Cross-Platform Multimodal Misinformation: Taxonomy, Characteristics and Detection for Textual Posts and Videos

Nicholas Micallef,[1] Marcelo Sandoval-Castañeda,[1] Adi Cohen,[2] Mustaque Ahamad,[3] Srijan Kumar,[3] Nasir Memon[4]

[1]New York University Abu Dhabi, Abu Dhabi, United Arab Emirates
[2]Memetica, New York, New York, United States
[3]Georgia Institute of Technology, Atlanta, Georgia, United States
[4]New York University, Brooklyn, New York, United States
{nicholas.micallef,marcelo.sc}@nyu.edu, cohenadico@gmail.com, {mustaque.ahamad,srijan}@gatech.edu, memon@nyu.edu

Abstract

Social media posts that direct users to YouTube videos are one of the most effective techniques for spreading misinformation. However, it has been observed that such posts rarely get deleted or flagged. Since multi-modal misinformation that leads to compelling videos has more impact than using just textual content, it is important to characterize and detect such textual post and video pairs to prevent users from becoming victims of misinformation. To address this gap, we build a taxonomy of how links to YouTube videos are used on social media platforms. We then use pairs of posts and videos annotated with this taxonomy to test several classification models built using cross-platform features. Our work reveals several characteristics of post-video pairs, in terms of how posts and videos are related to each other, the type of content they share, and their collective outcome. In addition, we find that traditional approaches to misinformation detection that rely only on text from posts miss a significant number of post-video pairs that contain misinformation. More importantly, we find that to reduce the spread of misinformation via post-video pairs, classifiers would be more effective if they are designed to use data and features from multiple diverse platforms.

Introduction

A common technique used to spread misinformation in social media platforms is to use cross-platform links to increase reach and evade detection by a single platform. For example, the tweet shown in Figure 1 does not contain explicit misinformation but it does include links that lead to a YouTube video with misinformation. In fact, Micallef et al. 2020 found that at least 7% of COVID-19 tweets about 5G direct to a YouTube video. Besides being fairly common (Yang et al. 2021; Knuutila et al. 2020), cross-platform links to such multi-modal misinformation have been shown to have more impact than using only textual content (Hameleers et al. 2020; Zannettou et al. 2018). Hence, to reduce the impact of post and video pairs that spread misinformation, it is important to detect their content and warn users. However, a very low number ($< 1\%$) of post-video pairs that contain misinformation get deleted or flagged (Knuutila et al. 2020). Moreover, it takes an average of 41 days for YouTube to delete COVID-19 related misinformation videos (Knuutila et al. 2020). This gives misinformation spreaders ample time to increase the reach of their videos by sharing them as links in posts on multiple social media platforms (Yang et al. 2021).

Figure 1: Example of post-video pairs.

A key impediment to detecting misinformation in post and video pairs is that standard text-based misinformation detection algorithms (Hossain et al. 2020; Micallef et al. 2020) are not effective and even impaired.[1] These classifiers only consider the text within the post, which might not be enough to determine whether the linked video contains misinformation. Moreover, these classifiers are designed to focus on content from a single platform, as social media platforms are mostly isolated from each other (Driscoll and Thorson 2015). This presents a challenge because coordinated misinformation campaigns leverage multiple platforms to evade detection and sustain themselves through complementary use of social media platforms (Horawalavithana, Ng, and Iamnitchi 2020; Wilson and Starbird 2020). Finally, classifiers (Medina Serrano, Papakyriakopoulos, and Hegelich 2020; Hou et al. 2019) may not be effective in detecting misinformation videos[1] because they do not consider the text written in the post that is sharing the video (e.g., the video might be factual, but the sharing post might be twisting the content of the video to spread misinformation). To address this gap in detecting misinformation in post and video pairs, and to characterize how these videos are used to spread misinformation, we test several classification models that use

[1]i.e., F1 score < 0.5, see Classifiers Setup and Experiments

cross-platform features (e.g., text in post from Facebook and video description from YouTube).

In addition to misinformation, we also consider counter-misinformation (e.g., COVID-19 is real and is not a normal virus that is being used to cover the effects of 5G) because previous work found that the inclusion of counter-misinformation messages can increase the effectiveness of misinformation detection (Micallef et al. 2020). Counter-misinformation is still an under-studied topic and including it in our work allows us to also characterize how videos are used to counter misinformation. In this way, this work starts to shed light on how platforms could leverage counter-misinformation post-video pairs to combat misinformation.

We use a data-driven approach by collecting data from three platforms (i.e., Twitter, Facebook, Reddit). This is motivated by our desire to address limitations of prior misinformation detection studies, which predominantly study individual platforms. We select these platforms due to their diverse properties and characteristics. For instance, Twitter allows only short-form content (up to 280 characters) and is not community-oriented because users follow individual accounts (or a group of people managing an account) rather than a whole community. In contrast, Reddit allows long form content with a limit of 40,000 characters per post. Moreover, Reddit is a community-oriented platform, in which members submit content that is accessible to other members, regardless of whether they follow each other or not. Facebook shares characteristics with both Twitter and Reddit. Similar to Reddit, Facebook allows long form content with a limit of up to 63,206 characters per post and has public groups, which users can join. The community characteristics of public groups are similar to Subreddits. In addition, Facebook has public pages, in which users follow a person or a group of people that manage a page. These pages are similar to following a Twitter account. Consequently, we study how post and video pairs are used on these three platforms. In this work, we make the following contributions:

- We develop a taxonomy of how YouTube videos are used to spread and counter misinformation by posting on other platforms. We find that most videos either support posts or are taken out of context. In the taxonomy, we categorize the type of content present in the video. We find that most videos contain explicit misinformation or explicit counter-misinformation, with only a few containing implicit misinformation.
- We found that posts are often created to drive traffic to newly uploaded videos rather than surfacing older videos. In the datasets analyzed, 56% of the posts appeared within 5 days from when the videos were uploaded.
- We develop and test several classifiers to detect misinformation in post-video pairs. To train the classifiers, we used features from 3,000 posts related to COVID and 5G (from Twitter, Facebook, and Reddit) and 991 YouTube videos hand-labeled by two researchers. We tested various standard and deep learning classifiers with different combinations of textual features (e.g., post & video title). We find that a multi-task learning classifier, composed of fine-tuned BERT connected to a Bi-LSTM, which takes post and video transcription features as input, achieves the best performance, with an F1 score and precision $>$ 0.74.
- Our analysis uncovers differences in the characteristics of post and video pairs that spread misinformation as opposed to those that counter it. We find that counter-misinformation post-video pairs are less commonly used, but seem to mostly be created to direct people to newly uploaded explicit counter-misinformation videos. In contrast, misinformation posts and videos are more frequent, direct users to both newly created and older videos, which mostly have explicit and sometimes implicit content.
- Our findings uncover that machine learning classifiers that only consider text from posts to identify misinformation miss a significant number of post-video pairs that contain misinformation, which could explain why post-video misinformation remains undetected (Knuutila et al. 2020). Moreover, collecting data from platforms with diverse properties and characteristics helped us uncover similarities and differences in how post and video pairs are used across platforms (see Table 4).
- Our experiments reveal that detection of post and video pairs that contain misinformation could be significantly improved by using cross-platform features and data from multiple platforms. Since we did not experience a decrease in classifier performance when using diverse data, it could be inferred that using data from multiple platforms not only strengthens the robustness of the classifier, it also decreases the time required to train misinformation detection models for new topics.

