Understanding Visual Memes: an Empirical Analysis of Text Superimposed on Memes Shared on Twitter

Yuhao Du, Muhammad Aamir Masood, Kenneth Joseph
Computer Science and Engineering, University at Buffalo
Buffalo NY, 14260
{yuhaodu,mmasood,kjoseph}@buffalo.edu

Abstract

Visual memes have become an important mechanism through which ideologically potent and hateful content spreads on today's social media platforms. At the same time, they are also a mechanism through which we convey much more mundane things, like pictures of cats with strange accents. Little is known, however, about the relative percentage of visual memes shared by real people that fall into these, or other, thematic categories. The present work focuses on visual memes that contain superimposed text. We carry out the first large-scale study on the themes contained in the text of these memes, which we refer to as *image-with-text* memes. We find that 30% of the image-with-text memes in our sample which have identifiable themes are politically relevant, and that these politically relevant memes are shared more often by Democrats than Republicans. We also find disparities in who expresses themselves via image-with-text memes, and images in general, versus other forms of expression on Twitter. The fact that some individuals use images with text to express themselves, instead of sending a plain text tweet, suggests potential consequences for the representativeness of analyses that ignore text contained in images.

Introduction

Knowledge of how information is shared and spread online has long been rooted in the study of large text corpora (Diesner, Frantz, and Carley 2005; Leskovec, Backstrom, and Kleinberg 2009; Hu and Liu 2012; Nguyen et al. 2019). However, information sharing online is increasingly multimodal, with users and platforms turning to other means of expression beyond text. These new forms of communication are both anecdotally and empirically consequential. Anecdotally, political actors are finding ways to use these forms of communication to express policy. On November 2, 2018, for example, President Donald Trump tweeted an image of himself with the text "Sanctions are coming," a play on *Game of Thrones*' iconic "Winter is coming" catch phrase and in reference to sanctions against Iran. To date, the tweet has attracted around 63,000 retweets and 195,000 favorites. Empirically, scholars have shown that mediums beyond text, in particular images and videos, have been used to promote extremism (Finkelstein et al. 2018), as new ways of expressing the self (Liu et al. 2016), and as a vehicle for the spread of misinformation (Gupta et al. 2013).

One particularly important form of expression beyond text is the *visual meme* (Xie et al. 2011), examples of which are shown in Figure 1. Visual memes have long been characterized in popular culture as a vehicle for humor (e.g., Grumpy Cat). However, recent work has shown that they are also a mechanism through which less trivial content, including misinformation and hate speech, is spread (Zannettou et al. 2018). Visual memes are also interesting in that many of them express text in ways that are not captured with standard data collection approaches. While this text is sometimes used as a supplement to other visual content in the image, there are many cases in which the other visual content is either a stock photograph meant only to supplement the emotional valence of the text (e.g. the left-most image in Figure 1) or a background upon which text is written (e.g. the right-most image in Figure 1). Thus, visual memes can therefore be viewed as an alternative way of sharing text.

These visual memes with text content—referred to here as *Image-with-text (IWT) memes*— thus occupy a unique middle ground between visual memes and more traditional forms of text data. The present work provides the first large-scale analysis of the text contained in IWT memes shared on social media. We do so using a dataset of over 7 million IWT memes shared by approximately 200,000 Twitter users for whom demographic data is available to us. With this data, we pose and answer three research questions. First, who opts to share text through IWT memes, relative to other forms of expression on Twitter—specifically, as compared to non-IWT memes, to images in general, and to all other forms of sharing? Second, what are the broad themes of the text contained in IWT memes shared on Twitter? Finally, how do Twitter users' demographics correlate with the themes of the IWT memes they share?

In order to address these research questions, a more fundamental question must first be answered—what, exactly, constitutes an IWT meme? To this end, we propose a more complete definition below and then develop and evaluate a pipeline to extract IWT memes from other kinds of images shared on Twitter. In sum, the present work makes the following major contributions:

- We carry out the first large-scale study of the textual content of visual memes shared on Twitter. Motivated by previous work, we engage in hand-coding to better understand the extent to which IWT memes are intended to be humorous, political, hateful, or to spread misinformation. We then use a topic model to identify the broad topics contained within the text of IWT memes. Amongst other observations, we find that almost 30% of memes that contain identifiable themes are political in nature.
- We find that certain demographics are more likely to express themselves via IWT memes, relative to memes without text, images in general, and text-only tweets. This observation suggests future work on content shared on Twitter may need to focus on multi-modal data to ensure it does not exclude critical voices.
- We develop a pipeline to identify IWT memes, and make it publicly available for others.[1]

Defining IWT Memes

Prior large-scale empirical work has operationalized the concept of a visual meme by relying on existing data sources, e.g., KnowYourMeme[2] (Zannettou et al. 2018; Coscia 2013). In these articles, anything within or similar to images in these databases is considered to be a meme, and anything outside is not. However, an initial analysis of the images containing text in our dataset suggested that many of them, including the right-most image in Figure 1,[3] were important and prevalent in our data but not found (or found to be similar to) images within existing meme databases. Consequentially, we take a different approach to identifying memes of interest. We define a class of memes we call IWT memes, collect annotated data for this definition, and construct and evaluate a classifier to extract these IWT memes from a larger dataset of images shared online.

Figure 1 gives three examples of IWT memes. These images fit two main requirements. First, we require that the text displayed is vital to understanding the image, i.e. someone could not understand the intention of the image without reading the superimposed text. The left-hand image in Figure 2, while a well known meme, is thus not an IWT meme. While tweets sharing this image have text associated with them, this text is not contained in the tweet body and therefore easily extracted using common data collection approaches. It is thus not of interest to the present work, which focuses on what text is shared *within* images.

Second, we require that the images are memes. We are interested in memes specifically, as opposed to all images with superimposed text, because of their potential for virality and importance in Internet culture. However, the definition of a meme is a contested topic; see, for example, the review work of (Díaz and Mauricio 2013) or the careful unpacking of Dawkins's (1976) original definition by Blackmore (2000; p. 4-8). In the present work, we focus on a more restricted definition of meme, having two conditions deriving from Díaz and Mauricio (2013). First, an image is only a meme if it is reasonable that the image could be *spread virally*, i.e., beyond a small, pre-ordained group. Thus, for example, images sharing information about a local community event, prominent in our data, were not considered memes. Second, an image is only a meme if the *structure or content of the image could reasonably be imitated, altered, and re-shared*. Thus, for example, pictures of specific items, like objects for sale, while perhaps intended to go viral, would not be considered memes. As the conditions under which images are considered IWT memes are clearly subjective, we engage in an extended annotation task described below.

