Social Meme-ing: Measuring Linguistic Variation in Memes

Anonymous ACL submission

Abstract

Much work in the space of NLP has used computational methods to explore sociolinguistic variation in text. In this paper, we argue that memes, as multimodal forms of language comprised of visual templates and text, also exhibit meaningful social variation. We construct a computational pipeline to cluster individual instances of memes into templates and semantic variables, taking advantage of their multimodal structure in doing so. We apply this method to a large collection of meme images from Reddit and make available the resulting SEMANTICMEMES dataset of 3.8M images clustered by their semantic function. We use these clusters to analyze linguistic variation in memes, discovering not only that socially meaningful variation in meme usage exists between subreddits, but that patterns of meme innovation and acculturation within these communities align with previous findings on written language.

1 Introduction

One objective in variationist sociolinguistics is to study how social factors contribute to differences in the way people use language. Work in natural language processing has followed this tradition, offering large-scale analyses of how language use is conditioned on geography, (Eisenstein et al., 2010; Hovy and Purschke, 2018; Demszky et al., 2021), community (Del Tredici and Fernández, 2017; Zhu and Jurgens, 2021b; Lucy and Bamman, 2021) and time (Hamilton et al., 2016). This work is important not only because language variation often exposes shortcomings in NLP tools, which are primarily developed for standard language varieties (Blodgett et al., 2016), but also because variation often embeds **social meaning**. We make inferences about people's social class, regionality, gender, and much more based on the way they talk (Campbell-Kibler, 2009; Zhang, 2005), and we strategically use language to actively construct and perform identities (Labov, 1963; Bucholtz and Hall, 2005).

Figure 1: Meme templates can be visually diverse, but often provide the same semantic function; in this case, all four templates show a scalar increase.

Most of this work has focused on lexical or morphosyntactic variation in written texts. However, language exists beyond text or speech. In face-to-face interaction, multimodality in language has been construed as features like co-speech gesture, facial expression or body movement (Perniss, 2018). In online communication, previous work has extended the term to include the interplay between images and text (Kress and Leeuwen, 2001; Zhang et al., 2021; Hessel et al., 2023). Understanding text in isolation is insufficient to understanding how we communicate online.

In the space of multimodal online language, memes are interesting for their compositionality. They consist of a base image (the **template**) as well as superimposed text (which we refer to as the **fill**). For example, the "Drake" template depicted in figure 2 serves the semantic function of expressing a preference relation between the fills. This same Drake template can be used to express preference relations between a range of fills; at the same time, multiple different templates can share the same or

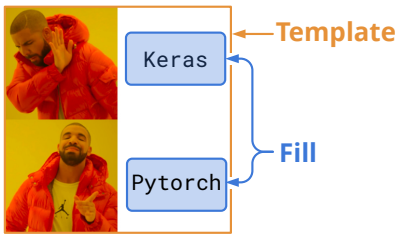


Figure 2: Memes are multimodal constructions where the base image **template** and additional text **fills** both have semantic value.

similar semantic function, as illustrated in fig. 1 for the function of "scalar increase." We refer to this set of functionally equivalent templates as a **semantic cluster**.

In this work, we follow the variationist sociolinguistics tradition by treating templates as *variants* and semantic clusters as *variables*, observing how social factors might contribute to the distribution among these variants. To conduct this analysis, we develop a method for identifying semantic clusters by exploiting the visual structure of meme templates and the linguistic structure of meme fills. We use this to create the SEMANTICMEMES dataset of 3.8M Reddit memes[1] grouped into semantic clusters and validated with a human evaluation. Finally, we use these semantic clusters to perform a series of case studies demonstrating their use in studying linguistic variation, linguistic innovation, and linguistic acculturation. We find that:

1. socially meaningful variation in template choice exists between subreddits;

2. subreddits that first introduce a new template continue to use it more than others; and

3. users who stay in a subreddit for longer tend to use templates distinctive to that subreddit.

These findings illustrate the ways in which memes function as multimodal acts of communication, and how methods from computational sociolinguistics can shed light on meaningful variation within them.

2 Methods

To study variation in meme use, we need to identify the meme variables that organize a collection of meme **instances**—the individual memes that are created and posted online by specific people at specific moments in time. We create a pipeline that visually clusters meme **instances** into **templates** (i.e., the same memes that differ by variation in fills) by exploiting the visual similarity between them; and linguistically clustering meme **templates** into **semantic clusters** by exploiting the similarity among the fills used in different templates. Fig. 3 provides an overview of the process, which involves first clustering instances into templates (§2.1), and then clustering templates into variables (§2.2).

2.1 Visually clustering instances

Our process starts with a set of meme instances, which we wish to group based on visual similarity; this process serves to group memes into their base templates as well as filter out many non-meme images. This is difficult due to the massive number of images as well as the amount of variation in zoom, crop, borders and other visual details. We lay out the steps of the process here, but provide further details and example images in Appendix A.

We first preprocess images to remove any solid color framing elements to isolate the base image, then follow Zannettou et al. (2018) and Morina and Bernstein (2022) in extracting templatized memes by running a perceptual hashing algorithm.

We then compute the pairwise Hamming distance between hashes that occur more than 10 times, discard any pairs where the distance was greater than a cut-off $d_{\text{max}} = 10$. We use the Leiden clustering algorithm to perform clustering (Traag et al., 2019).[2] The Leiden algorithm iteratively finds well-connected subgraphs; we construct a graph where image hashes were vertices and the edge weight was $e_{ij} = d_{\text{max}} - d_{ij} + 1$ for vertices i and j, where d_{ij} was the Hamming distance between them.

The clustering algorithm splits aggressively—instances with similar base images may be split across multiple templates due to variations in the zoom, crop, and borders. We find the next step, which clusters based on the fill text, serves as a remedy by placing many of these duplicate templates into the same semantic cluster. Appendix A contains examples of template clusters.

2.2 Linguistically clustering templates

Given a set of meme templates, we want to identify clusters of those templates that have a similar

[1] We make data and code under available under the MIT license at [anonymized]

[2] We found that using DBSCAN, as was done in prior work, resulted in many images being put into a single noisy cluster.

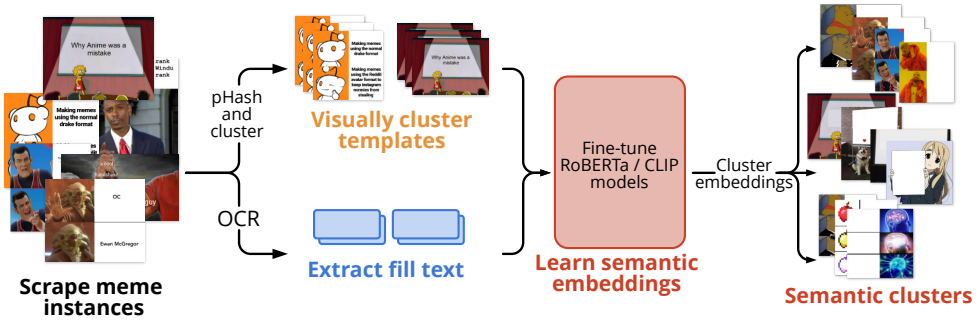


Figure 3: We group visually identical meme instances into templates, and extract the fills using OCR. This data is used to learn semantic embedding representations of templates, which we use to generate semantic clusters.

semantic function—i.e., that are used to assert a similar relation among the text in the fills (such as a comparison function exemplified by the Drake meme in fig. 2). These semantic clusters are the linguistic variables of analysis: discrete sets of variants which share a semantic function but vary in the social meanings they index.

We apply the key intuition that people will use certain templates to make certain classes of statements (comparison, declaration, surprise); as with any other language, fills that are "grammatical" for one template may be nonsensical in another. Templates that share similar sets of fills, then, may perform a similar function over them.

