COMMUNICATION & SOCIETY

Special issue

Virality, only the tip of the iceberg: ways of spread and interaction around COVID-19 misinformation in Twitter

Guillermo Villar-Rodríguez
https://orcid.org/0000-0001-7942-2879
guillermo.villar@upm.es
Univ. Politécnica de Madrid

Mónica Souto-Rico
https://orcid.org/0000-0002-9315-7861
msouto@inst.uc3m.es
Universidad Carlos III de Madrid

Alejandro Martín
https://orcid.org/0000-0002-0800-7632
alejandro.martin@upm.es
Univ. Politécnica de Madrid

Submitted
February 8th, 2022
Approved
March 7th, 2022

Communication & Society
ISSN 0214-0039
E ISSN 2386-7876
doi: 10.15581/003.35.2.239-256
www.communication-society.com

2022 – Vol. 35(2)
pp. 239-256

How to cite this article:
Villar-Rodríguez, G., Souto-Rico, M. & Martín, A. (2022). Virality, only the tip of the iceberg: ways of spread and interaction around COVID-19 misinformation in Twitter. *Communication & Society, 35*(2), 239-256.

Abstract

Misinformation has long been a weapon that helps the political, social, and economic interests of different sectors. This became more evident with the transmission of false information in the COVID-19 pandemic, compromising citizens' health by anti-vaccine recommendations, the denial of the coronavirus and false remedies. Online social networks are the breeding ground for falsehoods and conspiracy theories. Users can share viral misinformation or publish it on their own. This encourages a double analysis of this issue: the need to capture the deluge of false information as opposed to the real one and the study of users' patterns to interact with that infodemic. As a response to this, our work combines the use of artificial intelligence and journalism through fact-checked false claims to provide an in-depth study of the number of retweets, likes, replies, quotes and repeated texts in posts stating or contradicting misinformation in Twitter. The large sample of tweets was collected and automatically analysed through Natural Language Processing (NLP) techniques, not to give all the attention only to the posts with a big impact but to all the messages contributing to the expansion of false information or its rejection regardless of their virality. This analysis revealed that the diffusion of tweets surrounding coronavirus-related misinformation is not only a domain of viral tweets, but also from posts without interactions, which represent most of the sample, and that there are no big differences between misinformation and its contradiction in general, except for the use of replies.

Keywords

Misinformation, artificial intelligence, Twitter, COVID-19, Natural Language Processing (NLP).

1. Introduction

False information has always existed. There are known cases of the spread of falsehoods as old as the battle in which the Bedouins allowed themselves to be stopped in order to manipulate Ramses II (Cline, 2021), or the so-called Great Moon Hoax in 1835 where it was reported that life and culture had been found on the moon (Chorąś Michałand Demestichas *et al*., 2021). Also known are the political campaigns in the first and second World Wars to

demonise Germans in the first and to cast doubt on Nazi atrocities in the second (Neander & Marlin, 2010). Joseph Goebbels was the Nazi propaganda mastermind who made most use of false information (Herzstein, 1978).

Although it is an issue that has been present for a long time, the spread of social media has brought back to the table how easy it is to launch mass information that is real or not (Imran *et al*., 2015). There have been several elements that have reinforced the existence of misinformation in Online Social Networks (OSN): on the one hand, the need for immediate information that all social network users have today; On the other hand, the failure to check the profiles that users follow and the interactions (Viswanath *et al*., 2009).

Some authors (Said-Hung *et al*., 2021) demonstrated how the focus on false information grew exponentially in academia, but also how this issue is named differently. The popular expression *fake news*, leading the number of articles about false information, is also used as a weapon to attack certain media, regardless of its content (Said-Hung *et al*., 2021). Although fake news could be framed as falsehoods under the shape of news, there is political intent behind this name and lack of consensus in its definition (Salaverría *et al*., 2020). Alternatively, two terms arise to properly define this reality*: disinformation*, for the deliberate false information; and *misinformation*, for the non-intentional one (Said-Hung *et al*., 2021; Salaverría *et al*., 2020). As a result, misinformation has also been applied to describe falsehoods in general, regardless of the interests behind.

In 2020, the issue of how misinformation is generated, how it is produced and how far it goes has been brought to the forefront (Mottola, 2020). In fact, misinformation has become especially important since the start of the COVID-19 pandemic (Jwa *et al*., 2019), having a great impact on citizens who used social networks as their main source of information (Demestichas *et al*., 2021; Islam *et al*., 2020). Although Said-Hung *et al*. (2021) estimated the increasing number of papers about false information from 2020 to 2022, the irruption of coronavirus makes the impact of this topic on academia even bigger.

From a journalistic approach, fact-checking is the practice used to counter misinformation. According to Graves and Amazeen (2019), fact-checking conceived as an internal routine a century ago in the United States not to publish false facts has evolved to an external practice in order to report the degree of truth or falsehood of a statement based on evidence. This conception of external fact-checking, also born in the United States at the beginning of this century, emerged as a new journalistic genre to combat misinformation (Graves & Amazeen, 2019).

In 2015, the International Fact-Checking Network (IFCN) was born through Poynter Institute as an alliance of non-partisan fact-checking organizations that must follow a "Code of Principles," which standardizes the steps of this process, ranging from the focus on statements from all political ideologies and the analysis of academic or official data for the verification to the transparency of the methodology used (Graves & Amazeen, 2019). The members of this network are recognized with a signatory (Mantzarlis, 2018), granting readers' trustworthiness. However, all actors involved, especially public administrations have seen the need to nip false information in the bud in a fast and reliable way (Quandt *et al*., 2019). This impact affects all countries to the point of the creation of the IGC (International Great Committee) on Misinformation and Fake News (Choraś Michałand Demestichas *et al*., 2021), among other initiatives. In the context of COVID-19, the Associate Director of IFCN, Cristina Tardáguila, confirmed the unprecedented dimensions of the challenge of countering coronavirus-related misinformation (Brennen *et al*., 2020). For this reason, tools for automatic detection of false information are studied.

From a computer-science approach, artificial intelligence is the field of research with the greatest projection today in terms of understanding the user-technology relationship (Túñez-López, Fieiras-Ceide, & Vaz-Álvarez, 2021). Currently, information is not consumed and transmitted in the same way, specifically because of the inclusion of artificial intelligence

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

in journalism and social media (Carlson, 2015; Vállez & Codina, 2018). However, artificial intelligence can also help in the fight against misinformation, and technology, computer science and social sciences have come together not only to quickly detect falsehoods (Oshikawa, Qian, & Wang, 2018).

The problem of the spread of misinformation has been studied from different technical areas and points of view. From feature extraction, just as an example of how diversely this problem can be tackled, misinformation can be classified with methods based on linguistic features, deception modelling, clusters, models for prediction and non-text cues (Parikh & Atrey, 2018). Nowadays, the disciplines in the study of misinformation have been refined, using Natural Language Processing (NLP), reputation analysis, network analysis and image recognition (Hirlekar & Kumar, 2020).

Analysing the text without the linguistic context using NLP is a priori the most obvious direction, with traditional methods to collect the word count (TF) or their weighted count (TF-IDF), in accordance with the foundations under the bag of words (Harris, 1954) and the study of the relative frequencies of words (Jones, 1972). Nevertheless, from this starting point, further elements from the message such as punctuation marks, emojis, number of URLs, positive and negative words (Castillo, Mendoza, & Poblete, 2011) or the style of the author through the number of names, verbs, adjectives or adverbs, among others (Zhou & Zafarani, 2018), can improve the models and analyses chosen.

