Visual Persuasion in COVID-19 Social Media Content: A Multi-Modal Characterization

Mesut Erhan Unal  
meu6@pitt.edu  
Department of Computer Science  
University of Pittsburgh  
Pittsburgh, Pennsylvania, USA

Adriana Kovashka  
kovashka@cs.pitt.edu  
Department of Computer Science  
University of Pittsburgh  
Pittsburgh, Pennsylvania, USA

Wen-Ting Chung  
wtchung@pitt.edu  
Department of Psychology in Education  
University of Pittsburgh  
Pittsburgh, Pennsylvania, USA

Yu-Ru Lin  
yurulin@pitt.edu  
Department of Informatics and Networked Systems  
University of Pittsburgh  
Pittsburgh, Pennsylvania, USA

ABSTRACT
Social media content routinely incorporates multi-modal design to convey information and shape meanings, and sway interpretations toward desirable implications, but the choices and impacts of using both texts and visual images have not been sufficiently studied. This work proposes a computational approach to analyze the impacts of persuasive multi-modal content on popularity and reliability, in COVID-19-related news articles shared on Twitter. The two aspects are intertwined in the spread of misinformation: for example, an unreliable article that aims to misinform has to attain some popularity. This work has several contributions. First, we propose a multi-modal (image and text) approach to effectively identify popularity and reliability of information sources simultaneously. Second, we identify textual and visual elements that are predictive to information popularity and reliability. Third, by modeling cross-modal relations and similarity, we are able to uncover how unreliable articles construct multi-modal meaning in a distorted, biased fashion. Our work demonstrates how to use multi-modal analysis for understanding influential content and has implications to social media literacy and engagement.

CCS CONCEPTS
- Human-centered computing → Social network analysis.

KEYWORDS
multi-modal content analysis, multi-modal learning, content popularity, content reliability

1 INTRODUCTION
From campaigns to advertising, social media content routinely incorporates multi-modal design choices that combine texts and images to effectively convey information, shape meanings, and sway interpretations toward desirable implications. Compared to textual and linguistic analyses, how the different compositions of written
words and visual elements were created and disseminated on social media has not been sufficiently studied. This work situates in the context of prevalent online misinformation in the ongoing COVID-19 pandemic. Increased isolation and the anxiety about the pandemic drastically changed our lives – particularly, the increased use of social media can result in fast spreading of false content, make users more susceptible to misinformation, and create unique challenges to detect and debunk untruth [47]. This study attempts to reveal how the subtle multi-modal content elements are associated with the propagation of information from online news outlets that manipulate facts or shape misinterpretations.

In this work, we focus on two aspects of persuasive information, popularity and reliability. While inferring content reliability alone may seem enough to identify problematic content and prevent its spread, popularity is another aspect yet often overlooked. Besides allowing us to investigate content creation strategies to persuade the audience and propagate misinformation, estimating content popularity can also help with timely debunking and prevention of the spread of misinformation. For example, one can prioritize content estimated to become popular for manual fact-checking, when slow and costly expert evaluation is a part of the process.

Popularity and reliability of news articles shared on social media have been studied before as separate topics. Efforts on predicting popularity of news articles often rely on hand-crafted content features [1, 2, 34] and early engagement statistics [4, 27, 56]. Prior work focuses on textual content, and does not investigate in what way accompanying visuals contribute to popularity, even though modern media is often multi-modal. However, work in media studies and communication theory suggests images play a critical role in conveying meaning and are a powerful rhetoric tool [8, 30, 33]. In contrast to text, images are eye-catching and concisely paint a rich context. For instance, images can imply associations between people and qualities [18, 48], and use juxtaposition or contrast to suggest desirable properties or undesirable outcomes [55]. Because images are powerful, they can both make content popular, and also carry out an agenda and mislead. Since most news sources use special meta-tags to specify which image should be shown with the shared article on social media (e.g. Twitter), analyzing their target-specific imagery may help us better understand the relative contribution of visuals in COVID-19 (mis-)information on these platforms. However, to the best of our knowledge, no prior work examines popularity of COVID-19-related imagery.

Prior work on predicting reliability, on the other hand, focuses on detecting fake news using article content [13, 36] and social context features [29, 32, 39, 43, 52, 53, 57]. Nevertheless, detection methods that employ social context heavily rely on meta-data beyond the content itself. For example, network-based models (e.g. [32, 57]) utilize social network graphs which usually requires extensive data collection, pre-processing and computation efforts. Models that make use of user-based features (e.g. [43]) do not generalize well onto spreaders who have little to no previous social interaction. Finally, efforts that utilize multi-modal content (image and text) suffer from lack of interpretability, and fail to explain the link between reliability and high-level concepts in the input.