In the next section, we describe prior work that studied cross-platform information, developed multi-modal misinformation detection models and labeled post-video pairs. Then, we describe the processes used to collect data from three platforms, create a taxonomy of how posts use videos and annotate this data. Afterwards, we analyze the annotated post-video pairs and conduct experiments to test various configurations of cross-platform classifiers. Finally, we discuss the insights that emerged from our experiments and analysis, together with future directions.

Related Work

Cross-Platform Information Flow

In recent years, there has been much work investigating cross-platform information flow. For instance, Hunt, Wang, and Zhuang 2020 found that cross-platform sourcing was more frequent between Twitter and traditional web sites, such as news agencies, e.g., Washington Post, than between Twitter and social media platforms such as Instagram, YouTube, and Facebook. Another work that focused on studying white helmets (i.e., Syria Civil Defense) disinformation campaigns found that YouTube videos are heavily disseminated by coordinated campaigns from the same users (Horawalavithana, Ng, and Iamnitchi 2020) and that these coordinated campaigns sustain themselves through

consistent and complementary use of social media (Wilson and Starbird 2020). More recently, when studying pandemic misinformation on Twitter and Facebook, Yang et al. 2021 found that few Twitter accounts and Facebook pages exhibit a strong influence on each platform.

Prior research has also found that a prevalent cross-platform technique used by those who spread misinformation is to create posts to drive traffic to YouTube videos (Yang et al. 2021). Knuutila et al. 2020 also found that COVID-related misinformation videos attracted audiences by being shared on Facebook. Micallef et al. 2020 found a similar phenomenon occurring on Twitter for misinformation and counter-misinformation. This high prevalence of post and video pair content could be explained by multi-modal misinformation being found to be more credible and spreading further than textual misinformation (Hameleers et al. 2020; Zannettou et al. 2018). Only limited research has investigated characteristics of cross-platform post-video misinformation and counter misinformation, and how it is used on multiple platforms.

Multi-Modal Detection Algorithms

Several approaches have been explored to detect multi-modal misinformation (Alam et al. 2021; Agrawal, Gupta, and Narayanan 2017; Hou et al. 2019; Wang, Yin, and Argyris 2020). Unsupervised models have been used to study coherence between text and images. Yang et al. 2019 leveraged the credibility of users that spread multi-modal information to determine the trustworthiness of shared content. Due to limited availability of labeled multi-modal content, other work used semi-supervised methods to detect misinformation. For instance, Bansal et al. 2021 combined exogenous and endogenous signals with a semi-supervised co-attention network to detect COVID-19 misinformation. Supervised learning methods were used to detect multi-modal misinformation. Wang et al. 2018 developed an event adversarial neural network (EANN) to detect emerging misinformation. Qi et al. 2019 developed a multi-modal variational autoencoder to encode text and images to detect multi-modal misinformation. Recently, Qian et al. 2021 developed a supervised model that combines multi-modal context information and the hierarchical semantics of text, to leverage inter-modality and intra-modality relationships.

These classifiers have the limitation of catering only for text-image misinformation and ignoring text-video misinformation. In addition, they do not consider the relationship between different components (e.g., how posts use images). Hence, there is a gap in methods that characterize post-video pairs. We address this gap by testing different text-based classifiers, using standard and deep learning models, to leverage cross-platform features and the relationship between the components (e.g., how posts use videos) to detect post-video pairs that contain misinformation.

Labeling Post-Video Pairs

The development of a supervised classifier requires the manual labeling of data. Prior work labeled YouTube videos to study the cross-platform dynamics of events, such as the Occupy Movement (Thorson et al. 2013). Moreover, Washington Post developed a universal language to identify manipulated online videos (Kessler 2019), which includes missing context, deceptive editing, and malicious transformation. This labeling focuses on the tactics used in manipulated videos and does not consider other types of content that could be found in videos (e.g., videos that implicitly spread misinformation or videos that refute misinformation). To address this gap, in our research, we use an iterative approach to build a taxonomy of how YouTube videos are used to spread and counter misinformation on other platforms and the type of misinformation that these videos contain.

Data Collection

We began by collecting data on one of the most popular COVID-19 misinformation topics: *COVID-19 – 5G conspiracy theories* (Brennen et al. 2020). We select this topic because it is a polarizing topic that attracted plenty of controversy (Destiny 2021), which instigated violence,[2] and brought harm to society.[3] Below, we describe our process of finding posts on this topic that direct to YouTube videos, from three popular platforms on which misinformation is spread: Twitter, Facebook, and Reddit (Alam et al. 2021).

Twitter

We used a publicly available dataset of tweets[4] published by Micallef et al. 2020. The dataset was collected from January 21, 2020 to May 20, 2020 and contains English tweets related to *COVID-19 - 5G conspiracy theories* with both misinformation and counter-misinformation messages. It was constructed by selecting tweets that have at least one COVID-19 keyword (i.e., COVID-19 , covid, corona virus, coronavirus) along with the keyword 5G.

For our experiments, the dataset was filtered to include only tweets that have links to YouTube videos. This process reduced 50,835 tweets to 3,725 (i.e., 7%). For the taxonomy and annotation process, we randomly picked 2,000 tweets.

Facebook

We utilized the Academic version of CrowdTangle[5] to extract Facebook posts related to *COVID-19 - 5G conspiracy theories* (Team 2021). For consistency, we used the same keywords used in Micallef et al. 2020. The web-based tool includes a built-in feature that retrieves English posts that contain links to YouTube videos. Collecting posts from January 21, 2020 to May 20, 2020 returned less than 2,000 posts. Such a low number of posts was returned because CrowdTangle follows a strict set of guidelines to select the pages and groups to follow.[6] In addition, CrowdTangle does not return posts from personal profiles and posts from non-public groups. Since we wanted to analyze the same amount of posts that we collected for Twitter, we extend the data

[2] https://www.bbc.com/news/newsbeat-52395771

[3] https://www.bbc.com/news/uk-england-devon-58760598

[4] http://claws.cc.gatech.edu/covid_counter_misinformation.html

[5] https://www.crowdtangle.com/resources

[6] https://help.crowdtangle.com/en/articles/1189612-crowdtangle-api

collection timeline to December 31st, 2020. This process returned 1,209 posts from public Facebook pages and 3,260 from public Facebook groups. We randomly picked 2,000 of these posts for the taxonomy and annotation process.

Reddit

Using the keywords from Micallef et al. 2020, we utilized the Pushshift Reddit API (Baumgartner et al. 2020) to extract Reddit posts that contain links to YouTube videos related to *COVID-19 – 5G conspiracy theories*. We experienced several issues when collecting data from Reddit. The PushShift API is not robust enough to handle search queries that return large volumes of data. In addition, the returned data required additional manual filtering to retain relevant posts as searching for 5G returned a large volume of unrelated posts. For example, posts containing the substring 5G in links and usernames were returned. These challenges prevented us from collecting a volume of relevant data that is comparable to the other platforms. To address this concern, we improved our data collection by refining the search query to shorter time interval steps. This strategy, together with extending the data collection timeline to December 31st, 2020, enabled us to retrieve 10,304 posts from 2,351 different public subreddits. From this data we retrieved 2,000 posts that we use for the taxonomy and annotation process.

