IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 October 2023, accepted 10 October 2023, date of publication 13 October 2023, date of current version 23 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3324555

RESEARCH ARTICLE

Characterizing Discourse and Engagement Across Topics of Misinformation on Twitter

DOMINIKA NADIA WOJTCZAK[1], CLAUDIA PEERSMAN[2], LUISA ZUCCOLO[1], AND RYAN MCCONVILLE[3,4]

[1]School of Computer Science, University of Bristol, BS8 1QU Bristol, U.K.
[2]School of Computer Science Luisa-Bristol Medical School, University of Bristol, BS8 1QU Bristol, U.K.
[3]School of Engineering Mathematics and Technology, Human Technopole, 20157 Milan, Italy
[4]U.K. MRC Integrative Epidemiology Unit, University of Bristol, BS8 1QU Bristol, U.K.

Corresponding author: Dominika Nadia Wojtczak (lh20935@bristol.ac.uk)

ABSTRACT In recent years, online misinformation has become increasingly prevalent, leading to significant issues such as political polarisation and distrust of genuine information. Misinformation on social media platforms affects various aspects of society, including health and politics, and can take many forms, such as text and images. However, current studies mainly focus on analysing singular topics and modalities, without considering the heterogeneity of the issue. Our research aimed to examine the relationship between visual elements and engagement, as well as the relationship between sentiment analysis, hate speech, and bots on a variety of topics on the Twitter social media platform Twitter. We labelled 12,581 misinformation posts that were manually modelled into a topic hierarchy. We then analysed these posts, including their sentiments, the prevalence of hate speech, and bot activity on different topics. The results revealed that political misinformation tends to contain more hate speech than COVID-19 misinformation and that political misinformation also has a higher number of bots. Furthermore, the findings suggest that misinformation online with more than 40% negative sentences can have a high level of hate speech identified for both tweets and replies. This study provides detailed information on topics and the volume of misinformation on social media platforms, and the findings can be used to develop more advanced detection systems and support further analysis. Our findings can help policy makers understand what kind of online misinformation has been spreading on Twitter and how to plan campaigns to make users more aware of how to spot its various features in an online user-to-user Twitter environment.

INDEX TERMS Online misinformation, sentiment analysis, hate speech, Twitter, images, polarization.

I. INTRODUCTION

It is important to study online misinformation because it has become increasingly prevalent in the digital age, as people often consume and share false information more frequently than ever. By analysing the patterns of discourse and online misinformation, we can gain valuable insights into how it spreads, who is more susceptible to it, and how it can be effectively countered. Online social networks, such as Twitter, have become a crucial source of news for an increasing proportion of the population [1]. The proliferation of the Internet and social networks has led to a surge in news consumption on these platforms. However, the increasing sophistication of misinformation on social networks has emerged as a major challenge. Traditional media outlets have started to use social media to post news, which can lead to the amplification of misinformation [2]. For the purpose of this study, we define misinformation as false or misleading information, which also includes disinformation. This definition is consistent with the dataset used in our study, MuMiN [3], which is discussed in more detail in Section III.

Misinformation is a global problem that affects various aspects of society, such as public health (e.g., hesitancy in vaccines), politics (e.g., election interference) and social issues (e.g., abortion). Given the international, multilingual, and diverse topics involving misinformation, we manually

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

annotated the posts in MuMiN with the topics discussed for each post. Our aim was to gain deeper insight into the types of misinformation present in social networks.

Most modern social media platforms use user engagement metrics, such as likes, shares, and replies, to characterise how users engage with content. The structure of social networks can show echo chambers where users are grouped with peers who have similar opinions, resulting in exposure to similar posts [2]. As noted by Avram et al. [4], there is a strong relationship between lower levels of fact checking and higher levels of social engagement. Work on social interaction, transmission, and virality of posts in recent years [5], [6], [7], [8] has identified the main components of social influence as STEPPS (that is, social currency, triggers, practical value, public, and storeys), which provides a foundation to understand why users engage with particular content on Twitter.

Our approach to this issue draws on social psychology studies that have explored the relationship between brightness and popularity of online misinformation [9], [10]. These studies have produced contradictory results. The first research stream suggests that users are affected by different background colours, while the second stream shows a lack of relationship between users and colour. The research streams differ in terms of the topics examined. Therefore, we investigate these differences across different topics of online misinformation to determine similarities or differences with prior research.

In this study, we seek to identify misinformation topics on Twitter and investigate the relationship between user engagement metrics and visual characteristics in online misinformation posts. To do so, we manually annotated 12,581 tweets labelled as misinformation to identify topics. One of the largest categories of misinformation (27.1%) was related to US politics, with over 98% consisting of political misinformation. The second largest category was related to health, with misinformation about COVID-19 accounting for more than 96% of this category.

The study also examined the relationship between user engagement metrics and visual content in online misinformation posts. The MuMiN-small dataset, which contains approximately 1754 images labelled as misinformation, was analysed. Each image was labelled on the basis of its content and colour. The results show that users tended to engage more with online misinformation presented in the form of plain images without text than with other types of images.

In general, the study provides detailed information on the topics and volume of misinformation on social media platforms, and the findings can be used to develop more advanced detection systems and support further analysis.

The innovations and contributions of this work are:

1) **Multifaceted Analysis:** We have broken down misinformation into various categories, including topic, sentiment, hate speech, and bot activity. This thorough approach gives us a comprehensive overview of the misinformation landscape, particularly on Twitter.
2) **Topic Hierarchy:** We manually annotated a large sample of 12,581 tweets in order to create a topic hierarchy that would provide a structured way to comprehend the most widespread misinformation themes.
3) **Image Analysis**: The MuMiN-small dataset enabled us to conclude that plain images without text are more successful in stimulating engagement than other image types. This implies that visual misinformation could be a growing problem.
4) **Sentiment and Hate Speech Connection**:Our research has uncovered a powerful connection between negative sentiment and the presence of hate speech. This is especially important for websites that are striving to create a positive atmosphere.
5) **Bot Activity:** Highlighting the extensive bot activity in certain topics, especially US political misinformation, underscores the artificial amplification of certain narratives. This insight is valuable for platform developers and policymakers.

