2023 IEEE International Conference on Big Data (BigData)

DOI: 10.1109/BigData59044.2023.10386279

MERMAID: A Dataset and Framework for Multimodal Meme Semantic Understanding

Shaun Toh*, Adriel Kuek‡, Wen-Haw Chong†, Roy Ka-Wei Lee*
Singapore University of Design and Technology*, Singapore
Singapore Management University†, Singapore
DSO National Laboratories‡, Singapore
Email: shaun_toh@mymail.sutd.edu.sg, kyongjie@dso.org.sg, whchong.2013@phdis.smu.edu.sg, roy_lee@sutd.edu.sg

***Abstract*—Memes are widely used to convey cultural and societal issues and have a significant impact on public opinion. However, little work has been done on understanding and explaining the semantics expressed in multimodal memes. To fill this research gap, we introduce MERMAID, a dataset consisting of 3,633 memes annotated with their entities and relations, and propose a novel MERF pipeline that extracts entities and their relationships in memes. Our framework combines state-of-the-art techniques from natural language processing and computer vision to extract text and image features and infer relationships between entities in memes. We evaluate the proposed framework on a real-world meme dataset and establish the benchmark for the new multimodal meme semantic understanding task. Our evaluation also includes a low-resource setting, where we assess the applicability of our framework to low-resource settings, which is a common problem due to the high cost and lack of labeled data for relations in memes. Overall, our work contributes to the understanding of the semantics of memes, a crucial form of communication in today's society.**

***Index Terms*—Memes, Multimodal, Semantic Extraction**

I. Introduction

Memes have become an integral part of modern communication, with their viral nature allowing them to spread rapidly across social media platforms. However, despite extensive research on meme analysis, little work has been done on truly understanding and explaining the semantics expressed in multimodal memes. Understanding the semantics of memes is crucial for several reasons. Firstly, memes are a form of cultural expression that reflects the values and beliefs of a particular group or community. By analyzing and understanding the semantics of memes, we can gain insights into the cultural and societal issues that are important to people. Secondly, memes are used to spread information and opinions and can often significantly impact public opinion. Understanding the semantics of memes can help us understand the messages being conveyed and how they may influence people's opinions and beliefs.

Despite the widespread use and popularity of memes, research in this area has primarily focused on analyzing their spread and influence [1]–[3] rather than their meaning. There are also existing studies that have proposed machine learning models to analyze memes for downstream tasks such as detecting hateful content [4]–[8], performing sentiment analysis [9], and victim detection and role analysis [10]–[12]. However, no known study focuses on understanding and explaining the semantic content expressed in memes.

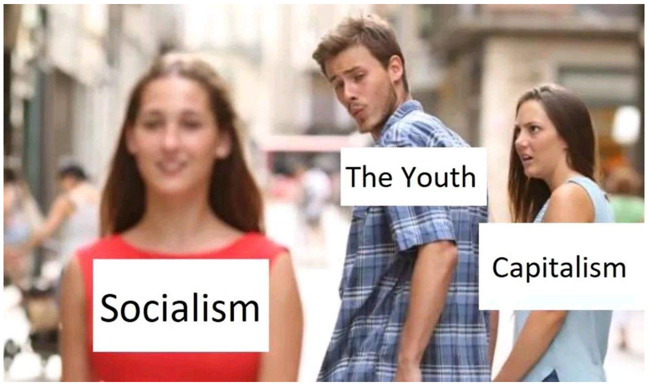


Fig. 1. An example of the viral "*Distracted Boyfriend*" meme used to express opinion about youth preferring socialism over capitalism.

Extracting semantics from memes is a challenging task. One of the key challenges is the need to account for the interaction between text and visual modalities. Memes often rely on the interplay between text and images to convey their intended meaning, and understanding this interplay is crucial to interpreting the semantics of memes. Another significant challenge in extracting the semantics of memes is the need to extract entities before inferring the relationships between them. Entities refer to objects, people, or concepts mentioned in memes. For instance, Figure 1 shows an example of the viral *Distracted Boyfriend*" meme. To understand this meme, we will first have to extract the three entities in this meme: *The Youth*", *Socialism*", and *Capitalism*". Next, we will infer the relationships among the entities. For example, we infer that *The Youth*" *favors Socialism*", and *Socialism*" is portrayed to be *superior* to *Capitalism*". Ideally, we hope to train a multimodal meme entity relations extraction model to perform entity extraction and relation inference in memes automatically. However, there is currently no known dataset that can facilitate the training of such a multimodal machine learning model.

To address this research gap, we constructed the **M**eme **E**ntity **R**elations for **M**ultimodal **A**utomated **I**nference **D**ataset

(MERMAID)[1], which consists of 3,633 memes from 39 viral meme templates, annotated with their entities and relations. MERMAID is a valuable resource for researchers and practitioners interested in studying the semantics of memes because it provides a comprehensive and diverse sample of memes that can be used for training and testing multimodal meme entity extraction and relation inference models.

We also propose a novel **M**eme **E**ntity **R**elation **F**ramework (MERF)[2] pipeline that extracts entities and their relationships in memes, making it the first of its kind in this domain. Our framework incorporates state-of-the-art techniques from natural language processing (NLP) and computer vision (CV) to extract the text and image features of memes and infer the relationships between the entities depicted in them. We designed various versions of MERF using existing multimodal and unimodal machine learning techniques and evaluated the models through extensive experiments. Through our experiments, we establish the benchmark for the new meme entity relation extraction task.

Contributions. We summarize our contributions as follows:

- We propose a new MERMAID dataset, which serves to improve semantic understanding of memes. The MERMAID dataset can also support training and testing of meme multimodal entity extraction and relation inference models.
- We propose a novel MERF pipeline that extracts entities and their relationships in memes. To the best of our knowledge, this is the first framework that proposes a framework to extract entities and relations in memes.
- We evaluate our proposed framework through extensive experiments on a real-world meme dataset. Our automatic and low-resource setting evaluations establish the benchmark for the new multimodal meme semantic understanding task.