Currently, NLP has moved towards neural networks to avoid the limitations the knowledge offered by more classical models (Sahoo & Gupta, 2021; Thota *et al*., 2018; Umer *et al*., 2020), generally by means of semantic embeddings that capture the meaning of the text In working with such systems, the authors have mixed simplified traditional approaches to news classification (Riedel *et al*., 2017), but the semantic embeddings of large pieces of texts have also demonstrated to counter misinformation by themselves (Anjali, Reshma & Lekshmy, 2019). Devlin *et al*. (2018) saw the emergence of transformers like BERT, providing embedding that also captures the linguistic context and thus boosting the capacity of detecting misinformation (Jwa *et al*., 2019).

Research against COVID-19 misinformation is not exempt from these methods. Transformer-based embeddings have demonstrated a key role in facing false information about coronavirus (Raha *et al*., 2021) to the point of even training transformer models with corpora about COVID-19 (Wani *et al*., 2021). This responds to the urge to combat misinformation as part of the fight of Artificial Intelligence against coronavirus (Nguyen *et al*., 2020; Shorten *et al*., 2021) and to the importance of NLP to address COVID-19 misinformation under these circumstances (Shorten *et al*., 2021).

However, technology makes it possible not only to collect the text of the misinformation but also the comments, or the account from which it is launched. In fact, misinformation is often initiated behind anonymous accounts (Xu *et al*., 2019) or even by bots (Himelein-Wachowiak *et al*., 2021) and algorithms can be fed with interaction variables such as the number of likes of a post to detect if it is true or misleading (Tacchini *et al*., 2017). This initially indicates that NLP is necessary but also an attention to the dynamics of the networks where falsehoods are harvested.

This leads to the concept of virality. In the conception of virality as popularity obtained from a user-to-user contagion like the spread of a virus, Goel *et al*. (2016) distinguish between two models of infection: broadcast models, where an only node infects the rest, and viral models, where each node infects some others, which can also infect, creating a cascade. Among other considerations, the authors discuss that popularity can be a mixture of virality and broadcasting through influential nodes and deep cascades (Goel *et al*., 2016).

The field of Social Network Analysis (SNA) goes in this direction by analysing the propagation of posts and the communities bolstering them. In this sense, Zubiaga *et al*. (2018) make an overview of how rumours are spread through networks and of the research that goes

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

to their origin or tracks their diffusion through keyword graphs. Furthermore, there are other approaches such as methods that identify the initiator and study its feedback (Sharma *et al*., 2019) or that analyse how content is propagated and checks the credibility of the headlines, the source, the comments and the disseminators (Tschiatschek *et al*., 2018).

Twitter is today's most widely used social network for political, economic and social communication (Hussain *et al*., 2021). Detecting false information here is complicated when they go viral, and everyone reproduces them. SNA shows that misinformation in Twitter can spread much further than true information according to the depth of the post's cascades, users involved in their diffusion and time of its dissemination (Vosoughi *et al*., 2018).

In the context of COVID-19, SNA studies have depicted the communities generated through the hashtags #FilmYourHospital (Ahmed *et al*., 2020a) and #5GCoronavirus (Ahmed *et al*., 2020b) to understand the flow of misinformation spreaders in Twitter. Even before the pandemic, research analyzed the networks of the debates about vaccination and concluded that vaccination-oriented decisions can be affected inside these circles (Bello-Orgaz, Hernandez-Castro, & Camacho, 2017).

However, the problem also arises when these publications do not follow this virality to be targeted by this type of research, given that misinformation can also be expanded without being viral. Whereas the viral and broadcast models compared by Goel *et al*. (2016) assume a unique message for the diffusion, the same false information can be expressed in different ways and disseminated simultaneously by independent nodes, and it is not always associated with hashtags that ease its search.

This paper expects to overcome this obstacle: whereas the innovations here explained for NLP mainly cover text classification, Transformers have also been successful in detecting untagged misinformation directly on Twitter, from the similarity among false statements and their tweets (Huertas-García *et al*., 2021a; Huertas-García *et al*., 2021b) to the particular analysis of their inference (Huertas-Tato *et al*., 2021). This would enable the analysis beyond a viral tweet or the use of a hashtag by a community of users. In other words, this project offers NLP at the service of the analysis of metrics that preceded additional SNA approaches to also assess the understanding of OSNs better, in this case Twitter.

Through this overview, two hypotheses are formulated through the following research questions that will be answered in the following sections:

- H1. What is the proportion of tweets with a few interactions and without interactions in comparison to the rest?
  - RQ 1.1. What is the proportion of tweets with a few interactions and without interactions in comparison to the rest?
  - RQ 1.2. What is the proportion of users contributing to tweets with a few interactions or with no interactions in comparison to the rest?
- H2. The amount of support or rejection of misinformation is different depending on whether the tweet disseminates or contradicts it.
  - RQ 2.1. To what extent are proportions different among tweets that disseminate misinformation to those that contradict it depending on their number of interactions?
  - RQ 2.2. To what extent are proportions different among users that disseminate misinformation to those that contradict it depending on the number of interactions of the tweets?

The metrics that count as interactions, the methods to determine the agreement or contradiction of a false claim and the categories that sort tweets from no interactions to many of them will be presented in the following sections. Whereas the sum of tweets (for questions 1.1. and 2.1.) corresponds to the raw counts to later compute the proportions, the sum of users contributing to it (to obtain the proportions for questions 1.2. and 2.2.) is extracted through the sum of the tweets (and thus, of the authors of the tweets) and of all their interactions (the

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

users that react to them). For this reason, we will refer to the sums and proportions for questions 1.2. and 2.2. as "weighted" in Methodology.

2. Methodology

The methodology of this paper consists of retrieving a large sample of tweets that state a false statement or contradict it, instead of only posts chosen because their metrics, such as retweets, are successful. In this way, we do not store the conversation in terms of its impact but rather those publications that also help to spread this type of misinformation.

In line with this goal, the mechanics in FacTeR-Check (Martín *et al*., 2021) describe how to extract tweets related to a given claim and filter them according to their position (support or denial) of the claim. The pipeline of this article is inspired by this recent research and comprises the following steps: 1) collection of the selected COVID-19 pieces of false information from fact-checkers; 2) creation of search queries composed of different representative keywords to retrieve tweets from Twitter API (Application Programming Interface; 3) a labelling process through Natural Language Processing (NLP) techniques to determine if every downloaded tweet supports or denies the input claim.

2.1. *Data*

Twitter was the network of choice for the research on dissemination of posts that state or contradict misinformation. This implies selecting pieces of misinformation that can be found in this platform. In order to achieve this, fact-checking news from the IFCN (International Fact-Checking Network) were extracted following this criteria: 1) They are COVID-19-related, taking into account claims about the measures against the pandemic and the existence of coronavirus; 2) They are obtained from fact-checkers that have a current signatory from the IFCN (International Fact-Checking Network) (Mantzarlis, 2018); 3) They are not only restricted to one language, although the resulting queries are written in English in order to capture tweets with misinformation in this language; 4) They are not a product of lack of context, which hardens the creation of the query and, thus, they can be summarised in a sentence and found in Twitter without any need of disambiguation. 5) They are not repeated, to have a varied dataset of misinformation to later build the final database.