To cluster templates using this principle, we extract the fill text from meme instances belonging to a template (§2.2.1), learn semantic representations for templates based on the distribution of text fills (§2.2.2), and cluster those representations (§2.2.3).

2.2.1 Extracting fill text

We extract text (along with the bounding boxes containing it) from meme instances using EasyOCR.[3] We use the order of the bounding boxes as a rough signal for the position and ordering of the text, but do not incorporate the bounding coordinates directly into the models described below.

Some meme templates contain text in the base image. To prevent these from trivializing the semantic embedding task, we remove bounding boxes with text that was identical in over 90% of the memes in a given template cluster.

2.2.2 Learning semantic embeddings

We examine four methods for learning semantic embeddings of memes, each described in more detail below: a RoBERTa (Liu et al., 2019) classifier fine-tuned to predict the template given the fill text; a CLIP model (Radford et al., 2021) fined-tuned on (fill text, image) pairs; the vector difference between fine-tuned and pretrained CLIP embeddings (CLIP-diff); and concatenating CLIP-diff and RoBERTa embeddings (Concat).

Text-only RoBERTa. In the text-only model, we fine-tune a RoBERTa model on a sequence classification task to predict a distribution over templates given the fill text as input. We separate text in different bounding boxes in a meme with a separator token when passing it into the model to impose a rough, linear notion of space.

After fine-tuning, we take the weights of the final classification layer $W \in \mathbb{R}^{768 \times N}$ as the embeddings, where N is the number of templates. Intuitively, RoBERTa is an encoder model that projects the fill text into a latent semantic space. The final classification layer can be thought of as a projection from that latent space into the discrete space of templates. Therefore, the transposition of the weight matrix can be viewed as a mapping from templates into the latent semantic space.

Multimodal CLIP. In learning the embeddings, the text-based RoBERTa model does not have direct access to the image features in the templates. We experiment with using both the image and text data by fine-tuning a CLIP model.

We fine-tune CLIP using a contrastive loss between the embedding of a meme instance and its fill text. To prevent the model from cheating by reading the text in the image, we sample a meme instance with different text but the same template. This fine-tuning step modifies the image embedding to align with fill text, which implicitly describes the semantic function of the meme, instead of with the pretraining dataset of image captions, which explicitly describe the contents of the image.

CLIP generates embeddings of meme instances.

[3] https://github.com/JaidedAI/EasyOCR

(a) Declaration

(b) Unpopular statement

(c) Surprise narrative

(d) Similarity / parts of a whole

Figure 4: Examples from semantic clusters generated from RoBERTa embeddings; visually diverse clusters emerge even for complex semantic functions like a surprise narrative.

To generate template embeddings, we randomly sample up to ten instances of a template as input for the image embedding module. We then compute the average image embedding of those instances. We don't embed the fill text for this step, since fill text greatly varies between meme instances that use the same template, but the image templates should be more or less visually identical.

CLIP-diff. It is possible that the fine-tuned model contains a notion of meme semantics that is in tension with the pretraining task of image captioning. To isolate the meme-specific knowledge learned by fine-tuning, we calculate the difference between the fine-tuned CLIP embedding of an image and the embedding from the base CLIP model.

CLIP-diff + RoBERTa. Finally, we concatenate the CLIP-diff and RoBERTa embeddings to incorporate both the visual features from CLIP as well as the semantics learned by the RoBERTa model.

2.2.3 Semantic clusters

To group templates into meme variables, we use Leiden clustering on the template representations from the embedding models. Following literature on using embeddings in later layers of language models (Timkey and van Schijndel, 2021), we first standardized the template embeddings before calculating the cosine similarity.

We construct an adjacency matrix from the top 50 nearest neighbors for each template embedding, weighting edges as a function of the ranked cosine similarity. We provide details about this process in Appendix B. We generate **semantic clusters** by running the Leiden algorithm on this graph.

3 SEMANTICMEMES Dataset

We used the pipeline described above to generate semantic clusters from a dataset of 27.9M images collected from Reddit (Baumgartner et al., 2020). We scraped image posts from the top 1000 most active subreddits with "meme" in name (e.g. `r/HistoryMemes`). Temporally, the dataset spans the decade between 2011 and mid-2021.

We fine-tuned both the RoBERTa and CLIP models for three epochs on memes whose template appeared at least 100 times in the dataset. We used an 80/10/10 split of train, dev and test data, ensuring there was no leakage of fill text between splits.

Using the pipeline with the RoBERTa model results in 784 semantic clusters spanning 6,384 templates and over 3.8M meme instances. Figure 4 shows some templates that appear in the same semantic cluster. The dataset includes posts to 655 subreddits by 908,917 users. We include examples and descriptive statistics for clusters generated with each of the embedding models in the appendix.

4 Evaluation

We evaluate the coherence and visual diversity of semantic clusters derived from each model using human judgment. We design an evaluation task in which annotators are presented with a pair of templates, and randomly vary if the templates are drawn from the same or different semantic clusters.

They are asked to evaluate whether the two templates are 1) semantically similar and 2) visually similar. We define semantic similarity as being able

Model	Precision	Visual-adjusted
RoBERTa	**0.78**	**0.44**
CLIP	0.65	-0.09
CLIP-diff	0.69	0.18
Concat.	0.70	0.30

Table 1: Comparison of cluster quality for different embedding models. CLIP-based models yield clusters that are biased towards visual features.

to reasonably substitute the text from one template into the other with minor changes. We define visual similarity to include sharing a similar art style or source (e.g., two different templates featuring Spongebob). We include example pairs in the appendix; one strong source of visual similarity (cf. Appendix Fig. 17) are sets of templates that are largely identical in their form but that exhibit slight variation in size, crop, and margins.

We collect judgments for the top ten semantic clusters from each model most commonly represented in our dataset as well as a random selection of ten clusters from each model. For each cluster, we sample 10 pairs. We find strong interannotator agreement (Krippendorff's α=0.75). From the human judgments, we calculate p_s (the probability that a pair of templates are semantically similar if they appear in the same cluster) and p_v (the equivalent measurement for visual similarity) for each model. We use precision as the evaluation metric because, to measure variation, it is more important each semantic cluster is semantically coherent and visually diverse, but less important that all relevant templates are surfaced within the same cluster.

Our goal in this work is to explore meaningful semantic variation across visually *diverse* memes, since memes that are visually similar (e.g., slight variations on the same template) have trivially similar semantics. Accordingly, we design a measure of *visually adjusted precision* based on Cohen's κ:

$$p_{\text{adj}} = \frac{p_s - p_v}{1 - p_v},$$

Intuitively, this metric represents the extent to which the semantic clusters agree with annotator judgments of semantic similarity while controlling for correlations with visual similarity. A negative score means the model clusters based on visual similarity instead of semantic coherence.

Table 1 presents the results of this evaluation. Introducing any visual features results in some clusters based on visual similarity instead of semantics; accordingly, RoBERTa clusters have the highest visually-adjusted precision (significant w.r.t to CLIP and CLIP-diff for a 95% bootstrap CI).

Semantic clusters provide a strong separation between content (the semantic cluster) and style (the choice of template within a semantic cluster). In other words, the choice of semantic cluster is *what* a user is trying to say, and the choice of a template within that cluster is *how* they are saying it. In the remainder of the paper, we use the clusters generated from the RoBERTa embeddings, which have the highest visual-adjusted precision, for our case studies on linguistic variation and change.