II. Related Work

A. Meme Analysis

In the digital age, memes have become a ubiquitous mode of communication, offering a unique way for people to express themselves. Across various disciplines, researchers have delved into the social and cultural impact of memes in online discourse, unlocking new insights into this fascinating phenomenon. From communication experts analyzing their effects on social dynamics [13]–[16], to computer scientists leveraging meme datasets for machine learning tasks, such as detecting hateful content [8], [17]–[23], the study of memes has become a hotbed of interdisciplinary research.

Some researchers have gone beyond simply analyzing existing memes, instead creating datasets focused on specific events, topics [24], or emotions [25]. These include the 2016 United States presidential election [26], memes related to COVID-19 [11], and even memes expressing fine-grained emotions [27], [28]. In one particularly innovative approach, researchers proposed a deep learning technique to classify the sentiment of memes, leveraging the spatial-domain correspondence between visual and textual elements [29].

Finally, some researchers have delved into the darker side of memes, investigating those that spread misinformation or attack marginalized groups. Qu et al. [30] collected memes from Reddit on topics such as COVID-19, labeling them with misinformation labels. By shining a light on these harmful memes, researchers hope to create a safer and more inclusive online environment.

While there are more studies on memes, most of them have centered around performing specific supervised classification tasks such as toxicity detection or sentiment classification, without performing in-depth analysis of the semantics communicated in the memes. In this study, we address the research gap by proposing a new meme analysis task, which aims to extract the entities and relationships among entities in memes to provide a more holistic understanding of the semantics communicated in memes. We construct the first meme dataset that annotates the memes' entities and relations. We further propose a novel meme semantic extraction framework that extracts entities and relations in memes. The constructed dataset and proposed semantic extraction framework will serve as valuable resources for researchers and practitioners interested in studying the semantics of memes.

B. Entity and Relation Extraction

In recent years, Named Entity Recognition (NER) has experienced significant progress, starting with the utilization of Conditional Random Fields and feature engineering [31]–[33]. Subsequent advances have been facilitated by the widespread use of neural networks, which range from LSTMs and CNNs [34], [35] to the present-day state-of-the-art large language models, such as BERT [36] and RoBERTa [37]. Despite the recent emphasis on enhancing NER performance for non-English languages [38]–[40], a significant research gap exists in the domain of generalized entity extraction based on linguistic and contextual cues, particularly in multimodal settings, such as meme analysis, where named entities frequently lack strict categorization or formal definitions.

Relation Extraction (RE) involves identifying and extracting the semantic relationship between two entities based on the context in which they appear. This task is typically accomplished using structured or unstructured text data, along with additional context, within various frameworks. Relationships are frequently categorized into predefined categories, enabling their broad applicability across multiple tasks. Early approaches to RE used kernels [41] and pattern-based techniques [42], [43], which continue to be developed in certain fields as recently as 2021 [44]. However, various neural network approaches have become widespread, with convolutional neural networks [45], [46] being replaced by recurrent neural network-based approaches [47], which have, in turn, been supplanted by graph-based neural networks [48]. The advent of transformers [49] has led to the use of pretrained large language models (LLMs) [36], [37], [50] as the backbone

[1] Link to the dataset will be released in cam-ready version

[2] Code is available here

of various RE methods due to their superior performance compared to earlier approaches.

The difficulty of obtaining well-labeled data for training large models has increased the interest in few-shot RE [51], [52]. Prototypical networks [53] and meta-learning [51], [54] are among the techniques employed, but developing robust models capable of generalizing to new tasks remains an open research question [55].

This paper proposes an extension of existing text-based NER and RE methods for extracting entities and relations from multimodal memes. To address the research gap, a framework is proposed for extracting unlabelled entities from memes using linguistic and contextual cues derived from meme templates and text box positioning. To extract relations, we annotated the relations between entities communicated in viral meme templates and trained a RE model to predict the relations among entities identified in memes.

III. Dataset

A. Data Construction

This section outlines the pipeline used for constructing the MERMAID dataset, which involves three main phases: *dataset collection*, *dataset processing*, and *dataset annotation*.

Data Collection. To begin, we retrieve a list of the top 39 most popular meme templates from ImgFlip [3]. Next, we retrieve 150 memes from the ImgFlip website for each identified popular meme template. This process results in a total of 5,850 memes.

Data Processing. To ensure that the collected memes align with our research objectives, we apply strict filtering criteria during the data processing phase. Firstly, we examine each meme to verify that it contains readable text. Secondly, we remove all duplicate memes and filter out any non-English memes. Next, we use the optical character recognition algorithm, EasyOCR[4], to extract text and bounding boxes from the remaining memes and discard any that do not contain text. Finally, we retain 3,633 memes that meet our filtering criteria, forming the final MERMAID dataset.

Data Annotation. We engaged three annotators who have knowledge of meme culture to label the entities and relations in the memes. The annotators consisted of two female undergraduates and one male postgraduate, all aged between 20 to 25.

Entity Annotation. The annotators were presented with the memes and the text extracted from them using EasyOCR, and were asked to identify the entity span within the extracted text. In addition to the entities explicitly mentioned in the meme text, we added a special entity, "MEME_CREATOR," to represent the meme's creator if the meme was found to express a direct opinion of the creator.

Relation Annotation. After identifying the entities in the memes, the annotators labeled the relationships between them. Unlike existing entity relation studies that focus on identifying predicates or property assignments for entity relations [56], [57], our entity relations aimed to understand the sentiments or opinions expressed in the memes. Specifically, we identified seven types of relation:

- Superior: Describes the relation between two entities when one is deemed to be superior to another. For instance, *Entity_A* is *superior* to *Entity_B*. Note that it is also implied that *Entity_B* is inferior to *Entity_A* but we do not explicitly label the "*inferior*" relation.
- Equal: Describes the relation between two entities when one is deemed to be equal to another. For instance, *Entity_A* is *equal* to be *Entity_B*.
- Upgrade: Describes the relation between two entities when one is a positive continuation version of the other. For instance, *Entity_A upgrades* to become *Entity_B*. Noted that it is also implied that *Entity_B* is a downgraded version of *Entity_A* but we do not explicitly label the "*downgrade*" relation.
- Affirm/Favor: Describes the relation between two entities when one is affirming a belief or favors the other. For instance, *Entity_A favors Entity_B*.
- Doubt/Disfavor: Describes the relation between two entities when one is doubts or disfavors the other. For instance, *Entity_A disfavors Entity_B*.
- Indifferent: Describes the relation when an entity expresses indifference towards the other entity while still acknowledging its existence. For instance, *Entity_A* is indifferent towards *Entity_B*.
- NULL: Describes no relation between two entities.