Each false information detected by fact-checkers is referenced through a claim stating that type of misinformation, because the process of Natural Language Inference needs this claim as input to filter the results rather than the news title that already debunks it. Successively, for each falsehood, a query has been designed with keywords, synonyms, variations, slang and logical operators towards the optimal search of this content in Twitter. For example, the claim "Vaccines contain fetal cells from abortions" has been transformed into "(corona-virus OR covid OR covid19 OR sars-cov-2 OR cov-19) (vaccine OR vaccines OR vaccinated OR vac OR vacs OR vax OR vaxes OR vaxxes OR vaxed OR vaxxed) (fetus OR fetuses OR fetal) cells."

Overall, the initial dataset is composed of 26 pieces of misinformation about the pandemic fact-checked by 13 different organisations to retrieve their related tweets. These fact-checking outlets are: AFP Fact-checking, Animal Político - El Sabueso, Aos Fatos, Check Your Fact, Chequeado, Colombiacheck, EFE Verifica, Facta, FactCheck.org, Full Fact, Newtral, Politifact and Re:Baltica. Although any other piece of misinformation matching our criteria could be also retrieved, these 26 claims have shown satisfactory results after their created queries were used for a manual search in Twitter. For this reason, only these 26 queries have been the input for the Twitter API to download the tweets for this work. The list of false claims can be grouped by topic as follows: eleven anti-vaccine claims and five denials of the pandemic, plus one that combines both topics; six mask-related claims; two false recipes/treatments, and one claim about the management of the pandemic. Additionally, six of them also attack relevant individuals or institutions.

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

Finally, filtered from an initial sample of 17,570 tweets, 2,837 examples between January 2020 and December 2021 with more than 99% of Entailment (1,735 tweets) or more than 99% of Contradiction (1,102) have been retrieved from 22 of the 26 queries of the pieces of false information through the Twitter API and stored in a MongoDB database. Retweeted posts have not been downloaded since this article only considers the metrics of the original tweets, including the number of retweets instead of further information from retweets themselves.

The examples of misinformation selected and the Twitter IDs that identify each post from this sample have been uploaded to the link http://aida.etsisi.upm.es/datasets/, in the section "Dataset with COVID-19 misinformation in English." This is because Twitter texts, metrics and users cannot be uploaded directly according to the Developer Policy of Twitter. This cited URL also contains the false claims chosen for the creation of queries and the tweets extraction.

2.2. *NLP Inference*

The NLP techniques used involve computational models able to determine if two sentences are related in terms of semantics and a Natural Language Inference (NLI) process that establishes if a sentence a, called hypothesis, is inferred with a sentence b, called premise, as input. In our scenario, the pair of sentences refers to a pair of a tweet and its false claim associated (through the query that has found it through Twitter API). For each pair, three probabilities are calculated based on its type of inference: 1) Entailment or, in our case, alignment between the input claim and the tweet; 2) Contradiction or negation of the false statement expressed by the tweet; 3) Neutrality between both sentences.

This process of assigning categories to tweets (Entailment, Contradiction or Neutrality) through probabilities is possible, like other NLP tasks, as a machine learning task in which the algorithms learn from a set of training data to establish patterns that result in the probabilities for the final labels. Martín *et al*. (2021) make a review of the datasets useful for training NLI-based models, which are composed of sets of two sentences with a label associated to them (one of the three mentioned). This review culminates in the explanation of cross-lingual NLI datasets, and the use of a Machine Learning architecture called Transformers, which justify their NLI implementation and our steps for this research.

Transformers, which also work through large corpora trained behind, are capable of encapsulating text semantics in vectors without losing the context in which they are shaped (Vaswani *et al*., 2017), following the research line of using state-of-the-art methods for falsehoods detection (Huertas-Tato *et al*., 2021) which makes the NLI step of FacTeR-Check (Martín *et al*., 2021) possible. This constitutes the final approach: a transformer model fine-tuned with a cross-lingual NLI dataset to evaluate if every downloaded tweet, encoded as a transformer-based vector, regardless of how differently it is expressed, matches the exact or the opposite meaning of the specific false information diffused (MacCartney, 2009).

3. Analysis

Once tweets with Entailment and Contradiction have been retrieved, these two will be the categories used for the comparison in each of the metrics. Depending on the quantity of that type of interaction, a tweet will fall into four groups that ease the contrast between Entailment and Contradiction. These indicators, also obtained through the Twitter API, are retweets, likes, replies and quotes. Moreover, repetitions of the same tweet have been added as the fifth variable for our study, which measures the number of posts that have the same texts (after excluding users, hashtags, and URLs). Finally, for each type of metric, the exploration is double: we examine the proportion of tweets for the four groups, but we also check the actual weight of them by summing all the interactions received in addition to the proportion of original tweets/creators.

An example that contextualizes this necessity of reporting the differences in the type of interaction among tweets can be the following: whereas the tweet "*Yep. We asked the Amish*

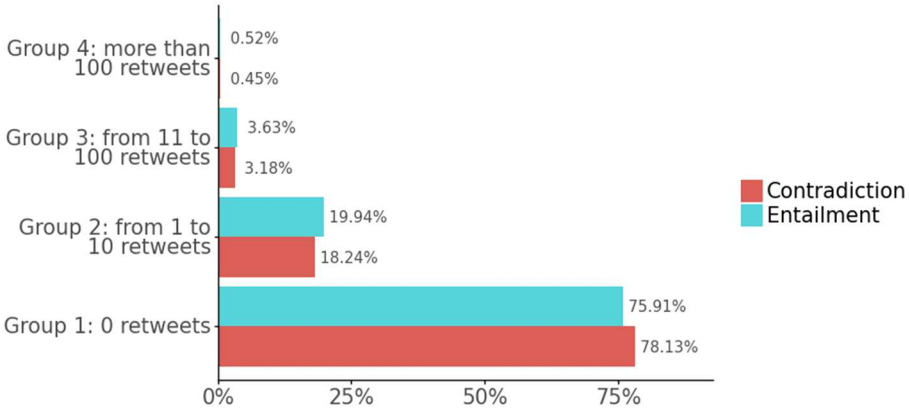
ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

why they are immune to COVID. They said, 'because we have no televisions'" had 239 retweets, 28 replies, 559 likes and 27 quotes when it was extracted, the tweet "*Recently someone asked the Amish people here in the US why Covid had not affected them. He said, 'Because we don't have TV'. That pretty much sums up the scamdemic*" had zero interactions (likes, retweets, quotes, replies) at the time of its retrieval. Both tweets refer to the same piece of misinformation (Entailment), but the response to them by Twitter users is the opposite.

3.1. *Retweets*

More than 75% of the content involving misinformation (Entailment) or rejecting it (Contradiction) does not include retweets. Regarding the remaining percent, less than 20% receive a moderate attention in terms of tweets, up to 10 retweets. Only around 3% of the tweets are relevant enough for users to stay between 11 and 100 retweets and only 1 of each 200 tweets (around 0.5%) surpasses these metrics.