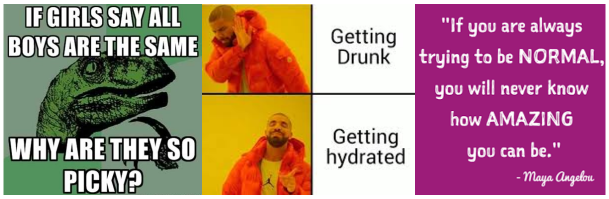
IF GIRLS SAY ALL BOYS ARE THE SAME
WHY ARE THEY SO PICKY?
Getting Drunk
Getting hydrated
"If you are always trying to be NORMAL, you will never know how AMAZING you can be."
- Maya Angelou

Figure 1: Samples of IWT memes

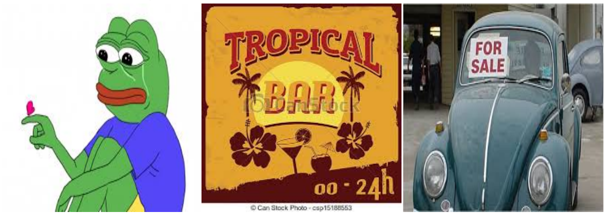

Figure 2: Samples of images that are not IWT memes

The IWT memes studied in this paper are therefore both a superset and a subset of what has been classified as a meme in prior work, most notably the work from Zannettou et al. (2018). IWT memes are a subset of those studied in previous work, most obviously, in that prior work considers memes that both do and do not contain text. However, in the data we study, we find that memes without superimposed text account for only around 1% of all images without text (approximately 300K out of 30M) and are 20 times less prevalent than IWT memes. The prevalence of IWT memes, relative to non-IWT memes, is in turn a natural function of our focus on a broader definition of what constitutes a meme. Thus, our analyses do not necessarily extend, but rather compliment the prior work in critical ways.

Literature Review

Our work relates to research on image content analysis and the study of other closely related forms of online memes. We review related work in these areas here.

Image Content Analysis

Semantic information contained in images shared online has attracted the attention of computer vision scholars and com-

[1] https://github.com/yuhaodu/TwitterMeme

[2] https://knowyourmeme.com/

[3] Note that these images, like all others here, are either extremely popular in our dataset or not from our data at all, but very similar to it, in order to protect privacy.

putational social scientists. Current studies have focused on two areas: improving computer vision techniques to allow machines to understand high-level information embedded in images, and using vision methods to understand sharing patterns of images posted on social media. We review the latter here, as it is more relevant to the present work.

Several studies have analyzed the content of images shared online using tools from computer vision. You et al. (2014) extracted features from images shared on Pinterest and demonstrated that these features could be used to predict users' gender. Hu, Manikonda, and Kambhampati (2014) used the Scale Invariant Feature Transform (SIFT) (Lowe 1999) to extract features from images, and discovered 8 types of popular image categories and 5 types of users on Instagram. Liu et al. (2016) used the features extracted from profile images of Twitter users to predict personality traits. And Garimella, Alfayad, and Weber (2016) used image recognition tools from Imagga.com [4] to get tags for images on Instagram. With these tags, they then predicted health outcomes in 100 U.S. counties.

This emerging research shows that images, as a prevalent communication tool on the social web, carry a significant amount of information relevant to a variety of research questions. Following the success of computer vision methods and their application in this area, we consider memes with superimposed text.

Analyses of Memes

Early quantitative work on text data studied short meme phrases and their variants in news article and blog posts, showing the interplay between news media and blogs (Leskovec, Backstrom, and Kleinberg 2009). Video memes have also been leveraged to track and monitor real-world events on Youtube (Xie et al. 2011). Bauckhage (2011) studied the temporal dynamics and infectious properties of several famous memes and conjectured that majority of them spread through homogeneous communities. JafariAsbagh et al. (2014) proposed a streaming pipeline for detection and clustering of memes in tweets and showed that clusters of memes can expose trending events.

With respect to image-based memes, Bharati et al. (2018) proposed a pipeline to detect the history of meme transformation. Zannettou et al. (2018) proposed a pipeline to use unsupervised methods to cluster images from social media and to then use data from Know Your Meme (OMS) to identify which of the clusters contained memes. Using a Hawkes process, they then modeled the dissemination of memes between several online communities, showing that fringe communities like 4chan's Political Incorrect board (/pol/) are influential meme disseminators. In the present work, we use their pipeline to identify non-IWT memes; i.e. memes that do not contain superimposed text. Beskow, Kumar, and Carley (2020) proposed a similar pipeline to classify online political memes and used graph learning to create an evolutionary tree of memes. Finally, Dubey et al. (2018) extracted features of memes using a pretrained deep neural network and optical character recognition. They then demonstrated that the extracted features can help predict meme virality.

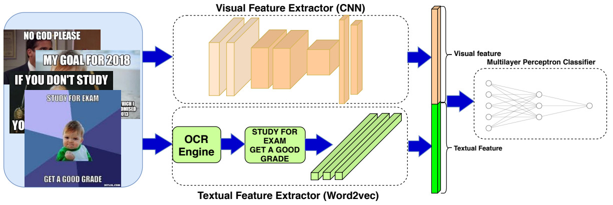
Visual Feature Extractor (CNN)
NO GOD PLEASE
MY GOAL FOR 2018
IF YOU DON'T STUDY
STUDY FOR EXAM
GET A GOOD GRADE
OCR Engine
STUDY FOR EXAM GET A GOOD GRADE
Visual feature
Multilayer Perceptron Classifier
Textual Feature
Textual Feature Extractor (Word2vec)

Figure 3: The overall structure of our multimodal neural network. First, input images are fed into pretrained neural networks to extract visual feature and textual features. We then concatenate these two feature vectors into a single multimodal feature representation and use a final neural network to perform classification

Similar to Dubey et al. (2018) and Beskow, Kumar, and Carley (2020), we extract visual and textual features from images and used them to identify IWT memes. As noted above, however, we focus on a broader class of memes than the prior work. Further complimenting prior work, we are the first to analyze the content of text superimposed on memes, and the first to link this content to demographic information about users.

Identification of IWT Memes

Given a dataset of images shared on social media, we develop a two-step pipeline to identify IWT memes. The first step of our pipeline is uses the Optical Character Recognition (OCR) engine Tesseract (Smith, Antonova, and Lee 2009) to filter out all images that do not contain text. The second uses a supervised classifier to distinguish, amongst the remaining images, those that are IWT memes from those that are not.

Brief details of the IWT meme classifier are given below with a focus on the training data and evaluation approach selected. Additional modeling details are given in the Appendix, as are details of our filtering strategy. With respect to filtering, we discuss an evaluation of the filter and its ability to retain images with text in general, and more importantly, to retain IWT memes. As described in the Appendix, on 100 randomly sampled images, we identify 27 IWT memes, of which 4 were dropped by the filter (a recall of 85%). The number of IWT memes we analyze thus is therefore a slight underestimate of the true number of IWT memes in the sample. However, investigation of the four IWT memes filtered out suggests that the images incorrectly rejected by our filter largely contain small, out-of-focus text. These images are important, but of less concern in the present work, where we focus primarily on the content of the text.

IWT Meme Classifier Model Overview

Figure 3 shows the overall architecture of our IWT meme classifier. Our approach is similar to Dubey et al. (2018). There are three major components: a visual feature extractor, a textual feature extractor, and a meme classifier.

[4] https://docs.imagga.com/#auto-tagging

To extract visual features, we apply the Residual Neural Network of 50 layers (ResNet50) (He et al. 2016) to the full image. To create textual features, we first use Tesseract to extract the text. We then take the element-wise average of the GloVe (Pennington, Socher, and Manning 2014) word embeddings for each word to create a single embedding for the entire text collection. Finally, we feed the combined image and text feature vectors into a three-layer neural network to make a final classification on whether or not the image is an IWT meme. Full details on the modeling approach are provided in the appendix.