5 Linguistic variation

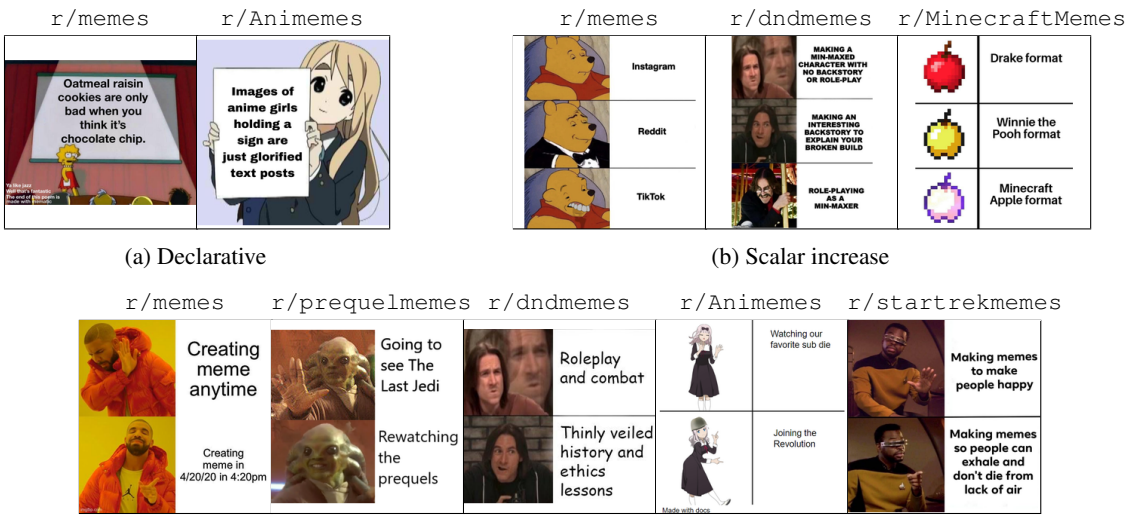
The sociolinguistic study of variation centers around the linguistic variable, which captures different ways of saying the same thing. The specific choice a speaker make varies systematically based on information such as the speaker's identity, their relationship to interlocutors, sociopragmatic context, among many other factors (Tagliamonte, 2006). Through variation, language conveys *social meaning* (Nguyen et al., 2021).

There is a rich body of work that aims to analyze linguistic variation computationally. Often, the focus is on lexical variation (Bamman et al., 2014; Zhang et al., 2017; Zhu and Jurgens, 2021a); semantic variation in online communities (Lucy and Bamman, 2021; Del Tredici and Fernández, 2018); or orthographic variation in online text (Eisenstein, 2015; Stewart et al., 2017). In our view of memes as language, we ask the same kind of question:

RQ1: Does the template choice within a semantic cluster vary systematically between communities?

Methods. The semantic clusters form our variable context, and set of templates within any given semantic cluster form a discrete set of choices with the same semantic value. We use the weighted log odds-ratio to compute the extent to which a template is specific to a given subreddit compared to all other subreddits, relative to the other templates in a semantic cluster (Monroe et al., 2017; Jurafsky et al., 2014). We find the templates that have a statistically significant association with a subreddit (z-score > 1.96); the semantic clusters these templates belong to are *in variation*: a community prefers one variant over the others in this cluster.

Results. We find 94 out of 784 semantic clusters exhibit statistically significant variation, span-



(a) Declarative

(b) Scalar increase

(c) Comparison

Figure 5: Subreddits exhibit variation in the preferred templates within a semantic cluster. All are statistically significantly overrepresented in their respective subreddits, $p < 0.05$.

ning 391 different templates. Figure 5 shows how functionally similar memes take different forms in different communities.

Speakers use language to construct their social identities (Bucholtz and Hall, 2005). We find that, not only do subreddits prefer certain variants of a template over others, but they choose templates that index into a localized cultural knowledge, making cultural allusions to characters or celebrities.

For example, the orange Drake template (fig. 5c, left) is used frequently in general purpose meme subreddits like `r/memes`, but alternatives are used in other subreddits. One version that is specific to `r/dndmemes` (which discusses the role playing game Dungeons and Dragons) replaces Drake with Matthew Mercer, a voice actor who stars in a popular Dungeons and Dragons web series (fig. 5c, middle).

Linguistic variants usually become associated with identities through a gradual process in which the association slowly permeates public awareness (Eckert, 2008). In general, a phonological variable does not inherently index any given identity. However, the multimodality of memes permits greater expressiveness—a meme in `r/Animemes` might use the anime art style, indexing into the aesthetic of that community explicitly.

6 Linguistic innovation

Equally as important as the study of synchronic linguistic variation is the study of diachronic linguistic change. Language change has been heavily studied in natural language processing (Rosenfeld and Erk, 2018; Martinc et al., 2020; Zhu and Jurgens, 2021b). We focus on understanding the innovation of meme templates within a semantic cluster.

RQ2: Do new meme templates co-exist with pre-existing templates in the semantic cluster, or does the most popular template monopolize the cluster?

When multiple templates that fulfill the same function appear, we expect there to be competition. Prior work has observed this competition between lexical choices, with two outcomes: new words replace old ones that serve the same function, but if similar words have discourse-relevant differences in meaning, they can coexist (Karjus et al., 2020).

Methods We measure the entropy of semantic clusters over time. If meme templates ultimately co-exist, we would expect entropy to increase; if a subset of templates dominate, we would expect the entropy to converge to a lower value.

For each semantic cluster, we group posts by the age of the semantic cluster in years at the time of posting. We define the "birth" of the cluster as when the first instance of a template in that cluster was posted. Within each year, we calculate the entropy of template distribution within each cluster.

It is possible that some clusters have low entropy early on due to data sparsity. To account for this, we filter to semantic clusters that have existed at least 5 years with at least 30 posts in all years, and resample with replacement within each year such

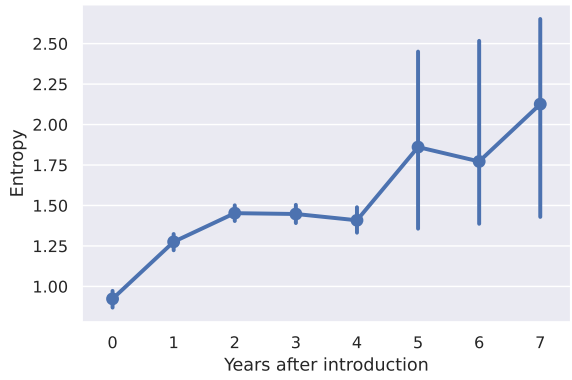


Figure 6: On average, semantic clusters diversify over time. Very old semantic clusters are rarer, leading to larger confidence intervals in later years.

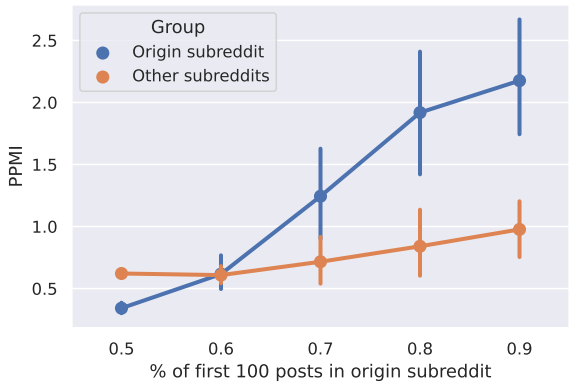


Figure 7: Communities that lead the introduction of a new template continue to use it more than others.

that every year has the same number of posts. Ultimately, we conduct our analysis over 146 semantic clusters that span over 950K posts.

Results Entropy steadily increases in the years following a semantic cluster's initial introduction (Figure 6). This suggests that no one meme template grows to become the de facto template for all users; there is steady variation. This is supported by our findings in Section 5 that there are socially meaningful differences between variants.

RQ3: Do new templates diffuse widely or occupy a niche?