Annotation Quality Control. To ensure the reliability of the dataset, we followed standard procedures for annotation quality control. Each meme was annotated by two annotators. If the disagreements contain similar opinions, i.e there are overlapping annotations, but additional or missing annotations from either of the two annotators, the overlapping annotations were considered correct labels for the sample. However, if there were disagreements with entirely different perspectives, a third annotator was brought in to provide an additional annotation for the meme. The overlapping annotations between at least two annotators were then considered the correct labels. In the extreme case where all three annotators had different opinions, the meme was flagged and removed from the dataset. We also held review discussions with the annotators to discuss the annotations with disagreements, allowing our annotators to receive feedback and improve their annotations. We computed Krippendorff's alpha as a metric for inter-annotator agreement. Our annotators achieved an average score of 0.602, signifying a high amount of systematic agreement in their annotations.

B. Dataset Analysis

We annotated a total of 8,876 unique entities in our dataset. Figure 2 presents the distribution of memes based on the number of unique entities annotated in each meme. We found that 48.4% of memes contain two unique entities. Notably, we observed that the majority of memes contain two or three unique entities, with only a small number of memes containing

[3] https://imgflip.com/memetemplates

[4] https://www.jaided.ai/easyocr/

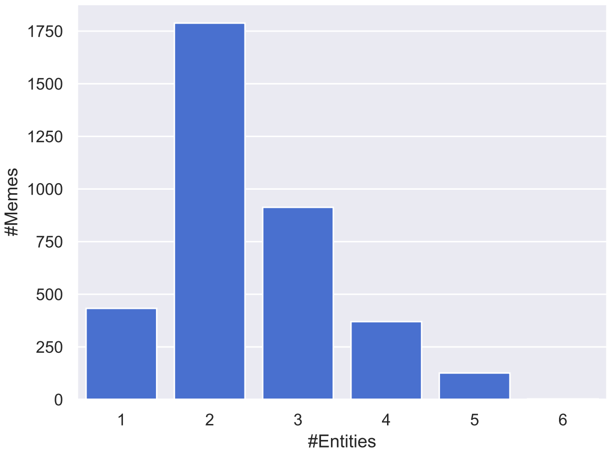


Fig. 2. Distribution of the number of entities in each meme.

TABLE I
DISTRIBUTION OF RELATIONS ANNOTATED IN MEMES

Relation Type	Train	Test	Total
Superior	1046	431	1477
Equal	163	69	232
Upgrade	582	231	813
Affirm/Favor	1258	506	1764
Doubt/Disfavor	1775	732	2507
Indifferent	211	94	305
NULL	6107	2574	8681

four or more unique entities. However, this does not imply lower complexity in comprehending the semantics of memes, as each identified entity can have multiple relationships with other entities.

Table I presents the distribution of relations annotated in memes. To ensure that all relations are adequately represented in our training set, we performed a 70:30 train-test split based on relations. We observe that the majority of memes express opinions about entities being superior or inferior to others or express doubt towards certain entities. However, we also note that the *NULL* relation is the most prevalent because many entities in memes have few relations with other entities, or no relation at all. For instance, in memes with three entities, *Entity_A* may have relations with both *Entity_B* and *Entity_C*, but *Entity_B* and *Entity_C* may have no relation with each other. It is important to note that the lack of explicit relations between entities does not necessarily indicate a lack of complexity in understanding the semantics of memes, as each identified entity can have multiple relationships with other entities.

IV. PROPOSED MODEL

To understand the semantics communicated in the memes, we propose the MERF pipeline to extract entities and their relationships in memes. Figure 3 illustrates the overall pipeline architecture of the proposed MERF pipeline. In the framework, the entity extraction module first performs entity extraction by taking the meme as an input and extract the entities communicated in the meme. Subsequently, the extracted entities and meme are input into the relation extraction model to predict the relations between two given entities. Note that relation prediction are performed pair-wise, i.e., the relation extraction takes in permutations of source and target entities as inputs in its relation prediction.

A. Problem Definition

There are two main sub-tasks in this study: (a) *entity extraction* and (b) *relation extraction*.

Entity Extraction. We consider the entity extraction task as a language modeling problem where the input is a textual sequence and the output is a textual span within the input textual sequence. The source sequence of the model is an input text $X = x_1, \cdots, x_n$ and the target sequence $E = e_1, \cdots, e_n$ is a span containing the target entities. We will discuss the details of the entity extraction process in Section IV-B.

Relation Extraction. The relation extraction task is defined as follows:

$$F_r : (E^s, Pos^s, E^t, Pos^t, Img) \mapsto R \tag{1}$$

where E^s and E^t refer to the source and target entities, respectively. Pos^s and Pos^t represent the position information of the source and target entities, respectively. Finally, Img refers to the meme image. The goal is to predict the relation $r \in R$ between source entity E^s and target entity E^t. We will discuss the details of relation extraction process in Section IV-C.

B. Entity Extraction

The entity extraction task aims to extract entities within an input textual span. We begin by applying EasyOCR to the meme to obtain separated spans of text and provides us with the positional information of extracted text boxes Pos^n. We then apply an encoder to obtain a contextual representation for the input span. We then feed the model into a Multi-layer Perception consisting of two token-level classifiers that predict the probabilities of each word being the start and stop of an entity. We utilize the flagged spans produced as inputs to our relation extraction model.

Text Encoder. For our experiments, we explore two large language models BERT [36] and RoBERTa [37] as our contextual representation encoders. Using BERT as an example, given an input span of text X^1, we first tokenize the sentence using the respective word-piece or byte-pair encoding vocabulary and concatenate a [CLS] token and [SEP] to the start and end of the sequence accordingly.

$$BERT([[CLS], X^1, X^2, ...X^n, [SEP]]) \rightarrow B^1, B^2 \cdots, B^n \tag{2}$$

This allows us to obtain our contextual representation for each of the input spans $[B^1, B^2 ...]$ which we feed into our multi-layer perceptron for entity span detection.