Fact-Checked 5G Claims

To determine whether a claim related to 5G is misinformation, counter-misinformation or unrelated, we compiled a false claims dataset by leveraging the work of fact-checkers. Specifically, we extracted 6,840 false statements fact-checked by the International Fact-Checking Network (IFCN) CoronaVirusFacts/DatosCoronaVirus alliance.[7] We then selected English statements related to *5G conspiracy theories* that state 5G technology is responsible for the spread of COVID-19 or that COVID-19 does not exist and people are getting sick due to 5G radiations. Our final list of fact-checked 5G claims consisted of 32 claims.

Taxonomy and Annotations

Starting with a sample of 300 randomly extracted Twitter posts that direct to YouTube videos, two researchers identified how the videos spread misinformation or counter-misinformation and how the posts use YouTube videos. To determine whether a social media post is a known false statement, the two researchers used the list of claims described above that were verified by IFCN fact-checkers. Specifically, the first iteration returned four categories for video classification (i.e., explicit, implicit, neutral, and others) and four categories for the relationship between a post and a video (i.e., support post, contradiction, unrelated, and out-of-context). In the second iteration, a second researcher went through the same 300 post and video pairs and either confirmed the provisional categories or suggested new categories (Verma and Patil 2021). The second researcher suggested an ambivalent category and to separate other videos into two categories (i.e., those related to the 5G topic and those that were unrelated). With respect to the relationship between a post and a video, the second researcher suggested a category which included those videos that answer posts, a category with videos that are related to post but do not support the post, and to change the 'taken out-of-context' to 'support post but taken-out-of context'. To converge into a common set of categories, a third iteration was conducted in which the two researchers went through the tweets together and discussed their categorizations (Teixeira et al. 2018; Verma and Patil 2021). During this iteration, the two researchers agreed to group two similar categories (i.e., instances where the post and video are posed as question and answer and instances where they both make the same arguments were combined into 'supports post') and to have six types of video content and five ways in which social media posts use YouTube videos to spread or counter misinformation (see second and third rows in Table 2). These general taxonomic groups captured all the data analyzed up to this stage. To verify the effectiveness of these categories created using 300 randomly selected Tweets, 400 randomly extracted posts from Facebook (200) and Reddit (200) were categorized by the same two researchers. This exercise did not require any changes to the defined categories. To further verify the effectiveness of the developed taxonomy, a professional fact-checker evaluated the categorization and confirmed that the taxonomy shown in Table 2 is complete and correct with respect to the data analyzed.

The above categories help to understand what kind of misinformation and counter-misinformation is posted in YouTube videos and how social media posts use these videos. Most YouTube videos contain "Explicit" misinformation or counter-misinformation, which is straightforward to categorize. Explicit misinformation is when a video contains clear statements which were verified to be false by an IFCN fact-checker. For example, a video which states that COVID-19 is like a normal flu and the real danger is 5G, which is what is making people sick. Videos categorized as "Implicit" did not directly refer to false information, but they indirectly referred to a claim that was found to be false by an IFCN fact-checker. For example, a video that discusses the installation of 5G antennas at the peak of the pandemic, without explicitly stating causation between COVID-19 and 5G. Other videos were "Neutral" because they only report on a claim that was found to be false by our fact-checking sources, without adding any arguments in favor or against. For example, a video which reports on the existence of a 5G/COVID-19 conspiracy theory without adding any value judgment. The YouTube videos that put both misinformation and messages refuting the false claims at the same level of credibility despite being fact-checked were categorized as "Ambivalent". For example, a video that presents 5G conspiracy theories and their rebuttals as equally valid and as a matter of opinion. We categorized other videos in two groups, "others related to topic" and "others unrelated to topic". Others related to topic are those videos that discuss topics related 5G and COVID-19 but do not mention a specific claim that was fact-checked to be false. For example, a video that talks about COVID-19 death toll. Others unrelated

[7] https://www.poynter.org/ifcn-covid-19-misinformation/

No	Post	Video Title & ID	Source	Classification
1	Very concerning, looks like there may be more to the whole #coronavirus thing than meets the eye. The reason why people can't believe these things ...they can't imagine such organized evil #5G	The BEST NEWS re Corona Virus you've heard all month! Kinda; ID: CtfqUtW_8AA	Twitter	*Post:* Misinformation; *Video:* Explicit; *Relationship:* Supports post; *Post-Video:* Misinformation
2	"The 5G story is complete and utter rubbish, it is nonsense" The <Country> government has addressed the conspiracy theories surrounding 5G and Coronavirus, watch here <URL>	The <Country> Government Addresses 5G Conspiracy Theories; ID: J-N7KsAgXnw	Twitter	*Post:* Countering; *Video:* Explicit; *Relationship:* Supports post; *Post-Video:* Countering
3	A ... video on 5g, cancer and coronavirus.	The Truth About 5G ft. MKBHD; ID: cw0A9FUTEKE	Reddit	*Post:* Ambiguous; *Video:* Explicit; *Relationship:* Supports Post; *Post-Video:* Countering
4	The Fourth Industrial Revolution (4IR) goes hand in hand with the rollout of 5G and the World Economic Forum's Great Reset. COVID-19 is paving the way towards acceptance of the New World Order.	What is the Fourth Industrial Revolution? ID: kpW9JcWxKq0	Facebook	*Post:* Misinformation; *Video:* Related to Topic; *Relationship:* Support but taken out-of-context; *Post-Video:* Misinformation
5	Coronavirus is a Cover Up for 5G sickness...	Can 5G radiation make you sick? What we found; ID: JjEwOAs2Kto	Facebook	*Post:* Misinformation; *Video:* Implicit; *Relationship:* Contradiction; *Post-Video:* Misinformation
6	The Threat of 5G <URL> via @YouTube ... Loss of privacy and it makes you susceptible to covid 19 !!!!!!!!	The Threat of 5G ID: AGkU7HmAAAc	Twitter	*Post:* Misinformation; *Video:* Implicit; *Relationship:* Related but does not Support; *Post-Video:* Misinformation
7	Covid and 5G connection. Watch quickly before it's removed....AGAIN! THE LARGEST GLOBAL COVER-UP IN HISTORY VIA 5G WHICH IS CAUSING CELL POISONING!	BlowerWhistle FoneVoda ... #YTWillRemoverize; ID: NmjR4Bodp2M	Facebook	*Post:* Misinformation; *Video:* Explicit; *Relationship:* Supports Post; *Post-Video:* Misinformation

Table 1: Examples of social media posts and labels assigned by the coders.

Feature	Classification
Post Classification	Misinformation, Countering, Others
Video Classification	Explicit, Implicit, Neutral, Ambivalent, Others related to topic, Others unrelated
Post-Video Relationship	Supports post, Related but does not support post, Contradiction, Unrelated, Supports but taken out-of context
Post-Video Classification	Misinformation, Countering, Others

Table 2: Manual Annotation

topic are videos that are not related to our research, such as the music video for Rick Astley's "Never Gonna Give You Up".

With respect to how YouTube videos are used, we find that in most instances videos "Support post" or answer a question posed in the social media post that links to them (see post 1 in Table 1). "Related but does not support post" are posts which use YouTube videos that are related to the posts but do not explicitly support the post (see post 6 in Table 1). We also had "Contradictions", in which the post contradicts the content of the video (see post 5 in Table 1). Plenty of instances have videos that "Supports but taken out-of-context" (see post 4 in Table 1). These are mostly implicit videos which are taken out-of-context to support false claims. Finally, we had posts and videos that are "Unrelated".