We note that there exists other information, both within the tweet itself and about the user, that could potentially be leveraged to build a meme classifier. However, incorporating these forms of information would lead our model to be platform-specific, whereas the current form is potentially platform-agnostic. Further, experiments suggested these additional features did not significantly increase model performance. We therefore do not consider these features in the present work.

Training Data and Approach

We use 18,583 negative samples—which include images both with and without text— and 23,710 IWT memes to train our model. We gather the set of IWT memes for training using three strategies. First, we download images from tweets with meme-related hashtags[5] sent by our panel of Twitter users, described below. We then manually filter these images, leaving us with 3,836 IWT memes. Second, we identify several well-known meme sharing accounts on Twitter (not in our panel), and collect 16,510 IWT memes by downloading images posted by these accounts. Finally, we manually label 16,947 randomly sampled images containing text in our Twitter dataset, from which we identify an additional 3,364 IWT memes. For negative samples, we take the 13,583 images identified through this manual labeling that we did not classify as IWT memes, and add 5,000 images without any superimposed text to balance the training dataset. While the latter images do not contain text, we find that including them considerably improves the performance of the classifier.

Validation and Test Data

Note that the procedure we use for *training* data produces noisy labels - not all of the images labeled as IWT memes are verified to be IWT memes, and the converse also holds. For *validation and testing*, we ensure that all data are manually labeled in a more rigorous annotation task. This process of using distance supervision for training and manually coded data for testing is common in computational social science, where larger datasets are needed to train models than to test them (Joseph, Wei, and Carley 2016).

To construct our validation and test sets, we first randomly sample 2,750 users from the panel data described below from whom no data in the training set is drawn. For each user, we then randomly extract one image, resulting in a dataset of 2,750 images. We then trained 10 graduate students in the task of identifying IWT memes. Each student annotated approximately 450 images, and each image was annotated by at least two annotators. Where annotators disagreed, the authors of the paper served as the final deciding vote. We measure inner-annotator agreement of our annotation task using Kripendorff's alpha (Krippendorff 1980), obtaining a value of 0.60. This result is inline or higher than results obtained in other subjective annotation tasks of data from social media platforms, such as hate speech (0.622) (ElSherief et al. 2018), toxic behavior (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015) (0.39) and personal attacks (0.45) (Wulczyn, Thain, and Dixon 2017).

Baseline Models

In order to provide useful comparison points for our classifier's performance, we develop a series of baseline models. The first, entitled *Meme(Image)*, classifies memes only using image features extracted from our visual feature extractor. Likewise, we construct *Meme(Text)* which classifies memes using only textual features extracted from our textual feature extractor. We also compare our algorithm with the meme classification pipeline used in Zannettou et al. (2018). Their method first uses DBSCAN (Ester et al. 1996) to cluster images. They then annotate images using data from KnowYourMeme; any image that is not in a cluster with a meme from the KnowYourMeme database is not considered to be a meme. Since the size of our test dataset is moderate, rather than millions of images, we set each image itself as an image cluster (i.e. we set the MinPts parameter of DBSCAN algorithm to be 1). We then use their pipeline in the same fashion as described in their paper. We call this comparable method from related work *ClusterMatching*.

Analysis of Twitter Data

The methods above provide us with tools to identify IWT memes. In this section, we provide details on the data and methods we use to address our three primary research questions about IWT meme sharing on Twitter.

Data

The IWT memes we use to address our three research questions are drawn from shares from a panel of around 490,000 Twitter users that have been linked to voter registration records and that were active[6] on Twitter in August of 2019. We use these records to ensure that the individuals we study are real people, and not, e.g., bots, and to provide demographic information. Critically, the demographic information we use is restricted by the often incomplete options presented to voters during registration. Despite this limitation, we chose to include analyses of these demographic variables, as they are important means of identifying differences in social media use that may have adverse impacts on traditionally marginalized populations.

[5] '#meme', #memes', '#dankmemes', '#funnymemes', '#memesdaily', '#MEMES', '#MEME', '#Meme', '#Memes', "#offensivememes", "#spicymemes", "#edgymemes"

[6] That is, these users had not been suspended or deleted, and did not have protected accounts

Starting in July of 2017, we conducted a bi-weekly crawl of the tweets shared by these users using the Twitter Search API. We collected up to the last 3,200 public tweets shared in the last two weeks. In July of 2018, we then extracted from this data all tweets containing at least one image. In total, this was 38M tweets. In order to identify IWT memes in these tweets, we apply our proposed two-step pipeline introduced before. Doing the first filtering step results in a dataset of 12M images. Applying the IWT meme classifier over these images resulted in a final dataset of 7,251,050 IWT memes shared by 202,038 Twitter users.

The methodology used to link users to voter registration records is similar to those used by an increasing number of scholars in prior work (Barberá 2016; Grinberg et al. 2019). Our approach begins with a large set of Twitter accounts (any user who shared a public tweet captured in the Twitter Decahose from 1/14-8/16, approximately 406M users) and a large set of voter registration records obtained from TargetSmart. The voter registration record included a comprehensive set of names and locations for U.S. adults. For each Twitter account, we link it to a particular voter registration record if the account and the record both have the exact same name and location,[7] if no other Twitter account without any location exists with the given name, and if the name is unique within a given U.S. city (or state, if a city cannot be identified). An individual's name and location are drawn exactly as they are given on Twitter and/or in the voter registration records. We only identify matches where these values match exactly, and only use locations that are identifiable in a standard gazeteer. Full details on our matching approach are available in other related publications (Joseph et al. 2019; Grinberg et al. 2019).

Three points are important with respect to the data described. First, the matching methodology we perform is conservative, matching only approximately .5% of the full set of Twitter accounts we begin with. Consequently, we believe the approach to be high precision - manual evaluation suggests accuracy rates above 90% (Grinberg et al. 2019). Second, our sample is biased in an important way - towards people who provide their real names and locations. While this does not bias the sample demographically relative to the general Twitter population, it may bias our study of IWT memes towards less virulent content. Finally, we note that the use of this data has been approved by Northeastern University's Institutional Review Board. On this point, we also note that we only attempt to match individuals who provide their actual name and location, individuals who, for any reason, modify their name in any way (including using a nickname) are excluded from collection. This falls within Twitter's Terms of Service, which states that linking Twitter data to offline sources is acceptable under reasonable expectations of privacy.

Methods

Assessing Demographics of Sharing We use a generalized additive regression model with a negative binomial link function to identify demographic factors associated with IWT meme sharing, holding an individual's overall sharing activity constant. A negative binomial model is appropriate for over-dispersed count data, and an additive model can be used to relax assumptions of linear dependencies between continuous independent variables and the outcome. The regression model is described in the context of our results below. In addition to this model, we also use the same independent variables and regression model to study factors associated with the sharing of two other kinds of images. These results help to contextualize IWT meme sharing in broader patterns of image use on Twitter.

First, we study factors associated with sharing *any* image on Twitter. Second, we study factors associated with sharing *non-IWT* memes—i.e., memes that do not contain text. We use the pipeline from Zannettou et al. (2018) to identify roughly non-IWT memes; see the Appendix for details.