Entity Span Detection. We aim to detect the spans of entities within each input text span. While there are instances where the entire span itself is a single entity, instances where

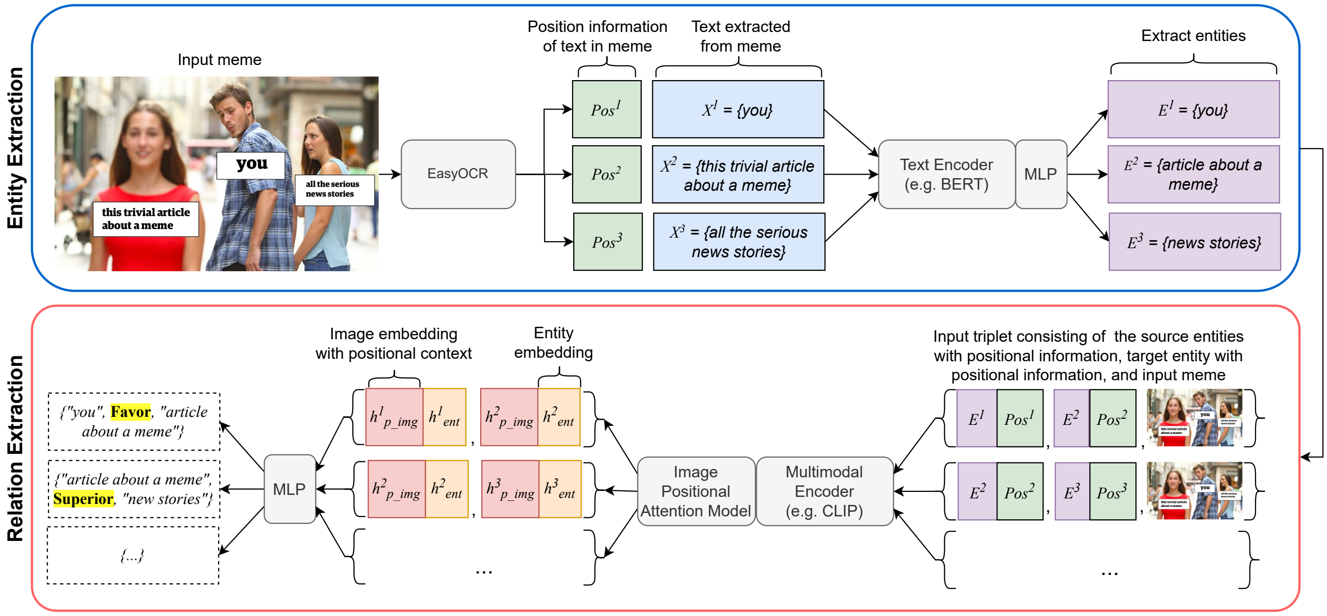

Entity Extraction
Input meme
you
this trivial article about a meme
all the serious news stories
EasyOCR
Position information of text in meme
Text extracted from meme
Pos1
Pos2
Pos3
X1 = {you}
X2 = {this trivial article about a meme}
X3 = {all the serious news stories}
Text Encoder (e.g. BERT)
MLP
Extract entities
E1 = {you}
E2 = {article about a meme}
E3 = {news stories}
Relation Extraction
{"you", Favor, "article about a meme"}
{"article about a meme", Superior, "new stories"}
{...}
MLP
Image embedding with positional context
Entity embedding
h1 p_img
h1 ent
h2 p_img
h2 ent
h3 p_img
h3 ent
...
Image Positional Attention Model
Multimodal Encoder (e.g. CLIP)
Input triplet consisting of the source entities with positional information, target entity with positional information, and input meme
E1
Pos1
E2
Pos2
E3
Pos3
...

Fig. 3. Architecture of MERF pipeline

multiple different entities are present in a single span are also common occurrences. We utilize the contextual span representations obtained from the encoder and feed it into an MLP layer, applying a softmax to obtain the probabilities of a token W^n being the first and last word of an entity, P^n_{start}. Similarly, we perform the same operation for predicting the probability of a token being the last word of an entity, P^n_{end}.

$$B^n = [W^1, W^2, \cdots, W^n] \tag{3}$$

$$P^n_{start} = softmax(MLP_{start}(W^n)) \tag{4}$$

$$P^n_{end} = softmax(MLP_{end}(W^n)) \tag{5}$$

We then define an objective function L_{Entity}, the sum of the binary cross entropy loss of both the start and stop entity span probabilities for each of the tokens.

$$L^n_{Entity} = L^n_{start} + L^n_{stop} \tag{6}$$

$$L^n_{start} = y^n_{start} \cdot logP^n_{start} + (1 - y^n_{start}) \cdot log(1 - P^n_{start}) \tag{7}$$

$$L^n_{stop} = y^n_{stop} \cdot logP^n_{stop} + (1 - y^n_{stop}) \cdot log(1 - P^n_{stop}) \tag{8}$$

where we take y^n_{start} and y^n_{stop} as the correctly labelled spans of an entity. Based off the detection of the entity start and stop signs, we are able to extract the detected entities E^n.

C. Relation Extraction

The relation extraction task aims to classify the relations between entities extracted from the entity extraction module. The task accepts an input triplet consisting of the source entity, target entity, and image of the meme. The positional information, Pos^n, of the source and target entities are also presented as input into the model. Therefore, the goal of the relation extraction is to predict the relationship between the input source and target entities.

Multimodal Encoder. We begin by taking the text of the extracted entities E^n from the entity extraction module and the meme. The extracted entities' texts are tokenized again according to the tokenization vocabulary unique to the multimodal encoder chosen for this module. In our experiments, we have explored utilizing CLIP [58], Data2vec [59], FLAVA [60], VisualBERT [61] and BLIP [62] as the multimodal encoder. While CLIP utilizes byte-pair encoding, BLIP, VisualBERT and FLAVA utilize BERT-based word piece encoding. The previous entity extraction task only required textual cues contained within the span to identify the spans containing an entity within the text boxes. However, inferring the directed relationships between entities requires context that can only be derived from cues contained within the image. Taking the image and individually extracted entity texts, we encode the images and text to obtain the contextual entity and image embedding.

$$Encoder_{text}(E^n) \rightarrow h^{ent}_n \tag{9}$$

$$Encoder_{Image}(I_{Meme}) \rightarrow I_{Emb} \tag{10}$$

Image Positional Attention. The image contextual entity embedding I_{Emb}generated by the multimodal encoder is fed into the Image Positional Attention Model along with the positional information Pos^n of the input entities to generate position-aware image embeddings for each of the entities.

