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 covey 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

ACM Reference Format:
Mesut Erhan Unal, Adriana Kovashka, Wen-Ting Chung, and Yu-Ru Lin. 2022. Visual Persuasion in COVID-19 Social Media Content: A Multi-Modal Characterization. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3487553.3524647

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524647

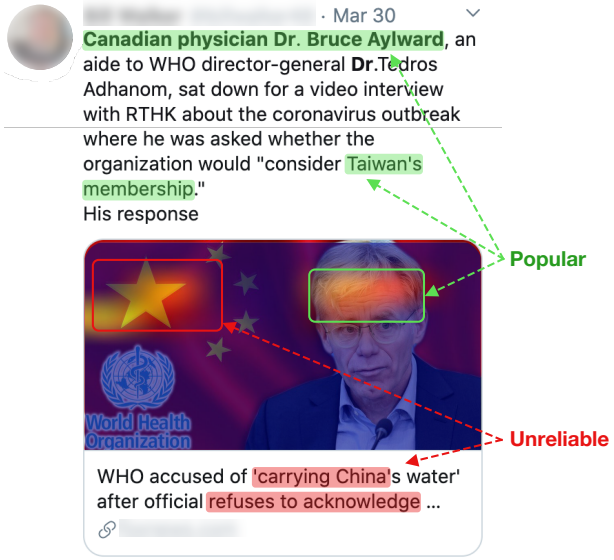

Figure 1: Our method performs article popularity and reliability classification using multi-modal cues. We highlight salient regions for the model's predictions using a gradient-based visualization technique [41]. In this example, our model associates the star in the Chinese flag, along with part of the title that has negative tone, with the tweeted article being *unreliable*. On the other hand, the forehead of a WHO officer (B. Aylward) and a part of the tweet text have been associated with the article being *popular*.

1 INTRODUCTION

From campaigns to advertising, social media content routinely incorporates multi-modal design choices that combine texts and images to effectively covey 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 visuals 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.

	Popular	Unpopular	Total
Red	934	1,149	2,083 (8.0%)
Orange	2,163	1,917	4,080 (15.7%)
Yellow	5,187	5,181	10,368 (39.8%)
Green	4,457	4,958	9,415 (36.1%)
Satire	80	32	112 (0.4%)
Total	12,821	13,237	26,058 (100%)

Table 1: Number of articles in our collection by domain coding [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.

	Popular	Unpopular	Total
Reliable	3,066 (.004, .010)	3,097 (.0, .0)	6,163
Unreliable	3,097 (.003, .008)	3,066 (.0, .0)	6,163
Total	6,163	6,163	12,326

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

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 on social media or that news articles choose to include. However, 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 image 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 drive 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):

$$\mathcal{P}_{A_i} = \frac{\sum_{t \in A_i} t_{\text{retweet}} + t_{\text{like}}}{[\sum_{t \in A_i} Q(t_{\text{author}})] + \lambda} \tag{1}$$

where t_{retweet} and t_{like} denote number of retweets/likes for tweet t in set A_i, Q the number of followers of a Twitter user, and λ 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 $\mathcal{P}$ greater than zero, and $\mathcal{P}$ for all *unpopular articles* is zero.

Choosing λ: Setting the right value for λ is important as it affects the calculated $\mathcal{P}$ values and thus the set of *popular articles*