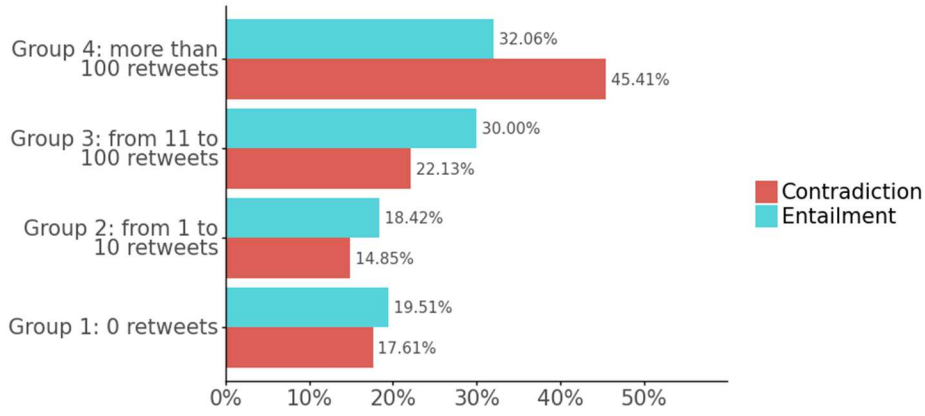
Figure 1: Proportion of shared tweets, grouped by their number of retweets.



Source: Own elaboration.

Concerning the differences found in this metric, the proportion of tweets with Entailment without retweets (75.91%) is slightly lower than the one with Contradiction (78.13%). This suggests that fewer attention may be paid to the tweets that reject a false claim, but without much difference in contrast to those that state it. Consequently, the proportion of tweets with Entailment is higher than the one of posts with Contradiction now they start to be retweeted, but in terms of maximum virality, the distance between these two types is minimum and, thus, not significant (0.52% for Entailment and 0.45% for Contradiction).

Figure 2: Proportion of users' posted tweets and their retweets, grouped by the tweets' number of retweets.



Source: Own elaboration.

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

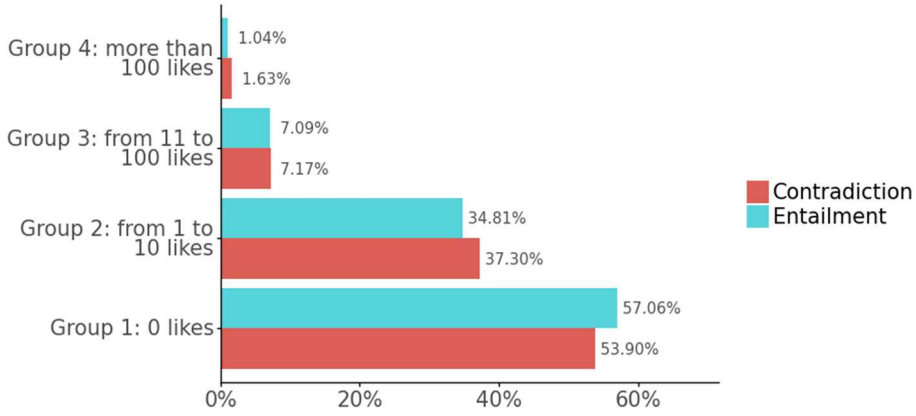
However, if we consider the real weight of these tweets in terms of retweets, given by the number of times a post has been shared (tweet and retweet), the statistics change. The tweets without retweets now only represent less than 20% of the total data, as well as those from 1 to 10 retweets. Now that not only tweets but also retweets are counted for the proportions, tweets with more than 10 retweets, and especially those with more than 100 retweets, have more weight in comparison to others.

With these aggregated sums for the proportions in the plot, the interpretations of the diffusion of posts with Entailment and Contradiction differ. The weight of tweets with 0 retweets that contradict misinformation (17.61%) is lower than the one from those that affirm it (19.51%). This lower weight of tweets with no associated retweets and that contradict the false information, although noteworthy, comes from the high proportion of users tweeting or retweeting posts that reject a false claim in the group with more than 100 retweets (45.41%), compared to the proportion of those that support that false information (32.06%). This means that the category of Entailment has more remaining percentage to be distributed in Groups 1,2 and 3 (67.94%) than the category of Contradiction (54.59%). Consequently, the percentages in these three groups are smaller in Contradiction than in Entailment.

3.2. *Likes*

The presence of likes already shows one difference with the metric of retweets: tweets without likes referring to the false claim or denying it are not much higher than 50%, in contrast to the figures higher than 75% from tweets without retweets. This indicates that there is more engagement with misinformation or its contradiction in terms of liking the post than of reposting it. Similarly, whereas tweets with more than 100 likes may be considered more important, they only represent less than 2% of a sample where most examples range from 0 to 10 likes.

Figure 3: Proportion of shared tweets, grouped by their number of likes.

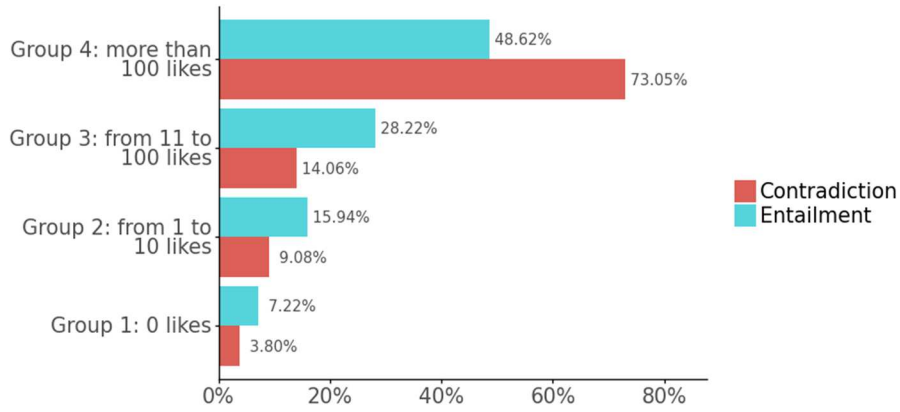


Source: Own elaboration.

The percentage of tweets without likes is higher in Entailment (57.06%) than in Contradiction (53.90%), meaning that the proportion of tweets with likes is bigger when tweets reject specific misinformation than when they are in favour of them. Moderate interactions are responsible for this: the proportion of tweets from 1 to 10 likes is higher when they contradict the false claim (37.30%) than when they just publish it (34.81%), but the difference of Entailment in comparison to Contradiction is not significant in the rest of the groups.

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

Figure 4: Proportion of users' posted tweets and their likes, grouped by the tweets' number of likes.



Source: Own elaboration.

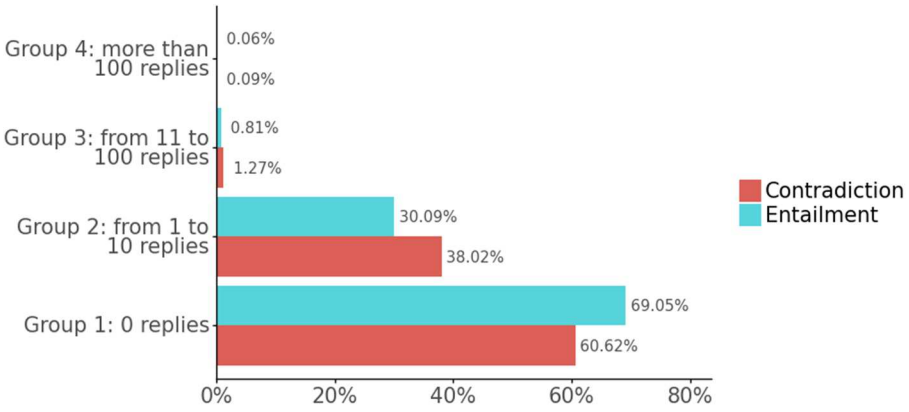
When we also consider the number of users that like a post in addition to the user that shares it, we obtain a second plot measuring the total weight of support under the shape of likes per group. In this case, the massive percentage of users that like certain tweets, much more than those that retweet them according to the resulting proportions, make the rest of the groups more invisible. In other words, a scarce number of tweets has more weight than all the tweets without likes, which corresponded to most of the sample.

