Fake News on the Covid-19 outbreak: a new metadata-based dataset for the analysis of Brazilian and British Twitter posts

Tuany Mariah Lima do Nascimento[1], **Laura Emmanuella Alves dos Santos Santana**[1], **Márjory Da Costa Abreu**[2]

[1]Universidade Federal do Rio Grande do Norte, lauraemmanuella@gmail.com

[2]Sheffield Hallam University, m.da-costa-abreu@shu.ac.uk

Abstract. *The dissemination of fake news is a problem that has already been addressed but by no means is solved. After the manipulation made by Cambridge Analytica which was based on classifying users by their political views and targeting specific political propaganda on the Brexit campaign, the Trump election and the Bolsonaro election, there is no doubt this issue can have a real impact on society. During a pandemic, any type of fake news can be the difference between life and death when the data shared can directly hurt the people who are believing in it. Moreover, there is also a new trend of using artificial robots to disseminate such news with a special target on Twitter which can be linked with political campaigns. Thus, it is essential that we identify and understand what kind of news is selected to be dressed as fake and how it is disseminated. This paper aims to investigate the dissemination of fake news related with Covid-19 in the UK and Brazil in order to understand the impact of fake news on public sector actions, social isolation and quarantine imposition. Those two case studies are well versed on the fake news dissemination. Our initial dataset of Twitter posts has focused on posts from four different cities (Natal, São Paulo, Sheffield and London) and has shown interesting pointers that will be discussed.*

1. Introduction

That we are living at exceptional times is undeniable. The coronavirus pandemic situation is a crisis of unprecedented consequences that will change our lives forever. One of the main helpers with knowing what is happening regarding the pandemic, what new treatments are point, what other countries are doing, etc, is the popularisation of social media, as Twitter, Facebook, Instagram, etc. Such tools could be used to make the public health recommendations wide spread, however, very often, these networks are used to publicise information that is not correct and have the simple goal of misinforming, being it for personal gain, lack of knowledge or malicious behaviour towards the other, and so on [Abd-Alrazaq et al. 2020].

Isolation/social distancing and vaccines have been considered the most efficient measure to prevent the collapse of health systems around the world. However, there has been an increase in the acceptance of the general public to comply with it due to the amount of fake news [Guitton 2020].

Moreover, there is an increasing lack of trust in the establishment, and any scientific information becomes irrelevant when you have governments that try to follow a "herd immunity" strategy or call Covid-19 as "only a mild flu". In some countries, such as the

USA, Brazil, the UK, and Sweden, there was still an attempt to undermine the severity of the illness and that gave a lot of fuel to the dissemination of fake news related to Covid-19. Thus, you have people sharing fake news messages to target such specific situations [Kouzy et al. 2020, DataLancet 2020].

Thus, this paper will present a new dataset of Twitter post from four different cities, two from Brazil (Natal and São Paulo) and two from the UK (London and Sheffield) with the aim of investigate the real impact of fake news, from defining how to differentiate the types of fake news to perform an initial statistical and social analysis on the selected dataset. Once we better understand this behaviour, we can better advise the government on what is the best format for the dissemination of guidance and scientific relevant information the next few weeks in the fight against the coronavirus.

2. Creating a Fake News Twitter-based dataset

The impact of fake news on social media is only growing as the political extremities continue to grow around the world. The human-based identification of fake news is quite straightforward. However, there has been significant discussion on how to analytically define such information in a way to make it viable to analyse it in an automatic way [Baines and Elliott 2020].

The most common definitions of fake news that can be described in the literature: *Misinformation* which is when an information is false, but not created with the intention of causing harm; *Disinformation* which is when an information is false and deliberately created to harm a person, social group, organization or country; and *Mal-information* which is when an information is based on reality but it is used to inflict harm on a person, organisation or country in an effort to ignite hatred of a particular ethnic group they are against.

Another important concept to understand regarding fake news is how the labeling of the data is performed. There are two main ways of doing it: content-based and metadata-based [Elhadad et al. 2020]. In the first instance, the labeling is done by performing natural language processing (NLP), also called sentiment analytics; the other approach is using any other information, such as number of mentions, sharing, emojis, friend linked, hashtags, etc.

Identifying who wrote such text is not easy and there are several possible ways of doing it, such as using machine learning algorithms that can investigate and perform predictions using the NLP and/or meta-data associated with it. Such tools can be used by Health Systems or other public organisations to adopt such intelligent models in order to create an automated way to promote greater security and reliability to the identification of fake news related with Covid-19 in social networks.

In our investigation, we have a few specificities that are important to be explained. The first one is the fact that our aim is to investigate if it is possible to identify similarities in the dissemination of fake news in two different countries (Brazil and UK) which have a certain history regarding fake news and started the actions against the virus in a similar way. The second one is that since we will deal with countries with different languages, it would be important to explore approaches that can be language independent, which is our case is the metadata-based investigation.