Using data collected from social media pertaining to the COVID-19 crisis, we attempt to characterize the elements of persuasion. In this work, “persuasion” refers to the communication tactics manifested as multi-modal (textual or visual) elements which articles use to reach their audiences and convey a particular message. We use popularity as a proxy measure of persuasiveness, and reliability relates to the agenda, i.e. the purpose of the persuasion (agenda to convey accurate or misleading information). We examine both popularity and reliability of COVID-related content, where “popularity” is captured by how frequent an article shared on social media, and “reliability” refers to the credibility of the online news outlets previously identified in prior work [11]. We seek to answer the following questions:

- **RQ1:** To what extent do textual and visual signals in a tweet predict the popularity and reliability of news articles shared on social media?
- **RQ2:** What textual and visual elements are predictive of the popularity and reliability of shared news? How can we identify the predictive signals?
- **RQ3:** How does the combination of textual and visual elements in unreliable and reliable sources differ?

To address these questions, we first develop a multi-modal approach using visual and textual cues from news-sharing tweets. We learn a shared feature space optimizing jointly for both popularity and reliability classification tasks, and use this space to visualize parts of the input that are salient (informative) for the model’s predictions, as well as to show how these salient parts change across two tasks and their classes. We finally formulate a cross-modal retrieval task to discover whether reliable and unreliable sources combine visual and textual elements differently to construct multi-modal meaning.

Our work is the first empirical study that analyzes the popularity and reliability aspects of multi-modal persuasive COVID-19-related content using a multi-task approach. Our method outperforms other multi-modal baselines on both popularity and reliability prediction tasks. We find that multi-modal data better enables detection of misleading or popular content, but the relative importance of visual and textual features varies: for instance, visual features are more important for reliability classification. One important finding is that unreliable content constructs multi-modal meaning in a biased and distorted fashion, as the results show that a multi-modal representation model trained on unreliable articles does not translate well to reliable ones. Finally, articles from unreliable sources often feature visual or mentions of national symbols, certain lab/medical equipment, charts, and comics. Our work can be used in high-school curricula to develop critical media literacy skills, to gauge bias in publicly funded news media, or to construct balanced presentation of news in search engines and social media feeds.

2 RELATED WORK

**Multi-modal learning on general data.** A plethora of recent work investigates the ways of integrating information from different modalities for tasks such as image captioning [20, 24, 60], but while captioning assumes the same objects are shown and mentioned, this is rarely the case in news articles where images and text serve complementary roles. We discuss multi-modal approaches for the tasks relevant to our problem setting, below.
We show our approach outperforms [21]'s.

Table 1: Number of articles in our collection by domain codings [11] and popularity. [11] uses Red, Orange, Yellow and Green to denote the tendency of news sources to elicit fake news and misinformation (Red being the highest) and Satire to denote self-describing satirical sources. Red, Orange and Green are used in experiments.

<table>
<thead>
<tr>
<th>Red</th>
<th>Orange</th>
<th>Yellow</th>
<th>Green</th>
<th>Satire</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Popular</td>
<td>934</td>
<td>2,163</td>
<td>5,187</td>
<td>4,457</td>
<td>12,821</td>
</tr>
<tr>
<td>Unpopular</td>
<td>1,149</td>
<td>1,917</td>
<td>5,181</td>
<td>4,958</td>
<td>13,237</td>
</tr>
<tr>
<td>Total</td>
<td>2,083 (8.0%)</td>
<td>4,080 (15.7%)</td>
<td>10,368 (39.8%)</td>
<td>9,415 (36.1%)</td>
<td>26,058 (100%)</td>
</tr>
</tbody>
</table>

Table 2: Number of articles by reliability and popularity in the experiment dataset. Descriptive statistics of popularity measure $P$ within each group reported as (mean, stdev).

<table>
<thead>
<tr>
<th>Reliable</th>
<th>Unreliable</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Popular</td>
<td>3,066 (.004, .010)</td>
<td>3,097 (.0, .0)</td>
</tr>
<tr>
<td>Unpopular</td>
<td>3,097 (.003, .008)</td>
<td>3,066 (.0, .0)</td>
</tr>
<tr>
<td>Total</td>
<td>6,163</td>
<td>6,163</td>
</tr>
</tbody>
</table>