Manual annotation of themes As an initial inquiry into the content of IWT memes, we sampled one images each from 500 random users in our case study dataset and annotated them for whether or not they contained humorous, political, conspiratorial, or hateful content. Our interest in humor was driven by popular stereotypes of visual meme content. Our interest in the other three categories was driven by recent work suggesting the importance of visual memes in the spread of these types of content (Zannettou et al. 2018). For hateful content, we used the definition provided by (Davidson et al. 2017). For conspiratorial content, we relied on the ideas presented in (Lazer et al. 2018), identifying expressions that were presented as fact but not easily verifiable. Categories of interest were non-exclusive, e.g., images could be both political and humorous.

The three authors of the paper annotated images over two rounds of coding. After the first round, the authors discussed their perceptions of the categories, and addressed any inconsistencies. The second round then required all annotators to independently code each image. Labels were assigned to each image based on a majority vote, and inter-annotator agreement was again measured using Krippendorf's alpha (Krippendorff 1980). Because we considered each label independently for each image (i.e. we allowed for multi-class categorization), decisions were binary and thus a majority could be established for each category and each image. This also means we assessed agreement for each category independently.

Inter-annotator agreement measures suggested that humor and politics were much easier to annotate than hateful or conspiratorial content. Krippendorf's alpha scores were 0.63, 0.76, 0.33, and 0.35 for humor, politics, hateful content, and conspiratorial content, respectively. The lower scores for hateful and conspiratorial content are still in line with other social media annotation tasks described above. They also are due in part to the small number of IWT memes in these categories . We therefore still find the categories to be useful for addressing our main research question, pertaining to overall proportions in the data. However, given the difficulty in annotation, we do not attempt any further efforts to calculate more detailed statistics, or to build classifiers based

[7] Approximately 61% of accounts list something in the location field

on this labeling.

Topic modeling to identify themes In order to extract topics from the text in IWT memes, we use the Biterm Topic Model (BTM) (Yan et al. 2013). The BTM is a generative model for un-ordered word-pairs co-occurring within a document (e.g. a tweet) and has been proven effective in identifying topics in short texts (Jónsson and Stolee 2015; Oliveira et al. 2018). Each IWT meme caption, whose length is short in general, is treated as document.

For preprocessing, we first seek to minimize the impact of spam accounts. To do so, we remove all memes shared by users who share more than 500 IWT memes. Second, to further improve the results of topic modeling, we then feed the classified IWT memes into the Text Detection function from Google Cloud's Vision API to extract captions.[8] Third, after obtaining the text content of each IWT meme remaining in our dataset, we then preprocess the text for the topic model by lower-casing, removing numbers, removing stopwords, and performing lemmatization using `spaCy`.[9]. Finally, images with only one word are removed. After all these procedures, we end up with 5,923,004 IWT meme captions from around 205k users. Each caption is treated as a short document and all of them form a document corpus.

We use the following parameters for the BTM: the number of topics $K = 20$, $\alpha = 50/20$, $\beta = 0.005$, and the number of iterations $n = 1000$. Note that we experimented with $K = 30, 40$ as well, and found that $K = 20$ had the highest level of coherence (Mimno et al. 2011). Other parameters were selected according to previous work (Oliveira et al. 2018).

Differentiating IWT Themes by Demographics We use three *Group Preference Difference (GPD)* measures to quantify differences across demographics in attention to the topics we extract. To do so, we first introduce some notation. The variable $y^{D}_{k,m}$ stands for the number of times individuals in demographic group D share meme m which has been assigned to topic k.[10] The variable n^{D}_{k} stands for the total number of memes relevant to topic k that are shared by individuals in demographic group D.

We first define the *Single Meme GPD*. The Single Meme GPD measures the extent to which a single meme m is more likely to be shared by one demographic group versus another. It is defined as the log-odds-ratio between two groups D_1 and D_2.[11] Because it is a log-odds ratio, the value of the Single Meme GPD ranges from negative infinity (if m is more likely to be shared by members of D_2) to infinity (more likely for D_1). As an example, if we were to calculate the MGPD score for a single meme disparaging President Donald Trump, the value would likely be highly positive if D_1 were Democrats and D_2 were Republicans.

The *Across Topic GPD* is used to measure the extent to which one *topic* is more frequently used by members of one demographic group versus another. The Across Topic GPD score for a given topic k between demographic groups D_1 and D_2 is given by:

$$ATGPD_k = \frac{n^{D_1}_{k}}{\sum_{l=1}^{K} n^{D_1}_{l}} - \frac{n^{D_2}_{k}}{\sum_{l=1}^{K} n^{D_2}_{l}} \qquad (1)$$

Where K is the number of topics. A higher Across Topic GPD score means that topic k is preferred by D_1, compared to D_2.

Finally, the *Within Topic GPD* is the overall extent to which different demographic groups share *different memes within a particular topic*. The Within Topic GPD is distinction from the Across Topic GPD in an important way; namely, the Within Topic GPD reveals a possible internal schism in which memes are shared, even when two demographic groups are focused on the same topic. As a simple example, assume that there are only two memes in a topic, m_1 and m_2, and that Democrats share only m_1, while Republicans only share m_2. Further, assume that m_1 and m_2 are shared at the same rate by Democrats and Republicans. In this case, the Across Topic GPD is 0, but the Within Topic GPD can be used to identify polarization in sharing between m_1 and m_2 within, say, a politically-oriented theme. As a measure for Within Topic GPD comparing D_1 and D_2, we simply use the Pearson correlation coefficient between the number of times each meme within the topic has been shared by members of the two demographic groups. This mirrors the approach taken by Joseph et al. (2019) for similar data.

Results

We first briefly provide results for our IWT meme classifier as compared to the baseline models we develop. We then address our primary research questions relating to who shared IWT memes, relative to other forms of expression on Twitter, the major themes within IWT memes, and how topical content of IWT memes varies across particular demographic subgroups.

Classifier Results

Table 1 gives results for our model, as well as the baselines, on the test data and shows that our model outperforms all baselines. Improvements over the *Meme(Image)* and *Meme(Text)* baselines indicate that it is indeed useful to combine features from both the image itself as well as the text extracted from IWT memes. We also find that the baseline we derive from prior work, *ClusterMatching*, does not perform well in comparison to any of the other models. This validates the need for a new approach, but critically, is due largely to the fact that we propose a different definition of meme than is used in the prior work. Nonetheless, our model's improvements over these baselines give us confidence that our analyses below are predominantly focused

[8] Google Cloud's API is more accurate than Tesseract. However, it is prohibitively expensive for massive corpora. We therefore chose to use Tesseract for filtering and classification, and applied the Google Cloud Vision API only for topic modeling. See the Appendix section on filtering for more details.

[9] https://spacy.io/

[10] Note that, for the purposes of these analyses, memes are assigned to the topic that had the highest value in the document's posterior distribution over topics.

[11] More specifically, as the Dirichlet and variance-smoothed log odds ratio introduced in Equation (22) in the widely used text scaled approach from (Monroe, Colaresi, and Quinn)

Algorithm	Recall	Precision	Accuracy	F1
Our Algorithm	**0.71**	**0.76**	**0.78**	**0.73**
Meme(Image)	0.66	**0.76**	0.77	0.71
Meme(Text)	0.63	0.66	0.72	0.64
ClusterMatching	0.09	0.65	0.61	0.16

Table 1: The results of different methods on test dataset.