II. RELATED WORK

The proliferation of social networks has led to a surge in the spread of misinformation on various topics, including health and politics. However, there is a notable gap in research on the spread of scientific and space-based misinformation on social media platforms, such as Twitter. Existing studies have focused primarily on a limited range of topics, with some examining user engagement metrics to identify discourse patterns.

For example, Bessi et al. [11] analysed the consumption of content on various conspiracy topics, such as the environment, health, diet, and geopolitics, through the lens of conspiracy theories and user engagement metrics, such as likes per post. This approach places greater emphasis on users' consumption behaviours and their connections to the wider community. On the contrary, our research aims to analyse the topics of scientific and space misinformation from a different perspective, with more focus on the topic itself than on users.

Other complementary studies have explored similar topics on various social media platforms such as TikTok, YouTube, Facebook, and Instagram [12], [13], [14], and [15]. However, engagement metrics, such as likes, tweets, and shares, have not been widely explored in the context of misinformation involving multiple modalities, such as images.

User engagement patterns can be an essential component in the analysis of online misinformation. As demonstrated by Bessi et al. [11], user interactions with social media posts can provide information on how different topics are consumed and shared. Ellison et al. [16] defined likes as positive interactions, retweets as expressions used to share posts with a wider audience, and comments as positive or negative interactions between different users in a post. Other studies have also analysed engagement metrics related to political and health misinformation [17], [18], [19]. However, these studies did not fully consider the impact of multiple modalities, such as images, on the spread of misinformation. Our motivation to categorise images according to brightness was based on previous research in psychology. In particular, Camgöz et al. [9] found that highly saturated and bright colours in the background, such as yellow, green, and cyan, have different effects on user attention. However, recent research by Chen et al. [10] has shown that this may not be the case for conspiracy videos on YouTube, which have lower colour variance and brightness than counter-conspiracy videos. Therefore, our research investigates how brightness and the type of visual misinformation affect users. In addition, we used a variety of features, including textual and visual features, to improve the detection methods of online misinformation on social media sites. Our methods include sentiment analysis, hate speech detection, and bot analysis to analyse the emotions behind posts on different topics. We also explore the connections between these three methods and the individual topics. Sentiment analysis has been widely used by social scientists (Medhat et al. [20], Zaeem et al. [21], Bhutani et al. [22]) to identify the discourse of posts on various social media platforms. Recent research has focused on the use of sentiment analysis and topic modelling to detect online misinformation on Twitter, as shown in studies conducted by Waheeb et al. [23] and Melton et al. [24]. In addition, interdisciplinary research has explored hate speech and bot analysis, such as Ferrara et al. [25].

Studies by Cinelli et al. [2], Kalantari et al. [26] and Giachanou and Rosso [27] have highlighted the role of anonymity and easy access to social networks in the spread and influence of online misinformation, particularly on polarising issues.

Previous studies have focused on identifying hate speech on various social media platforms, including Twitter and Facebook, using machine learning algorithms [28], [29], [30], [31]. Researchers have also analysed the characteristics of users who produce hateful content [32], [33] and those who are targeted by such speech [30].

Studies by Mathew et al. [32] and Ottoni et al. [34] investigated the impact of counter-reply speech and the prevalence of hatred, violence, and discriminatory bias in YouTube channels associated with right-wing content.

Ottoni et al. [34] conducted an analysis of right-wing channels on YouTube, focusing on detecting hatred, violence, and discriminatory bias. The authors observed that these channels often contain detailed content related to issues such as war and terrorism and also have a higher frequency of negative terms, including those related to aggression and violence. The findings of this study suggest that right-wing YouTube channels may be a source of online misinformation and may promote hateful and violent ideologies. A study conducted by Ottoni et al. [34] investigated the impact of external events on hate speech on Twitter and Reddit and found that violent extremism often leads to increased online misinformation, particularly among those advocating violence.

As businesses increasingly adopt digital technologies such as artificial intelligence,to gain a competitive advantage, they face a range of challenges. One such challenge is the proliferation of malicious social bots. Malicious botnets can be used to generate deception by programming bots to respond favourably to specific user profiles, as demonstrated in studies by Chu et al. [35] and Ferrara et al. [36]. Furthermore, malicious actors can use bots to cause harm, such as spreading anxiety and panic during emergencies such as the COVID-19 pandemic [37], damaging the reputation of a company, influencing political opinions [35], or disseminating rumours and fake news [37].

Bouvier [38] investigated the effect of the echo chamber on social networks, while Awan et al. [39] Verma et al. [40] have studied the effects of misinformation on politics during the pandemic, Broniatowski et al. [41] Sajinika et al. [42] have focused on the spread of online misinformation related to health topics. Pen [43] has conducted a comprehensive study of signal propagation across complex networks, while our research specifically focuses on the dynamics of misinformation spread on Twitter. Our research provides a multifaceted analysis, focusing on the interaction between visual elements, sentiment analysis, and bot-driven activities on Twitter. We manually annotated more than 12,000 tweets, uncovering nuanced insights such as the distinction in hate speech between political and COVID-19 misinformation. Our study not only corroborates the findings of current research, but also offers a granular, topic-centric perspective. The depth and breadth of our dataset demonstrate the validity and significance of our contributions to the academic discourse on online misinformation.

III. DATA

We chose the MuMin dataset as an appropriate dataset consisting of a large amount of posts, including both text and images, related to misinformation, across a range of topics. The MuMiN dataset [3] contains, in the largest version, 21 million Twitter posts pertaining to 26,000 Twitter threads connected to almost 13,000 fact-checked statements from 115 different organisations, covering a large number of topics, events, and domains, in 41 languages.

The dataset comprises three datasets such as MuMiN-small, MuMiN-medium, and MuMiN-large. The MuMiN-small dataset consists of just over 2 thousand claims, 4 thousand threads covering 8 million posts from over 600 thousand users. Of particular use for this study is that it contains just over 1000 images associated with either misinformation or factual claims.