Reliability and bias prediction. Predicting reliability of news articles on social media has seen interest in recent years, especially after the 2016 elections [35]. Some work requires manual fact-checking data from experts at the article-level granularity [42, 44], which is costly, slow and not scalable. Thus, [14] shifted the attention to source reliability. Following their approach, we use source-level reliability labels given in [11] for the articles in our dataset. Prior work has mostly examined cues from text and social context. [36] performs fake news detection using hand-crafted content features (e.g., number of paragraphs). [43] combines implicit (e.g., age, political orientation) and explicit (e.g., registration time, follower count) user features for fake news detection. Research efforts on bias prediction and persuasion in visual content is relatively recent and limited. [18, 19, 48, 59] examine how politicians’ portrayal can be used to predict personal qualities, electability, and bias of the news source. [48, 58] predict political ideology from images that politicians share. None of this work pertains to the COVID-19 crisis. The COVID-19 topic poses a challenge in that it is fairly narrow, thus the type of imagery will be limited, and the same images might often be reused and thus not be discriminative. Finally, multi-modal learning has also been used to analyze social media. [15, 16] fuse features and statistics from different modalities using an attention mechanism to perform rumor detection. [21] learn a feature space to capture explicit correlations between images and text by employing a multi-modal variational autoencoder. [54] learn event-agnostic multi-modal features for fake news detection performing event discrimination as an auxiliary task. [28, 51] utilize recent multi-modal transformer architectures to detect hateful memes. In contrast to these works, we use multi-modal cues in a multi-task setting to perform article popularity and reliability classification tasks, in the unique context of COVID-19 misinformation. Importantly, these works only perform classification, but do not examine the elements of misinformation. In other words, they do not explain which parts of images/text are important, do not reveal the associations between high-level visual concepts (e.g., a star) and reliability, and crucially, the different ways images and text are combined to convey meaning. We show our approach outperforms [21]'s.

Content popularity prediction. Some work models engagements on social media and number of page views [4, 27, 56]. Other methods purely rely on content, hypothesizing it is the ultimate driver for popularity. For example, [2] use textual features such as topic, sentiment and named entities mentioned in the article. [34] shows article titles reveal strong signals for popularity but its dataset is limited to two news sources. Some recent work investigates popularity of a specific type of content such as images [9, 61, 63] and videos [3, 17, 50]. Most of these works, except [3, 61], are uni-modal (visual) only, not multi-modal. Our work learns from multi-modal cues to predict article popularity, within a multi-task framework, from content only and no meta-data, using a dataset of 95 news sources. We experimentally compare against [3] and demonstrate superior performance.

3 DATASET

Our dataset is constructed using a list of pandemic-related tweets curated by Chen et al. 2020 [5], and reliability coding of news domains proposed in Grinberg et al. 2019 [11]. In their work, Grinberg et al. 2019 use red, orange, yellow and green to denote likelihood of news sources to spread misinformation, and satire to denote self-described satirical sources. After we retrieve tweet objects for tweets given in [5], we only keep tweets that include a link to one of the domains in [11]. After data collection, we obtain a set of articles $S = \{A_1, A_2, ..., A_N\}$ where each article $A_i$ is represented as a set of tweets which shared that particular article. Lastly, we crawl article URLs to retrieve their titles and images. We specifically check for twitter:title (og:title as fallback) and twitter:image (og:image as fallback) meta-tags since they are utilized by the news source to denote the title and the image to appear within a news-sharing tweet. We will share the URLs of images and the split into reliable/unreliable tweets and images as an extension of [5]'s dataset.

**Popularity labels**: The first task we want our model to perform is binary article popularity classification. Thus, we come up with a popularity measure which makes use of retweet and like counts of tweets that shared the same article (raw popularity), and follower counts of authors posted those tweets (audience size):

$$P_{A_i} = \frac{\sum_{t \in A_i} r_{\text{retweet}} + l_{\text{like}}}{\sum_{t \in A_i} Q(\text{title})} + \lambda \tag{1}$$

where $r_{\text{retweet}}$ and $l_{\text{like}}$ denote number of retweets/likes for tweet $t$ in set $A_i$, $Q$ the number of followers of a Twitter user, and $\lambda$ is a smoothing constant to prevent the score from being inflated when audience count is small. The top 20% articles are taken as popular and the bottom 20% are taken as unpopular. All popular articles have a popularity measure $P$ greater than zero, and $P$ for all unpopular articles is zero.