Specifically, we generate image embeddings with a positional context as follows:

$$Attn(Pos^n, I_{Emb}) = h^n_{p\_img} \tag{11}$$

Utilising attention [63] to generate position-aware image embeddings acts as a supervision signal to force the model to focus on regions of the image more relevant for classification. It also allows the model to derive any positioning that indicates hierarchical positioning within the meme.

Relation Classification. Classification of the relations between two detected entities within the same meme is performed by taking the position-aware image embedding $h^n_{p\_img}$ and the context-aware text embedding h^{ent}_n as inputs representing each of the entities and feeding it into an MLP layer before applying it into a sigmoid function to obtain the probabilities for each of the separate relationship classes.

$$\begin{aligned} V_{relation} = Sigmoid(MLP_{relation}([h^{Src}_{p\_img}, h^{Src}_{ent}], \\ [h^{Tar}_{p\_img}, h^{Tar}_{ent}])) \end{aligned} \tag{12}$$

where $h^{Src}_{p\_img}$ and h^{Src}_{ent} represents the position-aware image embedding and the context-aware text embedding of the source entity, $h^{Tar}_{p\_img}$ and h^{Tar}_{ent} represents the position-aware image embedding and the context-aware text embedding of the target entity, and $V_{relation}$ represents a vector containing all relationship probabilities between the two entities. Note that we have cast the relation classification as a multilabel classification task as two entities may have multiple relationships.

Finally, we define an objective function $L_{relation}$ to optimize the model. Similar to the entity extraction component, we utilize binary cross entropy due to the possibility of multiple relations between the entity pair being simultaneously true and define $L_{relation}$ as the sum of all the losses from every relation classification instance n R_n.

$$L_{relation} = \sum_{n}^{r} R_n \cdot logP^n + (1 - R_n) \cdot log(1 - P^n) \tag{13}$$

V. Experiments

In this section, we first provide a brief introduction to the evaluation setting followed by the configuration of text and multimodal encoders experimented in MERF pipeline. Next, we present a set of experiments conducted to evaluate MERF's entity relation extraction performance on multimodal memes. We also conduct low-resource setting experiments to understand MERF robustness on the meme entity relation extraction task.

A. Experiment Settings

Datasets. We evaluate MERF using our constructed MERMAID dataset. Specifically, we adopt a 60-40 train-test split on the MERMAID dataset. The split is stratified by the 39 meme templates such that all meme templates are observed in the training stage.

Evaluation Metrics. We report Precision, Recall, F1 and accuracy for our entity extraction task. We only consider an entity to have been correctly extracted if the model output matches the exact span of the gold-labelled entities in the MERMAID dataset. We report accuracy, micro and macro F1 scores for the relation extraction task. These evaluation metrics are utilized due to class imbalance for the relation type labels in MERMAID dataset.

TABLE II
Entity extraction results of MERF pipeline with various text encoders. Best performances are in bold.

Text Encoder	Precision	Recall	F1	Acc
BERT	0.933	**0.912**	0.923	**85.2**
RoBERTa	**0.943**	0.907	**0.925**	84.8

Text and Multimodal Encoders. The text and multimodal encoders are two key components in the MERF pipeline. In our experiments, we experiment with combinations of two text encoders and four multimodal encoders. For text encoders, we utilize pre-trained large-language models such as BERT-base [36] and RoBERTa-base [37]. For multimodal encoders, we utilize state-of-the-art pre-trained visual-language models such as Data2Vec [59], CLIP [58], FLAVA [60], VisualBERT [61] and BLIP [62]. Note that while we do not explicitly pre-trained these models, we have utilized their pre-trained weights in published in HuggingFace[5].

Implementation Details. We perform the data processing steps described in Section IV-B, where we extract the meme text and corresponding positional information using EasyOCR. We utilize the implementations and pre-trained weights in HuggingFace for the text encoders and multimodal encoders used in the MERF pipeline. For all configurations of MERF, we train with a batch size of 32 and empirically set the learning rate to $1e-5$ and set a weight decay of $1e-6$. We optimized MERF with the AdamW optimizer [64] and implement them in PyTorch.

B. Entity and Relation Evaluation

Entity Extraction Results. Table II shows the results of our experiments for entity extraction on the MERMAID dataset. Both BERT and RoBERTa perform similarly well in the entity extraction task. A possible reason for the models' high performance could be due to the nature of the dataset; most of the input text (i.e., text extracted from EasyOCR) only contain the entity itself. For instance, we noted that about 87% of the input text only contain the entity itself without additional words. The salient pattern makes it easier for the model to learn and detect that the entity span would cover the entire input text. Nevertheless, we also noted that the strong performance suggests that both models are also able to extract entities from samples where the input text contain words besides the entities.

Relation Extraction Results The results of the different combinations of text and multimodal encoders on the relation extraction task are presented in Table III. Our analysis revealed that, in contrast to the entity extraction task, most models did not perform as well on the relation extraction task, indicating

[5] https://huggingface.co/

TABLE III
Relation extraction results of MERF pipeline with various text-multimodal encoders. Best performances are in bold.

Text-Multimodal Encoder	**Micro-F1**	**Macro-F1**	**Acc.**
BERT-Data2vec	0.699	0.560	64.4
BERT-CLIP	**0.704**	0.558	**65.1**
BERT-FLAVA	0.539	0.248	45.3
BERT- BLIP	0.524	0.135	43.8
BERT-VisualBERT	0.317	0.0867	23.1
RoBERTa-Data2vec	0.685	0.548	62.6
RoBERTa-CLIP	0.701	**0.561**	64.3
RoBERTa-FLAVA	0.252	0.526	43.5
RoBERTa-BLIP	0.516	0.140	41.9
RoBERTa-VisualBERT	0.237	0.0755	15.3

TABLE IV
Low-resource setting evaluation results of MERF (BERT-CLIP) with various k training samples from five complex templates. Scores reported are Micro-F1 scores

Template	k-training samples				
	0	5	10	20	Max
Cuphead-Flower	0.639	0.643	0.728	0.781	0.816
Hide-the-Pain-Harold	0.071	0.277	0.292	0.581	0.68
Mother-Kid-Drowning	0.384	0.282	0.389	0.540	0.679
Running-Away-Balloon	0.651	0.546	0.745	0.739	0.825
Soyboy-Vs-Yes-Chad	0.246	0.058	0.198	0.3352	0.349

its complexity and challenge. To effectively perform relation extraction in memes, a model needs to comprehend and consider multiple possibilities simultaneously. The contextual cues provided by the template and positioning of the entities within the meme can indicate the most common types of relationships that the meme creator aims to convey. However, the presence of various entities in the context can alter the meaning of the meme entirely.