The topics in the data set were automatically assigned based on clustering, using DBSCAN, embeddings of the claims text. Of these 26 clusters were identified, however, we note that given the automated machine learning-based nature of the topic clustering, the clusters upon manual inspection were imperfect and typically not fine-grained

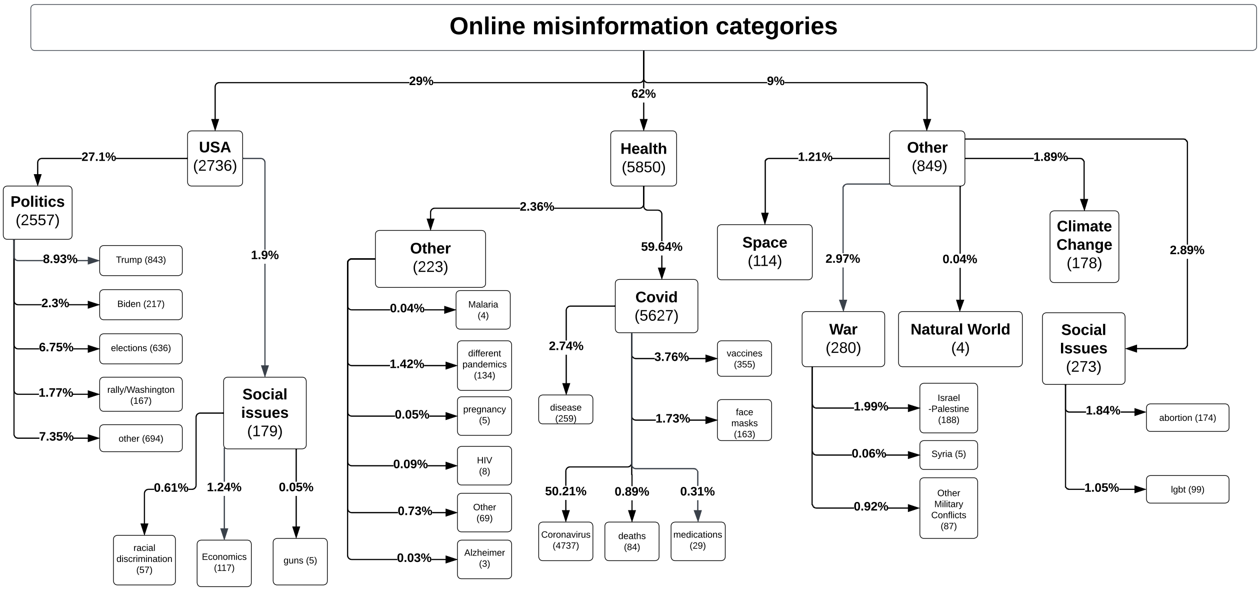


FIGURE 1. Classification of the topics.

enough for our purpose. On this basis, we were motivated to manually annotate the topics ourselves to ensure that the topics were coherent for our analysis.

IV. ANALYSIS AND RESULTS

A. RQ1: WHAT ARE THE TOPICS OF MISINFORMATION?

Research Question 1 aimed to identify misinformation topics on Twitter. Although the MuMiN platform provided keyphrase annotations for each misinformation post, these annotations were not sufficiently detailed for the hierarchical topic analysis required for this study. Therefore, we manually annotated 12,581 tweets labelled as misinformation using MuMiN. Rather than labelling the tweet directly, categories were assigned based on the claim to which the tweet was referring, providing a rich set of information for topic assignment.

Figure 1 shows the results of the topic annotation. The largest category of misinformation involved US politics, with more than 98% consisting of political misinformation. The second largest category of misinformation was related to health, with the COVID-19 subcategory comprising over 96% related to health. The highest level of the hierarchy showed that 9% of misinformation was neither related to the United States nor health and instead was categorised as ‘other‘ with the main subcategories being war, space and climate change.

Further analysis revealed that the main subcategories of the US category were politics (including topics relating to Trump, Washington political rally (United States Capitol attack on 6 January), election fraud, and Biden) and social issues (including topics relating to economic policies, gun legislation, and racial discrimination). The health category consisted mainly of COVID-19 and pandemic related issues, such as vaccination or face masks, but other issues such as HIV, Malaria, or Alzheimer’s were identified. The remaining identified topics were classified into the category Other due to their small size, including War, Space, Natural World, Climate Change, and Global Social Issues.

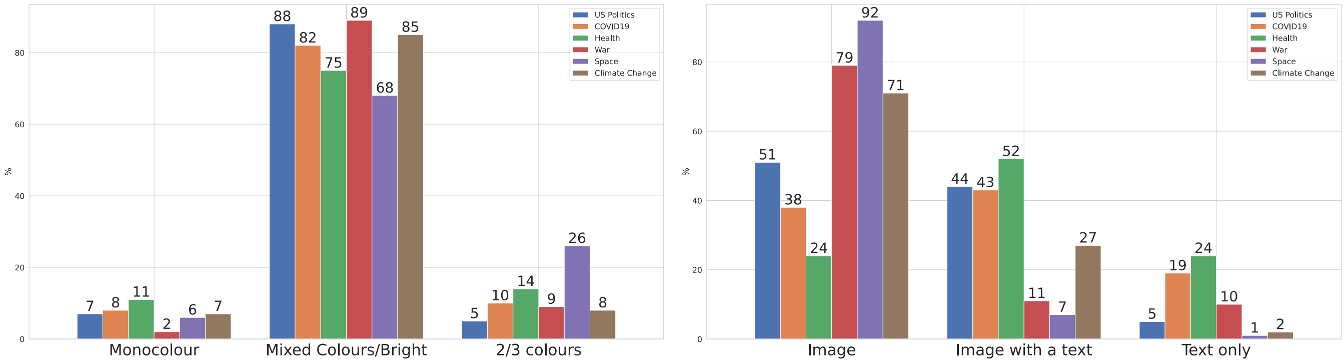
First, we tried different well-known topic modelling approaches, but were unable to successfully apply them to our dataset. They did not provide appropriate annotations based on tweet data. By manually annotating the misinformation hierarchy, this study provides a detailed insight into the topics and volume of misinformation on social media platforms. These labels can be used to develop more advanced detection systems and support further analysis. Research quantifying the amount of misinformation on different topics on any platform is scarce, but our findings align with the limited research available.

