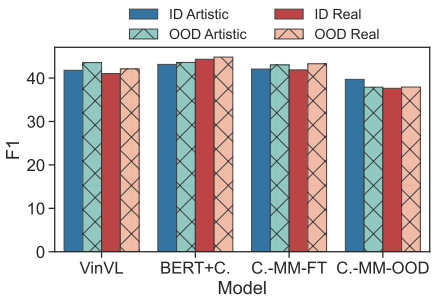
(a) Text distribution shifts



(b) Text distribution shifts (large train)



(c) Visual distribution shifts



(d) Visual distribution shifts (large train)

Figure 7: Comparison of evaluation results on distribution shifts using different splits. Memotion 2 dataset. In (c) and (d) ID Artistic means train and test on Artistic visual style data. OOD Artistic means train on Artistic and test on all *except* Artistic style data. Similar for ID/OOD Real.

Model	**EM ↑**	**HL ↓**
BERT	13.73	29.42
DeBERTa	12.55	29.89
ConvNeXt-LP	8.25	31.67
ConvNeXt-FT	17.06	25.21
CLIP-CNN-LP	21.35	23.31
CLIP-CNN-FT	27.71	19.55
VinVL	21.51	22.82
BERT + CLIP	22.37	23.04
CLIP-MM-LP	22.50	22.60
CLIP-MM-FT	**30.58**	**18.21**

Table 10: Exact Match score and Hamming loss (averaged over 4 runs) of benchmark models in the standard evaluation setting.

	FigMemes (F1)		
Tested on	**ID**	**OOD 1**	**OOD 2 (w/o Text)**
VinVL	36.63	40.86	27.94
BERT + CLIP	39.13	40.28	32.19
CLIP-MM-LP	37.86	40.82	33.18
CLIP-MM-FT	39.74	43.00	29.15
CLIP-MM-OOD	32.03	35.39	30.38
	Memotion 2 (F1)		
Tested on	**ID**	**OOD 1**	**OOD 2**
VinVL	40.66	34.75	37.77
BERT + CLIP	45.08	42.00	44.97
CLIP-MM-LP	44.44	41.73	42.77
CLIP-MM-FT	43.36	38.7	40.59
CLIP-MM-OOD	39.46	35.14	35.34
	MAMI (F1)		
Tested on	**ID**	**OOD 1**	**OOD 2**
VinVL	44.42	26.66	36.52
BERT + CLIP	48.82	35.00	40.68
CLIP-MM-LP	52.24	42.25	45.58
CLIP-MM-FT	50.42	36.68	42.81
CLIP-MM-OOD	48.38	30.22	38.32

Table 11: Benchmark results (averaged over 4 runs) when trained and evaluated on different topic clusters.

	FigMemes (F1)							
Trained on	Artistic				Real			
Tested on	***Artistic***	**Real**	**Infographic**	**Mixed**	**Artistic**	***Real***	**Infographic**	**Mixed**
VinVL	38.44	29.09	34.90	37.87	30.86	35.98	35.14	35.07
BERT + CLIP	42.89	31.06	34.44	39.11	30.29	37.65	25.66	34.68
CLIP-MM-LP	39.33	31.30	36.97	41.93	30.26	35.25	30.27	34.22
CLIP-MM-FT	39.96	28.70	33.22	37.98	25.95	36.02	32.75	30.38
CLIP-MM-OOD	37.00	26.02	28.76	32.58	19.31	31.91	21.09	18.87
	Memotion 2 (F1)							
Trained on	Artistic				Real			
Tested on	***Artistic***	**Real**	**Infographic**	**Mixed**	**Artistic**	***Real***	**Infographic**	**Mixed**
VinVL	41.77	41.69	-	43.46	42.19	40.79	-	42.37
BERT + CLIP	48.46	47.32	-	50.19	43.95	43.88	-	45.26
CLIP-MM-LP	43.53	43.57	-	43.68	43.98	43.17	-	45.20
CLIP-MM-FT	45.02	43.89	-	44.72	39.49	38.54	-	40.69
CLIP-MM-OOD	39.54	38.73	-	37.47	37.50	35.45	-	37.76
	MAMI (F1)							
Trained on	Artistic				Real			
Tested on	***Artistic***	**Real**	**Infographic**	**Mixed**	**Artistic**	***Real***	**Infographic**	**Mixed**
VinVL	26.36	29.82	-	23.18	21.74	43.20	-	28.43
BERT + CLIP	28.72	37.56	-	30.40	25.99	48.38	-	32.88
CLIP-MM-LP	30.64	37.36	-	29.48	38.38	53.70	-	40.18
CLIP-MM-FT	29.39	32.48	-	28.83	26.48	50.57	-	36.75
CLIP-MM-OOD	27.48	26.23	-	20.35	30.07	51.46	-	33.44