We observed that variants of the MERF model utilizing Data2vec and CLIP as multimodal encoders outperformed the variants that utilize FLAVA and BLIP. The difference in performance is likely due to the difference in pretraining datasets. CLIP and Data2vec have training data containing image-text pairs directly from the internet, while BLIP and FLAVA utilize captioned images grounded in real-world captioning examples, such as the COCO or LAION datasets, making them less suitable for inferring context not grounded directly in the content of the image itself. VisualBERT performs the worst, suffering not just from the utilization of only images grounded in real world caption examples, but also a low pretraining dataset size.

C. Low-Resource Evaluation

Memes are a fast-evolving communication medium; there are always new unseen memes and we may not have the resources to annotate a large number of memes for training MERF. Therefore, to examine the practicality and robustness of MERF pipeline, we conduct further relation extraction experiments in low-resource settings. Specifically, we first selected five complex meme templates with higher number of relations (i.e., meme templates with more than seven relations). Next, we vary the training samples of these five selected templates. For instance, we randomly retain k training samples for each of the five selected templates in the training set. In our experiment, we set $k = \{0, 5, 10, 20, max\}$, where $k = 0$ refers to no training samples from the five selected templates (i.e., unseen), while $k = max$ refers to the including all samples in the original training set. It should be noted that the number of samples for non-selected templates remained unchanged in the training set. Finally, we utilized the same test set used in the experiments conducted in Section V-B, reporting the micro-F1 scores for each of the five selected complex meme templates.

The experimental results of our best-performing MERF configuration (i.e., BERT-CLIP) in a low-resource setting are presented in Table IV. It can be observed that the Micro F1 scores of our model in the $k = 0$ setting, which pertains to unseen memes, are exceptionally high. Upon further analysis of the predictions made by our model, we noted that MERF tends to predict a "*NULL*" relation between all entities. This is primarily attributed to the insufficient number of training samples for the unknown meme types.

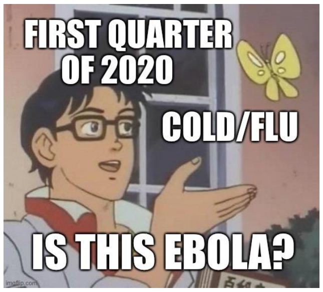
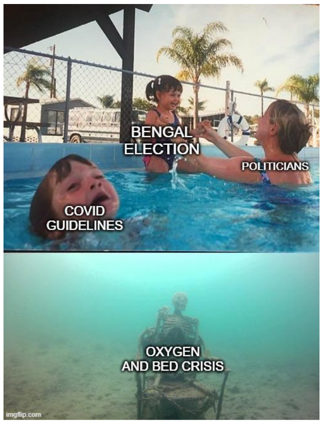
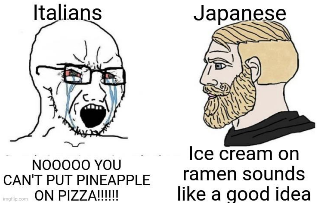
It is worth mentioning that an increase in the number of k samples for the low-resource setting leads to a performance dip initially. This can be attributed to the model's inclination towards categorizing relations between entities detected within the model, rather than indicating only a "*NULL*" relation. However, after the initial dip, there is a general improvement in the model's performance as the number of samples increases. Moreover, our analysis revealed that the performance difference between $k = 20$ and $k = max$ is relatively insignificant. This observation suggests that training MERF with only 20 samples of a specific template yields relatively good performance for relation extraction of the memes from the template. The promising outcomes of our experiments indicate that MERF has the potential to facilitate meme semantic extraction in low-resource settings.

D. Case Studies

In this section, we conduct case studies to qualitatively analyze the strenths and limitations of our proposed MERF pipeline. Table V shows three sample memes from MERMAID and the entities and relations extracted using MERF (BERT-CLIP). The meme on the left is based on the template "*Is this a butterfly?*", communicating that in the first quarter of 2020, many people overacted and view cold/flu as dangerous as Ebola. MERF is able to correctly extract the three entities in the meme and their relations with each other. For instance, MERF predicts an "*Equal*" relation between the two extracted entities: "*Cold/Flu*" and "*Ebola*".

MERF is also able to handle complex memes with more entities and relations. For instance, the middle meme in Table V is a satire ridiculing that the politicians are more interested in the elections than the COVID-19 pandemic issues. The meme contains four entities and five relations, and MERF is able to correctly exact them. Specifically, in relation extraction, MERF is able to correctly identify the preference of the entity "*Politicians*" within the meme, which is "*Bengal*

TABLE V
EXTRACTED ENTITIES AND RELATIONS FROM THREE SAMPLE MEMES. THE INCORRECT OR MISSING EXTRACTIONS ARE HIGHLIGHTED IN RED.

Meme	 FIRST QUARTER OF 2020 COLD/FLU IS THIS EBOLA? imgflip.com	 BENGAL ELECTION POLITICIANS COVID GUIDELINES OXYGEN AND BED CRISIS imgflip.com	 Italians Japanese NOOOOO YOU CAN'T PUT PINEAPPLE ON PIZZA!!!!!! Ice cream on ramen sounds like a good idea
Extracted Entities	First quarter of 2020 Cold/Flu Ebola	Bengal Elections Politicians Covid Guidelines Oxygen and Bed Crisis	Italians Japanese Ice cream on ramen Pineapple on Pizza

	{ Cold/Flu, *Equal*, Ebola }
	{ First Quarter of 2020 *Indifferent*, Cold/Flu }
	{ First Quarter of 2020 *Indifferent*, Ebola }
		{ Bengal Elections Superior, Covid Guidelines }
		{ Covid Guidelines Superior, Oxygen and Bed Crisis }
		{ Politicians Doubt/ Disfavor Oxygen and Bed Crisis }
		{ Politicians Indifferent Covid Guidelines }
			{ Japanese, Superior Italians }
			{ Japanese, Superior Ice Cream on Ramen }
			{ Italians, Null Ice Cream on Ramen }

Elections", "*Covid Guidelines*", and "*Oxygen and Bed Crisis*", in descending order.