When designing our dataset, we have decided to collect the Twitter posts from two cities in Brazil (Natal and São Paulo) and two in the UK (Sheffield and London). They were chosen based on their sizes in relation to the countries population as well as on their stage on the coronavirus pandemic at the time. Also, based on the current situation and on the specificities of the application investigated, we have adopted the following labels for our dataset:

1. **TRUE**: This group will include posts (which could be RT, Retweet is the republishing of a Tweet, without additional comments) from media outlets (media that disseminate the news) recognized as reliable/responsible source or posts with information confirmed by reliable media outlets
2. **FAKE**: This group will include posts that are based on a true news, but are accompanied by comments with ideological tendencies (Mal-information); Posts with news which comments can involve prejudices against groups (Disinformation), posts with news that did not have a reliable source (Misinformation) and will include posts (which could be RT, Retweet is the republishing of a Tweet, without additional comments) that configure comments or promises from political personalities
3. **NNCOVID**: This group will include posts that are not news, but only personal comments/opinions about Covid-19

To build our database, we collect the most retweeted tweets in each city, as we believe that what is being retweeted is something important and that others should know. These tweets were collected from April 8 to November 30, 2020 with the tags coronavirus, coronavirus, covid-19, covid19 in addition to the specific language of each city using the public twitter API accessible with a developer account. In total, 2.135.882 tweets were collected. The features considered for the composition of the base were: 1) features of the tweet: *favourites, retweet, is favourites, is retweet, is retweeted* 2) User features: *User - Time Zone, User - Statuses, User - Followers, User - Friends, User - Favorites, user-location* We enriched the dataset by creating another 19 features from the set of initial attributes, so that we could have more information about the news from the respective cities. The features created were: *Morning, Afternoon, Night, Dawn, Hashtag, Url, Emojis, is_hashtag_tweet, is_mentios_tweet, is_url_tweet, is_mentions_user, is_url_user, is_hashtag_user, hashtag_user, emojis_user, url_user, mentions_user, country, city.*

Our initial process for classification of the posts involved ranking the most shared samples in each city, which in our case were the cases of retweets. We have manually labelled 1000 samples for each city, from April April 8 to 15, 2020, in order to keep consistency as well as to evaluate the relevance of our proposed classes.

3. Initial findings and statistical analysis using metadata: Brazil vs. United Kingdom

Our main aim is to have a large and significant dataset that is scraped across several months during the pandemic so we will be able to make statistical founded assumptions about our findings.

This data collections is happening as we go, but our first findings and the results presented in this paper were scraped in the period from April 8 to November 30, 2020 and the tags used were: coronavirus, coronavírus, covid-19, covid19, plus the specific language of the city. The total number of Twitter posts from this initial search were 216.534. This first analysis is being done by selecting the 1000 most retweeted posts in each city with the assumption that what is being retweeted is something important that others should know. The distribution of posts across all the proposed labels can be seen in Figure 1.

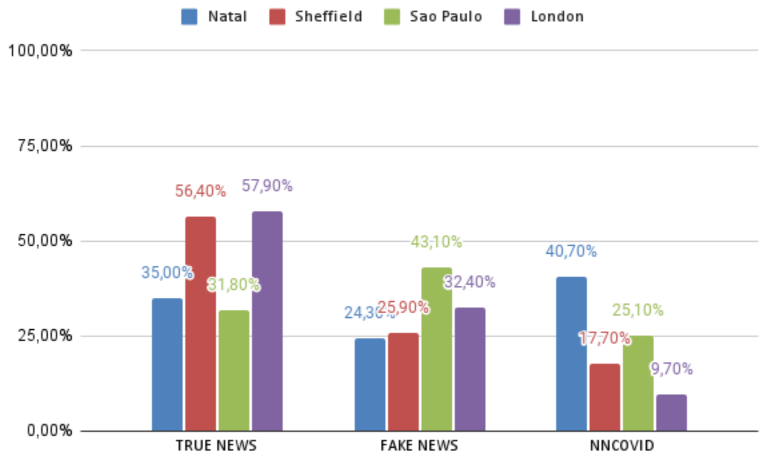

Figura 1. Distribution of the Twitter posts by label and city

It is quite striking at a first look the main difference in concentration of TRUE in the cities of UK and Brazil. In London, the largest city of the UK in size and population has a prevalence of 57,80% being retweeted as when we compare it with SP, which is the largest city in Brazil in population, we can see a close to a quarter of TRUE only. As London and SP are seen as cities with a similar cosmopolitan a varied population, with similar levels of educations, the impact of what is being disseminated is very clear. We can also see a similar behaviour when we look at the concentration of TRUE retweets from Natal and Sheffield, being the first with 35% and the latter with 56%. When observing the posts related with our definition of Fake News (Misinformation, Disinformation and Mal-information), we can see a closer behaviour with London and Sheffield with 32.4% and 25.9% and Natal and SP with 24.3% and 43.1%. However, SP can be seen as an outlier. It is interesting to note that by adding the posts of political figures like FAKE, we can see that in the city of São Paulo they have more than 50% of their posts higher than the other classes.