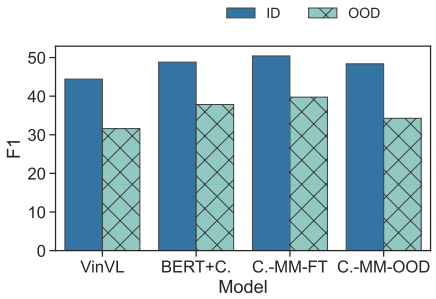
Table 12: Benchmark results (averaged over 4 runs) when trained and evaluated on different visual categories.

Dataset	**Train**	**Val.**	**ID**	**OOD 1**	**OOD 2**
Memotion 2	5000	150	1897	1043	410
MAMI	5000	150	1869	3284	697

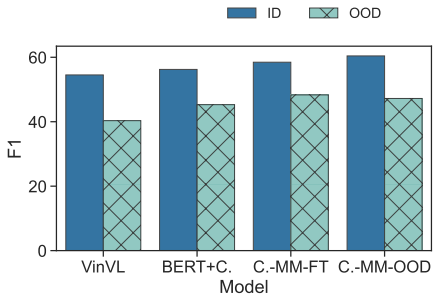
Table 13: Statistics on text-based distribution-shifts splits (large train size).

Dataset	**Style**	**Train**	**Val.**	**Test**
Memotion 2	Artistic	900	150	696
	Real	5000	150	1330
	Infographic†	-	-	1
	Mixed	-	-	273
MAMI	Artistic	900	150	913
	Real	5000	150	3505
	Infographic†	-	-	4
	Mixed	-	-	378

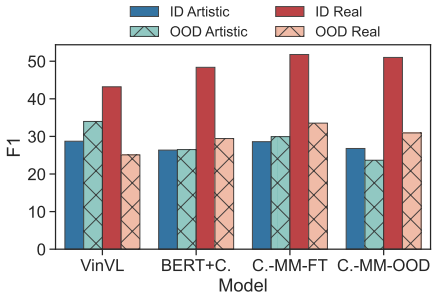
Table 14: Statistics on visual-based distribution-shifts splits (large train size). †: Infographic is excluded from evaluation for Memotion 2 and MAMI due to its data size.



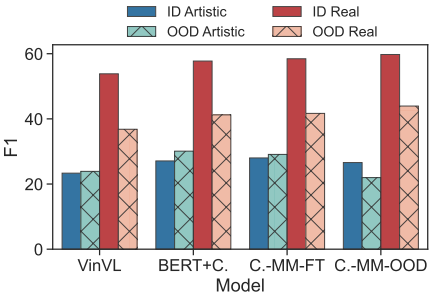
(a) Text distribution shifts



(b) Text distribution shifts (large train)



(c) Visual distribution shifts



(d) Visual distribution shifts (large train)

Figure 8: Comparison of evaluation results on distribution shifts using different splits. MAMI dataset. In (c) and (d) ID Artistic means train and test on Artistic visual style data. OOD Artistic means train on Artistic and test on all *except* Artistic style data. Similar for ID/OOD Real.