Manual Annotation

The aim of the annotation task was to generate the ground-truth for our analysis and the classifier. Specifically, we annotated posts, YouTube videos, post and video relationships and post-video classifications using the labels listed in Table 2. For Post and Post-Video Classification, we borrow the same categories used by Micallef et al. 2020 in their work. For Video Classification and Post-Video Relationship, we used the categories that emerged from the exercise that we described in the previous subsection.

Annotation process: To hand-label posts, 6,000 posts (i.e., 2,000 from Twitter, 2,000 from Facebook, and 2,000 from Reddit) were annotated by two researchers. Inter-rater agreement was measured by Kappa score (Schuster 2004) on a random sample of 600 posts (i.e., 200 from Twitter, 200 from Facebook, and 200 from Reddit) (Micallef et al. 2020). The inter-rater agreement returned 0.898 for posts classification, 0.871 for video classification, 0.873 for post-video relationship, and 0.937 for post-video classification. These

kappa scores show substantial agreement among the two coders. This high agreement is related to the two coders being involved in the definition of the taxonomy.

This process was challenging and time consuming (i.e., took 2 months to complete) as social media posts and YouTube videos are noisy, and their interpretation can be subjective. To ensure high-quality data, we conducted this process over several iterations. After each iteration, a meeting was held in which the two researchers discussed challenging and dubious pairs (Zubiaga et al. 2015). To assure high-quality of labeled data, those few instances where the coders did not agree were eliminated.

Specifically, to annotate the posts and videos, the two coders used the following process. They started by examining just the text of the post and asked the question: *Does this post refer to a known false statement related to 5G and COVID-19 ?* If the answer is yes, it was determined whether the post is supporting the false claim or refuting it. For instance, post 1 in Table 1 supports the verified false claim which states that people are getting sick for reasons other than the coronavirus, so it was annotated as a misinformation post. To determine whether the post was a known false statement, a list of claims that were verified by IFCN fact-checkers was used (see Fact-Checked 5G Claims in Data Collection section). When a post refutes a false claim, such as when people categorically state that 5G is not related to the COVID-19 pandemic (see post 2 in Table 1), it was annotated as a counter-misinformation post. If the post was neither misinformation nor counter-misinformation, such as post 3 in Table 1, which is just referring to a video about 5G, cancer and coronavirus, we annotated the post as other.

Next, the linked YouTube video was watched. If the video had already been deleted from YouTube, an attempt was made to retrieve it from four other video repositories: waybackmachine,[8] bitchute,[9] kzclip,[10] ruplayers.[11] If the video was still not found, the social media post that had linked to the video was discarded. If the video was found, it was annotated using one of the categories listed in row 2 of Table 2.

After annotating posts and videos, we examined the relationship between a post and video pair and classified it using one of the categories listed in row 3 of Table 2. Finally, we considered both the post and the video and asked the same question as before, this time also considering the contents of the video: *Does this post-video pair refer to a known false statement related to 5G and COVID-19 ?* Table 1 lists some examples of our annotations.

Analysis of Cross-Platform Post-Video Pairs

The process described in the previous section returned 1,000 posts each from Twitter, Facebook, and Reddit. This considerable reduction in posts from the randomly picked selections is due to the deletion of many videos by YouTube which could not be retrieved. These videos were not uploaded on other video repositories and could not be retrieved from there either. Other researchers have also experienced this problem (Yang et al. 2021). The key finding from the annotated posts-video pairs (see Figure 2) is that considering only what is written in the posts misses a significant number of posts containing misinformation or counter-misinformation ($p < 0.0001$). Thus, cross-platform information needs to be considered to more effectively determine whether post-video pairs contain misinformation.

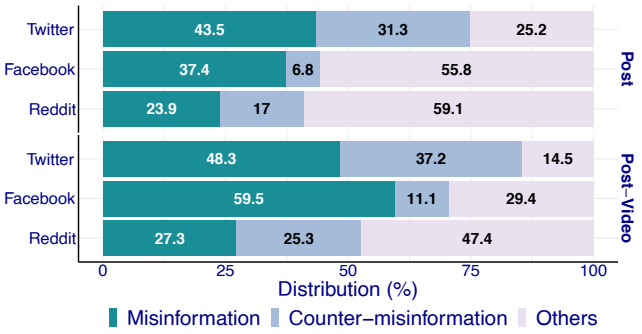

Figure 2: Grouping posts and post-video pairs into Misinformation, Counter-misinformation, and Others.

We also analyzed the dates when videos are posted and corresponding social media posts that link to them are created. Our findings show that most posts and videos are published within a narrow time window. 56% of the posts were published less than 5 days after the videos were uploaded, while only 23% of the posts directed to videos that were older than a month. This finding indicates that it is more likely that posts are being created to drive traffic to newly uploaded videos rather than surfacing older videos.

Platform Effect

We investigated the effect of platforms on how and which YouTube videos were used to spread misinformation. When studying the distribution of post-video pairs across platforms, we find a connection between the platform and the type of posts ($p < 0.0001$ using chi-squared test). Using Crammer's V test we find a moderate positive correlation ($\rho = 0.38$). This correlation could be explained by Facebook having more post-video pairs which contain misinformation compared to the other platforms (59.5% vs 48.3% & 27.3%). In addition, Reddit (27.3% & 25.3%) and Twitter (48.3% & 37.2%) have a more balanced distribution of post-video pairs that contain misinformation and counter-misinformation, with Reddit having a disproportionate amount of others (47.4% vs 14.5% & 29.4%). This finding indicates that some platforms might be used to spread more post-video pairs that contain misinformation than others.

Next, we studied how posts were used to spread videos. We found that there is a relationship between the platform and how the videos are used in posts ($p < 0.0001$ using chi-squared test). However, using Crammer's V test we found only a small effect ($\rho = 0.15$). The most predominant relationship is the sharing of videos that directly support the contents of the post (see Figure 3). The differences are in the use of out-of-context videos, which is uncommon on

[8] https://archive.org/web/
[9] https://www.bitchute.com/
[10] https://kzclip.com
[11] https://ruplayers.com/

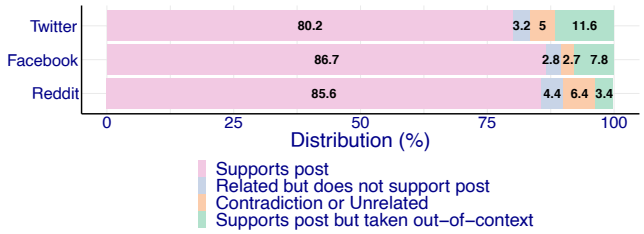

Figure 3: Grouping post-video relationship in Supports post, Related but does not support post, Contradiction or Unrelated, and Supports post but taken out-of-context.

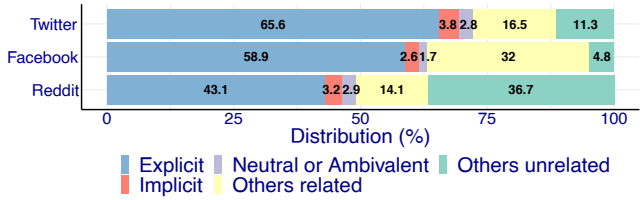

Figure 4: Grouping videos in Explicit, Implicit, Neutral or Ambivalent, Others related, and Others unrelated.