To answer the remaining research questions, we will characterise misinformation on different topics in terms of their discourse, the frequency of bots, and the use of different types of media.

B. RQ2: HOW DO USER ENGAGEMENT METRICS RELATE TO VISUAL FEATURES IN POSTS?

This research question aimed to explore whether online misinformation posts have different features on different topics and to investigate the relationship between user engagement metrics and visual content. As social media platforms such as Twitter allow the use of images in posts, this study aimed to measure the effectiveness of images in misinformation.

To achieve this, the MuMiN-small dataset, which contains approximately 1754 images labeled as misinformation, was analyzed. Each image was labeled based on its content and color using the approach proposed by Camgöz et al. [9]. For



(a) (b)

FIGURE 2. Comparative analysis of image distribution. This figure presents a bar chart comparing (a) the brightness and colour of images and (b) type of image across different topics. Figure 2(a) reveals that all topics exhibit high brightness and colour intensity. Figure 2(b) focuses on the same analysis, but specifically on the distribution of topics across various types of images, highlighting the variations observed within each specific category.

content, the images were manually labeled as containing only text, only an image, or both text and an image. Images were also labeled based on the number of colors in the image (one, two, or three colors, or greater than three colors).

The relationship between image type, colour, and user engagement metrics (ie retweets, replies, and quotes) was investigated to understand how the type and colour of images affect user engagement with social posts. 2b shows the analysis of the types of images analysed for the topics. We differentiated three different types of images such as: image, image with text, and text only. The results showed that plain images without text had the highest level of retweets (64% of retweeted posts contained this type of image) and had the highest relation in terms of frequency to quotes and replies. These findings suggest that users tend to engage more with online misinformation presented in the form of plain images than other types of images.

An analysis of the brightness and colour of images in different categories of misinformation online is presented in 2a. The results showed that mono-colour pictures (which contain two or three colours only) were much less frequent and had a similar frequency on different topics of online misinformation with an average of 6.83% (ranging from 2% for War to 11% for Health-Other category).

Figure 2b shows the frequency of posts that contain different types of images in different categories. The results showed that the highest values appear for plain images with an average of 59.1% (with the highest values of 92% for space and 71% for war). The images with text had an average value of 30.6% with similar higher values for the US Politics and Health categories (with an average value of 46.3%) and much lower values for the rest of the categories (with an average value of 15%). Text-only images had the lowest frequency among all images with a value of 10.1%.

Manual annotation analysis revealed that most of the images in the analysed dataset could be classified as plain images with many bright colours. However, the Health category differed from other categories, including Health-Covid19. Most of the images followed the pattern of being simple and colourful. However, this was not the case for the health (other) category where there were high values for different categories (i.e., image with text, text only, or mono-colour and 2/3 colours). Hence, the most popular image for a health (other) category would be a colourful image containing text. However, the analysis was limited to only 1754 images, with a small proportion of 356 images classified into the health category, which may limit the generalisability of the findings.

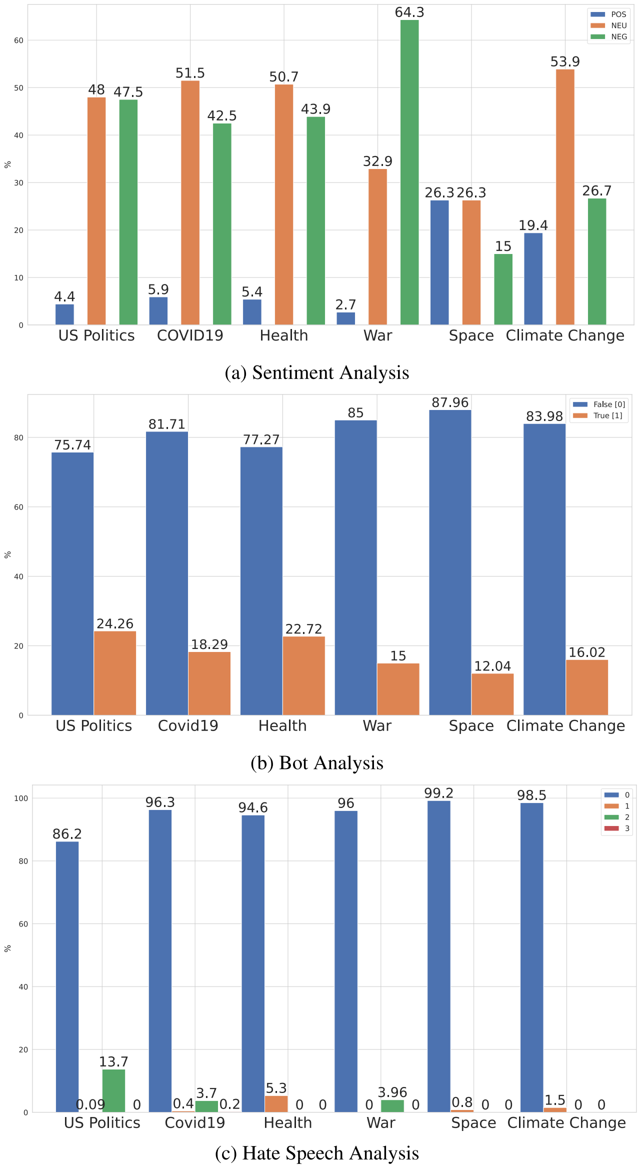
In general, the findings suggest that plain images without text are more effective in promoting user engagement with online misinformation. The colour and content of images in online misinformation vary across different topics, with the health category showing distinct patterns compared to other categories.