**Choosing $A$**: Setting the right value for $\lambda$ is important as it affects the calculated $P$ values and thus the set of popular articles.
(top 20%). One should expect that, with an appropriate choice of \( \lambda \), the distribution of audience size in popular articles and in articles that gained some popularity (i.e. \( P > 0 \)) should be similar as the former is a subset of the latter. Otherwise, the chosen \( \lambda \) could be favoring articles with small/large audience as having higher \( P \). After experimenting with different values, we set \( \lambda \) to 10\(^4\) as it makes these two article sets’ audience distributions similar.\(^1\)

Reliability labels: Data points are also assigned binary reliability labels. To this end, domain codings in our data collection need to be collapsed into two categories: reliable and unreliable. After a careful review of [11]’s domain labeling strategy, we strip visual and text features from the same article are embedded in a metric space to understand how image-text composition varies in un/reliable articles. Note: The feature extraction module does not share weights between the two architectures.

As the language used in article titles is likely different than in tweets (e.g. tweets are more informal), we hypothesize these two should not share the word embedding space. We train two separate Word2Vec [31] models offline using article titles (\( \phi \)) and tweet texts (\( \psi \)). Both Word2Vec models embed a token into a 128-D space (\( \phi, \psi : X \rightarrow R^{128} \)). Finally, we represent titles and tweet texts as a sequence of Word2Vec embeddings, preserving token order and padding with \( \theta \in R^{128} \) to the length of the longest sequence. Our model employs [22]’s Text-CNN architecture on top of these 128-D representations. Concisely, our textual feature extractors (\( G, H \)) employ 1-D filters of size \( [3, 5, 7] \), 128 filters for each. We apply max-pooling over filter outputs, resulting in one scalar per filter, and feature extractors \( G : A^{title} \rightarrow R^{384} \) and \( H : A^{tweet} \rightarrow R^{384} \).

We compare to alternative text representations in Sec. 5. For images, we employ ResNet-50 [12] pre-trained on ImageNet [7] as feature extractor \( (F : A^{image} \rightarrow R^{2048}) \). We concatenate text and image modalities to perform T1 (popularity) and T2 (reliability prediction) using two classification branches (Fig. 2) and a multi-task binary cross-entropy loss:

\[
L(\theta) = L_{T_1}(\theta) + L_{T_2}(\theta)
\]

\[
L_{T_1}(\theta) = -\sum y_p \log(\hat{y}_p) + (1 - y_p) \log(1 - \hat{y}_p)
\]

\[
L_{T_2}(\theta) = -\sum y_r \log(\hat{y}_r) + (1 - y_r) \log(1 - \hat{y}_r)
\]

where \( y_p \in \{0, 1\} \) and \( y_r \in \{0, 1\} \) denote ground-truth popularity and reliability labels respectively, and \( \hat{y}_p = p(y'_p = 1 | \theta) \) and \( \hat{y}_r = p(y'_r = 1 | \theta) \) denote predictions.

The intuition for using convolutions for text is that popularity and reliability may be inferable from local patterns in the text. Thus, learning convolutional filters that match these patterns may be easier than modeling the entire text autoregressively. We show in Sec. 5 that our method outperforms both [3] and [21], which use bi-directional LSTM for text encoding. Convolutions also facilitate our interpretation of pattern importance.

We train our model with an initial learning rate of \( 1 \times 10^{-4} \) and decrease it by \( \times 0.1 \) if validation loss does not improve in the last

---

\(^1\)The audience distribution is highly skewed (mean: 430,381, median: 13,824 within the initial 69,591 articles, and mean: 668,180, median: 52,743 among the 48,562 articles that gained at least one like or retweet.)
Table 3: Comparison of classification performance (mean accuracy, ± standard error) between our multi-task architecture and other baselines. The best method is shown in bold, and the second-best is underlined.

<table>
<thead>
<tr>
<th>Method</th>
<th>T1: Popularity</th>
<th>T2: Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMAGE+Doc2Vec-FUSION</td>
<td>63.1% (±0.10)</td>
<td>61.9% (±0.10)</td>
</tr>
<tr>
<td>IMAGE+Doc2Vec-GRU</td>
<td>65.0% (±0.08)</td>
<td>63.4% (±0.09)</td>
</tr>
<tr>
<td>Bielski &amp; Trzcinski [3]</td>
<td>69.8% (±0.09)</td>
<td>73.1% (±0.09)</td>
</tr>
<tr>
<td>Khattar et al. [21]</td>
<td>70.3% (±0.03)</td>
<td>69.2% (±0.09)</td>
</tr>
<tr>
<td>Ours (single-task)</td>
<td>70.8% (±0.08)</td>
<td>78.0% (±0.09)</td>
</tr>
<tr>
<td>Ours (multi-task)</td>
<td>71.2% (±0.08)</td>
<td>78.0% (±0.08)</td>
</tr>
</tbody>
</table>

Table 3 summarizes the results. We observe that learning task-specific document representations (as done by [3], [21] and our method), instead of using task-agnostic document embeddings (Doc2Vec is trained on our data but in unsupervised fashion), leads to better exploitation of the textual modality and stronger performance for both tasks. Our method is the best single-task method for both tasks, outperforming prior art, in part due to the use of convolutions (discussed previously). The success of our model addresses RQ1 and indicates that popularity and reliability can indeed be estimated from content alone (textual and visual features) with reasonable accuracy, without needing to rely on meta-data (network features). We also observe our proposed multi-task approach improves T1 accuracy by 0.4%, indicating that even though these two tasks seem unrelated, optimizing them jointly enables learning more informative feature representations.