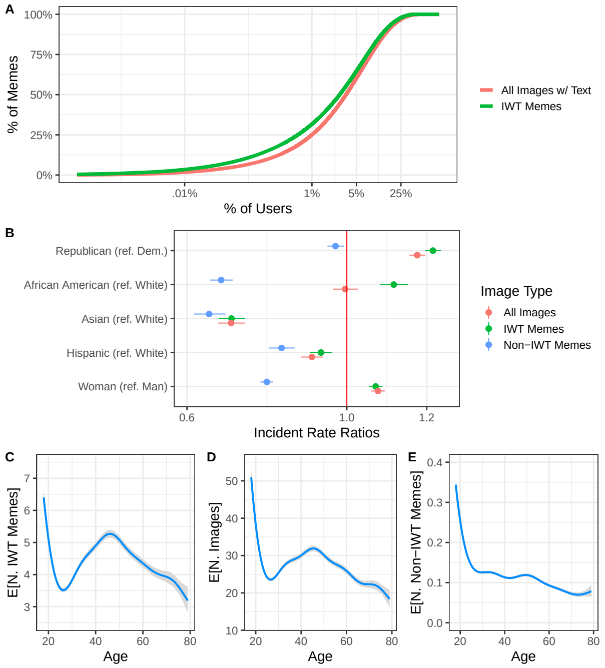

Figure 4: A) On the x-axis is the number of IWT memes (green) and total images (red) shared, and the y-axis the percent of all users who share fewer or more than that number of memes or total images. B) The x-axis gives the incident risk ratio for sharing IWT memes (green), any image (red), and non-IWT memes (blue) for various demographics (y-axis) in comparison to reference categories (given in the label on the y-axis). C) The x-axis represents the age of an individual, the y-axis the expected number of IWT memes shared by that individual. D) and E) are the same as C), but for all images and non-IWT memes, respectively. Results for B-E) are drawn from a negative binomial additive regression, described in the text.

on the phenomena of interest, and that our publicly available classifier can be used by other scholars, on other datasets, to study IWT memes as well.

Who shares memes?

Figures 4A-C provide a general characterization of the extent to which Twitter users in our panel share IWT memes relative to other kinds of content, and how this varies by demographics. Figure 4A) shows an empirical CDF of the distribution of shares of IWT memes (green), and images in general (red) for all 490,000 individuals in our dataset. IWT meme sharing, and image sharing in general, are concentrated in a few individuals. Only 41% of the users in our panel shared any IWT memes, and 4.6% of the users account for 50% of all meme shares. However, the sharing of IWT memes, and images in general, is considerably less concentrated than other phenomena recently studied on Twitter, such as the spreading of fake news (Grinberg et al. 2019). Whereas 80% of fake news was shared by .1% of the population studied in (Grinberg et al. 2019), it takes 12.3% of the population we study in order to reach 80% of the content shared.

Figure 4B) (green estimates) and C) show a curvilinear association between IWT meme sharing and age, that sharing more IWT memes is associated with being a self-identified Republican, African American (relative to Whites), or woman, and that self-identified Asian and Hispanic individuals share fewer IWT memes than self-identified white users. Note that in addition to the variables presented in Figures 4B-C, we include a control for the total number of tweets sent. Results can therefore be understood as the extent to which individuals share IWT memes, *holding the number of statuses they share overall to be constant*.

As such, Figure 4B) shows that, holding the total number of shares constant, men share slightly fewer IWT memes than women, African Americans share approximately 10% more memes than others, and Republicans share almost 20% more memes than Democrats. Figure 4C) shows that, controlling for the total number of statuses sent on Twitter, IWT meme sharing, as a proportion of overall sharing, peaks for individuals under 20 and, roughly, between 40-50.

Figure 4 also addresses the question of whether these same demographic shifts hold for other kinds of image sharing. As noted above, we use the same regression model and independent variables to predict the number of *overall* image shares for panel members (red estimates in Figure 4B, age results in Figure 4D) and the number of *non-IWT meme* image shares (blue estimates in Figure 4B, age results in Figure 4E). We find that in general, demographics that share IWT memes also share more images in general. There are two exceptions to this. First, African Americans share more IWT memes, but not more images in general. Second, individuals aged 40-50 see a slightly higher increase in IWT meme usage relative to their increase in overall image sharing. These demographics are therefore more likely to express text through images—relative to a standard text tweet—*and* more likely to send an IWT meme, relative to any other kind of image.

Finally, we find that demographic patterns in who shares IWT memes versus who shares non-IWT memes are largely inconsistent. Non-IWT memes are shared almost equally by Democrats and Republicans, and are more likely to be shared by white, male Americans aged 18-20. These non-IWT memes, in contrast to IWT memes, thus fit more consistently with the popular narrative of meme-sharing by young, internet-saavy males (Haddow 2016).

What are the topics of shared memes?

Our manual annotation of 500 random IWT memes from our dataset revealed that both humor and politics were prevalent, and that both misinformation and hateful content were not often contained in IWT memes. Humor was most preva-

Label	Top Five Words
Insurance & Health	**state health tax pay care**
Race & Gender	**right woman white black American**
Terrorism & Guns	**gun kill shoot police attack**
Political Figure	**Trump president news Clinton vote**
Food	chicken eat food cheese chocolate
Education	student work learn school support
Spam	like reply tweet retweets follow
Weather	mph weather wind storm forecast
Spanish	que los por con del
Music/Art	star book music story movie
Religion	god love life man lord
Celebrity	taylor michael james swift feat
Book Ads	book paludan vampire bestselling author
Activity	new city park march school
Sport	game win team season nfl
Unlabeled 1	free new time use home order
Unlabeled 2	not people love know like
Unlabeled 3	day year time not family
Unlabeled 4	like not water time day
Unlabeled 5	not like be get say

Table 2: Five of the top ten most important words from the BTM topic model. We use the probability by which certain topic generates words in topic modeling to measure the importance of words in that topic.

lent category, represented in 27.7% of IWT memes (95% bootstrapped confidence interval [23.8, 32.0], followed by political content (16.6%, [13.4,20.2]), conspiratorial content (1.7%, [0.8,3.4]) and hateful content (0.6%, [0.1,2.0]). Half of the IWT memes labeled conspiratorial were also political, including mis-attributed quotes and memes linking President Trump to unfounded conspiracies. All three of the observed hateful IWT memes were political as well, two broadly targeting individuals with liberal political views and one targeting a specific Republican politician. Further, in only one case was it necessary to observe a background image in order to discern the target of the hateful content. This suggests that text-based hate speech classifiers may be useful in identifying hateful content in IWT memes. However, our own attempts at applying the model from Davidson et al. (2017) resulted almost exclusively in false positives, precluding further analyses. However, given the targets of the hateful content, it remains unclear whether or not IWT memes are a primary vehicle for hate directed at traditionally marginalized groups.

Our topic model revealed similar conclusions with respect to the prevalence of political content. Table 2 lists 5 of the 10 most important words for each topic, and shows the variety of themes present in the IWT memes shared by our panel.[12] To identify topic names, two of the papers' authors independently reviewed the top five words from each topic, as indicated by the posterior probabilities from the topic model, and at least 25 random IWT memes from each topic. Images were assigned to the topic that they were most likely to be associated with according to the posterior probabilities from the topic model. The authors then discussed the names they identified for each topic and attempted to resolve differences. Topics that could not be agreed upon were labeled "Unlabeled". While these topics may represent important themes, we here choose a conservative labeling approach in order to emphasize coherent topics.