Language change is often socially motivated; a community can opt to use a particular variant to distinguish themselves from others (Trudgill, 1986; Giles and Powesland, 1975). Thus, we might expect meme templates to be most specific to the subreddits in which they were first introduced.

Methods We measure the extent to which template variants are ultimately specific to the subreddits that originated them.

We filter our dataset to templates which occur at least 200 times. For each template, we identify a set of "seed posts", which we define as the first 100 posts using the template. We then filter to templates with a subreddit that comprises the majority of the seed posts, which we call the "origin subreddit."

We modify the method from (Zhang et al., 2017) to measure the specificity of a template-subreddit pair by using the positive pointwise mutual information (PPMI) between templates and the subreddits in which they are used, matching other work in NLP (Church and Hanks, 1990; Jurafsky and Martin, 2009). Formally, we calculate

$$\text{PPMI}(t; s \mid c) = \max\left(\log \frac{P(t \mid s, c)}{P(t \mid c)}, 0\right),$$

where $P(t \mid s, c)$ is the probability of template t appearing in subreddit s and semantic cluster c, $P(t \mid c)$ is the probability of template t in that cluster globally, and templates are only compared against others within the same semantic cluster. We calculate the PPMI over non-seed posts to measure the specificity of a template after its introduction.

Results For each template, we compare the PPMI for origin subreddits with the average PPMI of all other subreddits. Figure 7 shows a significant positive correlation between the proportion of seed posts that originated in the origin subreddit and the eventual specificity of template. These results support previous findings that lexical innovations succeed when filling in a social niche (Altmann et al., 2011; MacWhinney, 1989).

7 Linguistic acculturation

Finally, we study how users alter their meme posting habits as they spend more time in a subreddit. Previous work on linguistic acculturation show that users adopt more community-specific language as they become enculturated within a community (Danescu-Niculescu-Mizil et al., 2013; Srivastava et al., 2018). We can ask a similar question here:

RQ4: Do veteran users in a subreddit use more community-specific templates?

Methods To answer this question, we measure the average specificity of a user's posts in successive months after they enter a community. We once

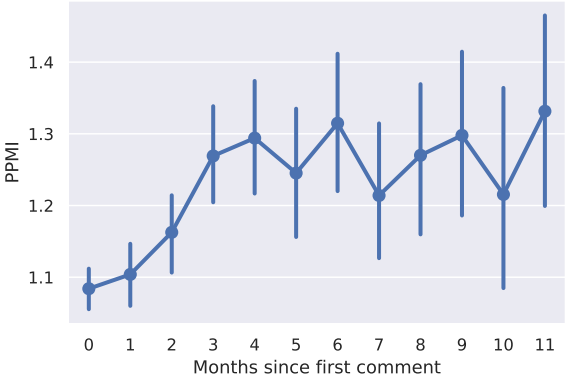


Figure 8: Veteran (acculturated) users employ more subreddit-specific meme templates.

again calculate the PPMI of templates as a measure of specificity; this time, we calculate the value over the full range of the dataset.

For each user in a subreddit, we bin their posts by 30-day windows starting with their first post in the subreddit (i.e., for each month after they joined), and compute the average PPMI of their posts for that time period. We filter the dataset to users with at least 10 lifetime posts and subreddits with at least 30 such users. To prevent extremely popular subreddits from unduly influencing the results, we sample up to 100 users from each subreddit to compute the average across all subreddits. This yields a total of 3,174 users in 130 subreddits.

Results We find that acculturated users use templates that are more specific to the communities in which they post (Pearson's $r = 0.074, p < 0.001$), shown in Figure 8. Our finding aligns with existing literature on linguistic acculturation as well as theories in new media that memes are cultural capital. The "correct" use of memes can demonstrate a user's assimilation into a shared language and identity (Nissenbaum and Shifman, 2017).

8 Related work

Prior work on memes in NLP and social computing has largely focused on two tasks: meme understanding and modeling how memes originate and spread. Our work offers novel methods and perspectives at the intersection of these areas of research.

Meme understanding encompasses a number of discrete tasks, including classifying if memes convey harmful messages (Kiela et al., 2021; Qu et al., 2022), labeling emotion (Mishra et al., 2023), and detecting humor (Tanaka et al., 2022) or figurative speech (Liu et al., 2022) within them. While these can generally be framed as classification tasks, other work generates open-ended explanations of visual humor using large multimodal language models (Hwang and Shwartz, 2023; Hessel et al., 2023). Our work complements this existing body of research by inferring semantic variables in an unsupervised approach, leveraging the implicit structure within memes by modeling modeling template semantics separately from the fills.

In modeling the internal structure of memes, our work draws on existing research examining the relationship between fills and templates to match semantic roles to entities within harmful memes (Sharma et al., 2023a) and mapping fill text to explanatory background information (Sharma et al., 2023b). We hope that our method of construing templates as semantic predicates can contribute to this body of work.

In the social computing space, another line of research focuses on understanding how memes originate (Morina and Bernstein, 2022) and spread across platforms (Zannettou et al., 2018). These treat meme templates as discrete tokens. We model template semantics, which have the granularity to enable analysis of variation and social meaning. Qu et al. uses CLIP to understand how memes evolve as they spread. While they use the text in comments to model the high-level concepts indexed by particular variants, we use the fill text of memes to model low-level template semantics.

9 Conclusion

In this paper, we analyze memes as a form of language subject to the same kinds of sociolinguistic variation as other modalities, such as written language and speech. We propose a new approach to understanding meme semantics, taking advantage of the multimodal structure of memes to learn semantic representations of templates from an unlabeled dataset. We use this method on a large dataset of memes scraped from Reddit, and demonstrate that it yields coherent, visually diverse clusters of semantically similar memes. We make these clusters and the code publicly available for future research. Finally, we use these clusters to study language variation and change in subreddits. We show that variations between meme template are socially meaningful and memes often share usage patterns with the textual language that has been studied in the past. We find that memes can be rich resources for understanding social language use.

10 Ethical considerations

The data used in this work was collected from Reddit in 2021 and is publicly available. To preserve the right to be forgotten, we release only the post IDs of the posts we used and the labels from the semantic clustering process. There may be offensive, hateful, or sexual messages present in the memes and comments in this dataset.

The models we trained are also publicly available. We use them only to better understand the semantics of memes. We do not train any generative models, and warn against training generative models on the data without careful consideration of how to mitigate the toxic, offensive, or otherwise harmful outputs that might be generated.

11 Limitations

We note several limitations to this work. First, we only study memes posted to Reddit meme communities, which are topically-focused and primarily English-speaking. One should be cautious in extrapolating these results to other settings in which memes are used. However, our data pipeline and models are platform agnostic—the semantic clusters can be generated from any set of memes. By making our code and models public, we hope to encourage other researchers to replicate and extend our analysis to other settings.

The meme clustering pipeline is also imperfect. As we note in the paper, the visual clustering is overly aggressive, resulting in the same base image being split into multiple template clusters. Although we show that the semantic clustering step mostly addresses this issue, improving the visual clustering could yield more precise analysis. Additionally, there are edge cases where a small visual modification changes the semantic meaning (e.g., sometimes the order of panels is reversed). The data pipeline does not always identify these visual differences.

Finally, the time series analyses are limited by data sparsity in earlier years—this is due in part to a smaller Reddit user-base, but also because many images have been deleted or removed since they were first posted. Though it is unlikely that this natural decay is systematic in a way that would significantly bias our estimates, it nonetheless reduces the precision of our analysis.

References

Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. 2011. Niche as a Determinant of Word Fate in Online Groups. *PLOS ONE*, 6(5):e19009.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *ArXiv*, abs/2001.08435.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7(4-5):585–614.