RQ2: Predictive signals from texts and images. We conduct another experiment to identify which source(s) of information are useful in predicting article popularity and source reliability. We use the single-task version of our architecture, i.e. Ours (single-task), to see each input’s effect separately for each task. Results in Table 4 show that tweet text is the most important source of information for popularity classification, while title and image are significantly weaker (see appendix for hashtag/mention effect experiment). One possible explanation could be that articles may share very similar titles and images regardless of popularity as all of them are related to the same topic, COVID-19. For example, images that portray the US President holding a news conference can be found on both sides of popularity.

On the other hand, while the article title is the most important input for source reliability, all inputs carry useful signals. Adding tweets to the inputs improves performance over title only by 3.6%, and adding the image adds an additional 3.4% in accuracy. These results may indicate news sources have a unique way of conveying information through images and titles, and this distinction persists among user-generated content shared along with articles.

Experiments in this section answer RQ2, concluding that tweet text and article title are the most important sources of information for T1 and T2, respectively.
Table 4: Importance of inputs for popularity (T1) and reliability (T2). The method with the best accuracy is bolded, second-best is underlined, and third-best is italicized.

<table>
<thead>
<tr>
<th>Method</th>
<th>T1: Popularity</th>
<th>T2: Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image only</td>
<td>54.2% ± .008</td>
<td>62.2% ± .009</td>
</tr>
<tr>
<td>Title only</td>
<td>54.8% ± .008</td>
<td>71.0% ± .009</td>
</tr>
<tr>
<td>Tweet only</td>
<td>70.6% ± .008</td>
<td>67.2% ± .010</td>
</tr>
<tr>
<td>Title + Tweet only</td>
<td>70.7% ± .009</td>
<td>74.6% ± .008</td>
</tr>
<tr>
<td>Image + Title + Tweet</td>
<td>70.8% ± .009</td>
<td>78.0% ± .009</td>
</tr>
</tbody>
</table>

Table 5: Top-10 tweet tokens in each task class.

<table>
<thead>
<tr>
<th>Top-10 Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Popular</td>
</tr>
<tr>
<td>Boris, hanks, mail, johnson, vp, declares</td>
</tr>
<tr>
<td>Unpopular</td>
</tr>
<tr>
<td>Wuhan, smartnews, toll, yahoo, positive, chinese, worldtruthtv, research, report, lines</td>
</tr>
<tr>
<td>Reliable</td>
</tr>
<tr>
<td>Stay, social, hong, home, face, wearing, workers, que, safe, care</td>
</tr>
<tr>
<td>Unreliable</td>
</tr>
<tr>
<td>Wire, caller, mail, donald, hedge, aag, president, mike, bernie, white</td>
</tr>
</tbody>
</table>

Figure 3: Per-token attention scores in example titles, scores sorted in descending order of unreliability importance. Figure best viewed in color, zoom. See appendix for original titles.

Visualizing important regions. One advantage of having a multi-task architecture is that one can pinpoint important parts of the inputs for each task within the same model, because the exact same input representation is used to perform different tasks. In this work, we combine Grad-CAM [41] and SmoothGrad [45] to visualize important regions for the model’s predictions and show how these regions change across tasks and their classes (popular/not, reliable/not).

Grad-CAM uses gradient information to build class-discriminative localization maps. It calculates an importance score for each feature map by performing global average pooling on back-propagated gradients and then takes linear combinations of forward feature maps using their importance scores. To prevent rapid gradient fluctuations within local structures, SmoothGrad computes a stochastic approximation to Gaussian smoothing by averaging gradients for multiple noisy versions of the input. As our feature extractors for textual inputs are also CNNs, we use the same technique to visualize important parts of the input text.