These concentration of what can be considered FAKE is a clear sign of the behaviour of the population of SP in relation to veracity of the news being shared. This might be a consequence of the political situation where there has been a wide open social media-based war ever since the coup suffered by President Dilma in 2016 and after that impact of the active creation of what has been seen as an epidemic in the production of fake news [Davis and Straubhaar 2020].

When we analyse the type of information being retweeted, based on the user, the hashtags and the way the user is performing the action of sharing that information, for instance, if it is just sharing a real piece of news or if it is sharing something with an ideological-based comment, we also see very clear and different behaviours:

- **London**: The main shared posts from political figures are associated with the opposition ($DrRosena$ and $Keir_Starmer$, for example). Another important feature is the that most of the TRUE came from well known news agencies, such as the BBC. Finally, most of the Misinformation was posted by a single user ($richardhorton1$) and it was related with attacks against China.
- **Sheffield**: The far majority of retweeted posts were made from big news outlets, political figures (without comments) and NHS (health staff). A small percentage was related with charity work. And the main political figures to have posts shared were from the opposition ($natalieben$, $DouglasJSheff$ and $LabourEuropean$, for instance). The news agencies mostly shared were Guardian and The Star.
- **SP**: The main characteristics of the posts from this city is the fact that most of the retweetes (around 85%) were made by government figures ($BolsonaroSP$, $CarlaZambelli38$, SF_Moro, $jairbolsonaro$, $AbrahamWeint$ and $CarlosBolsonaro$) and very few were made by the opposition ($samiabomfim$). The news agencies most shared were $folha$, $estadao$ and $g1$. The disinformation is all related with racist comments directed to China (as being the creator of the virus).
- **Natal**: Following the same trend as SP, but at a smaller scale, we can see a large prevalence of posts being retweeted by government figures ($rogeriosmarinho$, $GeneralGirao$ and $fabiofaria5555$) with a small prevalence of opposition ($natbonavides$). A good amount of local news agencies were shared ($tribunadonorte$) with some of the national agencies being cited as well (uol e $estadao$). On a different note, in this city, the sharing of scientific sources is quite large ($ufrnbr$, $laishuol$ and $metropoledigi$).

From the UK data, it is possible to note a majority of posts questioning the scientific-based advice the government is following in order to make its decisions. On the other hand, in Brazil, the posts are mostly related with questioning the veracity of the pandemic and the virus itself, as well as the use of specific medicine, even though there is not scientific indication for using it. More worryingly is the fact that these allegations are mostly made by members of the government.

Finally, the majority of retweets related with Fake News have a very low incidence of # and emojis in comparison with the other categories, but they have a high quantity of mentions, almost always attacking an opposition or an enemy. This behaviour is even clearer with checking the political figures posts, which have almost 80% of mentions.

These are important findings when we aim to understand what makes a post that has a FAKE inclination popular as well as what is the impact of such shared information in the actions the government and population was (is) taking in order to fight the virus transmission.

4. Conclusion

The types of the misinformation can be directed to most diverse aspects including new miraculous medication, diet that will protect you from the virus, and the worst of all, the question if the virus really exists with the intent of misleading society.

Our very first statistical analysis have pointed out to the real possibility of using metadata and of having a differentiated set of classes for the classification of fake news using Twitter posts and overcoming the issue of different languages that is so common when dealing with content-based sentiment analysis.

Our approach involves: Collecting and classifying a new misinformation dataset from Twitter data that will include fake news and real correct information. This is being done by using the expertise of specialists; The proposition of a systematic evaluation of how to identify fake news related to Covid-19, as well as the features or metadata associated with the most common fake news in the dataset.

As future works analysing which existing semi-supervised machine learning models can work better with the metadata used and thus, automatically perform the labeling of the rest of the dataset, based on the tweets initially labelled; and, finally, an analysis of how to correlate the information about death and confirmed cases in both specific places once these numbers are consolidated considering cases registered in hospitals as well as in the community in the same period of the Twitter data.

5. Acknowledgement

This work has been supported by the Brazilian Funding Agency CAPES.

Referências

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., and Shah, Z. (2020). Top concerns of tweeters during the covid-19 pandemic: Infoveillance study. *J. Med Internet Res*, 22(4):e19016.

Baines, D. and Elliott, R. (2020). Defining misinformation, disinformation and malinformation: An urgent need for clarity during the covid-19 infodemic. Discussion Papers 20-06, Department of Economics, University of Birmingham.

DataLancet (2020). Covid-19 in brazil: ”so what?”.

Davis, S. and Straubhaar, J. (2020). Producing antipetismo: Media activism and the rise of the radical, nationalist right in contemporary brazil. *International Communication Gazette*, 82(1):82–100.

Elhadad, M., Li, K., and Gebali, F. (2020). A novel approach for selecting hybrid features from online news textual metadata for fake news detection. In Barolli, L., Hellinckx, P., and Natwichai, J., editors, *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, pages 914–925, Cham. Springer International Publishing.

Guitton, M. (2020). Cyberpsychology research and covid-19. *Computers in human behavior*, page P106357.

Kouzy, R., Jaoude, J., Kraitem, A., Alam, M., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E., and Baddour, K. (2020). Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12.