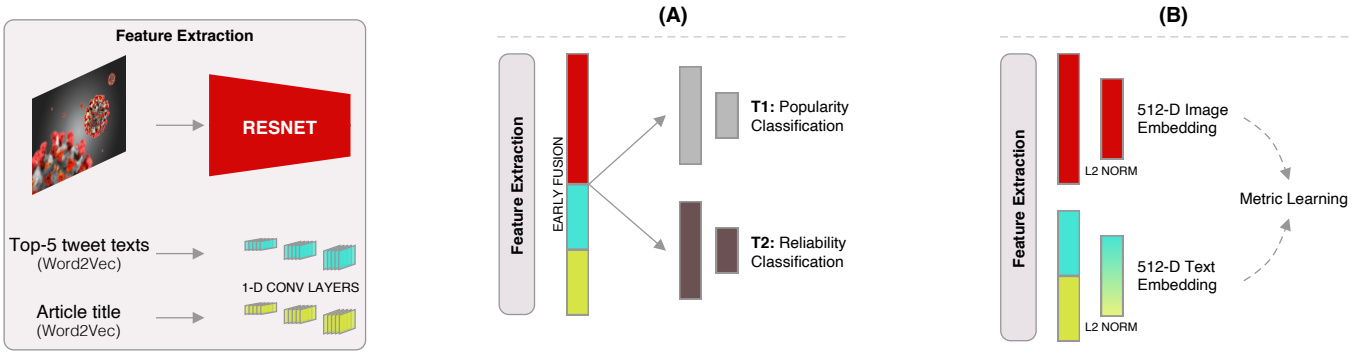

Figure 2: (A) Multi-task architecture for multi-modal popularity and reliability classification. Features from both modalities extracted through convolutional layers are fused to perform both tasks simultaneously. (B) Cross-modal relation modeling. Visual and text features from the same article are embedded 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.

(top 20%). One should expect that, with an appropriate choice of λ, the distribution of audience size in *popular articles* and in articles that gained some popularity (i.e. $\mathcal{P} > 0$) should be similar as the former is a subset of the latter. Otherwise, the chosen λ could be favoring articles with small/large audience as having higher $\mathcal{P}$. After experimenting with different values, we set λ 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 yellow and satire sources out as they cannot be perfectly associated with eliciting misinformation. We consider articles from green sources as *reliable*, and articles from either red or orange sources as *unreliable*. Lastly, we undersample reliable articles to balance our experiment dataset, and split it into fixed train/val/test with 70/10/20 ratio. Even after undersampling, our dataset is still much larger than [62] (2,017 vs 12,326 articles). Table 1 shows the number of articles that fall into each category in our data collection (before reliability label assignment) and Table 2 shows descriptive statistics of the experiment dataset. We use the latter in the classification experiments to answer RQ1&2, and a subset of the initial data collection (Table 1) in the cross-modal relation experiments to answer RQ3.

4 MODELS

Popularity and reliability classification. We describe our multi-task architecture (see Fig. 2A) to perform the binary popularity classification (**T1**) and source reliability classification (**T2**) tasks simultaneously given inputs:

- A_i^{title}: Title of the article in the generated preview,
- A_i^{tweet}: Concatenated user-generated content of top-5 tweets (retweet+like) sharing the article; we oversample if $|A_i| < 5$,
- A_i^{image}: Image of the article in the generated preview.

[1]The audience distribution is highly skewed (mean: 430,381, median: 13,824 within the initial 69,591 articles, and mean: 686,180, median: 52,743 among the 48,562 articles that gained at least one like or retweet.

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 (ϕ) and tweet texts (ψ). 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 $\vec{0} \in R^{128}$ 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 $(\mathcal{G}, \mathcal{H})$ 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 $\mathcal{G} : A^{title} \rightarrow R^{384}$ and $\mathcal{H} : A^{tweet} \rightarrow R^{384}$. We compare to alternative text representations in Sec. 5. For images, we employ ResNet-50 [12] pre-trained on ImageNet [7] as feature extractor ($\mathcal{F} : A^{image} \rightarrow R^{2048}$). 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:

$$\mathcal{L}(\theta) = \mathcal{L}_{T_1}(\theta) + \mathcal{L}_{T_2}(\theta) \tag{2a}$$

$$\mathcal{L}_{T_1}(\theta) = -\sum y_p \log(\hat{p}_p) + (1 - y_p) \log(1 - \hat{p}_p) \tag{2b}$$

$$\mathcal{L}_{T_2}(\theta) = -\sum y_r \log(\hat{p}_r) + (1 - y_r) \log(1 - \hat{p}_r) \tag{2c}$$