C. RQ3A: HOW DOES THE DISCOURSE BETWEEN USERS DIFFER ON DIFFERENT TOPICS OF MISINFORMATION?

Measuring attitudes toward posts relies on the engagement metrics of social media users, such as ‘likes’, ‘retweets’, and ‘replies’, drive their interactions with content [44]. Prior research has examined the main linguistic characteristics of content, as social interactions on social networks are based on language to express personality characteristics [45]. As such, sentiment analysis plays an important role in the analysis of online misinformation.

In this study, we focused on sentiments at the sentence level and classified each post as positive, neutral, or negative. We employed the transformer-based Twitter-XLM-roBERTa-base model to establish tweet sentiment for the identified topics. This is a multilingual language model trained on nearly 200 million tweets from eight datasets that encompass more than 30 different languages.

To investigate the distribution of online hate speech in relation to online misinformation, we used an English-only hate speech classifier for the social media content. Hate speech is typically defined as biased, aggressive, and malicious rhetoric directed at a person or group based on actual or perceived innate traits [46], [47].



(a) Sentiment Analysis

(b) Bot Analysis

(c) Hate Speech Analysis

FIGURE 3. Comparative analysis of sentiment, hate speech and bot distribution.

The hate speech classifier model used in this study was trained on a dataset of 103,191 YouTube comments and pre-trained using the BERT language model. Each post was classified into one of the four categories: acceptable [0], unsuitable [1], insulting [2], and violent [3].

1) RESULTS

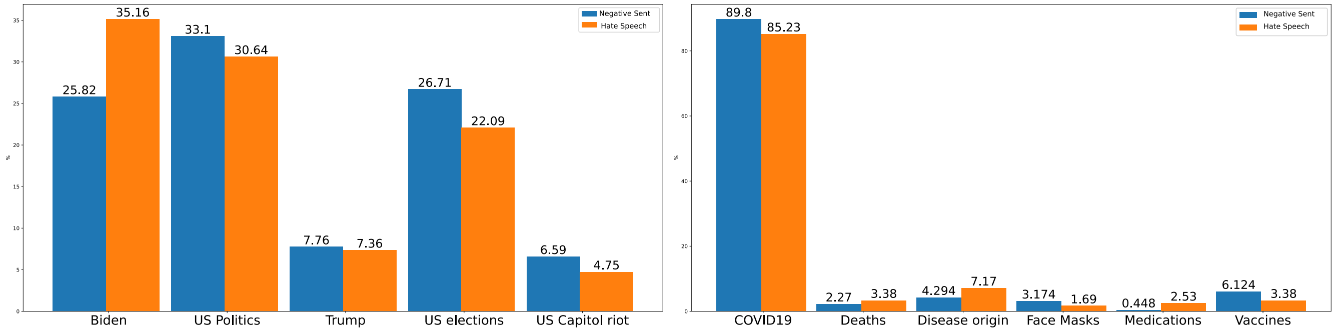
We applied both sentiment and hate speech models to our data, and our findings demonstrated significant disparities between the topics identified in RQ1. As shown in Figure 1, COVID-19 and political misinformation were the most prevalent online misinformation categories, accounting for 91% of all instances. We found that US political misinformation and war had the highest percentages of negative tweets and comments, with values of 47.5% and 64.32%, respectively. On the contrary, misinformation about politics, health, and war had the lowest levels of positive sentiment, with values of 4.4%, 5.5%, and 2.7%, respectively, compared to topics related to space and climate change, with values of 26.1% and 19.4% of tweets, respectively. Our hate speech analysis (3c) showed that political misinformation featured more offensive sentences, with a value of 13%, than any other category, by 10%. War also had a similar number of offensive sentences, with a value of 3.96%, as did Health and Health-COVID, with values of 5.3% and 3.66%, respectively. We selected these topics due to their high percentage of posts and replies. Our findings suggest that online misinformation with more than 40% negative sentences, as identified in our sentiment analysis, can have a high level of hate speech identified for both tweets and replies. Our findings are consistent with those of Hswen et al. [48], who found a strong relationship between online misinformation and hate speech in their research on anti-Asian sentiment during COVID-19. Hswen et al. [48] findings also showed that users expressed a high level of hate speech towards Asian culture and society after being exposed to particular online misinformation on Twitter, which confirms our findings. Based on our results, we confirm that there are differences between online misinformation and topics, especially regarding US political misinformation.

Figure 4 shows a more granular analysis of the two most popular topics, Politics and COVID-19. The majority of COVID-19 negative and offensive statements are classified in the coronavirus category, which contained general tweets and replies about the coronavirus without stating any particular theme. Conversely, political misinformation contains a high percentage in both categories.An intriguing occurrence was observed in the descriptive analysis of the Trump topic, with offensive statements being the most frequent among the other categories. However, the negative statements are much lower than those of other political issues and US elections. It seems that users were much more polarised, but less negative, for Trump compared to different subtopics. On the other hand, US elections and other political issues had the highest scores for negative sentences. Both Biden and the US rallies have had far fewer negative and offensive statements.

Overall, our results indicate that online misinformation is prevalent, especially in the context of politics and COVID-19, and can be associated with high levels of negative sentiments and hate speech. These findings underscore the need to pay greater attention to online misinformation and its potential harmful effects.

D. RQ3B: DOES THE LEVEL OF BOT ACTIVITY DIFFER ON DIFFERENT TOPICS OF MISINFORMATION?

The aim of this research question is to investigate whether the level of bot activity differs for different topics of online misinformation. Social bots refer to software-controlled accounts programmed to actively participate in social media platforms with the intention of influencing public opinion.



(a) US Politics

(b) COVID-19

FIGURE 4. Comparison of the sentiment analysis and hate speech detailed distribution for US politics and COVID-19 subcategories of the online misinformation. This figure compares the sentiment analysis and hate speech distribution for sub-topics of two online misinformation categories 'US Politics' and 'COVID19'. In the US Politics category (4a), negative sentiment prevails for most of the analysed subtopics. On the contrary, Figure 4(b) shows less negativity. This analysis highlights distinct sentiment and hate speech patterns within each subcategory of online misinformation.