Kathryn Campbell-Kibler. 2009. The nature of sociolinguistic perception. *Language Variation and Change*, 21(1):135–156.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 307–318, New York, NY, USA. Association for Computing Machinery.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Marco Del Tredici and Raquel Fernández. 2018. Semantic Variation in Online Communities of Practice. *arXiv:1806.05847 [cs]*.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476.

Jacob Eisenstein. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2):161–188.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA. Association for Computational Linguistics.

Howard Giles and Peter F. Powesland. 1975. Accommodation Theory. In *Speech Style and Social Evaluation*, Speech Style and Social Evaluation, pages 232–239. Academic Press, Oxford, England.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A Dataset for Captioning and Interpreting Memes. ArXiv:2305.13703 [cs].

Dan Jurafsky, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.

Andres Karjus, Richard A. Blythe, Simon Kirby, and Kenny Smith. 2020. Communicative need modulates competition in language change.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. ArXiv:2005.04790 [cs].

Gunther R. Kress and Theo Van Leeuwen. 2001. *Multimodal Discourse: The Modes and Media of Contemporary Communication*. Arnold; Oxford University Press.

William Labov. 1963. The Social Motivation of a Sound Change. WORD, 19(3):273–309.

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022. FigMemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Li Lucy and David Bamman. 2021. Characterizing English variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556.

Brian MacWhinney. 1989. Competition and lexical categorization. In Roberta Corrigan, Fred R. Eckman, and Michael Noonan, editors, *Current Issues in Linguistic Theory*, volume 61, page 195. John Benjamins Publishing Company, Amsterdam.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Shreyash Mishra, S. Suryavardan, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Memotion 3: Dataset on Sentiment and Emotion Analysis of Codemixed Hindi-English Memes. ArXiv:2303.09892 [cs].

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2017. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, 16(4):372–403.

Durim Morina and Michael S. Bernstein. 2022. A web-scale analysis of the community origins of image memes. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1).

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612, Online. Association for Computational Linguistics.

Asaf Nissenbaum and Limor Shifman. 2017. Internet memes as contested cultural capital: The case of 4chan's /b/ board. *New Media & Society*, 19(4):483–501.

Pamela Perniss. 2018. Why We Should Study Multimodal Language. *Frontiers in Psychology*, 9.

Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. 2022. DisinfoMeme: A Multimodal Dataset for Detecting Meme Intentionally Spreading Out Disinformation.

Yiting Qu, Xinlei He, Shannon Pierson, Michael Backes, Yang Zhang, and Savvas Zannettou. 2023. On the Evolution of (Hateful) Memes by Means of Multimodal Contrastive Learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 293–310. IEEE Computer Society.

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.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484, New Orleans, Louisiana. Association for Computational Linguistics.

Shivam Sharma, Atharva Kulkarni, Tharun Suresh, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023a. Characterizing the entities in harmful memes: Who is the hero, the villain, the victim? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2149–2163, Dubrovnik, Croatia. Association for Computational Linguistics.

Shivam Sharma, Ramaneswaran S, Udit Arora, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023b. MEMEX: Detecting explanatory evidence for memes via knowledge-enriched contextualization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5272–5290, Toronto, Canada. Association for Computational Linguistics.

Sameer B. Srivastava, Amir Goldberg, V. Govind Manian, and Christopher Potts. 2018. Enculturation Trajectories: Language, Cultural Adaptation, and Individual Outcomes in Organizations. *Management Science*, 64(3):1348–1364.

Ian Stewart, Stevie Chancellor, Munmun De Choudhury, and Jacob Eisenstein. 2017. #Anorexia, #anarexia, #anarexyia: Characterizing online community practices with orthographic variation. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4353–4361.

S.A. Tagliamonte. 2006. *Analysing Sociolinguistic Variation*. Key Topics in Sociolinguistics. Cambridge University Press.

Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. 2022. Learning to evaluate humor in memes based on the incongruity theory. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93, Gyeongju, Republic of Korea. Association for Computational Linguistics.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

V. A. Traag, L. Waltman, and N. J. van Eck. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233. Number: 1 Publisher: Nature Publishing Group.

Peter Trudgill. 1986. *Dialects in Contact*. Number 10 in Language in Society. B. Blackwell, Oxford, UK ; New York, NY, USA.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, page 188–202, New York, NY, USA. Association for Computing Machinery.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. 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.

Justine Zhang, William L. Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community Identity and User Engagement in a Multi-Community Landscape. ArXiv:1705.09665 [physics].

Qing Zhang. 2005. A Chinese yuppie in Beijing: Phonological variation and the construction of a new professional identity. *Language in Society*, 34(3):431–466.

Jian Zhu and David Jurgens. 2021a. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 279–297.

Jian Zhu and David Jurgens. 2021b. The structure of online social networks modulates the rate of lexical change. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2201–2218.

A Details on visually clustering templates

A.1 Preprocessing

One common meme layout that would caused issues in the template clustering step was a text frame around the source image, where there is a border around a source image, as well as some text above or below (see Figure 9a for an example).

For each image, we use a rectangular kernel to detect potential text patches, replace those patches with the background color, and finally identify the bounding box for the remaining source image without any excess borders. Figure 9 walks through the steps visually.

A.2 Image hashing

We create a 64-bit perceptual hash for each preprocessed image in the dataset. The preprocessed images are only used for the hashing step; all other steps use the original image. Figure 10 shows examples of images whose preprocessed versions have the same hash.

A.3 Hash clustering

We first compute pairwise Hamming distance between all the hashes. Then, we discard any pairs with a Hamming distance greater than 10. Then we construct a network of hashes, where edges of the graph are calculated as $11 - d_{ij}$ for Hamming distance d_{ij} between the ith and jth hashes before finally using the Leiden algorithm to cluster hashes. Figure 11 shows the top 18 most heavily represented hash clusters, with 4 sampled images from each.

We use the Leiden algorithm with the Constant Potts Model (CPM) as the quality function; we use a density of 1.0, but experiments with other density values (0.01, 0.1, 10) yielded qualitatively similar or worse results. The algorithm results in aggressively split clusters, where each cluster is coherent, but there are some memes that share a base template but are split between two clusters (e.g. Winnie the Pooh appears twice in Figure 11, among others).

We find that these duplicate hash clusters are often merged when we generate the semantic clusters.

B Details on the semantic clusters

B.1 Training details

The RoBERTa and CLIP models were fine-tuned using the Huggingface transformers library (Wolf et al., 2020). The RoBERTa model was fine-tuned on an NVIDIA A5000 GPU for 3 epochs (27 hours) with a learning rate of 1e-6. The CLIP model was fine-tuned with mixed-precision on 8 NVIDIA A6000 GPUs for 3 epochs (27 hours) with a learning rate of 4e-6. There was no hyperparameter tuning.

B.2 Calculating edge weights.

To prepare the data for clustering, we constructed a weighted adjacency matrix by keeping only the top 10 nearest neighbors for each template embedding. We calculated the weight as

$$w_a(b) = \lambda^{r_a(b)},$$

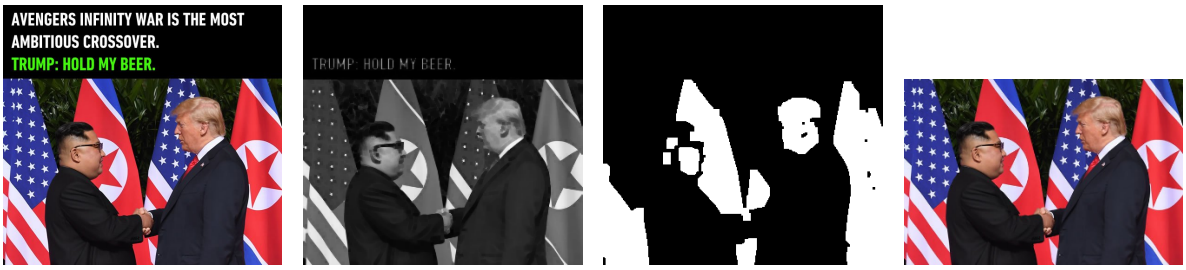
where $w_a(b)$ is the weight of the edge between templates b and a, $r_a(b)$ is the rank of the cosine similarity between a and b, and λ is a discount factor (we set this to 0.9).