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

The intuition for using convolutions for text is that popularity and reliability may be inferrable 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 1×10^{-4} and decrease it by $\times 0.1$ if validation loss does not improve in the last

	T1: Popularity	T2: Reliability
IMAGE+DOC2VEC-FUSION	63.1% (± .010)	61.9% (± .010)
IMAGE+DOC2VEC-GRU	65.0% (± .008)	63.4% (± .009)
BIELSKI & TRZCINSKI [3]	69.8% (± .009)	73.1% (± .009)
KHATTAR ET AL. [21]	70.3% (± .032)	69.2% (± .009)
OURS (SINGLE-TASK)	<u>70.8%</u> (± .008)	**78.0%** (± .009)
OURS (MULTI-TASK)	**71.2%** (± .008)	**78.0%** (± .008)

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.

four epochs. We use early stopping to terminate training if the validation loss does not improve in the last six epochs. We use the Adam [23] optimizer with default parameters of $\beta_1 = 0.9, \beta_2 = 0.999$.

Cross-modal relation modeling. We next describe our architecture (Fig. 2B) for learning a cross-modal embedding space wherein a paired (belonging to the same article) visual and textual data resides closer than an unpaired one. This embedding enables analysis of the link between modalities in terms of the message they convey, and the different ways in which multi-modal meaning is constructed in articles with different labels. We employ an ImageNet pre-trained ResNet-50 followed by a linear transformation as the image embedding branch ($\mathcal{F} : A^{image} \rightarrow R^{512}$) and two Text-CNNs followed by a concat and a linear transformation as the text embedding ($\mathcal{G} : A^{title} \times A^{tweet} \rightarrow R^{512}$). Outputs of these branches are then L2-normalized to place embeddings on the surface of a 512-D unit hypersphere. To optimize our model, we minimize an N-pairs loss [46]:

$$\mathcal{L} = \sum_{A_i, A_j \in \text{minibatch}, i \neq j} \mathcal{L}_{trip}(A_i, A_j) \tag{3a}$$

$$\mathcal{L}_{trip}(A_i, A_j) = [\|\mathcal{F}(A_i^{image}) - \mathcal{G}(A_i^{title}, A_i^{tweet})\|^2 - \|\mathcal{F}(A_i^{image}) - \mathcal{G}(A_j^{title}, A_j^{tweet})\|^2 + \alpha]_+ \tag{3b}$$

where $\mathcal{L}_{trip}$ denotes the triplet loss [40] commonly used for learning cross-modal representations. For each article in a minibatch, we take the article image (A_i^{image}) as anchor, paired text (A_i^{title}, A_i^{tweet}) as positive and all other article texts (A_j^{title}, A_j^{tweet}) from minibatch as negatives (hence *N-pairs*), and accumulate the loss for each negative that violates the margin α. We use the same hyperparameters and training strategy as for popularity and reliability classification, and set the margin α to 0.5.

5 EXPERIMENTS

We describe the experiments conducted in order to answer our research questions with empirical evidence.

RQ1: Multimodal prediction of popularity and reliability. The first experiment aims to verify the appropriateness of the architecture we use, by comparing it with several other multi-modal, single-task baselines described below. We train two instances for each baseline, one for each task.

- **IMAGE+DOC2VEC-FUSION**: Uses 128-D Doc2Vec [26] embeddings for article title and tweet texts, then fuse them with the image feature for classification, similar to [48].
- **IMAGE+DOC2VEC-GRU**: Employs two GRU [6] cells, Title-GRU and TweetGRU, as write function for messages passed from document embeddings to the image feature, then uses the average final GRU states to classify.
- **BIELSKI & TRZCINSKI [3]**: A popularity classification method that uses self-attention on visual and textual features before fusion.
- **KHATTAR ET AL. [21]**: A reliability classification method that learns cross-modal correlations at the bottleneck layer of a multi-modal variational autoencoder.

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.