Using this approach, we identified topics spanning a broad range of cultural facets, from food to weather to sports. We also identified four topics containing politically relevant themes, bolded in Table 2. Figure 5A) shows the percentage of memes that best align with each topic, and B) the aggregate percentage of political content across all four political memes.[13] Figure 5A) shows that almost 50% of the IWT memes in our data come from topics that we could not label with a specific theme. Consequently, work remains to understand the extent to which other dimensions of meaning beyond thematic content within the text may provide insights into IWT meme sharing patterns. However, Figure 5A) also shows that the remaining shares are distributed widely across named topics, insinuating the diversity in the topics of shared IWT memes. In Figure5B), we see that around 30% of the memes not belonging to a unlabeled topic are political in nature, or around 15% of all memes.

Results in this section provide the first empirical evidence of the diversity of topics in IWT memes shared on Twitter. The popular notion of memes as simply a tool for conveying irreverent humor is at odds with both our manual and automated analyses of meme content. Political content accounts for around 15-20% of all memes shared. Consequently, further analyses of how the texts of these memes are used in political contexts are warranted. At the same time, both of our analyses suggest that those studying the ideological content of memes should recognize that a minority of the IWT memes shared by our panel are political, and that hateful or conspiratorial content is relatively rare. These findings accord with prior work on misinformation, at least, which suggests that real people rarely shared misinformation via URLs on Twitter (Grinberg et al. 2019).

Who shares memes on which topics?

Our final research question involves digging into demographic differences in IWT meme sharing. Motivated by our observation that Republicans and African American share more IWT memes relative to other demographics, controlling for other forms of expression on Twitter, we further evaluate how these two demographic groups vary in the IWT memes they sent.

Figure 6 explores differences between Democrats and Republicans in the topics they share on Twitter, showing that Democrats are more likely, on average, to share political memes (higher Across Topic GPD for these topics), but that within the political topics, Democrats and Republicans differ widely on the specific memes they share (political topics have lower Within Topic GPD). The y-axis of Figure 6 shows that despite similar sharing patterns on a-political top-

[12] Our analysis did not remove duplicates, which constituted 23% of our data. Topics identified without duplicates qualitatively matched those with duplicates, thus we present only results from our original analysis here. Results for de-duplicated data can be found alongside the code release for this paper.

[13] For a further demographic breakdown of topic use, see Table 3 in the Appendix.

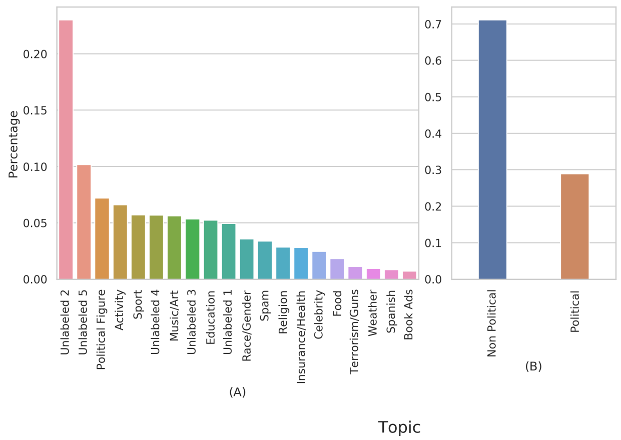

Figure 5: (A) Shows the percentage of each topic across all IWT memes in our dataset. The x-axis represents the label of each theme. The y-axis the percentage of corresponding theme. (B) Shows the percentage of political and non political meme after removing memes in Noise topics.

ics like food and sports, Republicans and Democrats share very different sets of political memes.

Our explanation for this is reflected in the examples extracted from the Political Figure topic shown in Figure 7. Figure 7 represents typical memes shared most heavily by Democrats (left-most), most heavily by Republicans (right-most), and those shared more or less equally between Democrats and Republicans. Our analyses of these memes, and others within the topic, suggest that the most heavily polarized memes were about Donald Trump or Hillary Clinton, and that those in the middle were typically related more to policy. Our observations suggest that memes that were most ideologically distinct with respect to sharing patterns seem to have emphasized attacks of or support for individuals, rather than particular policies.

Assessing sharing patterns of self-identified African Americans relative to other demographic groups using the same Across Topic GPD and Within Topic GPD analysis did not reveal obvious differences in topical focii or within topic differences. We therefore carried out a manual investigation of the 50 memes that had the highest Single Meme GPD scores calculated between African American and the other self-identified race/ethnicity groups in Figure 4B). These memes were most likely to be shared by self-identified African Americans, relative to other self-identified race/ethnicities. Two authors spent time assessing emergent characteristics of the images, looking for commonalities between them.

After independent analysis and discussion, the two authors agreed that the most important characteristic shared by the images was the representation of black individuals in the background image. Using another round of manual coding, where authors came to an agreement on each image, we found that 75.6% of the IWT memes that were still available online in August of 2019 (31 of 41) contained a background image of an African American.

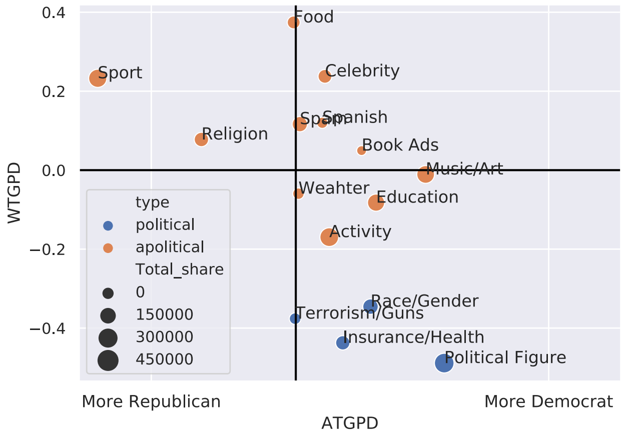

Figure 6: This plot shows Within Topic GPD (WTGPD) and Across Topic GPD (ATGPD) scores of each labeled theme between Republicans and Democrats. Each topic is represented by a dot. Blue dots represent politically-related topics, orange dots represent non-political topics. The size of each point is proportional to the number of shares of the memes assigned to the topic. The x-axis represents the Across Topic GPD score, indicating the extent to which all memes relevant to the topic are shared more by Democrats (to the right of the black vertical line) versus more by Republicans (to the left of the black vertical line). The y-axis represents the Within Topic GPD score, indicating the extent to which Democrats and Republicans share different memes within that particular topic. Topics below the black line are those where Republicans and Democrats, on average, share different sets of memes within the topic. Above the line, memes within the topic more frequently shared by Democrats are the same memes more frequently shared by Republicans.

In some cases, the text of the image was then used to convey comments on experienced racism (e.g., an IWT meme retweeted 13,000 times picturing Lebron James commenting that “being black in America is tough”[14]). Others emphasized the successes of black athletes and/or celebrities[15]. In general, we therefore find that IWT memes may be an important means by which positive black identity is constructed, and by which the consequences of being black in America are emphasized on Twitter. As with the use of emojis (Robertson, Magdy, and Goldwater 2018), the ways in which race and identity are expressed within IWT memes implies ways in which the study of race, identity, and self-expression online must move beyond traditional approaches that analyze text alone.