For article titles (Fig. 3), we observe that sentence fragments which can be associated with oppression (e.g. “censoring” and “suppress” in [e, g]), conspiracy (e.g. “china falsified”, “secretly” and “spying,” in [a, c, d]), decline in economic activity (e.g. “shares crash” and “sales crash” in [f, h]) or ridiculing and portraying COVID-19 as a hoax (e.g. “billion-jillion” and “might lower” in [b, i]) become important for classifying an article as unreliable. On the other hand, China-related tokens are linked to unpopularity (e.g. “Wuhan”, “China”, “Chinese” in [c, f, g]). Interestingly, our model puts very little attention on title when classifying an article as reliable or popular, and relies on other inputs.

Next, Fig. 4 shows smoothed Grad-CAM output for 18 article images. For each image, from top to bottom, we show important regions for classifying an article as popular, unpopular, reliable, and unreliable, respectively. In the top row, we show images with Chinese flag [a-d], charts [e-g], and comics [h-i]. We observe that stars in Chinese flag are used to predict these images coming from unreliable sources [a-d]. In [e-g], charts are consistently associated with unreliability and often with popularity, signaling that unreliable sources use chart visuals while talking about economic impact of the pandemic and these visuals attract the audience. Similarly in [h-i], comics are associated with being both popular and unreliable, revealing another successful strategy used by unreliable sources to make their articles more noticeable when shared on Twitter.

In the second row of Fig. 4, we show images with 3-D models of coronavirus [j-k], pipettes and needles [l-o], and large texts [p-r], all associated with being unreliable. In [l-o], however, pipettes and needles are also tied to popularity, probably because the types of unreliable articles these images can belong to (e.g. anti-vaccine, COVID being lab-made) draw people’s attention more easily.

Finally in Table 5, we report the 10 tweet tokens with largest average attention score in each task class. Results show that while prevention-related tokens are associated with the shared article being reliable, political tokens are mostly tied to being unreliable. It is also clear that certain emojis indicate article popularity.

RQ3: Difference in cross-modal relationship between reliable and unreliable domains. The social media posts we examine construct meaning from multiple modalities, i.e. tweet, title and...
A combination of signs and symbols were observed in images from unreliable sources, e.g., national symbols [a-d], charts [e-g], comics [h-i], 3-D models of coronavirus [j-k], pipettes and needles [l-o], and large texts [p-r].

We next examine how the textual and visual components relate to each other, and how their relationship differs between reliable and unreliable samples. We learn two separate cross-modal embedding spaces (using Fig. 2(b) but different training data) for each domain: one using only reliable (green) and another using only unreliable articles (red, Table 1). These models allow us to compare similarity across modalities (e.g., find the text that most closely matches an image). Rather than absolute performance of these models for cross-modal retrieval, we are interested in how they generalize across domains. If a model trained on domain A performs poorly when the test domain is switched from A to B, this may be because domain A contains a distortion or bias the model can exploit.

Table 6 shows the results. Regardless of which domain we train on, performance is inflated when the training and test domains are the same, and drops when testing on a different domain (drop shown in the last column). However, this performance drop is much larger when training on red (unreliable) articles—performance drops drastically when the test domain switches from red to green, i.e. the model does not generalize to the green (reliable) domain. On the other hand, cross-domain performance decrease is much smaller for the model trained on green articles. Thus, the image-text association in the unreliable domain is much less general compared with that in the reliable domain. In other words, the image-text association in the unreliable domain is more biased. This finding relates to RQ3. We complement it with another measurement and discussion in the next section.

We chose K-way retrieval to test generalization performance, as in [49], for the following reason. Semantic discrepancy between image and text of an article is generally large (e.g. an article image with people wearing masks can be paired with several different texts), so a retrieval quality metric used for semantically well-aligned modalities (e.g. image and its caption), namely Recall@K, is not suitable to assess performance. In K-way retrieval, for a query image, we
within-domain homogeneity. We measure how coherent the distribution of tweets, titles and images are in reliable and unreliable sources using maximum mean discrepancy (MMD) [10]. Given two sets of observations $X = \{x_1, x_2, ... , x_N\}$ and $Y = \{y_1, y_2, ... , y_M\}$ drawn i.i.d. from two distributions $p$ and $q$ respectively, empirical estimate of MMD is computed:

$$MMD^2[X,Y] = \frac{1}{N(N-1)} \sum_{i}^{N} \sum_{j}^{N} k(x_i, x_j) + \frac{1}{M(M-1)} \sum_{i}^{M} \sum_{j}^{M} k(y_i, y_j) - 2 \frac{1}{NM} \sum_{i}^{N} \sum_{j}^{M} k(x_i, y_j)$$

where we use the Laplace kernel, $k(x, x') = e^{-\alpha \|x-x'\|}$, in our experiments. We randomly sample 2N articles from each domain (reliable or unreliable), divide them into two N-sized sets and calculate MMD between these two sets, both of which are from the same domain. We represent article images with their 2048-D features extracted from a ResNet-50 pre-trained on ImageNet, and text inputs with 128-D Doc2Vec embeddings. We repeat the sampling process 250 times for each N value, and report average MMD.