	T1: Popularity	T2: Reliability
IMAGE ONLY	54.2% (± .008)	62.2% (± .009)
TITLE ONLY	54.8% (± .008)	*71.0%* (± .009)
TWEET ONLY	*70.6%* (± .008)	67.2% (± .010)
TITLE + TWEET ONLY	70.7% (± .009)	74.6% (± .008)
IMAGE + TITLE + TWEET	**70.8%** (± .008)	**78.0%** (± .009)

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.

	Top-10 Tokens
Popular	boris, 👏, hanks, mail, johnson, vp, 🤣, 🚨, declares, 💥
Unpopular	wuhan, smartnews, toll, yahoo, positive, chinese, worldtruthtv, research, report, lines
Reliable	stay, social, hong, home, face, wearing, workers, que, safe, care
Unreliable	wire, caller, mail, donald, hedge, aag, president, mike, bernie, white

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

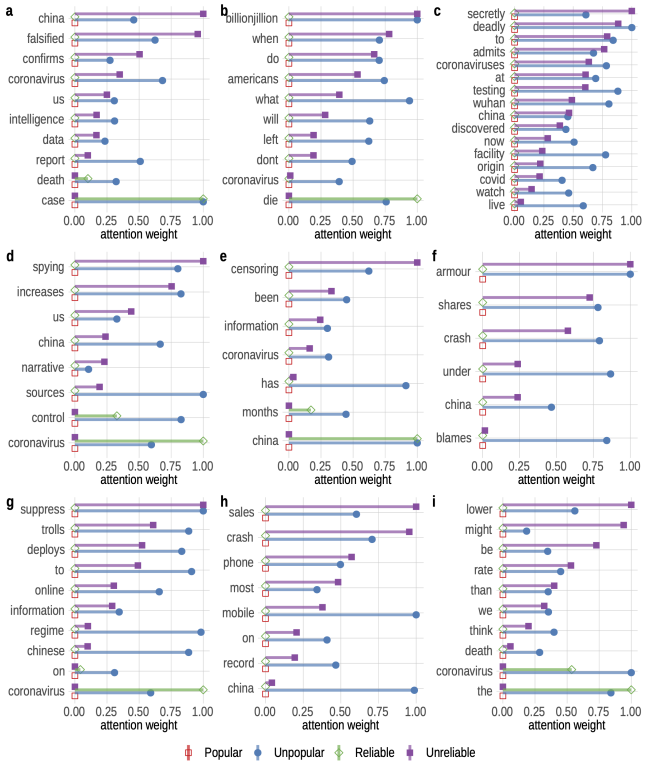

Figure 3: Per-token attention scores in example titles, scores sorted in descending order of unreliability importance. Figure best viewed in color, zoom. See appendix for original titles.