In addition to evaluating the performance of MERF on memes, we also aimed to analyze instances of incorrect classification made by the model. As presented in the right-most example of Table V, there exists a meme with both incorrect entity and relation extraction. Specifically, MERF is only able to identify three out of four entities present within the meme, namely "*Italians*", "*Japanese*", and "*Ice cream on ramen*". The model fails to detect the last entity "*Pineapple on Pizza*", which leads to an error propagation where the model does not attempt to classify relations for this undetected entity. Furthermore, the model incorrectly identifies the relationships between "*Japanese*" and "*Ice cream on ramen*", labelling the relationship as "*Superior*" instead of "*Affirm/Favor*". In addition, the model indicates a null relationship between "*Italians*" instead of "*Doubt/Disfavor*". We postulate that this is primarily due to the non-standard expression found only in memes, which may not have provided sufficient textual and positional cues for successful entity and relation extraction.

VI. CONCLUSION

In this paper, we proposed the new tasks of entity and relation extraction in multimodal memes. To facilitate the new tasks, we constructed the MERMAID dataset, which serves to improve semantic understanding of memes. The MERMAID dataset can also support the training and testing of meme multimodal entity extraction and relation inference models. We designed a novel MERF pipeline that extracts entities and their relationships in memes. To the best of our knowledge, this is the first framework that proposes a framework to extract entities and relations in memes. We evaluated our proposed framework through extensive experiments on a real-world meme dataset. Our automatic and low-resource setting evaluations establish the benchmark for the new multimodal meme semantic understanding task. Our experimental results demonstrated a promising step forward towards multimodal entity and relation extraction and are crucial in the understanding of meme semantics under a highly contextualized and ambiguous social media landscape. For future works, we will explore the conceptual adaptation into non-template memes to uncover a generalization possibility by large-scale pretraining on relation extraction. We will also design new approaches that can better model and reason the relations between entities in memes.

REFERENCES

[1] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–30, 2018.

[2] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, "Detecting and tracking the spread of astroturf memes in microblog streams," *arXiv preprint arXiv:1011.3768*, 2010.

[3] S. Mahbub, E. Pardede, A. Kayes, and W. Rahayu, "Controlling astroturfing on the internet: a survey on detection techniques and research challenges," *International journal of web and grid services*, vol. 15, no. 2, pp. 139–158, 2019.

[4] Y. Zhou, Z. Chen, and H. Yang, "Multimodal learning for hateful memes detection," in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2021, pp. 1–6.

[5] W. Zhang, G. Liu, Z. Li, and F. Zhu, "Hateful memes detection via complementary visual and linguistic networks," *arXiv preprint arXiv:2012.04977*, 2020.

[6] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, and T. Chakraborty, "Momenta: A multimodal framework for detecting harmful memes and their targets," *arXiv preprint arXiv:2109.05184*, 2021.

[7] T. Deshpande and N. Mani, "An interpretable approach to hateful meme detection," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 723–727.

[8] R. K.-W. Lee, R. Cao, Z. Fan, J. Jiang, and W.-H. Chong, "Disentangling hate in online memes," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5138–5147.

[9] C. Sharma, D. Bhageria, W. Scott, S. Pykl, A. Das, T. Chakraborty, V. Pulabaigari, and B. Gamback, "Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor!" *arXiv preprint arXiv:2008.03781*, 2020.

[10] S. Sharma, M. S. Akhtar, P. Nakov, and T. Chakraborty, "DISARM: Detecting the victims targeted by harmful memes," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1572–1588. [Online]. Available: https://aclanthology.org/2022.findings-naacl.118

[11] S. Pramanick, D. Dimitrov, R. Mukherjee, S. Sharma, M. Akhtar, P. Nakov, T. Chakraborty *et al.*, "Detecting harmful memes and their targets," *arXiv preprint arXiv:2110.00413*, 2021.

[12] S. Fharook, S. Sufyan Ahmed, G. Rithika, S. S. Budde, S. Saumya, and S. Biradar, "Are you a hero or a villain? a semantic role labelling approach for detecting harmful memes." in *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 19–23. [Online]. Available: https://aclanthology.org/2022.constraint-1.3

[13] J. Danung and L. H. Attaway, "All your media are belong to us: An analysis of the cultural connotations of the internet meme," *Literature, culture and digital media*, vol. 17, 2008.

[14] R. Chandler, "Meme world syndrome: A critical discourse analysis of the first world problems and third world success internet memes," 2013.

[15] B. E. Wiggins and G. B. Bowers, "Memes as genre: A structurational analysis of the memescape," *New media & society*, vol. 17, no. 11, pp. 1886–1906, 2015.

[16] K. Chen, A. Feng, R. Aanegola, K. Saha, A. Wong, Z. Schwitzky, R. K.-W. Lee, R. O'Hanlon, M. De Choudhury, F. L. Altice *et al.*, "Categorizing memes about the ukraine conflict," in *International Conference on Computational Data and Social Networks*. Springer, 2022, pp. 27–38.

[17] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2611–2624, 2020.

[18] L. Mathias, S. Nie, A. Mostafazadeh Davani, D. Kiela, V. Prabhakaran, B. Vidgen, and Z. Waseem, "Findings of the WOAH 5 shared task on fine grained hateful memes detection," in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 201–206.

[19] J. Zhu, R. K.-W. Lee, and W. H. Chong, "Multimodal zero-shot hateful meme detection," in *14th ACM Web Science Conference 2022*, 2022, pp. 382–389.

[20] M. S. Hee, R. K.-W. Lee, and W.-H. Chong, "On explaining multimodal hateful meme detection models," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3651–3655.