Figure 5 shows how within-domain MMD changes for different values of N; small MMD indicates large homogeneity. For $N = 1, 000$, t-test results show that the image pool of reliable articles is more homogeneous than of unreliable articles ($t(498) = -7.46, p < 0.01$). We found no significant difference in homogeneity between tweet pools of reliable and unreliable articles ($t(498) = 1.17, p = 0.24$), and between their title pools ($t(498) = -0.34, p = 0.74$). Thus, unreliable sources are not more homogeneous than reliable ones, indicating their bias has another cause.

Homogeneity of reliable/unreliable content. In the previous section, we found that unreliable content is more biased and generalizes worse than reliable content. One hypothesis is that this bias is due to homogeneity of the unreliable content (i.e. the same ideas being propagated, so embeddings trained on these do not generalize to other data). We test this hypothesis by measuring within-domain homogeneity. We measure how coherent the distributions of tweets, titles and images are in reliable and unreliable sources using maximum mean discrepancy (MMD) [10]. Given two sets of observations $X = \{x_1, x_2, ... , x_N\}$ and $Y = \{y_1, y_2, ... , y_M\}$ drawn i.i.d. from two distributions $p$ and $q$ respectively, empirical estimate of MMD is computed:

$$MMD^2[X,Y] = \frac{1}{N(N-1)} \sum_{i}^{N} \sum_{j}^{N} k(x_i, x_j) + \frac{1}{M(M-1)} \sum_{i}^{M} \sum_{j}^{M} k(y_i, y_j) - 2 \frac{1}{NM} \sum_{i}^{N} \sum_{j}^{M} k(x_i, y_j)$$

where we use the Laplace kernel, $k(x, x') = e^{-\alpha \|x-x'\|}$, in our experiments. We randomly sample 2N articles from each domain (reliable or unreliable), divide them into two N-sized sets and calculate MMD between these two sets, both of which are from the same domain. We represent article images with their 2048-D features extracted from a ResNet-50 pre-trained on ImageNet, and text inputs with 128-D Doc2Vec embeddings. We repeat the sampling process 250 times for each N value, and report average MMD.

Figure 5 shows how within-domain MMD changes for different values of N; small MMD indicates large homogeneity. For $N = 1, 000$, t-test results show that the image pool of reliable articles is more homogeneous than of unreliable articles ($t(498) = -7.46, p < 0.01$). We found no significant difference in homogeneity between tweet pools of reliable and unreliable articles ($t(498) = 1.17, p = 0.24$), and between their title pools ($t(498) = -0.34, p = 0.74$). Thus, unreliable sources are not more homogeneous than reliable ones, indicating their bias has another cause.

6 CONCLUSION & DISCUSSION

We examined the elements of multi-modal information and misinformation on social media. We showed that the popularity and reliability of an article can be inferred with good accuracy from visual and textual content alone, without relying on expensive network or user features. We measured the impact of the visual and textual channels, as well as which segments within them (regions in images, words in tweets and titles) most contribute to the persuasive power of the articles. For instance, national symbols and conspiracy-related words become important for classifying an article as unreliable. We showed unreliable articles use imagery-text associations very differently to construct multi-modal rhetoric. This has an important implication in relevant downstream tasks: general-purpose image datasets and models cannot be readily used for combating misinformation in multi-modal content unless accounting for the bias. Our work is a step towards understanding misinformative COVID-19-related content and demonstrate that there are differential patterns of textual and visual elements in online misinformation, which suggests media literacy educators and online platforms should look at multiple modalities that shape user experience and meaning in the shared media content.

One major drawback of our approach is that it is not able to associate important regions with high-level semantic concepts. This requires a vocabulary of these concepts which is very hard to construct considering our diverse dataset. It is currently not feasible to compute a table like Table 5 for visual tokens, i.e. some frequency-based statistic over common patterns appearing in images. Unfortunately, the state-of-the-art computer vision methods are insufficient for this task in the space of COVID-related persuasion. One strategy for extracting visual tokens could be to run an off-the-shelf object detection model on article images, then count how frequently each object category is attended to by each of our four task classes. However, we found that even large-vocabulary detection models perform poorly on our data, and miss important categories (e.g. medical equipment, flags, banners, etc.). Alternatively, to avoid the need for semantic labels, we have experimented with clustering of visual inputs, but semantic/topical similarity and visual similarity are quite distinct, and visual similarity models (and clustering) do not capture the theme of each image. For example, images of a couple performing partner stunt at a park, a store front, and a government building are grouped together. Because computing semantically-aware representations for the specific domain of COVID imagery is a full-fledged ML task, we leave it as future work.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support from NSF #2027713 and AFOSR awards.
REFERENCES