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

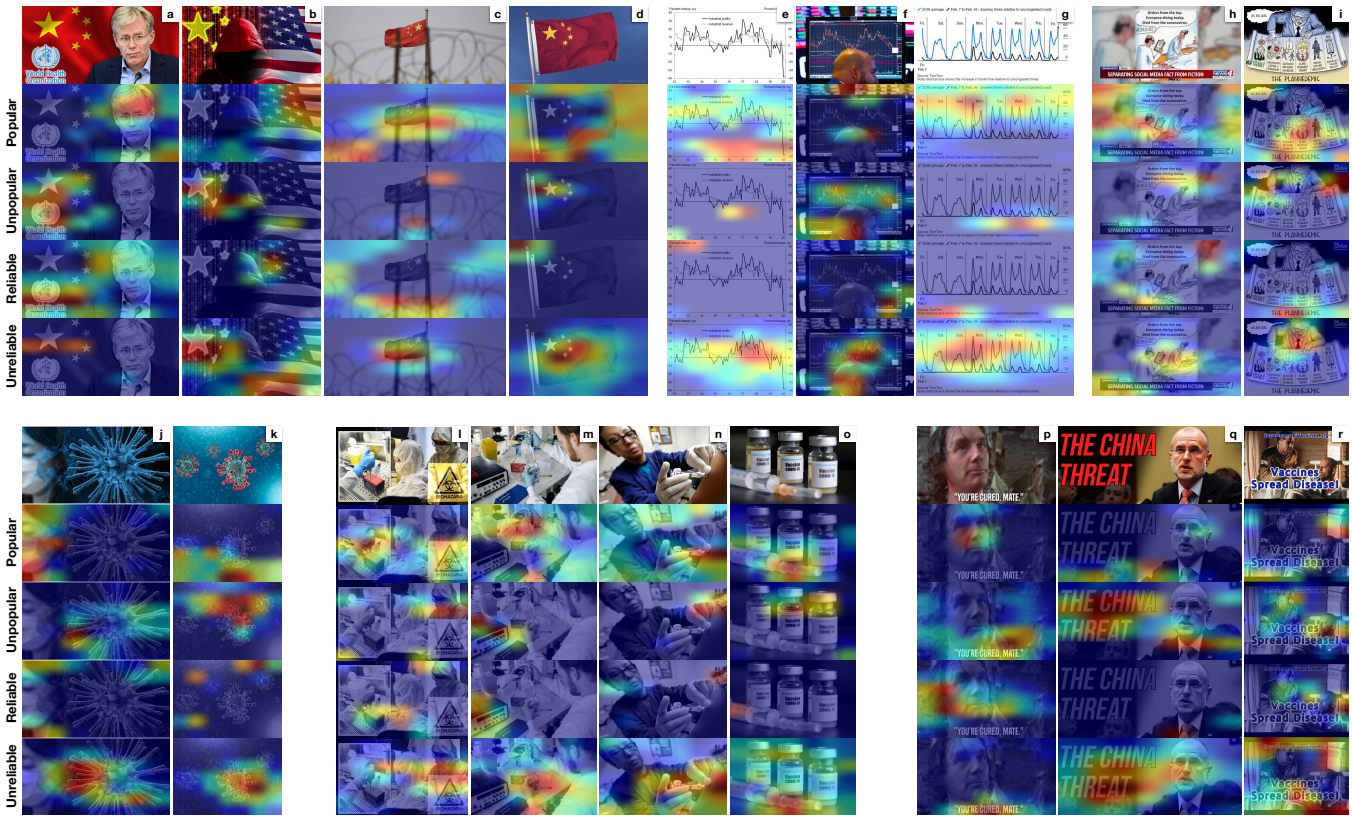

Figure 4: Salient image regions for predicting different classes (popular/not, reliable/not), highlighted with smoothed Grad-CAM. A combination of signs and symbols were observed in images from unreliable sources, e.g., national symbols [a-d], charts [e-g], comics [h-i], 3-D models of coronavirus [j-k], pipettes and needles [l-o], and large texts [p-r].

image. 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.

		Test Domain		
		Red	Green	Cross-domain diff.
Train Domain	Red	**3-way:** .516 **5-way:** .363 **10-way:** .215	**3-way:** .471 **5-way:** .317 **10-way:** .174	−.045 (−8.72%) −.046 (−12.67%) −.041 (−19.07%)
	Green	**3-way:** .489 **5-way:** .338 **10-way:** .198	**3-way:** .493 **5-way:** .346 **10-way:** .207	−.004 (−0.81%) −.008 (−2.31%) −.009 (−4.35%)

Table 6: K-way cross-modal retrieval test results. Numbers in parentheses indicate relative gain/loss for cross-domain testing.