To detect bots, we used a transformer-based model [49], trained on 229,573 users and 33,488,192 tweets. We chose this bot detection model because it employs multi-modal community-based detection measures and semantic analysis, making it more effective than other methods.

1) RESULTS

Figure 3c illustrates the differences in bot distribution across the investigated topics. Our findings indicate that US political misinformation has been associated with a higher level of bot usage (24.26%) compared to other topics, with an average value of 17%. Our manual data inspection revealed that the data contained social media posts related to the George Floyd protests in Minneapolis, US presidential elections, and post-presidential Capitol riots. Furthermore, we observed slightly higher positive bot scores for both categories of health, COVID-19, and Other Health topics, with values of 18.29% and 29.72%, respectively. In contrast, the categories related to space and war had the lowest positive bot values of 12.04% and 15%, respectively.

V. DISCUSSION

We present one of the first evaluations that assesses the similarities and differences between a diverse set of misinformation topics on Twitter, while taking into account the nature of the content and how users engage with it.

This research extends previous studies that have explored different types of online misinformation and how their spread online. We examine various topics of online misinformation, which is a significant improvement over previous research that has focused on a narrower range of topics.

Our study considers online discourse to be characterised by sentiment, hate speech, and the role of bots on many topics. Previous studies have mainly concentrated on a single on-line discourse metric and a single subject. Therefore, we focused our work on three main metrics of online discourse across different topics. One of our main findings is that political misinformation had one of the highest values for negative sentiment (the war category had the highest value of 64%, but there was a small sample of analysed tweets in this category) and had the highest value in the remaining categories for bot presence and offensive statements, accordingly, compared to an average of 17%, 6%, 11%.

We identified the highest number of bots within both political and health categories, which can also indicate a high level of online disinformation that is not considered in our research. The health topic is much less negative than political misinformation. Prior research in regards to analysing engagement and discourse across different topics is limited, as previous studies focused more on finding individual features of posts or the spread of online misinformation. However, by examining various topics and their similarities and differences for both visual and textual content, we improved the current ML or AI models by adding additional data.

Our research confirms the theory that the most popular circulated images are multicoloured and contain both text and images. This postulates the complexity of online misinformation and shows that detection models must be prepared for diverse inputs that can differ in many ways at a fine-granular level. Additionally, our data suggest that political and war-related misinformation has higher negative sentiment, offensive statements, and bot presence, compared to other topics. Health-related misinformation also had a high level of bot presence, but with lower negative sentiment and offensive statements.

Our findings can help policy makers understand what kind of misinformation has been spreading on Twitter and how to plan campaigns to make users more aware of online misinformation and how to spot its various features. Future research could explore how user decision changes over time when exposed to misinformation on different topics. This is an important area for further investigation, as online disinformation threatens both the foundations of fundamental democratic structures and the health of society. Hostile environments with a high level of tweets and replies posted by bots, whether coming from internal or foreign organisations, can cause people to doubt their beliefs. Therefore, it is crucial to address the problem of high bot presence in online social media, especially in topics that are more susceptible to online misinformation.

VI. LIMITATIONS

We have identified several limitations that should be taken into account when interpreting the results. Firstly, the external validity of the findings is questionable, as the dataset focused on specific historical milestones, which may not be applicable to other contexts. For example, tweets about Covid-19 may not accurately reflect discussions about health in general, and tweets about the US elections may not accurately represent conversations about politics in general. Furthermore, the manual labelling method used to address RQ1 is difficult to replicate, and the sample size mentioned in Section B, RQ2, limits the generalisability of the findings. Furthermore, the model for detecting hate speech was initially trained on comments sourced from YouTube, not Twitter, which could affect the applicability of the model when transferred from one platform to the other, as the text sizes, user behaviours, and the overall nature of interactions on these two social platforms are significantly different.

VII. CONCLUSION

In this study, we sought to understand and characterise how users engage with misinformation on a wide range of topics on the popular Twitter social media platform, as well as the relationship between visual elements and engagement. We annotated approximately 13 thousand misinformation posts into a topic hierarchy to understand the prevalence of different topics within the misinformation domain. From this, we proposed and answered research questions surrounding user engagement with different types of misinformation, as well as the levels of hate speech, bot activity, and sentiment types. Furthermore, to acknowledge the role and popularity of different types of media on Twitter, we manually labelled a number of images based on the visual format of the image and contrasted this between misinformation topics. Our results show that misinformation related to the United States and health dominated the misinformation landscape of the collected data. Furthermore, we found that political misinformation had the highest levels of hate speech and bot activity, while COVID-19 related misinformation was the most negative.

REFERENCES

[1] Y. Kwong, "The dynamics of mainstream and internet alternative media in Hong Kong: A case study of the umbrella movement," *Int. J. China Stud.*, vol. 6, pp. 273–295, Dec. 2015.

[2] M. Cinelli, A. Pelicon, I. Mozetič, W. Quattrociocchi, P. K. Novak, and F. Zollo, "Dynamics of online hate and misinformation," *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, Nov. 2021.

[3] D. S. Nielsen and R. McConville, "MuMiN: A large-scale multilingual multimodal fact-checked misinformation social network dataset," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2022, pp. 3141–3153.

[4] M. Avram, N. Micallef, S. Patil, and F. Menczer, "Exposure to social engagement metrics increases vulnerability to misinformation," 2020, *arXiv:2005.04682*.

[5] J. Berger and K. L. Milkman, "What makes online content viral?" *J. Marketing Res.*, vol. 49, no. 2, pp. 192–205, Apr. 2012.

[6] J. Berger and K. Milkman, "Social transmission, emotion, and the virality of online content," *Wharton Res. Paper*, vol. 49, no. 2, pp. 1–52, 2010.

[7] J. Berger and K. L. Milkman, "Social transmission and viral culture," Dept. Marketing, Univ. Pennsylvania, Philadelphia, PA, USA, Tech. Rep. 10-114, 2010.