We chose to weight by a function of ranked similarity instead of cosine similarity directly because we found the cosine similarity was often low even for the embeddings of semantically equivalent templates, resulting poor recall and many small clusters. Using the weighted ranking, we get more templates per cluster without introducing too many false positives.

B.3 Outputs

There are 784 semantic clusters generated from the RoBERTa embeddings—table 2 shows statistics for cluster sizes for all of the models. We include more extensive examples of the clusters in appendix D.

The distribution of clusters is highly skewed. Figure 12 shows the distribution of cluster sizes in our Reddit dataset, for the RoBERTa embeddings. The largest 103 clusters account for 50% of the posts the dataset. The largest 10 account for 12%.



(a) Original image (b) Remove text (c) Remove text artifacts (d) Trim excess borders

Figure 9: Example of the image preprocessing steps, described in Section 2.1

(a) 8763e2636178d897

(b) 9465a9596e1a66bc

Figure 10: Examples of groups of meme instances with the same perceptual hash when processed.

Figure 11: Sample images from the 18 most common perceptual hash clusters; each row contains two clusters with four sampled images.

Model	# Clusters	Avg. Size
RoBERTa	784	8.7
CLIP	657	10.4
CLIP-diff	617	11.1
Concat.	685	10.0

Table 2: Count and average sizes of semantic clusters generated from each embedding model.

C Model evaluation details

Human annotators were presented with pairs of images with the following instructions:

> You will be looking at pairs of memes; for each pair, you will be answering two yes/no questions.
>
> 1. is this pair semantically similar (can you conceivably copy / paste the text of one into the other with minor changes and have it still make sense)
> 2. is this pair visually similar (do they have the same characters, art style, etc? e.g. two harry potter memes. If the layout is the same but the charac-

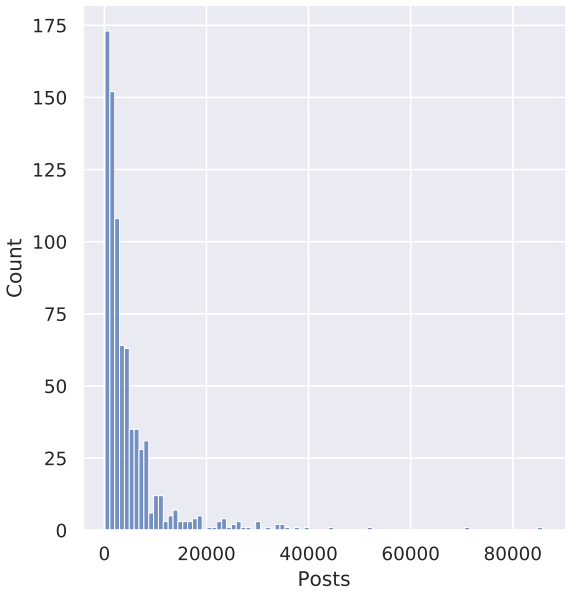


Figure 12: Distribution of semantic cluster coverage

ters are different, you should mark it as not visually similar)

The annotators were three students employees whose job responsibilities included such annnotations. They were recruited at a large public university in the United States and paid above the minimum wage for the geographic region they're in. They were informed that their annotations would be used to evaluate different model outputs.

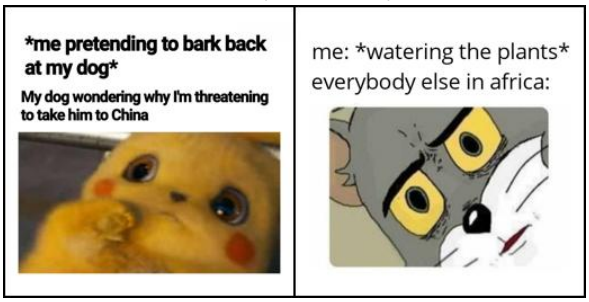
Figure 13 includes examples of some of the image pairs presented to annotators, as well as the expected judgments for semantic and visual similarity.

We calculate metrics on the model judgments over the set of all annotated pairs (across models). This not only allows us to evaluate on a larger set of annotations, but also helps highlight differences between models.

We found that CLIP-based methods would include clusters that were visually similar but semantically different. Figure 14 shows an example of a cluster generated with CLIP embeddings that contains meme templates that have different semantic functions, but almost all contain characters from the *Star Wars* franchise. This cluster had an visually adjusted precision score of -2.08.

The failure case of over-indexing on visual similarity is not restricted to creating semantically incoherent clusters with visual similarity. The CLIP-diff and, to a greater extent, Concat embeddings were good at surfacing less-used variants of tem-

(a) Semantically and visually similar



(b) Semantically similar, visually different

(c) Semantically different, visually similar

(d) Semantically and visually different

Figure 13: Example meme pairs for annotation

plates. However, they do so at the expense of splitting into several stylistically delineated semantic clusters, which is detrimental to our desired analysis on variation. Figure 15 shows how templates from the large RoBERTa cluster for declarative templates are divided into several Concat clusters.

D Further semantic cluster examples

Figures 16,17,18,19 contain more examples of semantic clusters generated using the different embedding models. Content warning: though we tried to filter for toxic speech, these memes may still contain offensive content.