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

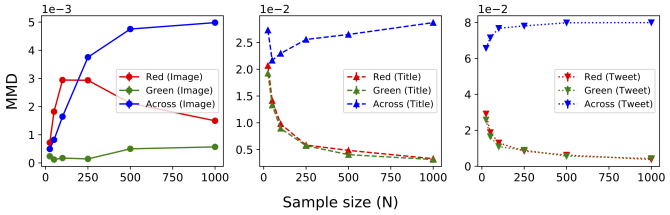

Figure 5: MMD within green, within red, and between green and red articles w.r.t. sample size, for image inputs (left), titles (middle), and tweets (right). Discrepancy within green article images is significantly smaller than it is within red article images. On the other hand, we find no significant difference in within-domain discrepancy for other input types.

choose paired text as positive, and randomly sample $K-1$ article texts as negatives, then check whether the positive is the closest to the query image among K. All models get the same negative set for the same query. The green training set is undersampled to match the size of the red training set.

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:

$$MMD^2[X,Y] = [\frac{1}{N(N-1)}\sum_{i}^{N}\sum_{j\neq i}^{N}k(x_i,x_j) + \frac{1}{M(M-1)}\sum_{i}^{M}\sum_{j\neq i}^{M}k(y_i,y_j) - \frac{2}{NM}\sum_{i}^{N}\sum_{j}^{M}k(x_i,y_j)] \quad (4)$$

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**.

Findings in this and the previous section answer our **RQ3** and show that unreliable and reliable articles construct meaning in different ways. However, generalization performance of a metric learning (embedding) model trained on unreliable articles is much worse than the one trained on reliable articles, indicating unreliable articles are distorted and biased. This bias is not because unreliable articles are more homogeneous (less diverse and broad) than reliable ones.

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 image-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

[1] Ioannis Arapakis, B. Barla Cambazoglu, and Mounia Lalmas. 2014. On the Feasibility of Predicting News Popularity at Cold Start. In *Lecture Notes in Computer Science*. Springer International Publishing, 290–299.

[2] Roja Bandari, Sitaram Asur, and Bernardo Huberman. 2012. The pulse of news in social media: Forecasting popularity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 6. 26–33.

[3] Adam Bielski and Tomasz Trzcinski. 2018. Understanding multimodal popularity prediction of social media videos with self-attention. *IEEE Access* 6 (2018), 74277–74287.

[4] Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. 2014. Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*.

[5] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance* 6 (2020).

[6] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*. 1724–1734.

[7] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[8] Charles Forceville. 2002. *Pictorial metaphor in advertising*. Routledge.

[9] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. 2015. Image popularity prediction in social media using sentiment and context features. In *Proceedings of the 23rd ACM International Conference on Multimedia*.

[10] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13, 1 (2012), 723–773.

[11] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. *Science* 363, 6425 (2019), 374–378.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[13] Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[14] Benjamin D Horne, William Dron, Sara Khedr, and Sibel Adali. 2018. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Companion Proceedings of the The Web Conference*.

[15] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*.

[16] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia* 19, 3 (2016), 598–608.

[17] Peiguang Jing, Yuting Su, Liqiang Nie, Xu Bai, Jing Liu, and Meng Wang. 2017. Low-rank multi-view embedding learning for micro-video popularity prediction. *IEEE Transactions on Knowledge and Data Engineering* 30, 8 (2017), 1519–1532.

[18] Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. 2014. Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[19] Jungseock Joo, Francis F Steen, and Song-Chun Zhu. 2015. Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE International Conference on Computer Vision*.

[20] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[21] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *The World Wide Web Conference*.

[22] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

[24] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *CoRR* abs/1411.2539 (2014). http://arxiv.org/abs/1411.2539

[25] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4. *International Journal of Computer Vision* 128, 7 (2020), 1956–1981.

[26] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*.

[27] Dongliang Liao, Jin Xu, Gongfu Li, Weijie Huang, Weiqing Liu, and Jing Li. 2019. Popularity prediction on online articles with deep fusion of temporal process and content features. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[28] Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A Multimodal Framework for the Detection of Hateful Memes. arXiv:2012.12871 [cs.CL]

[29] Mayank Meghawat, Satyendra Yadav, Debanjan Mahata, Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. 2018. A multimodal approach to predict social media popularity. In *IEEE Conference on Multimedia Information Processing and Retrieval*.