Reddit (3.4%), but seen more on Facebook (7.8%) and Twitter (11.6%). This finding indicates that out-of-context videos are the second most common way of presenting misinformation, especially on microblogging platforms, such as Twitter.

We also studied the type of content used in the YouTube videos that are linked via posts published on other platforms. We find a relationship between the platform and the type of content ($p < 0.0001$ using chi-squared test). Using Crammer's V test we find a weak effect ($\rho = 0.2$). This weak effect could be related to Reddit having considerably more unrelated videos than the other two platforms (36.7% vs 11.3% & 4.8%). In addition, this weak effect could be related to Facebook having considerably more 'other videos related to topic but are not misinformation' (32%) than Twitter (16.5%) and Reddit (14.1%) (see Figure 4). Our findings indicate that although platforms might spread different types of non-misinformation videos, they are still used similarly to spread explicit misinformation videos.

Misinformation vs Counter-Misinformation

To understand whether post-video pairs that spread misinformation have different characteristics from those that counter it, we started by analyzing the dates when videos were posted and posts created. We found that more counter-misinformation posts seem to be created to drive traffic to videos, since 65% of the posts are created within 5 days from when the video is uploaded. Misinformation has a lower percentage (49%). These differences are significant ($p < 0.0001$ using Wilcoxon pairwise comparisons). This finding indicates that more misinformation spreaders point to videos that have been around for a longer time. This could be explained by the finding that, in 70% of posts that direct to out-of-context videos, the videos were uploaded more than a month before the post being created.

Next we investigated whether there are differences in the way that misinformation and counter-misinformation posts use YouTube videos. We find that there is a relationship between the type of posts and how the videos are used ($p < 0.0001$ using chi-squared test). Crammer's V test shows that this is a weak correlation ($\rho = 0.21$). Both misinformation and counter misinformation messages predominantly share videos that support the post (82% and 85% respectively). The differences are in the use of out-of-context videos which are uncommon when countering misinformation (4%), but are more frequent when spreading misinformation (14%).

Finally, we analyzed the connection between the type of content in YouTube videos and the type of the social media post that links to it. We did not find a relationship between the type of post and the type of content in the video ($p > 0.05$ using chi-squared test). This finding shows that misinformation and counter misinformation posts use different types of videos to fulfill their purposes.

Overall, our analysis uncovers various differences in the way that posts use videos to spread misinformation as opposed to when used to counter misinformation.

Classifiers Setup and Experiments

After characterizing the collected post-video pairs, we conducted several experiments to automate the detection of misinformation, by testing various classifiers. Due to our dataset having limited data for some categories in the taxonomy (see Figures 2, 3, and 4), we grouped together some categories to provide a more balanced input to our classifiers. For Post and Post-Video classification, we included all classes. For Video classification we used Explicit, Implicit, and Others. For Post-Video relationship classification, we used Supporting, Out-of-Context, and Others.

Classification Features

We used a mix of textual features extracted from posts and videos. We extracted the following features since they have been used as the standard features in recent literature (Chen et al. 2021; Khan et al. 2021; Islam et al. 2020) and could be retrieved from all videos present in our datasets: **Post text**: included the entire text of the post from Twitter, Facebook, while for Reddit, we concatenate the title and the body together as a single piece of text; **Video title** and **Video description**: extracted from YouTube metadata; and **Video transcription**: extracted the closed captions from video or generated them using Google Cloud's Speech-to-Text services.

Baselines

We start our experiments by evaluating the effectiveness of single-source traditional misinformation classifiers on our dataset of social media posts that direct to YouTube videos. This allows us to investigate the effectiveness of these models in classifying post-video pairs based on the post text only. As a baseline for post classification, we used the model from Micallef et al. 2020 because it is publicly available and classifies counter-misinformation and misinformation posts.

To find single-source models that achieve reasonable performance on our data, we also trained a simple neural network on the data from Micallef et al. 2020 for post classification, and a second neural network on the video data that we extracted from tweets, as a baseline for video classification. For all models, we use BERT for text representation, since prior work found it to be the most effective representation to detect misinformation (Islam et al. 2020; Khan et al. 2021). Representations were generated after removing URLs, hashtags, Twitter mentions, and emojis (Micallef et al. 2020).

Our results show the limited performance of models trained on a single data source when detecting misinformation in post-video pairs. The logistic regression model from Micallef et al. 2020 achieves an F1 score of only *0.49* on Twitter posts which direct to YouTube videos.

Model	**F1** (σ)	**Prec.** (σ)	**Rec.** (σ)
Post			
Naive Bayes	0.61 (0.04)	0.62 (0.04)	0.61 (0.04)
AdaBoost	0.71 (0.08)	0.73 (0.08)	0.71 (0.09)
SVM	0.67 (0.06)	0.68 (0.06)	0.68 (0.06)
Logistic Reg.	0.72 (0.06)	0.74 (0.04)	0.72 (0.06)
Random For.	0.64 (0.07)	0.67 (0.05)	0.66 (0.06)
Neural Net.	0.68 (0.08)	0.70 (0.07)	0.69 (0.07)
Co-Attention	0.65 (0.11)	0.70 (0.10)	0.66 (0.10)
Bi-LSTM	0.77 (0.08)	0.79 (0.09)	0.77 (0.08)
Video			
Naive Bayes	0.50 (0.10)	0.58 (0.06)	0.49 (0.10)
AdaBoost	0.58 (0.11)	0.60 (0.11)	0.61 (0.10)
SVM	0.72 (0.10)	0.71 (0.10)	0.75 (0.09)
Logistic Reg.	0.67 (0.10)	0.68 (0.08)	0.67 (0.11)
Random For.	0.66 (0.09)	0.72 (0.08)	0.68 (0.08)
Neural Net.	0.70 (0.10)	0.70 (0.10)	0.71 (0.10)
Co-Attention	0.64 (0.10)	0.65 (0.12)	0.65 (0.08)
Bi-LSTM	0.76 (0.07)	0.78 (0.06)	0.75 (0.08)
Relationship			
Naive Bayes	0.73 (0.08)	0.80 (0.10)	0.69 (0.07)
AdaBoost	0.72 (0.07)	0.68 (0.09)	0.77 (0.07)
SVM	0.72 (0.07)	0.65 (0.08)	0.80 (0.05)
Logistic Reg.	0.78 (0.05)	0.80 (0.07)	0.80 (0.03)
Random For.	0.70 (0.07)	0.66 (0.11)	0.78 (0.06)
Neural Net.	0.78 (0.09)	0.78 (0.10)	0.80 (0.08)
Co-Attention	0.75 (0.08)	0.81 (0.05)	0.72 (0.12)
Bi-LSTM	0.81 (0.08)	0.89 (0.12)	0.75 (0.08)
Post-Video			
Naive Bayes	0.62 (0.07)	0.66 (0.06)	0.61 (0.08)
AdaBoost	0.52 (0.10)	0.54 (0.09)	0.55 (0.09)
SVM	0.73 (0.07)	0.70 (0.09)	0.77 (0.05)
Logistic Reg.	0.71 (0.08)	0.72 (0.08)	0.73 (0.08)
Random For.	0.59 (0.10)	0.63 (0.07)	0.64 (0.10)
Neural Net.	0.68 (0.12)	0.70 (0.11)	0.69 (0.13)
Co-Attention	0.62 (0.11)	0.68 (0.11)	0.63 (0.10)
Bi-LSTM	0.75 (0.06)	0.80 (0.07)	0.73 (0.06)

Table 3: 10-Fold Cross Validation on all tasks using Twitter text and YouTube Video Titles and Descriptions.