How the proportions of our groups change completely when likes are added to the plot is even more evident in the category of Contradiction. In this type of tweets, a small percentage of posts with more than 100 likes (1.63%) changes into 73.05% of all the weight with this sum of interactions. According to these relative numbers, this relevant support in Group 4 significantly reduces the percentage of Contradiction in the rest of the subsets, whereas likes in Entailment are more distributed.

3.3. *Replies*

In line with the results from the preceding metrics, most tweets belong to the group with zero replies (more than 60%). However, parallelly to the use of likes, the percentage of tweets with interactions must be considered (more than 30% of the tweets have from 1 to 10 replies), in comparison to the smaller figures that tweets with at least a retweet gave in the first chart. Unsurprisingly, replies may not be as prone as retweets or likes to surpass certain numbers, and this is shown through Groups 3 and 4.

Figure 5: Proportion of shared tweets, grouped by their number of replies.

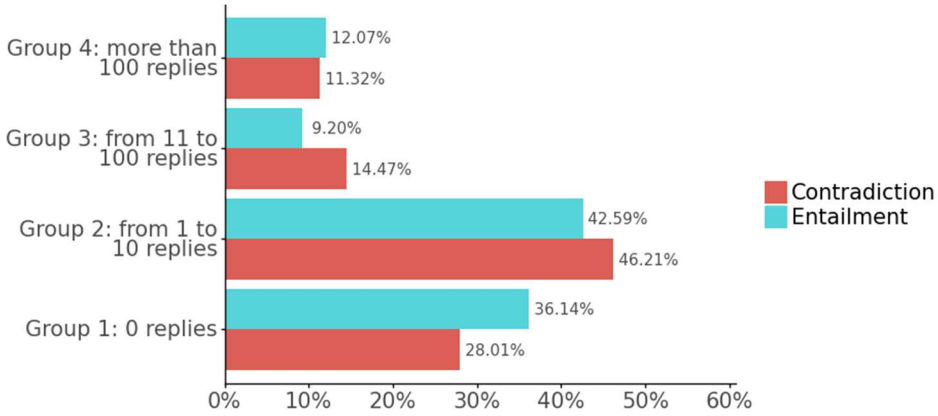


Source: Own elaboration.

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

In this case, the contrast between Contradiction and Entailment is more relevant than in the previous figures. Whereas 69.05% of the tweet's spreading misinformation does not have replies, only 60.02% of the tweets that contradict it do not receive this type of response, and for this reason the percentage of tweets with Entailment in Group 2 corresponds to 30.09% but the proportion of those with Contradiction is 38.02%. This suggests that users tend to reply more often when the content of a tweet is rejecting the false claim rather than stating it.

Figure 6: Proportion of users' posted tweets and their replies, grouped by the tweets' number of replies.



Source: Own elaboration.

With the aggregated sum of users involved in the direct conversation with the original tweets, either by creating the posts or by replying to them, the importance of the small proportion of tweets with many replies arise, especially in the subset of posts with more than 100 replies. However, the weights of tweets with no replies and with a small number of replies, from 1 to 10, still surpass the proportion of Group 4 despite their abundant responses.

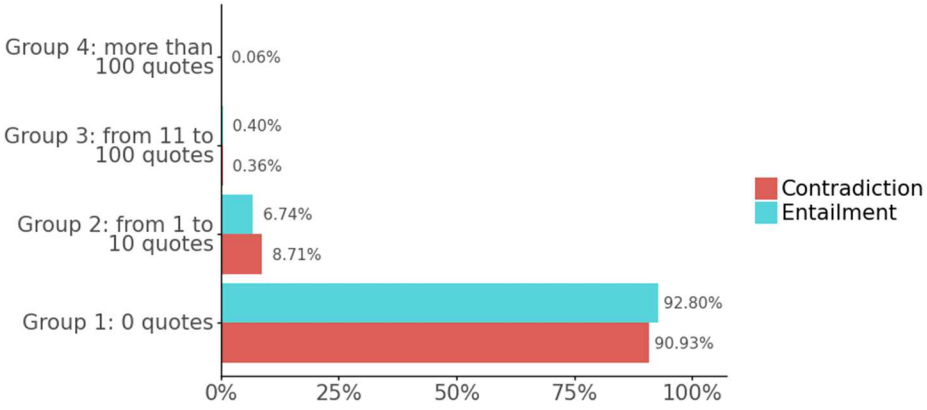
The findings about users frequently answering more to contradictions of false claims (46.21% from Group 2 and 14.47% from Group 3) in comparison to those targeting misinformation per se (42.59% from Group 2 and 9.20% from Group 3) remain visible in Group 1, but in this case, not only does the plot indicate that there are more tweets with Contradiction answered, but that the number of answers is also proportionally more in each category in comparison with Entailment. Group 4 is the exception to this rule, but with similar percentages for both groups (12.07% for Entailment and 11.32% for Contradiction).

3.4. *Quotes*

The metric of quotes reveals how in general users do not usually share misinformation or its contradiction by citing it with their own content, representing tweets without this type of reaction more than 90% of the sample. However, it was previously shown that the act of sharing through retweets was not common either for the misinformation spread or contradicted (more than 75% without retweets), but only for some cases that called users' attention. The rest of the proportion goes for tweets with a moderate number of quotes (Group 2), and rarely users go beyond (Groups 3 and 4) when tweets are related to misinformation.

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

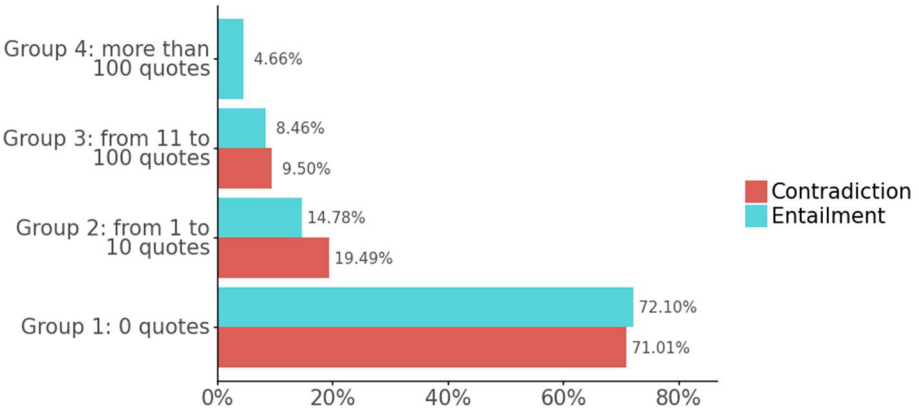
Figure 7: Proportion of shared tweets, grouped by their number of quotes.



Source: Own elaboration.