A APPENDICES

In this section, we include an extra experiment to identify if hashtags/mentions have an effect on tweet texts being highly predictive for popularity, sample detection results for a set of article images outputted by a state-of-the-art object detector that builds on well-cited models [37, 38] and trained on a large dataset [25] with 600 object categories, and original titles for the examples presented in Figure 3.

Importance of hashtags and mentions. Having seen that tweet texts carry the most information for popularity classification, one can suspect that certain mentions and/or hashtags could be correlated with popularity and the model might be learning to exploit this correlation instead of focusing on the underlying message tweet authors are trying to convey. For example, could it be that articles shared by tweets mentioning "@realDonaldTrump" are mostly popular and the model just looks for this cue ignoring the rest? We perform another experiment to see whether the model exploits such hashtag and/or mention cues. In this experiment, the single-task version of our model is trained to perform popularity classification using tweet text as the only input. We train separate word embedding models for each experiment, removing corresponding elements from training data, as some words might only appear with certain hashtags or mentions, and this would greatly affect where that particular word will be embedded in the learned Word2Vec space.

Table 7 shows trimming mentions and hashtags out of tweet does not have a drastic effect on popularity classification performance. Surprisingly, removing hashtags slightly improves performance. One possible reason could be that common hashtags appearing on both sides of popularity (e.g. #COVID19, #coronavirus) may increase the noise and make the task harder. On the other hand, removing mentions from tweets causes a slight decrease in performance: this causes a loss of contextual information, since referring to people with their Twitter handles is common practice (e.g. "[@realDonaldTrump | @POTUS] holds a press conference ..." instead of "President Trump holds a press conference ...").

Sample of detected objects in article images. One possible strategy to extract visual tokens out of article images would be to apply an off-the-shelf object detector with a large object dictionary and group regions attended by our classification model based on their labels assigned by the detector. However, we have seen that even a state-of-the-art object detector fails to detect object categories that seem important in COVID context. Figure 6 shows objects detected by a state-of-the-art YOLO variant trained on the largest object detection dataset in the literature, namely Open Images.

Original titles for examples presented in Figure 3. In Figure 3, we presented 9 example of how attention is distributed among title and cause loss of context. We include original token ordering and attention score, which breaks the original token ordering and cause loss of context. We include original article titles of these examples, in the same order they appear in the figure (a-i).

Sample of detected objects in article images. One possible strategy to extract visual tokens out of article images would be to apply an off-the-shelf object detector with a large object dictionary and group regions attended by our classification model based on their labels assigned by the detector. However, we have seen that even a state-of-the-art object detector fails to detect object categories that seem important in COVID context. Figure 6 shows objects detected by a state-of-the-art YOLO variant trained on the largest object detection dataset in the literature, namely Open Images.

Original titles for examples presented in Figure 3. In Figure 3, we presented 9 example of how attention is distributed among title tokens for each task. However, we sorted tokens based on unreliability attention score, which breaks the original token ordering and cause loss of context. We include original article titles of these examples, in the same order they appear in the figure (a-i).

a. REPORT: US Intelligence Confirms China Falsified Coronavirus Death, Case Data
b. What Will the Left Do When a Billion-Jillion Americans Don’t Die of Coronavirus?
d. Sources: China increases spying on US to control coronavirus narrative
e. China Has Been Censoring Coronavirus Information for Months
f. Under Armour Shares Crash, Blames China
g. Chinese Regime Deploys 1,600 Online Trolls To Suppress Information On Coronavirus
h. China Mobile Phone Sales Crash Most On Record
i. The Coronavirus Death Rate Might Be Lower Than We Think
Figure 6: Detection results for a subset of article images. Solid boxes and attached labels denote the object instances and their semantic classes detected by a YOLO variant, and dashed boxes denote object instances we expect to be detected. We observe that flag instances were not detected [A, C, F], even though flag was one of the classes in the object detector’s vocabulary. Similarly, it fails to detect goggles [B], gloves [B, G] and bottles [E], as well as most of the person instances [B] (all in the vocabulary). Although both needle [E] and pipette [G] are out of dictionary classes, we anticipated them to be detected given there are contextually and visually similar objects in the detector’s dictionary.