Our single-source neural networks achieves an F1 score of *0.45* on post classification and *0.58* on video classification. These findings underscore the need for developing models that can leverage cross-platform features. Next, we develop models that can be more effective in classifying post-video pairs which contain misinformation or counter-misinformation.

Setup of Classifiers

In our experiments, we use five standard classification models: Naive Bayes, AdaBoost, SVM, Logistic Regression, and Random Forest (Khan et al. 2021). In addition, we use the following deep learning models which have become state of the art in recent misinformation detection research (Islam et al. 2020): a single-layer neural network, a dEFEND-based co-attention model (Shu et al. 2019), and a Bi-LSTM classifier (Khan et al. 2021). We follow the same procedure for generating representations detailed in the Baselines section.

For most models, we concatenate the BERT-generated 768-dimensional vector representations of text from post and video sources into a single vector. For the co-attention model, the inputs are traditional BERT representation for post text and sentence BERT for textual video features. The Bi-LSTM classifier takes the BERT tokens directly as input.

Experiment 1: Twitter Dataset

We compare the effectiveness of the previously defined single-source baselines with models that take cross-platform text features: YouTube (e.g., video titles and descriptions) as well as the source platform (i.e., tweet post). Results in Table 3 show that all cross-platform models outperform the baseline models, since for the Post classification task they all obtain an F1 score > 0.5. For Video classification, all models except Naive Bayes outperform the baselines as well. Models also achieve reasonable performance (i.e., F1 score $>$ 0.59) for the other classification tasks: Post-Video Classification and Post-Video Relationship. Another main outcome from this first experiment is that the Bi-LSTM classifier is the best performing model for all classification tasks, followed by SVM and Logistic Regression.

Experiment 2: Three Platforms Datasets

To investigate the effect of all classifications when having data from three different platforms, we trained and tested the top three performing models from the first experiment (i.e., Logistic Regression, SVM, and Bi-LSTM) using three datasets together (i.e, Twitter, Facebook, and Reddit). We find that on three out of the four tasks, using the three datasets achieves similar performance as using only the Twitter dataset. The one exception being Post-Video Relationship, where there is a significant increase in performance produced by including the three datasets (see Figure 5). This is an important finding since a more diverse dataset did not deteriorate the performance of the classifiers.

Experiment 3: Video Textual Features

Since a YouTube video contains several textual features, we investigated which features were more effective in classifying post-video pairs. We trained and tested models on all

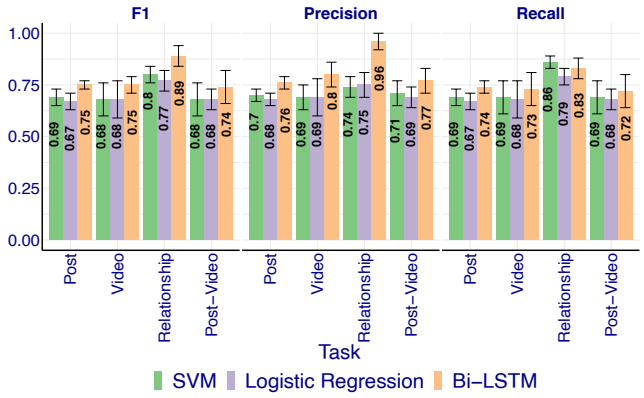

Figure 5: 10-Fold Cross Validation performance on all tasks using datasets from three social networks.

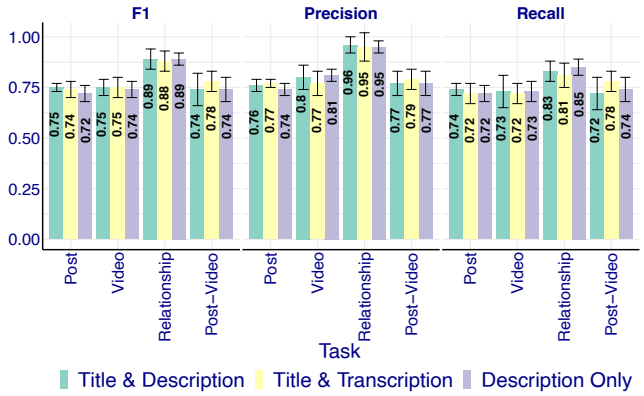

Figure 6: Performance of Bi-LSTM's 10-Fold Cross Validation using three datasets with different text video features.

three datasets using video descriptions, video titles & descriptions, and title & video transcriptions (i.e., YouTube-generated captions). Figure 6 shows the results for Bi-LSTM, the best performing model from our previous experiments, when using three combinations of video textual features. Title & description and title & transcription achieve comparable performance ($p > 0.05$) in all tasks except for Post classification, where description is more effective.

Experiment 4: Multi-Task Learning

To leverage the improved performance obtained by the post-video relationship classifier when using three datasets (see Experiment 2), we investigated whether using multi-task learning can increase effectiveness. We used the same model for all tasks with four different classification heads, one for each task (Zhang and Yang 2021), where the loss for each task is calculated independently. As Title & Description and Title & Transcription yielded comparable results (see Figure 6), in this experiment we investigate these setups.

Results in Figure 7 show that the use of multi-task learning achieves increased or similar performance on all tasks using textual features, with the exception of Post classification, where there is a deterioration in performance when

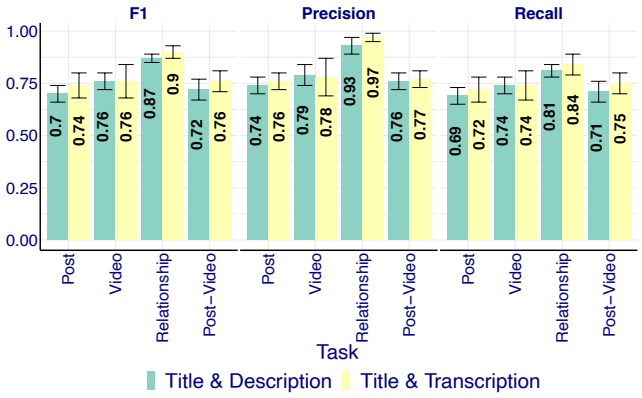

Figure 7: Performance of Bi-LSTM's 10-Fold Cross Validation using multi-task learning.

using Title & Description. For the case of Title & Transcription, the model achieves superior performance to the Bi-LSTMs in Figure 5 on all tasks. Overall, Title & Transcription video features along with multi-task learning yield our best performing model.

Discussion

Our findings reveal that for post-video pairs, considering only text from social media posts misses a considerable amount of misinformation and counter-misinformation. This could explain why only a small number of post-video pairs are deleted or flagged as misinformation (Knuutila et al. 2020). Hence, our research suggests that to reduce misinformation, platforms that allow cross-platform links should not limit themselves to features that are extracted from their platforms, but should consider enhancing their algorithms to use features from other platforms. We acknowledge that it is not straightforward to forge collaborative arrangements among platforms, since plenty of factors (e.g., business interests, costs, etc.) must be considered to establish such relationships. Nevertheless, misinformation can be combated more effectively when social media platforms establish such collaborations and detecting post-video pairs that contain misinformation could be of benefit to multiple stakeholders.