[8] J. Berger, *Contagious: Why Things Catch On*. New York, NY, USA: Simon & Schuster, 2016.

[9] N. Camgöz, C. Yener, and D. Güvenç, "Effects of hue, saturation, and brightness: Part 2: Attention," *Color Res. Appl.*, vol. 29, no. 1, pp. 20–28, Feb. 2004. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/col.10214

[10] K. Chen, S. J. Kim, Q. Gao, and S. Raschka, "Visual framing of science conspiracy videos: Integrating machine learning with communication theories to study the use of color and brightness," *Comput. Commun. Res.*, vol. 4, no. 1, pp. 98–134, Feb. 2022.

[11] A. Bessi, F. Zollo, M. D. Vicario, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Trend of narratives in the age of misinformation," *PLoS ONE*, vol. 10, no. 8, Aug. 2015, Art. no. e0134641.

[12] L. Shang, Z. Kou, Y. Zhang, and D. Wang, "A multimodal misinformation detector for COVID-19 short videos on TikTok," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 899–908.

[13] H. O.-Y. Li, A. Bailey, D. Huynh, and J. Chan, "YouTube as a source of information on COVID-19: A pandemic of misinformation?" *BMJ Global Health*, vol. 5, no. 5, May 2020, Art. no. e002604.

[14] P. M. Massey, M. D. Kearney, M. K. Hauer, P. Selvan, E. Koku, and A. E. Leader, "Dimensions of misinformation about the HPV vaccine on Instagram: Content and network analysis of social media characteristics," *J. Med. Internet Res.*, vol. 22, no. 12, Dec. 2020, Art. no. e21451.

[15] A. Heydari, J. Zhang, S. Appel, X. Wu, and G. Ranade, "YouTube chatter: Understanding online comments discourse on misinformative and political YouTube videos," 2019, *arXiv:1907.00435*.

[16] N. B. Ellison, C. Steinfield, and C. Lampe, "The benefits of Facebook 'friends': Social capital and college students' use of online social network sites," *J. Comput.-Mediated Commun.*, vol. 12, no. 4, pp. 1143–1168, Jul. 2007.

[17] M. L. Kornides, S. Badlis, K. J. Head, M. Putt, J. Cappella, and G. Gonzalez-Hernadez, "Exploring content of misinformation about HPV vaccine on Twitter," *J. Behav. Med.*, vol. 46, pp. 239–252, Jul. 2022.

[18] A. Jamison, D. A. Broniatowski, M. C. Smith, K. S. Parikh, A. Malik, M. Dredze, and S. C. Quinn, "Adapting and extending a typology to identify vaccine misinformation on Twitter," *Amer. J. Public Health*, vol. 110, no. S3, pp. S331–S339, Oct. 2020.

[19] J. Penney, "It's my duty to be like 'this is wrong': Youth political social media practices in the Trump era," *J. Comput.-Mediated Commun.*, vol. 24, no. 6, pp. 319–334, Oct. 2019.

[20] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014.

[21] R. N. Zaeem, C. Li, and K. S. Barber, "On sentiment of online fake news," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Dec. 2020, pp. 760–767.

[22] B. Bhutani, N. Rastogi, P. Sehgal, and A. Purwar, "Fake news detection using sentiment analysis," in *Proc. 12th Int. Conf. Contemp. Comput. (IC3)*, Aug. 2019, pp. 1–5.

[23] S. A. Waheeb, N. A. Khan, and X. Shang, "Topic modeling and sentiment analysis of online education in the COVID-19 era using social networks based datasets," *Electronics*, vol. 11, no. 5, p. 715, Feb. 2022.

[24] C. A. Melton, O. A. Olusanya, N. Ammar, and A. Shaban-Nejad, "Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence," *J. Infection Public Health*, vol. 14, no. 10, pp. 1505–1512, Oct. 2021.

[25] E. Ferrara, S. Cresci, and L. Luceri, "Misinformation, manipulation, and abuse on social media in the era of COVID-19," *J. Comput. Social Sci.*, vol. 3, no. 2, pp. 271–277, Nov. 2020.

[26] N. Kalantari, D. Liao, and V. G. Motti, "Characterizing the online discourse in Twitter: Users' reaction to misinformation around COVID-19 in Twitter," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 4371–4380.

[27] A. Giachanou and P. Rosso, ''The battle against online harmful information: The cases of fake news and hate speech,'' in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.* New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 3503–3504, doi: 10.1145/3340531.3412169.

[28] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, ''Deep learning for hate speech detection in tweets,'' in *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp. 759–760.

[29] T. Davidson, D. Warmsley, M. Macy, and I. Weber, ''Automated hate speech detection and the problem of offensive language,'' in *Proc. Int. AAAI Conf. Web Social Media*, 2017, vol. 11, no. 1, pp. 512–515.

[30] J. Qian, M. ElSherief, E. Belding, and W. Yang Wang, ''Hierarchical CVAE for fine-grained hate speech classification,'' 2018, *arXiv:1809.00088*.

[31] F. D. Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, ''Hate me, hate me not: Hate speech detection on Facebook,'' in *Proc. 1st Italian Conf. Cybersec. (ITASEC)*, 2017, pp. 86–95.

[32] B. Mathew, N. Kumar, R. Ravina, P. Goyal, and A. Mukherjee, ''Analyzing the hate and counter speech accounts on Twitter,'' 2018, *arXiv:1812.02712*.

[33] M. Horta Ribeiro, P. H. Calais, Y. A. Santos, V. A. F. Almeida, and W. Meira Jr., '''Like sheep among Wolves': Characterizing hateful users on Twitter,'' 2017, *arXiv:1801.00317*.

[34] R. Ottoni, E. Cunha, G. Magno, P. Bernardina, W. Meira Jr., and V. Almeida, ''Analyzing right-wing YouTube channels: Hate, violence and discrimination,'' in *Proc. 10th ACM Conf. Web Sci.*, May 2018, pp. 323–332.