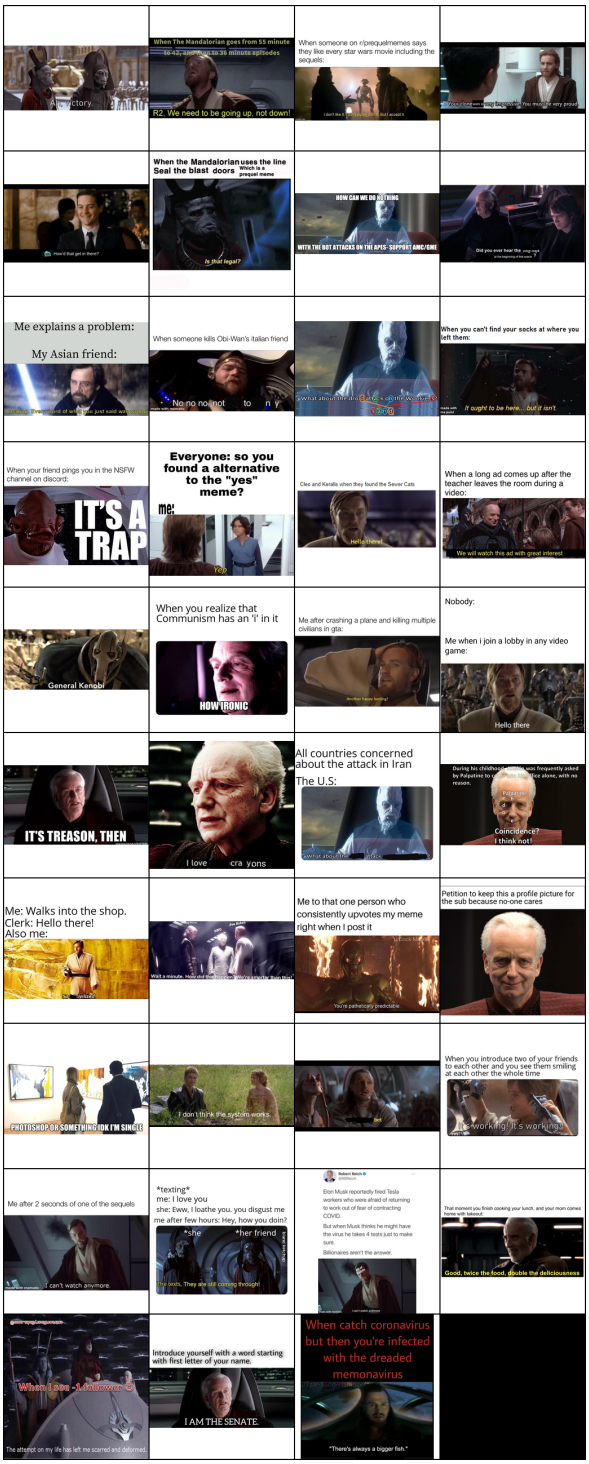
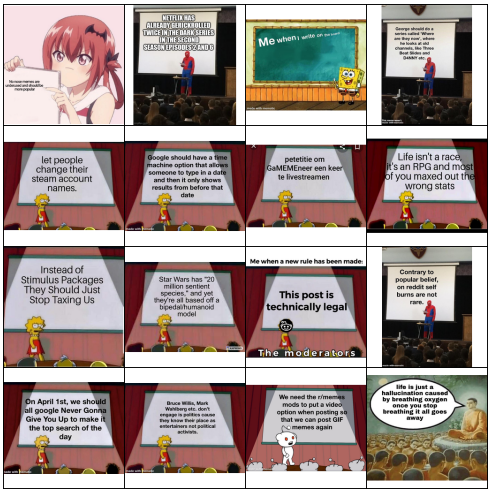
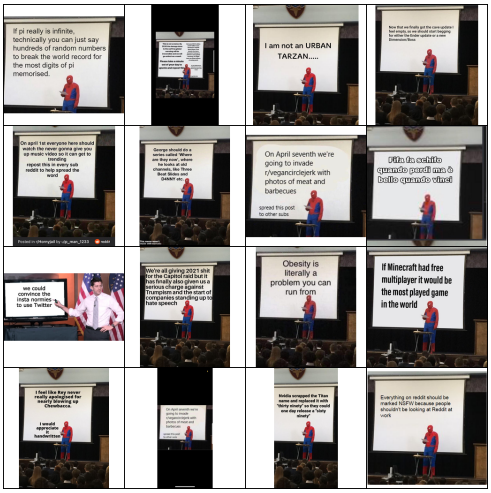


Figure 14: This Concat cluster has a low visual-adjusted precision.



(a) RoBERTa Cluster 0

(b) Concat Cluster 2



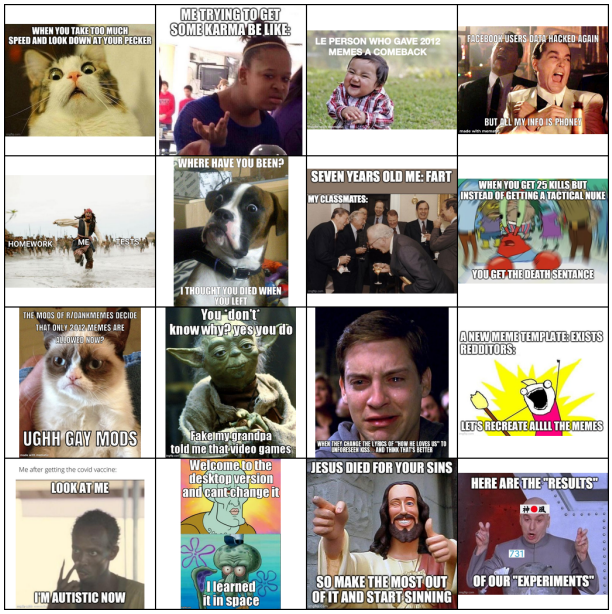
(c) Concat Cluster 31

(d) Concat Cluster 134

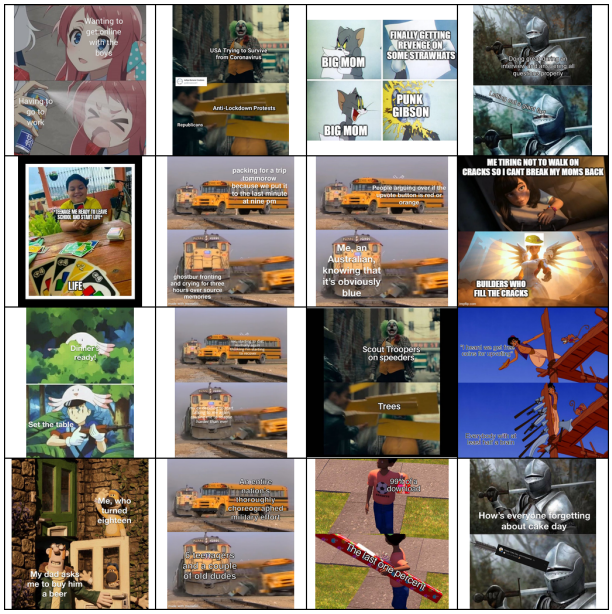
Figure 15: Templates that are clustered together by RoBERTa appear in stylistically delineated semantic clusters in the Concat clusters. Displayed are a sample of up to 16 templates from each cluster.