Cross-platform misinformation classifiers could help fact-checkers gain more insights about how debunked videos are used on multiple platforms, which could make the fact-checking process less time-consuming (Micallef et al. 2022). Also, content moderators could benefit by having a better prioritization, which takes into account how videos are being used on multiple platforms. At the moment this does not seem to happen, since prior work found that on average it takes 41 days to remove COVID-19 misinformation videos from YouTube (Knuutila et al. 2020). More importantly, social media users could benefit from such a classifier by having more context about the relationship between the post and video, but also the type of misinformation that the video contains. This added context could help users make better decisions, which could prevent them from clicking on the links that direct them to misinformation videos.

Another important finding from our research is that regardless of the platform, misinformation spreaders predominantly use explicit YouTube videos that support posts to spread misinformation. Implicit, neutral and ambivalent videos were used similarly on the examined platforms. These findings indicate that for the studied topic, the type of videos that are used to spread/counter misinformation and how the videos are used in these posts are mostly not affected by the platform on which they are shared. This explains why the evaluated classifiers were most effective when using data from multiple platforms (see Experiments 2 & 4). Despite collecting data from platforms that have different characteristics, our classifiers did not experience a decrease in performance. This implies that using data from multiple platforms could strengthen the robustness of the cross-platform classifier as it is being trained with more diverse data (Feng et al. 2021). Moreover, when training the classifier to learn a new topic, rather than relying on one source to collect data, it would be faster and similarly effective to collect data from multiple sources. Faster training of a classifier to detect misinformation in post-video pairs is important because at least 50% of posts are created within 5 days from video upload. Hence, having classifiers that learn to detect misinformation at a faster pace can further reduce the spread of impactful post-video misinformation.

Combining post-video pairs from multiple platforms could also improve the effectiveness and speed of early warning and detection of misinformation algorithms. Faster and more accurate early detection of post-video pairs could allow social media platforms to take down these post-video pairs. In addition, they could also be flagged earlier to professional fact-checkers, who often use virality as a metric for selecting claims worthy of their attention (Micallef et al. 2022). Hence, combining post-video pairs from multiple platforms could provide fact-checkers with early warnings about the misinformation posts that might be trending soon. This would reduce the significant lag that currently exists between producing fact-checking outcomes and the time of origin of the underlying misinformation, which allows misinformation to spread and cause widespread damage while fact-checking work is still in progress.

Our work also reveals some cross-platform differences (see Table 4). For instance, we find that Facebook is used significantly more to spread post-video misinformation, while countering post-video pairs are not as common. Although this finding could be affected by our choice of topic, it is important to also consider that Facebook has a considerably larger user-base (i.e., more than 1.15 billion vs Twitter's and Reddit's 500 million). Thus, our finding could be explained by the fact that misinformation spreaders consider the potential reach of their posts when choosing platforms. Our findings also show that Twitter and Reddit are used significantly more than Facebook to share videos that refute misinformation. This is interesting because these two platforms have contrasting characteristics. Our findings are consistent with past work (Micallef et al. 2020) that demonstrated a considerable number of Twitter users who refuted false claims related to COVID-19 fake cures. Although further research is required to investigate whether Twitter and Reddit are popular platforms to share refuting non-COVID posts, our research indicates that using these platforms to refute COVID misinformation appears to be a trend. More research is also required to investigate why these platforms are frequently used to share refuting messages. An implication of this finding is that social media platforms that seek to implement advanced counter-messaging to address important societal issues, such anti-vaccine rhetoric, might need to make use of other platforms to retrieve countering messages that refute posts that use specific strategies (e.g., "Brave Truthteller") (Hughes et al. 2021), if they find that there are not enough of these posts on their platform.

Cross-Platform Similarities
• Considering only posts misses a significant number of misinformation and counter-misinformation posts.
• 56% of posts are posted within a month from when a video is created. This strategy is more common with counter-misinformation videos.
• Only 23% of posts direct to videos older than a month, 70% of which are out-of-context videos used to spread misinformation.
• Explicit videos that support posts are the most common strategy used to spread misinformation.
• Strategy of using implicit, neutral or ambivalent videos is rare, but similar across platforms.
Cross-Platform Differences
• Facebook seems to be the preferred platform to spread misinformation in post-video pairs.
• Countering post-video pairs are more common on Twitter and Reddit.
• Considerable amount of Facebook and Twitter posts use videos out-of-context, which is a rare practice on Reddit.

Table 4: Similarities and differences in how post and video pairs are used across platforms.

Limitations and Future Work

Our work has some limitations that could be addressed in the future. In this research, we only focus on YouTube videos that were not deleted or could be retrieved from other repositories. In future work, one can extend our monitoring to capture recent posts, which are less likely to direct to deleted videos. This approach would allow us to collect and use comments in our classifiers. Another limitation is that we studied one misinformation topic. This topic was instrumental to understand how videos are used in posts to spread misinformation or counter it. Despite this limitation, we believe our major findings about the use of cross-platform features should hold for other topics. However, there could be variations across areas that are targets of misinformation (e.g., elections). Further research is required to investigate whether the classifiers proposed in this research achieve similar effectiveness even with different misinformation topics.

For Facebook and Reddit, we had to extend our data collection to include additional months, otherwise we would not have retrieved enough posts for our analysis and classi-

fication. This limitation should not impact our findings in a major way because previous work has shown that behavioral shifts take longer to materialize and misinformation tends to resurface multiple times over a period of time (Shin et al. 2018). Moreover, we still had a considerable time overlap between the three platforms (i.e., Jan 2020 till May 2020). Finally, our research does not study the language used in the post and video in detail because the aim of this research is to characterize cross-platform post-video misinformation and counter-misinformation, and how post-video pairs are used on multiple platforms. Further research is required to study the language used in post-video pairs to examine whether important relationships could be extracted between the social media post language and the video content.

Conclusion

In this work, we characterized how YouTube videos are used in social media posts on different platforms to spread misinformation or refute it, and how multiple platforms are leveraged to increase the reach of these videos. Our work reveals that regardless of the platform, misinformation spreaders mostly use similar tactics to spread post-video pairs that contain misinformation (i.e., explicit videos that support their posts). In addition, our work shows that misinformation detection models that use only text from posts miss a significantly high number of post-video pairs that contain misinformation (i.e., $< 50\%$). We address this problem by demonstrating that classifiers that use features from multiple platforms can significantly improve detection of misinformation in post-video pairs. Consequently, in addition to benefiting various stakeholders such as fact-checkers, content moderators, social media users, our work can also contribute to the reduction of the spread of high impact post-video pairs that contain misinformation.

Acknowledgments

This research has been supported in part by NSF IIS2027689, NSF ITE2137724, New York University Abu Dhabi, Georgia Tech IDEaS, Adobe, and Microsoft Azure. Thank you to Bing He for his assistance with the initial setup of the experiments. Facebook data is provided courtesy of CrowdTangle.

References

Agrawal, T.; Gupta, R.; and Narayanan, S. 2017. Multimodal detection of fake social media use through a fusion of classification and pairwise ranking systems. In *2017 25th European Signal Processing Conference (EUSIPCO)*, 1045–1049.