[35] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, ''Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?'' *IEEE Trans. Depend. Sec. Comput.*, vol. 9, no. 6, pp. 811–824, Nov. 2012.

[36] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, ''The rise of social bots,'' *Commun. ACM*, vol. 59, no. 7, pp. 96–104, Jun. 2016.

[37] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, ''Sentiment analysis of COVID-19 tweets by deep learning classifiers—A study to show how popularity is affecting accuracy in social media,'' *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 106754.

[38] G. Bouvier, ''From 'echo chambers' to 'chaos chambers': Discursive coherence and contradiction in the #MeToo Twitter feed,'' *Crit. Discourse Stud.*, vol. 19, no. 2, pp. 179–195, 2022.

[39] I. Awan, P. Carter, H. Sutch, and H. Lally, ''Online extremism and Islamophobic language and sentiment when discussing the COVID-19 pandemic and misinformation on Twitter,'' *Ethnic Racial Stud.*, vol. 46, no. 7, pp. 1407–1436, May 2023.

[40] G. Verma, A. Bhardwaj, T. Aledavood, M. De Choudhury, and S. Kumar, ''Examining the impact of sharing COVID-19 misinformation online on mental health,'' *Sci. Rep.*, vol. 12, no. 1, p. 8045, May 2022.

[41] D. A. Broniatowski, D. Kerchner, F. Farooq, X. Huang, A. M. Jamison, M. Dredze, S. C. Quinn, and J. W. Ayers, ''Twitter and Facebook posts about COVID-19 are less likely to spread misinformation compared to other health topics,'' *PLoS ONE*, vol. 17, no. 1, Jan. 2022, Art. no. e0261768.

[42] H. Sajinika, S. Vasanthapriyan, and P. Wijeratne, ''Twitter sentiment analysis and topic modeling for online learning,'' in *Proc. 3rd Int. Conf. Adv. Res. Comput. (ICARC)*, Feb. 2023, pp. 262–267.

[43] P. Ji, J. Ye, Y. Mu, W. Lin, Y. Tian, C. Hens, M. Perc, Y. Tang, J. Sun, and J. Kurths, ''Signal propagation in complex networks,'' *Phys. Rep.*, vol. 1017, pp. 1–96, May 2023.

[44] R. Jaakonmäki, O. Müller, and J. V. Brocke, ''The impact of content, context, and creator on user engagement in social media marketing,'' in *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, vol. 50, Jan. 2017, pp. 1152–1160.

[45] J. A. Fuhse, ''The meaning structure of social networks,'' *Sociol. Theory*, vol. 27, no. 1, pp. 51–73, 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9558.2009.00338.x

[46] R. Faris, A. Ashar, U. Gasser, and D. Joo, ''Understanding harmful speech online,'' Berkman Klein Center, Cambridge, MA, USA, Tech. Rep. 2016-21, 2016.

[47] R. Cohen-Almagor, ''Fighting hate and bigotry on the internet,'' *Policy Internet*, vol. 3, no. 3, pp. 1–26, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.2202/1944-2866.1059

[48] Y. Hswen, X. Xu, A. Hing, J. B. Hawkins, J. S. Brownstein, and G. C. Gee, ''Association of '#COVID19' versus '#Chinesevirus' with anti-Asian sentiments on Twitter: March 9–23, 2020,'' *Amer. J. Public Health*, vol. 111, no. 5, pp. 956–964, May 2021, doi: 10.2105/AJPH.2021.306154.

[49] S. Feng, H. Wan, N. Wang, J. Li, and M. Luo, ''TwiBot-20: A comprehensive Twitter bot detection benchmark,'' in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2021, pp. 4485–4494.

DOMINIKA NADIA WOJTCZAK received the M.Sc. degree in computer science from the University of Bath. She is currently pursuing the Ph.D. degree with the ESPRC Centre for Doctoral Training in Cyber Security, University of Bristol. Her research interests include online misinformation, social network analysis, hate speech, and behavioral science.

CLAUDIA PEERSMAN is currently a Senior Research Associate with the Cyber Security Research Group, University of Bristol. Her research interests include text mining and cyber security and focuses on developing new tools and techniques to support law enforcement agencies in their investigations pertaining to cyber crime. More particularly, her work has focused on automatically detecting criminal media on P2P networks and identifying deceptive users in online social media.

LUISA ZUCCOLO received the Graduate degree in physics, the M.Sc. degree in epidemiology from the London School of Hygiene and Tropical Medicine, and the Ph.D. degree in genetic epidemiology from the University of Bristol, U.K., with Prof. George Davey Smith. She is an epidemiologist with expertise in causal inference applied to population health. During the Graduate degree, she obtained a fellowship from the University of Turin, Italy, in cancer epidemiology and surveillance. Then, she moved to the University of Bristol, and she was awarded a Predoctoral Fellowship from the U.K. Medical Research Council. Then, she was awarded a second MRC Fellowship in population health science and epidemiology, after which, in 2018, she secured a tenured position with the University of Bristol. Her past research includes the causal effects of alcohol on health, in particular of prenatal alcohol exposure, using methods and designs that improve causal inference. More recently, she has focused on maternal and child health, researching barriers to and effects of prolonged breastfeeding, the impact of COVID-19 on fertility and pregnancy outcomes, and misinformation around public health messaging on social media.

RYAN MCCONVILLE received the Ph.D. degree from the Centre for Secure Information Technologies (CSIT), Queen's University Belfast, in 2017. He was appointed as a Lecturer in data science, machine learning and AI with the Intelligent Systems Laboratory and the Department of Engineering Mathematics, University of Bristol, in September 2019. He researched large scale unsupervised machine learning for complex data with Queen's University Belfast. He has worked with inter-disciplinary academic and industrial partners on numerous projects, including large-scale fraud detection and large-scale pervasive personal behavior analysis for clinical decision support. His research interests include unsupervised machine learning, deep learning on multimodal and complex data with applications to social network analysis, recommender systems, healthcare, and cybersecurity.