[14] https://twitter.com/bleacherreport/status/870046378315046912

[15] https://twitter.com/sloanestephens/status/909149510215000064, retweeted 4,000 times

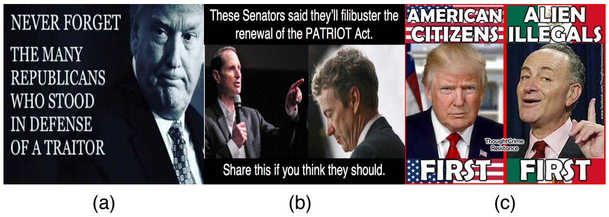
NEVER FORGET
THE MANY REPUBLICANS WHO STOOD IN DEFENSE OF A TRAITOR
These Senators said they'll filibuster the renewal of the PATRIOT Act.
Share this if you think they should.
AMERICAN CITIZENS
FIRST
ThoughtCrime Resistance
ALIEN ILLEGALS
FIRST
(a)
(b)
(c)

Figure 7: Three samples of political figure memes. (a) is shared only by Democrats. (b) is shared by a similar number of Democrats and Republicans. (c) is only shared by Republicans

Conclusion

The core contributions of present work are threefold. First, we provide the first large-scale analysis of who shares IWT memes, relative to other forms of expression on Twitter, and the topical focus of text content extracted from these images. Second, we provide an analysis of the relationship between the demographics of users and their meme sharing patterns. Finally, as a function of our primary research questions, we develop an accurate and publicly available classifier to identify IWT memes in other datasets.

These contributions speak to two broader issues of interest to computational social scientists. First, we find that IWT memes, as a non-traditional form of sharing text, are more heavily used by African Americans, even relative to non-IWT memes and images in general. Similar to the findings of Blodgett, Green, and O'Connor (2016), who found that the standard preprocessing step of removing non-English text may marginalize Black voices, we find that not analyzing text superimposed on images may create similar, albeit less severe and harder to tackle, issues with representation. While not a marginalized population, similar issues with representation exist for Republican users in our dataset, although largely for image sharing in general.

Second, in a similar vein, we find that although political, hateful, and conspiratorial content account for a minority of shares of IWT memes, real people do use them for this purpose. Consequently, as others have noted (e.g. (Zannettou et al. 2018)), future work is needed to adapt our understanding of these problems beyond shares of news URLs and text data into the visual meme domain. More specifically, we find evidence for a potentially novel use of political memes - specifically, manual evaluation of highly polarized political memes suggests that they focus largely on supporting and/or attacking well known political figures, relative to any discussion about particular policies.

There are several limitations in our research. First, we use binary descriptions for both gender and race/ethnicity. While these values represent self-identified expressions, binarization of these variables can nonetheless be problematic. Second, we focus on a particular form of meme that may be overly general or too specific for other research questions. Third, we do not study the interplay between tweet text and the shared image, and thus may miss various subtexts under which IWT memes are shared. Finally, as emphasized above, we study a particular subset of users (those linked to voter registration records) on a particular social media platform (Twitter). Much remains to be done to extend our findings to other platforms on which memes have been studied and are often shared.

Online memes have become an important way to spread information, identity, and ideology. We look forward to future work leveraging more advanced computer vision methods to understand this exciting form of social media content.

Acknowledgements

This work was supported by a Google Cloud Platform Research Credit grant. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

Appendix

Details on Image Filtering

Due to the high cost of the more accurate Google Cloud Vision API, we use the open-source tool Tesseract to filter out images without text. To understand the possible consequences of using this less accurate tool, we carry out an evaluation. Specifically, we sample 100 images and carry out two annotation tasks. In the first, we annotate each image according to whether or not it contains text, in the second, whether or not it is an IWT meme.

We find that the Google API is 100% accurate, keeping 59/59 of images with text and 27/27 IWT memes. In contrast, Tesseract retains 38/59 of the images with text, and 23/27 (85%) IWT memes. The difference between the two tasks stems from the fact that the Cloud Vision API is much more sensitive to small, non-focal text. As noted above, these images are not critical for the present work.

Additional Classifier Details

As noted above, to extract image-based features, we first run the image through ResNet50. On the top of the last layer of ResNet50, we add a fully connected layer to adjust the dimension of visual feature output to p. Thus, given an image input I, the extracted visual feature $F_v \in \mathbb{R}^p$ for the input is $F_v = W_{pf} \cdot F_{v_{res}}$. Here, $F_{v_{res}}$ stands for the visual feature extracted by ResNet50. W_{pf} represents the weight of the added fully connected layer.

Also as noted above, to identify text-based features, we first use Tesseract to extract unigram word tokens from the image. We then translate each word into a d dimensional vector in a shared look-up table $T \in \mathbb{R}^{|V| \times d}$, where V is the vocabulary. Thus given one sentence $I_c = (w_1, w_2, ..., w_n)$ that is superimposed on the input image, where w_i represents the ith word in the sentence, we can get the corresponding word embedding representation for this sentence as $E_c = (t_1, t_2, ..., t_n)$, where $t_i \in \mathbb{R}^d$. In order to get a representation for all superimposed text, we calculate the element-wise average of the vector representation of each word. This sentence representation will serve as the textual feature of input images. Denote the textual feature output as $F_t \in \mathbb{R}$, the operation of element-wise average can be expressed as $F_t = \frac{(\sum_{i=1}^{n} t_i)}{n}$.

Finally, given the textual feature F_t and visual feature F_v of the input image, we use a deep neural network to perform classification. We first concatenate two feature vectors to form a singe multi-modal feature representation denoted by $F_r = F_v \oplus F_t \in \mathbb{R}^{d+p}$. The concatenated vector is then fed into 3 fully connected layers. In order to add nonlinearity and avoid overfitting for neural network, we add an ReLU activation function and a batch normalization layer respectively after each fully connected layer except last one. Finally, we feed the output of the last fully connected layer into a sigmoid activation function which outputs the probability of an image being an IWT meme. Denote the probability that the ith image is an IWT meme as o_i, the operation of classifier can be expressed as:

$$o_i = \sigma_2(W_3 \cdot \sigma_1(W_2 \cdot (\sigma_1(W_1 \cdot \sigma_1(F_t))))) \qquad (2)$$

Where W_i represent the parameters of the ith fully connected layer. σ_1 stands for combined operation of ReLU activation and batch normalization. σ_2 is the sigmoid function.

Note that our approach therefore makes use of transfer learning, which has become a common practice in computer vision (Dubey et al. 2018) and natural language processing (Felbo et al. 2017), by initializing weights of our classifier with pretrained models. We initialize the weights of the image feature extractor by a model pretrained on ImageNet (Deng et al. 2009), and the weights in the textual feature extractor with word embeddings from the GloVe model (Pennington, Socher, and Manning 2014). We then fine-tune all parameters by minimizing the cross entropy between the predicted probabilities and the target labels. We choose to use Adam (Kingma and Ba 2015) optimization method with a learning rate of 0.00001 and a batch size of 32.

Identifying Non-IWT Memes

To identify non-IWT memes, we being with the set of all images filtered out by the first step of our pipeline. We then feed the remaining images into the pipeline from Zannettou et al. (2018), introduced as *ClusterMatching* baseline before. In this step, from the remaining images, *ClusterMatching* is able to pick out non-IWT memes which are documented in KnowYourMeme. For the clustering process in *ClusterMatching*, we treat each image itself as an image cluster.

Topics for Different Demographic Groups

Table 3 presents the percentage of all memes shared by various demographic groups from each topic. Note that columns in the table sum to 100%.