Alam, F.; Cresci, S.; Chakra borty, T.; Silvestri, F.; Dimitrov, D.; Martino, G. D. S.; Shaar, S.; Firooz, H.; and Nakov, P. 2021. A survey on multimodal disinformation detection. *arXiv preprint arXiv:2103.12541*.

Bansal, R.; Paka, W. S.; Nidhi; Sengupta, S.; and Chakraborty, T. 2021. Combining Exogenous and Endogenous Signals with a Semi-supervised Co-attention Network for Early Detection of COVID-19 Fake Tweets. In *Advances in Knowledge Discovery and Data Mining*, 188–200. Cham: Springer International Publishing.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*.

Brennen, J. S.; Simon, F.; Howard, P. N.; and Nielsen, R. K. 2020. Types, sources, and claims of Covid-19 misinformation. *Reuters Institute*, 7.

Chen, K.; Kim, S. J.; Gao, Q.; and Raschka, S. 2021. Visual Framing of Science Conspiracy Videos: Integrating Machine Learning with Communication Theories to Study the Use of Color and Brightness. arXiv:2102.01163.

Destiny, T. 2021. Conspiracy theories about 5G networks have skyrocketed since COVID-19. *The Conversation*, 19.

Driscoll, K.; and Thorson, K. 2015. Searching and Clustering Methodologies: Connecting Political Communication Content across Platforms. *The ANNALS of the American Academy of Political and Social Science*, 659(1): 134–148.

Feng, S. Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; and Hovy, E. 2021. A survey of data augmentation approaches for nlp. *preprint arXiv:2105.03075*.

Hameleers, M.; Powell, T. E.; Meer, T. G. V. D.; and Bos, L. 2020. A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media. *Political Communication*, 37(2): 281–301.

Horawalavithana, S.; Ng, K. W.; and Iamnitchi, A. 2020. Twitter Is the Megaphone of Cross-platform Messaging on the White Helmets. In *Social, Cultural, and Behavioral Modeling*, 235–244.

Hossain, T.; Logan IV, R. L.; Ugarte, A.; Matsubara, Y.; Young, S.; and Singh, S. 2020. COVIDLIES: Detecting COVID-19 Misinformation on Social Media. In *Proc. of the 1st Workshop on NLP for COVID-19 at EMNLP 2020*.

Hou, R.; Perez-Rosas, V.; Loeb, S.; and Mihalcea, R. 2019. Towards Automatic Detection of Misinformation in Online Medical Videos. In *2019 International Conference on Multimodal Interaction*, ICMI '19, 235–243. New York, NY, USA: Association for Computing Machinery.

Hughes, B.; Miller-Idriss, C.; Piltch-Loeb, R.; White, K.; Creizis, M.; Cain, C.; and Savoia, E. 2021. Development of a Codebook of Online Anti-Vaccination Rhetoric to Manage COVID-19 Vaccine Misinformation. *medRxiv*.

Hunt, K.; Wang, B.; and Zhuang, J. 2020. Misinformation debunking and cross-platform information sharing through Twitter during Hurricanes Harvey and Irma: a case study on shelters and ID checks. *Natural Hazards*, 103: 861–883.

Islam, M. R.; Liu, S.; Wang, X.; and Xu, G. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1): 1–20.

Kessler, G. 2019. Introducing The Fact Checker's guide to manipulated video. *The Washington Post*.

Khan, J. Y.; Khondaker, M. T. I.; Afroz, S.; Uddin, G.; and Iqbal, A. 2021. A benchmark study of machine learning

models for online fake news detection. *Machine Learning with Applications*, 4.

Knuutila, A.; Herasimenka, A.; Au, H.; Bright, J.; Nielsen, R.; and Howard, P. N. 2020. COVID-related misinformation on Youtube.

Medina Serrano, J. C.; Papakyriakopoulos, O.; and Hegelich, S. 2020. NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube. In *Proc. of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: ACL.

Micallef, N.; Armacost, V.; Memon, N.; and Patil, S. 2022. True or False: Studying Work Practices of Professional Fact-Checkers. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1).

Micallef, N.; He, B.; Kumar, S.; Ahamad, M.; and Memon, N. 2020. The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. In *2020 IEEE International Conference on Big Data (Big Data)*, 748–757.

Qi, P.; Cao, J.; Yang, T.; Guo, J.; and Li, J. 2019. Exploiting Multi-domain Visual Information for Fake News Detection. In *2019 IEEE International Conference on Data Mining (ICDM)*, 518–527.

Qian, S.; Wang, J.; Hu, J.; Fang, Q.; and Xu, C. 2021. Hierarchical Multi-Modal Contextual Attention Network for Fake News Detection. In *Proc. of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, 153–162. New York, NY, USA: Association for Computing Machinery.

Schuster, C. 2004. A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, 64(2): 243–253.

Shin, J.; Jian, L.; Driscoll, K.; and Bar, F. 2018. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, 83: 278–287.

Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. 2019. DEFEND: Explainable Fake News Detection. In *Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, 395–405. New York, NY, USA: ACM.

Team, C. 2021. *CrowdTangle. Facebook, Menlo Park, California, United States*. List ID: [1517853, 1518092].

Teixeira, C. R. G.; Kurtz, G.; Leuck, L. P.; Tietzmann, R.; de Souza, D. R.; Lerina, J. A. M. F.; Manssour, I. H.; and Silveira, M. S. 2018. Humor, Support and Criticism: A Taxonomy for Discourse Analysis about Political Crisis on Twitter. In *Proc. of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. New York, NY, USA: Association for Computing Machinery.

Thorson, K.; Driscoll, K.; Ekdale, B.; Edgerly, S.; Thompson, L. G.; Schrock, A.; Swartz, L.; Vraga, E. K.; and Wells, C. 2013. YouTube, Twitter and The Occupy Movement. *Information, Communication & Society*, 16(3): 421–451.

Verma, P.; and Patil, S. 2021. Exploring Privacy Aspects of Smartphone Notifications. In *Proc. of the 23rd International Conference on Mobile Human-Computer Interaction*, MobileHCI '21. New York, NY, USA: Association for Computing Machinery.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, 849–857. New York, NY, USA: ACM.

Wang, Z.; Yin, Z.; and Argyris, Y. 2020. Detecting Medical Misinformation on Social Media Using Multimodal Deep Learning. *IEEE Journal of Biomedical and Health Informatics*, 1–1.

Wilson, T.; and Starbird, K. 2020. Cross-platform disinformation campaigns: lessons learned and next steps. *Harvard Kennedy School Misinformation Review*, 1(1).

Yang, K.-C.; Pierri, F.; Hui, P.-M.; Axelrod, D.; Torres-Lugo, C.; Bryden, J.; and Menczer, F. 2021. The COVID-19 Infodemic: Twitter versus Facebook. *Big Data & Society*, 8(1).

Yang, S.; Shu, K.; Wang, S.; Gu, R.; Wu, F.; and Liu, H. 2019. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of AAAI conference*, volume 33, 5644–5651.

Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018. On the Origins of Memes by Means of Fringe Web Communities. In *Proc. of the Internet Measurement Conference 2018*, IMC '18, 188–202. New York, NY, USA: ACM.

Zhang, Y.; and Yang, Q. 2021. A Survey on Multi-Task Learning. *IEEE Transactions on Knowledge and Data Engineering*, 1–1.

Zubiaga, A.; Liakata, M.; Procter, R.; Bontcheva, K.; and Tolmie, P. 2015. Towards detecting rumours in social media. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.