(a) Cluster 0

(b) Cluster 1



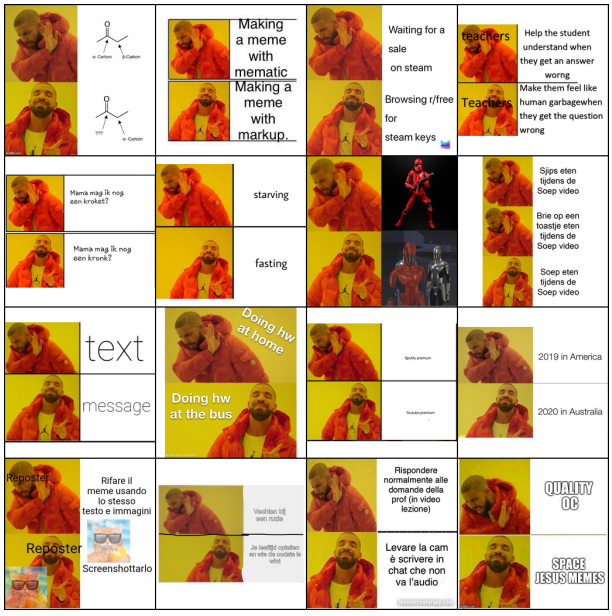
(c) Cluster 10



(d) Cluster 30

Figure 16: Samples from RoBERTa clusters

(a) Cluster 0

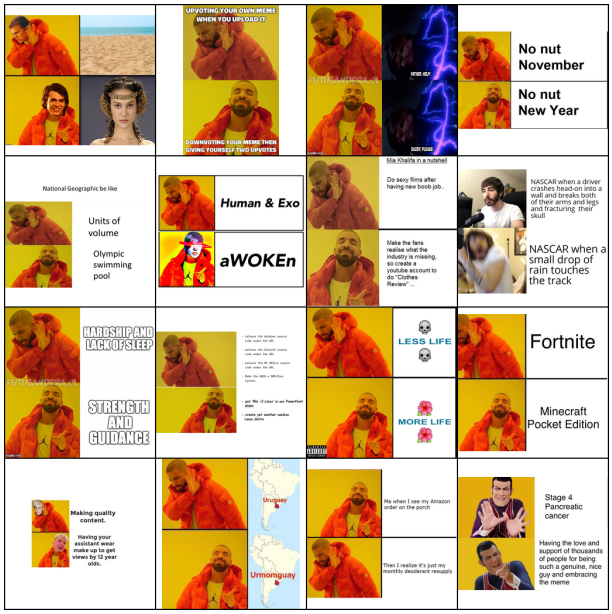


(b) Cluster 1

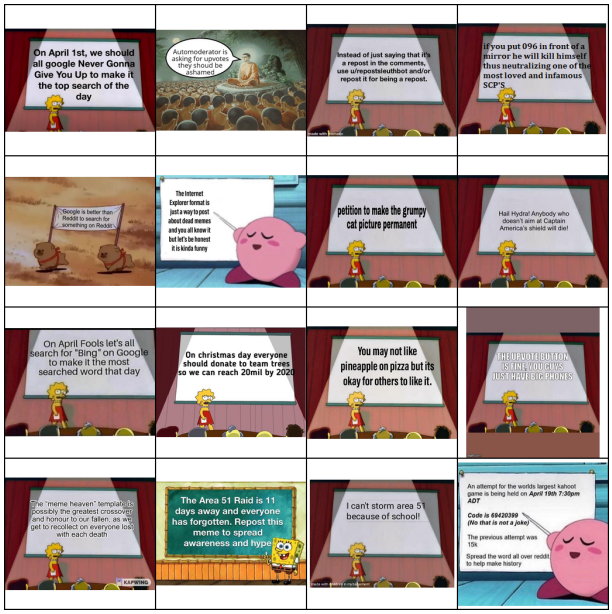
(c) Cluster 36

(d) Cluster 23

Figure 17: Samples from CLIP clusters

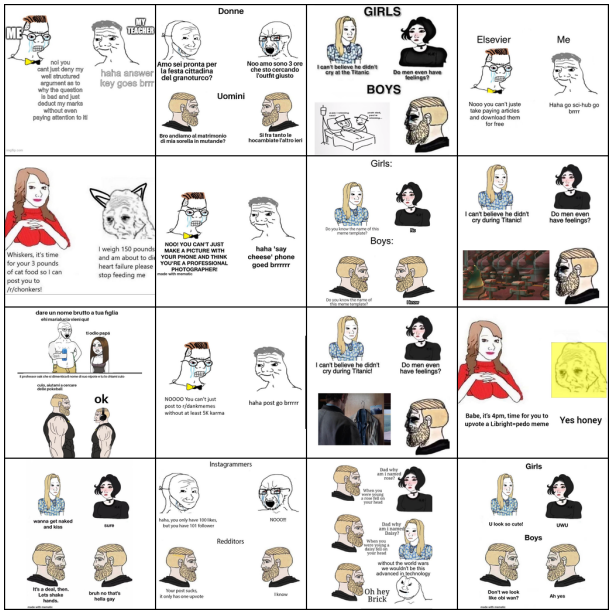


(a) Cluster 0



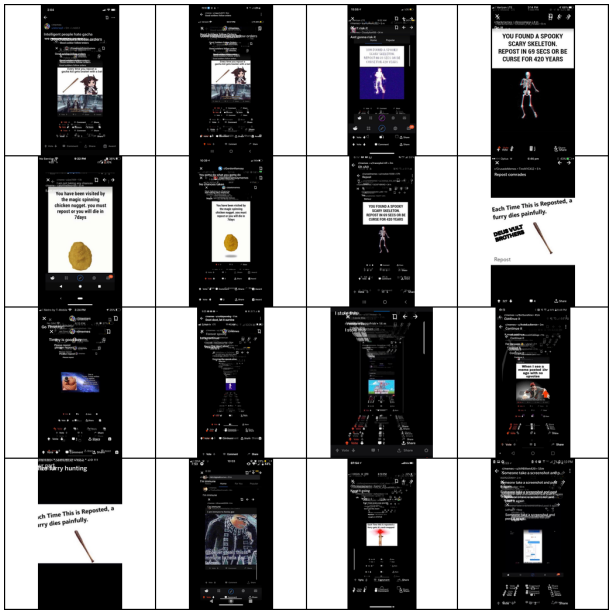
(b) Cluster 1

(c) 32

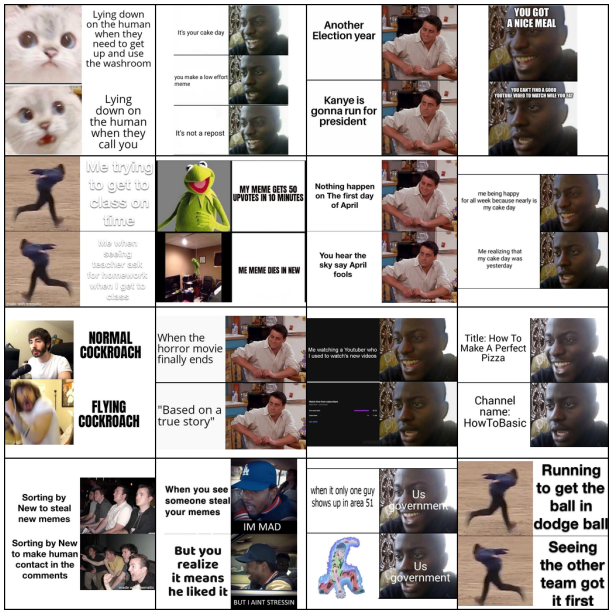


(d) Cluster 22

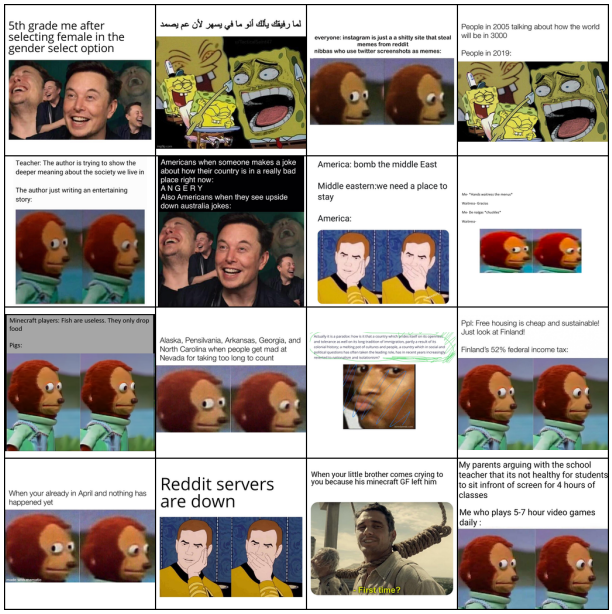
Figure 18: Samples from CLIP-diff clusters



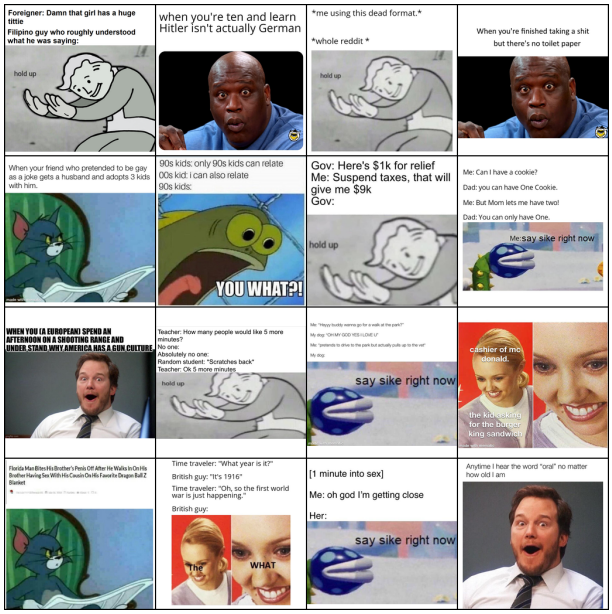
(a) Cluster 0



(b) Cluster 1



(c) Cluster 15



(d) Cluster 29

Figure 19: Samples from CLIP-diff + RoBERTa clusters