[30] Paul Messaris. 1997. *Visual persuasion: The role of images in advertising*. Sage.

[31] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

[32] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673* (2019).

[33] John O'Shaughnessy and Nicholas O'Shaughnessy. 2004. *Persuasion in advertising*. Routledge.

[34] Alicja Piotrkowicz, V. Dimitrova, Jahna Otterbacher, and K. Markert. 2017. Headlines Matter: Using Headlines to Predict the Popularity of News Articles on Twitter and Facebook. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[35] David Pogue. 2017. How to Stamp Out Fake News. *Scientific American* 316, 2 (2017), 24–24.

[36] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[38] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.

[39] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the ACM on Conference on Information and Knowledge Management*.

[40] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[41] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*.

[42] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data* 8, 3 (2020), 171–188.

[43] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2019. The Role of User Profiles for Fake News Detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 436–439.

[44] Jeremy Singer-Vine. 2016. Fact-Checking Facebook Politics Pages. https://github.com/BuzzFeedNews/2016-10-facebook-fact-check/tree/master/data.

[45] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[46] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*.

[47] Yan Su. 2021. It doesn't take a village to fall for misinformation: Social media use, discussion heterogeneity preference, worry of the virus, faith in scientists, and COVID-19-related misinformation beliefs. *Telematics and Informatics* 58 (2021), 101547. https://doi.org/10.1016/j.tele.2020.101547

[48] Christopher Thomas and Adriana Kovashka. 2019. Predicting the Politics of an Image Using Webly Supervised Data. In *Advances in Neural Information Processing Systems 32*.

[49] Christopher Thomas and Adriana Kovashka. 2020. Preserving Semantic Neighborhoods for Robust Cross-modal Retrieval. In *Proceedings of the European Conference on Computer Vision*.

[50] Tomasz Trzciński, Paweł Andruszkiewicz, Tomasz Bocheński, and Przemysław Rokita. 2017. Recurrent neural networks for online video popularity prediction. In *International Symposium on Methodologies for Intelligent Systems*.
[51] Riza Velioglu and Jewgeni Rose. 2020. Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge. arXiv:2012.12975 [cs.AI]
[52] Ke Wang, Mohit Bansal, and Jan-Michael Frahm. 2018. Retweet wars: Tweet popularity prediction via dynamic multimodal regression. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
[53] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
[54] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
[55] Judith Williamson. 1978. *Decoding advertisements*.
[56] Bo Wu and Haiying Shen. 2015. Analyzing and predicting news popularity on Twitter. *International Journal of Information Management* 35, 6 (2015), 702–711.
[57] Liang Wu and Huan Liu. 2018. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*.
[58] Nan Xi, Di Ma, Marcus Liou, Zachary C Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. 2020. Understanding the Political Ideology of Legislators from Social Media Images. In *Proceedings of the International AAAI Conference on Web and Social Media*.
[59] Jeewoo Yoon, Jungseock Joo, Eunil Park, and Jinyoung Han. 2020. Cross-Domain Classification of Facial Appearance of Leaders. In *International Conference on Social Informatics*.
[60] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[61] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. 2018. User-Guided Hierarchical Attention Network for Multi-Modal Social Image Popularity Prediction. In *Proceedings of the 2018 World Wide Web Conference*.
[62] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*. 3205–3212.
[63] Alireza Zohourian, Hedieh Sajedi, and Arefeh Yavary. 2018. Popularity prediction of images and videos on Instagram. In *4th International Conference on Web Research*.

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 ...").

	T1: Popularity
Tweet w/o hashtags	**71.0%** (± .009)
Tweet w/o mentions	70.0% (± .008)
Tweet w/o hashtags+mentions	70.2% (± .007)
Tweet only	70.6% (± .008)

Table 7: Change in popularity classification accuracy when hashtags and mentions are stripped out of tweet text.

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?
c. Origin of COVID-19 Discovered? China Now Admits To Secretly Testing Deadly Coronaviruses At Wuhan Facility — Watch Live
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.