Although the differences between Entailment and Contradiction are not as relevant as the ones from replies, the tendency may be the same: tweets rejecting false claims will be quoted at least by an individual (8.71% belonging to Group 2) in more cases than tweets spreading those claims (6.74% in Group 2), although generally this type of interaction is not seen in most tweets (92.80% with Entailment and 90.93% with Contradiction in Group 1).

Figure 8: Proportion of users' posted tweets and their quotes, grouped by the tweets' number of quotes.



Source: Own elaboration.

The plot for the weighted proportions with the summed accounts that quoted tweets does not reveal great differences with the mere proportion of tweets, since most content still falls into Group 1. As expected, the number of users responsible for tweets in Groups 3 and 4 make these two categories have some more impact.

Percentages are not surprising either in the comparison between Entailment and Contradiction in contrast to the previous plot. It can be seen the proportion of reactions with Contradiction (19.49%) belonging to the Group from 1 to 10 quotes is clearly higher than the one with Entailment (14.78%), but all the reactions for tweets stating misinformation in Group 4 (4.66%) balance the percentages of tweets with this sort of interaction in both types.

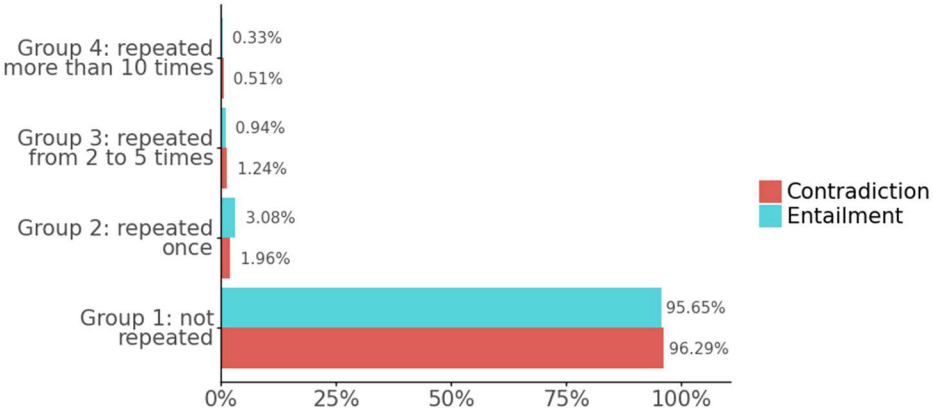
3.5. *Repetitions*

The plot about the number of repetitions in tweets demonstrates how this practice is not as usual (more than 95% of tweets are not repeated) if we analyse this on posts related to misinformation. Nevertheless, it is necessary to highlight that the rest of the proportion is not

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

only restricted to tweets repeated once, but also for tweets that are repeated more times (even more than 10 times). This manifests the intention of this type of practice.

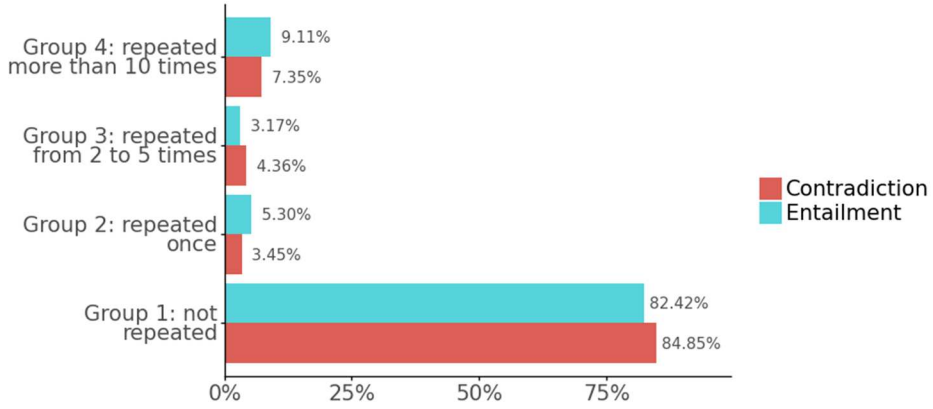
Figure 9: Proportion of shared tweets, grouped by their number of repetitions.



Source: Own elaboration.

Important differences between Entailment or Contradiction are not appreciated. The distance between the percentage of tweets without repetitions containing misinformation (95.65%) and of those that reject it (96.29%) is small. In the case of the posts repeated only once, the proportion of Entailment (3.08%) is slightly higher than the one of Contradiction (1.96%).

Figure 10: Proportion of users' posted tweets and their repetitions, grouped by the tweets' number of repetitions.



Source: Own elaboration.

However, the relevance of this practice is better seen when the total number of repetitions is added to the count of the original tweets. By looking at the figures of Group 1, now it can be perceived how more than 15% percent of the data on Twitter about misinformation is cloned content from existing tweets to increase the impact of the message, and the half of it belongs to tweets that have 10 or more clones.

With these weighted proportions, now the differences between Entailment and Contradiction are more visible. In Group 1, tweets contradicting false claims (84.85%) surpass those that state them (82.42%) in relative numbers. More than 9% of all the tweets spreading these false claims per se are clones of at least 10 more tweets, but also the 7% of all the tweets from this group contradicting these claims demonstrates how this strategy is followed to disseminate both types of posts.

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

4. Discussion

When we approach the subject of misinformation transmitted through social networks, the first thing we can think is that these falsehoods are initiated from an account that receives a lot of interaction from the rest of the community. In other words, that false information always has many likes or retweets so that it spreads and disperses. This research has shown how false information also spreads through other methods.

On the one hand, our work demonstrates that most information shared does not usually have a lot of interactions, but because users only give their massive support to a minimal number of tweets. This has been observed when the high proportions of tweets without likes and retweets lose relevance with the aggregated sum of practitioners of this type of reaction in addition to the creators of the primary content.

This states that viral misinformation and its viral rejections in Twitter are only a tiny part of all the content spread surrounding misinformation (answering to Research question 1.1.), but the number of users that collaborate in the spread make this little proportion stand out from the vast proportion of tweets without interactions (in response to Research question 1.2.). This does not mean, though, that communities that do not manifest their support by these means are not affected by the high percentage of tweets without interactions, which suggest that the fight against false information is also beyond virality. This ecosystem of viral and not viral misinformation-related tweets discovered throughout this paper confirms the first hypothesis.

In contrast to likes and retweets, the aggregated sums of users that reply, quote or repeat a tweet (copy and paste its text for another tweet) do not change the plot towards a major importance of tweets with this type of interactions, and the weight of users that publish tweets without success in these metrics still represents the largest proportion by far. This reflects those actions in favour or against a piece of misinformation previously shared may be more often lazy: a product of massively pressing the buttons Retweet and Like instead of a result of proactive elaborated answers through replies, quotes and repetitions, which require an extra effort.

The emergence of misinformation due to the act of cloning the false claims inside a tweet is manifested through the proportion (more than 15%) of tweets, but the number of clones does not make the group of posts with more than 10 repetitions outstanding. This indicates that bots, at least conceived as accounts repeating specific content, might not always be as predominant for this specific issue as expected, in contrast to previous research about COVID-19 misinformation (Himelein-Wachowiak *et al*., 2021). This goes in line with Vosoughi *et al*.'s research (2018) concluding that responsibility for falsehoods getting deeper in an OSN is mainly human, not machine-based. Moreover, the proportion of repetitions of tweets contradicting misinformation is also considerable, but it can also be because users directly share the headline of a news article that counters a piece of misinformation and its URL.