[21] R. Cao, R. K.-W. Lee, W.-H. Chong, and J. Jiang, "Prompting for multimodal hateful meme classification," *arXiv preprint arXiv:2302.04156*, 2023.

[22] R. Cao, M. S. Hee, A. Kuek, W.-H. Chong, R. K.-W. Lee, and J. Jiang, "Pro-cap: Leveraging a frozen vision-language model for hateful meme detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5244–5252.

[23] M. S. Hee, W.-H. Chong, and R. K.-W. Lee, "Decoding the underlying meaning of multimodal hateful memes," *arXiv preprint arXiv:2305.17678*, 2023.

[24] N. Prakash, H. Wang, N. K. Hoang, M. S. Hee, and R. K.-W. Lee, "Promptmtopic: Unsupervised multimodal topic modeling of memes using large language models," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 621–631.

[25] M. S. Hee, A. Kumaresan, N. K. Hoang, N. Prakash, R. Cao, and R. K.-W. Lee, "Matk: The meme analytical tool kit," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 9689–9692.

[26] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, and P. Buitelaar, "Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text," in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 32–41.

[27] C. Sharma, D. Bhageria, W. Scott, S. PYKL, A. Das, T. Chakraborty, V. Pulabaigari, and B. Gambäck, "SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor!" in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 759–773. [Online]. Available: https://aclanthology.org/2020.semeval-1.99

[28] N. Prakash, M. S. Hee, and R. K.-W. Lee, "Totaldefmeme: A multi-attribute meme dataset on total defence in singapore," in *Proceedings of the 14th Conference on ACM Multimedia Systems*, 2023, pp. 369–375.

[29] S. Pramanick, M. S. Akhtar, and T. Chakraborty, "Exercise? i thought you said 'extra fries': Leveraging sentence demarcations and multi-hop attention for meme affect analysis," *ArXiv*, vol. abs/2103.12377, 2021.

[30] J. Qu, L. H. Li, J. Zhao, S. Dev, and K.-W. Chang, "Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation," *arXiv preprint arXiv:2205.12617*, 2022.

[31] B. Settles, "Biomedical named entity recognition using conditional random fields and rich feature sets," in *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, 2004, pp. 107–110.

[32] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[33] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 188–191. [Online]. Available: https://aclanthology.org/W03-0430

[34] N. Limsopatham and N. H. Collier, "Bidirectional lstm for named entity recognition in twitter messages," 2016.

[35] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns," *Transactions of the association for computational linguistics*, vol. 4, pp. 357–370, 2016.

[36] J. Devlin, M. Chang, K. Lee, and 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, NAACL-HLT 2019*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186.

[37] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[38] P. Liu, Y. Guo, F. Wang, and G. Li, "Chinese named entity recognition: The state of the art," *Neurocomputing*, vol. 473, pp. 37–53, 2022.

[39] I. Budi and R. R. Suryono, "Application of named entity recognition method for indonesian datasets: a review," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 969–978, 2023.

[40] Y. An, X. Xia, X. Chen, F.-X. Wu, and J. Wang, "Chinese clinical named entity recognition via multi-head self-attention based bilstm-crf," *Artificial Intelligence in Medicine*, vol. 127, p. 102282, 2022.

[41] R. Mooney and R. Bunescu, "Subsequence kernels for relation extraction," *Advances in neural information processing systems*, vol. 18, 2005.

[42] P. Pantel and M. Pennacchiotti, "Espresso: Leveraging generic patterns for automatically harvesting semantic relations," in *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics*, 2006, pp. 113–120.

[43] S. Blohm and P. Cimiano, "Using the web to reduce data sparseness in pattern-based information extraction," in *Knowledge Discovery in Databases: PKDD 2007: 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007. Proceedings 11*. Springer, 2007, pp. 18–29.

[44] S. Deepika and T. Geetha, "Pattern-based bootstrapping framework for biomedical relation extraction," *Engineering Applications of Artificial Intelligence*, vol. 99, p. 104130, 2021.

[45] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, 2014, pp. 2335–2344.

[46] T. H. Nguyen and R. Grishman, "Relation extraction: Perspective from convolutional neural networks," in *Proceedings of the 1st workshop on vector space modeling for natural language processing*, 2015, pp. 39–48.

[47] D. Sorokin and I. Gurevych, "Context-aware representations for knowledge base relation extraction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1784–1789.

[48] Y. Zhang, P. Qi, and C. D. Manning, "Graph convolution over pruned dependency trees improves relation extraction," *arXiv preprint arXiv:1809.10185*, 2018.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[50] A. Conneau and G. Lample, "Cross-lingual language model pretraining," *Advances in neural information processing systems*, vol. 32, 2019.

[51] B. Dong, Y. Yao, R. Xie, T. Gao, X. Han, Z. Liu, F. Lin, L. Lin, and M. Sun, "Meta-information guided meta-learning for few-shot relation classification," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1594–1605. [Online]. Available: https://aclanthology.org/2020.coling-main.140

[52] X. Han, H. Zhu, P. Yu, Z. Wang, Y. Yao, Z. Liu, and M. Sun, "Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation," *arXiv preprint arXiv:1810.10147*, 2018.

[53] W. Wen, Y. Liu, C. Ouyang, Q. Lin, and T. Chung, "Enhanced prototypical network for few-shot relation extraction," *Information Processing & Management*, vol. 58, no. 4, p. 102596, 2021.

[54] M. Qu, T. Gao, L.-P. Xhonneux, and J. Tang, "Few-shot relation extraction via bayesian meta-learning on relation graphs," in *International conference on machine learning*. PMLR, 2020, pp. 7867–7876.

[55] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," 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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3816–3830. [Online]. Available: https://aclanthology.org/2021.acl-long.295

[56] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning, "Position-aware attention and supervised data improve slot filling," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 35–45. [Online]. Available: https://aclanthology.org/D17-1004

[57] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. O. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz, "Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals," *arXiv preprint arXiv:1911.10422*, 2019.

[58] 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.

[59] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 1298–1312. [Online]. Available: https://proceedings.mlr.press/v162/baevski22a.html

[60] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela, "Flava: A foundational language and vision alignment model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 638–15 650.

[61] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[62] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[63] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.