Label	**Male**	**Female**	**Democrat**	**Republican**
Insurance & Health	3%	2.6%	3.1%	2.4%
Race & Gender	3.7%	3.5%	4.0%	3.0%
Terrorism & Guns	1.2%	0.9 %	1.1%	1.1%
Political Figure	7.9%	6.6%	8.0%	6.0%
Food	1.4%	2.1 %	1.8%	1.8%
Education	4.8%	5.6%	5.7%	4.6%
Spam	3.5%	3.3 %	3.4 %	3.4%
Weather	1.3%	0.7 %	0.9 %	0.9%
Spanish	0.8%	0.8 %	0.9%	0.6%
Music/Art	5.3%	5.8 %	6.3%	4.5%
Religion	2.4%	3.2 %	2.3%	3.6%
Celebrity	2.8%	2.2%	2.6%	2.2%
Book Ads	2.9%	1.0%	1.1%	0.1%
Activity	7.2%	6.1%	6.8%	6.3%
Sport	9.5%	2.7%	4.6%	7.3%
Unlabeled 1	5.3%	4.6%	5.1%	4.8%
Unlabeled 2	18.6%	26.4%	22.0%	24.6%
Unlabeled 3	4.7%	5.9%	5.0%	5.9%
Unlabeled 4	5.6%	5.7%	5.6%	5.8%
Unlabeled 5	10.0%	10.1 %	9.6%	10.9%

Table 3: The first column shows the annotated themes. For the second to fifth columns, the top row defines the different demographic groups, and the rest rows show the sharing percentage of memes relevant to different themes.

References

Barberá, P. 2016. Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data.

Bauckhage, C. 2011. Insights into internet memes. In *ICWSM*.

Beskow, D. M.; Kumar, S.; and Carley, K. M. 2020. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing & Management* 57(2):102170.

Bharati, A.; Moreira, D. M.; Brogan, J.; Hale, P.; Bowyer, K. W.; Flynn, P. J.; Rocha, A.; and Scheirer, W. J. 2018. Beyond pixels: Image provenance analysis leveraging metadata. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* 1692–1702.

Blackmore, S. 2000. *The Meme Machine*. OUP Oxford.

Blodgett, S. L.; Green, L.; and O'Connor, B. 2016. Demographic dialectal variation in social media: A case study of African-American English. *EMNLP'16*.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial behavior in online discussion communities. In *ICWSM*.

Coscia, M. 2013. Competition and success in the meme pool: a case study on quickmeme.com. *CoRR* abs/1304.1712.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *arXiv preprint arXiv:1703.04009*.

Dawkins, R. 1976. 1989. the selfish gene.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

Díaz, C., and Mauricio, C. 2013. Defining and characterizing the concept of Internet Meme. *CES Psicología* 6(2):82–104.

Diesner, J.; Frantz, T. L.; and Carley, K. M. 2005. Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different". *Comput. Math. Organ. Theory* 11(3):201–228.

Dubey, A.; Moro, E.; Cebrian, M.; and Rahwan, I. 2018. Memesequencer: Sparse matching for embedding image macros. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, 1225–1235.

ElSherief, M.; Kulkarni, V.; Nguyen, D.; Wang, W. Y.; and Belding-Royer, E. M. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *ICWSM*.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. 226–231. AAAI Press.

Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; and Lehmann, S. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *EMNLP*.

Finkelstein, J.; Zannettou, S.; Bradlyn, B.; and Blackburn, J. 2018. A quantitative approach to understanding online antisemitism. *CoRR* abs/1809.01644.

Garimella, V. R. K.; Alfayad, A.; and Weber, I. 2016. Social media image analysis for public health. In *CHI*.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363(6425):374–378.

Gupta, A.; Lamba, H.; Kumaraguru, P.; and Joshi, A. 2013. Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, 729–736. New York, NY, USA: ACM.

Haddow, D. 2016. Meme warfare: How the power of mass replication has poisoned the US election. *The Guardian*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hu, X., and Liu, H. 2012. Text analytics in social media. In *Mining Text Data*.

Hu, Y.; Manikonda, L.; and Kambhampati, S. 2014. What we instagram: A first analysis of instagram photo content and user types. In *ICWSM*.

JafariAsbagh, M.; Ferrara, E.; Varol, O.; Menczer, F.; and Flammini, A. 2014. Clustering memes in social media streams. *Social Network Analysis and Mining* 4:1–13.

Jónsson, E., and Stolee, J. 2015. An evaluation of topic modelling techniques for twitter.

Joseph, K.; Swire-Thompson, B.; Masuga, H.; Baum, M. A.; and Lazer, D. 2019. Polarized, Together: Comparing Partisan Support for Trump's Tweets Using Survey and Platform-Based Measures. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 290–301.

Joseph, K.; Wei, W.; and Carley, K. M. 2016. Exploring patterns of identity usage in tweets: A new problem, solution and case study. In *Proceedings of the 25th International Conference on World Wide Web*, 401–412. International World Wide Web Conferences Steering Committee.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Krippendorff, K. 1980. Content analysis: An introduction to its methodology.

Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; and Rothschild, D. 2018. The science of fake news. *Science* 359(6380):1094–1096.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, 497–506. New York, NY, USA: ACM.

Liu, L.; Preotiuc-Pietro, D.; Samani, Z. R.; Moghaddam, M. E.; and Ungar, L. H. 2016. Analyzing personality through social media profile picture choice. In *ICWSM*.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, 1150–1157 vol.2.

Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 262–272. Stroudsburg, PA, USA: Association for Computational Linguistics.

Monroe, B. L.; Colaresi, M. P.; and Quinn, K. M. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16(4):372–403.

Nguyen, D.; Liakata, M.; DeDeo, S.; Eisenstein, J.; Mimno, D.; Tromble, R.; and Winters, J. 2019. How we do things with words: Analyzing text as social and cultural data. *arXiv:1907.01468 [cs]*.

Oliveira, L. S.; de Melo, P. O. S. V.; Amaral, M. S.; and Pinho, J. A. G. 2018. When politicians talk about politics: Identifying political tweets of brazilian congressmen. In *ICWSM*.

Know your meme. https://knowyourmeme.com/. Accessed: 2010-09-30.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Robertson, A.; Magdy, W.; and Goldwater, S. 2018. Self-representation on twitter using emoji skin color modifiers.

Smith, R.; Antonova, D.; and Lee, D.-S. 2009. Adapting the tesseract open source ocr engine for multilingual ocr. In *Proceedings of the International Workshop on Multilingual OCR*, MOCR '09, 1:1–1:8. New York, NY, USA: ACM.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *WWW*.

Xie, L.; Natsev, A.; Kender, J. R.; Hill, M.; and Smith, J. R. 2011. Visual memes in social media: Tracking real-world news in youtube videos. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, 53–62. New York, NY, USA: ACM.

Yan, X.; Guo, J.; Lan, Y.; and Cheng, X. 2013. A biterm topic model for short texts. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, 1445–1456. New York, NY, USA: ACM.

You, Q.; Bhatia, S.; Sun, T.; and Luo, J. 2014. The eyes of the beholder: Gender prediction using images posted in online social networks. In *2014 IEEE International Conference on Data Mining Workshop*, 1026–1030.

Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, 188–202. New York, NY, USA: ACM.