In response to Research question 2.1., The most notable difference between Entailment and Contradiction for the non-weighted proportions analysed can be found in replies. There are, by far, more contradictions of falsehoods than falsehoods per se, suggesting that misinformation is, paradoxically, less questionable or debatable than any type of content that contradicts it. The rest of differences from comparisons are not big enough to depict a gap between misinformation and its denial but, among them, two considerations may be more remarkable: in relative numbers, falsehoods retweeted are slightly higher than their contradictions, but Contradiction liked surpasses Entailment liked. This means that the support to misinformation is manifested through their diffusion, but their denial is embraced more silently, without being directly spread.

Nevertheless, answering to Research question 2.2., the figures with the aggregated sum of all users around the creation of the tweet and the type of interaction selected give a layer

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

of complexity. For example, both the retweets and the likes from Contradiction clearly surpass Entailment when all these interactions are added to the plot, in contrast to the small disparities found in the non-weighted proportions. This would show, then, that misinformation deniers also spread viral content. However, considering our sample and the very small percentages for tweets in Group 4 (without the weighted sum), differences among groups should be treated carefully, given that the popularity of an only tweet from this group may distort the whole proportion. This needs to check this type of weighted percentages more deeply (for future approaches about Research question 2.2.) and the similarities in some metrics for non-weighted proportions (as shown for Research question 2.1) indicate that we can confirm the second hypothesis in the case of parameters such as replies but not in general.

In any case, our analysis shows a change in the perception of the spread of false information. Our plots show that misinformation is disseminated with posts with few or no interactions in general rather than only through virality. In contrast to Vosoughi *et al.* (2018) and the fact of misinformation spreading through cascades quicker and further than false information, this study questions this behaviour as the only way of transmitting misleading content or its denial. Specifically, whereas Vosoughi *et al.* (2008) approach did not focus on the diffusion of all the content about the same claim without interactions, the refined extraction of our sample demonstrated that the wave of misinformation and its contradiction is also built under a massive number of tweets with no retweets or other metrics.

This change of focus made through our research leads to a contribution for a more refined detection of misinformation and for its tracking. The results of our paper expect to be closer to the real radiography of OSNs, like Twitter in this case, allowing journalists and fact-checkers to check the real flow of certain false claims beyond an only tweet that has gone viral. Given the overwhelming amount of false information challenging fact-checkers' routines (Brennen *et al.*, 2020; Grave & Amazeen, 2019), this may be useful to concentrate even more efforts against pieces of misinformation circulating in the shadow of viral tweets.

All in all, despite the importance of our results, three limitations can be addressed in future lines of study: firstly, the lack of a richer list of false claims, which could be solved by the automated creation of queries, as proposed by Martín *et al.* (2021) in FacTeR-Check, to fasten the process of the obtention of pieces of misinformation circulating in Twitter; secondly, the absence of the analyses of false information through its type of content to check the percentage of tweets with and without interactions grouped by the sort of false claim retrieved, in accordance with Vosoughi *et al.*'s (2018) findings stating that political misinformation succeeds more in the dissemination that other types of false information; finally, the necessity of other Twitter indicators, such as the number of followers of each node / user to compute "virality" coefficients (along with other factors, since viral tweets also belong to users with a few followers).

5. Conclusions

Overall, this paper shows that although attention is given to relevant tweets in terms of their type of reaction in the platform (retweets, likes, replies, quotes, repetitions), most tweets about misinformation do not have interactions beyond clicking on those posts and, thus, the first hypothesis is confirmed. This implies that: 1) not only is false information and its contradictions spread through virality, but they are also disseminated in communities without the need of receiving any interaction to scale in the ONS; 2) A certain text with misinformation or its contradiction may not be relevant by itself but by further characteristics of the tweet where it is posted, since the majority of analysed posts do not receive any type of response; 3) users mentioning false information may do not always interact with the tweets that have given it virality, suggesting that they publish this type of content because they have consumed it elsewhere.

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

Accordingly, the vast proportion around misinformation without any interactions invites analysis of the deluge of misinformation through methodologies like ours, which can extract the exact content that states or contradicts false claims, and that ignores if a tweet has been viral or not. This implies an alternative to previous research tracking the trajectory of the content of a post by only looking at the cascade generated by it (Vosoughi *et al*., 2018), since our results show that there is a considerable percentage of content unaware of the most viral cascades that may also have an effect in ONS.

Fortunately, despite the differences and patterns observed between Entailment and Contradiction, and the awareness of how misinformation can be encoded in jargon not initially captured by our queries, this article demonstrates that the mechanisms of the false information spread are not so different from the diffusion of posts that reject it. Although the original sample has many more posts of misinformation (1,735) than of its contradiction (1,102), the proportions of each type of tweets reveal that the fight against misinformation is also covered at all levels, from users that post primary content from elsewhere on Twitter without interactions to those that make it viral through their reactions. However, the second hypothesis cannot be totally rejected, because the use of replies has shown to be diverse depending on the type of tweets, encouraging future research of further indicators that separate Entailment from Contradiction. In any case, knowing that the dissemination of both types of content is not so dissimilar sheds some light in the fight against misinformation, a battle of forces that are now not far from each other in terms of their trajectory in online social networks, according to our results.

This work has been supported by the research project CIVIC: "Intelligent characterisation of the veracity of the information related to COVID-19", granted by BBVA Foundation Grants for Scientific Research Groups SARS CoV-2 and COVID-19, by the Spanish Ministry of Science and Innovation under FightDIS (PID2020-117263GB-100) and XAI-Disinfodemics (PLEC2021-007681) grants, by Comunidad Autónoma de Madrid under S2018/TCS-4566 grant, by European Commission under IBERIFIER - Iberian Digital Media Research and Fact-Checking Hub (2020-EU-IA-0252), by "Convenio Plurianual with the Universidad Politécnica de Madrid in the actuation line of Programa de Excelencia para el Profesorado Universitario" and by the research project DisTrack: "Tracking disinformation in Online Social Networks through Deep Natural Language Processing", granted by Barcelona Mobile World Capital Foundation.

References

Ahmed, W., Seguí, F. L., Vidal-Alaball, J., Katz, M. S. & others. (2020). Covid-19 and the "film your hospital" conspiracy theory: social network analysis of twitter data. *Journal of Medical Internet Research*, *22*(10), e22374.

Ahmed, W., Vidal-Alaball, J., Downing, J., Seguí, F. L. & others. (2020). COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data. *Journal of Medical Internet Research*, *22*(5), e19458.

Anjali, B., Reshma, R. & Lekshmy, V. G. (2019). Detection of counterfeit news using machine learning. *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, *1*, 1382-1386.

Bello-Orgaz, G., Hernandez-Castro, J. & Camacho, D. (2017). Detecting discussion communities on vaccination in twitter. *Future Generation Computer Systems*, *66*, 125-136.

Brennen, J. S., Simon, F. M., Howard, P. N. & Nielsen, R. K. (2020). *Types, sources, and claims of COVID-19 misinformation*. Oxford: University of Oxford.

Carlson, M. (2015). The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digital Journalism*, *3*(3), 416-431.

Castillo, C., Mendoza, M. & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th International Conference on World Wide Web*, 675-684.

Choraś Michałand Demestichas, K., Giełczyk, A., Herrero, Á., Ksieniewicz Pawełand Remoundou, K., Urda, D. & Woźniak, M. (2021). Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study. *Applied Soft Computing*, *101*, 107050.

Cline, E. H. (2021). 1177 BC. In *1177 BC*. Princeton University Press.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Goel, S., Anderson, A., Hofman, J. & Watts, D. J. (2016). The structural virality of online diffusion. *Management Science*, *62*(1), 180-196.

Graves, L. & Amazeen, M. A. (2019). Fact-checking as idea and practice in Journalism. In *Oxford Research Encyclopaedia of Communication*. Retrieved from https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-808

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146-162.

Herzstein, R. E. (1978). *The war that Hitler won: The most infamous propaganda campaign in history*. New York: Putnam Publishing Group.

Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H. A. & others. (2021). Bots and misinformation spread on social media: Implications for COVID-19. *Journal of Medical Internet Research*, *23*(5), e26933.

Hirlekar, V. V. & Kumar, A. (2020). Natural language processing based online fake news detection challenge –a detailed review. *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 748-754.

Huertas-García, Á., Huertas-Tato, J., Martín, A. & Camacho, D. (2021a). CIVIC-UPM at CheckThat! 2021: integration of transformers in misinformation detection and topic classification. *CLEF (Working Notes)*, 520-530.

Huertas-García, Á., Huertas-Tato, J., Martín, A. & Camacho, D. (2021b). Countering Misinformation Through Semantic-Aware Multilingual Models. *International Conference on Intelligent Data Engineering and Automated Learning*, 312-323.

Huertas-Tato, J., Martín, A. & Camacho, D. (2021). SILT: Efficient transformer training for inter-lingual inference. arXiv preprint arXiv:2103.09635

Hussain, A., Tahir, A., Hussain, Z., Sheikh, Z., Gogate, M., Dashtipour, K., Ali, A. & Sheikh, A. (2021). Artificial intelligence –enabled analysis of public attitudes on Facebook and Twitter toward Covid-19 vaccines in the United Kingdom and the United States: Observational study. *Journal of Medical Internet Research*, *23*(4), e26627.

Imran, M., Castillo, C., Díaz, F. & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, *47*(4), 1-38.

Islam, A. K. M. N., Laato, S., Talukder, S. & Sutinen, E. (2020). Misinformation sharing and social media fatigue during COVID-19: An affordance and cognitive load perspective. *Technological Forecasting and Social Change*, *159*, 120201.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*.

Jwa, H., Oh, D., Park, K., Kang, J. M. & Lim, H. (2019). exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences*, *9*(19), 4062.

MacCartney, B. (2009). *Natural language inference*. Standford, CA: Stanford University.

Mantzarlis, A. (2018). Fact-checking 101. *Journalism, Fake News \& Disinformation: Handbook for Journalism Education and Training*, 85-100.

Martín, A., Huertas-Tato, J., Huertas-García, Á., Villar-Rodríguez, G. & Camacho, D. (2021). FacTeR-Check: Semi-automated fact-checking through Semantic Similarity and Natural Language Inference. *ArXiv Preprint ArXiv:2110.14532*.

Mottola, S. (2020). Las *fake news* como fenómeno social. Análisis lingüístico y poder persuasivo de bulos en italiano y español. *Discurso \& Sociedad*, *3*, 683-706.

Neander, J. & Marlin, R. (2010). Media and Propaganda: The Northcliffe Press and the Corpse Factory Story of World War I. *Global Media Journal: Canadian Edition*, *3*(2).

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Hsu, E. B., Yang, S. & Eklund, P. (2020). Artificial intelligence in the battle against coronavirus (COVID-19): a survey and future research directions. *ArXiv Preprint ArXiv:2008.07343*.

Oshikawa, R., Qian, J. & Wang, W. Y. (2018). A survey on natural language processing for fake news detection. *ArXiv Preprint ArXiv:1811.00770*.

Parikh, S. B. & Atrey, P. K. (2018). Media-rich fake news detection: A survey. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 436-441.

Quandt, T., Frischlich, L., Boberg, S. & Schatto-Eckrodt, T. (2019). Fake news. *The International Encyclopedia of Journalism Studies*, 1-6.

Raha, T., Indurthi, V., Upadhyaya, A., Kataria, J., Bommakanti, P., Keswani, V. & Varma, V. (2021). Identifying COVID-19 fake news in social media. *ArXiv Preprint ArXiv:2101.11954*.

Riedel, B., Augenstein, I., Spithourakis, G. P. & Riedel, S. (2017). A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *ArXiv Preprint ArXiv:1707.03264*.

Sahoo, S. R. & Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, *100*, 106983.

Said-Hung, E., Merino-Arribas, M. A. & Martínez, J. (2021). Evolución del debate académico en la Web of Science y Scopus sobre *unfaking news* (2014-2019). *Estudios sobre el Mensaje Periodístico, 27*(3), 961-971.

Salaverría, R., Buslón, N., López-Pan, F., León, B., López-Goñi, I. & Erviti, M.-C. (2020). Desinformación en tiempos de pandemia: tipología de los bulos sobre la Covid-19. *El Profesional de La Información (EPI)*, *29*(3).

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M. & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(3), 1-42.

Shorten, C., Khoshgoftaar, T. M. & Furht, B. (2021). Deep Learning applications for COVID-19. *Journal of Big Data*, *8*(1), 1-54.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S. & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. *ArXiv Preprint ArXiv:1704.07506*.

Thota, A., Tilak, P., Ahluwalia, S. & Lohia, N. (2018). Fake news detection: a deep learning approach. *SMU Data Science Review*, *1*(3), 10.

Tschiatschek, S., Singla, A., Gómez Rodríguez, M., Merchant, A. & Krause, A. (2018). Fake news detection in social networks via crowd signals. *Companion Proceedings of the The Web Conference 2018*, 517-524.

Túñez-López, J.-M., Fieiras-Ceide, C. & Vaz-Álvarez, M. (2021). Impact of Artificial Intelligence on Journalism: transformations in the company, products, contents and professional profile. *Communication & Society*, *34*(1), 177-193.

Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G. S. & On, B.-W. (2020). Fake news stance detection using deep learning architecture (CNN-LSTM). *IEEE Access*, *8*, 156695-156706.

Vállez, M. & Codina, L. (2018). Periodismo computacional: evolución, casos y herramientas. *Profesional de La Información*, *27*(4), 759-768.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Viswanath, B., Mislove, A., Cha, M. & Gummadi, K. P. (2009). On the evolution of user interaction in Facebook. *Proceedings of the 2nd ACM Workshop on Online Social Networks*, 37-42.

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256

Vosoughi, S., Roy, D. & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151.

Wani, A., Joshi, I., Khandve, S., Wagh, V. & Joshi, R. (2021). Evaluating deep learning approaches for covid19 fake news detection. *CONSTRAINT@AAAI*, 153-163.

Xu, K., Wang, F., Wang, H. & Yang, B. (2019). Detecting fake news over online social media via domain reputations and content understanding. *Tsinghua Science and Technology*, *25*(1), 20-27.

Zhou, X. & Zafarani, R. (2018). Fake news: A survey of research, detection methods, and opportunities. *ArXiv Preprint ArXiv:1812.00315*, *2*.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, *51*(2), 1-36.

ISSN 2386-7876 – © 2022 Communication & Society, 35(2), 239-256