Culture e Studi del Sociale-CuSSoc
ISSN: 2531-3975
Editor-in-Chief: Emiliana Mangone

# *Digital Methods to Study (and Reduce) the Impact of Disinformation*

MIRIAM DI LISIO & DOMENICO TREZZA

**Come citare / How to cite**

DI LISIO, M., & TREZZA, D. (2021). Digital Methods to Study (and Reduce) the Impact of Disinformation**.** *Culture e Studi del Sociale*, 6(1), Special, 143-156.

Disponibile / Retrieved from http://www.cussoc.it/index.php/journal/issue/archive

**1. Affiliazione Autore / Authors' information**
University of Naples Federico II, Italy

**2. Contatti / Authors' contact**
Miriam Di Lisio**:** miriam.dilisio[at]unina.it
Domenico Trezza: domenico.trezza[at]unina.it

**Articolo pubblicato online / Article first published online:** October 2021



Informazioni aggiuntive / Additional information

Culture e Studi del Sociale

# *Digital Methods to Study (and Reduce) the Impact of Disinformation*

*Miriam Di Lisio & Domenico Trezza*[1]

University of Naples Federico II, Italy
E-mail: miriam.dilisio[at]unina.it - domenico.trezza[at]unina.it

**Abstract**
Social media have democratized communication but have led to the explosion of the so-called "fake news" phenomenon. This problem has visible implications on global security, both political (e.g.the QANON case) and health (anti-Covid vaccination and No-Vax fake news). Models that detect the problem in real time and on large amounts of data are needed. Digital methods and text classification procedures are able to do this through predictive approaches to identify a suspect message or author. This paper aims to apply a supervised model to the study of fake news on the Twittersphere to highlight its potential and preliminary limitations. The case study is the infodemic generated on social media during the first phase of the COVID-19 emergency. The application of the supervised model involved the use of a training and testing dataset. The different preliminary steps to build the training dataset are also shown, highlighting, with a critical approach, the challenges of working with supervised algorithms. Two aspects emerge. The first is that it is important to block the sources of bad information, before the information itself. The second is that algorithms could be sources of bias. Social media companies need to be very careful about relying on automated classification.

**Keywords**: Digital methods, Fake news, Supervised classification, Text analysis.

## Introduction

Social networks have gradually transformed the contemporary scenario. With the birth of the Internet, there has been a process of democratization of knowledge: it's no longer necessary to resort to the opinion of the expert since everyone is transformed into broadcasters, everyone can produce and share content and distribute their own vision of "world wide" reality without filters and control (Quattrociocchi, Vicini, 2016, p. 22).Not all the fake news that transits the web, however, is disclosed with the intent to misinform, not all those who spread them know that they are contributing to the sharing of unreliable information. There is a difference between disinformation and misinformation: while disinformation concerns the voluntary sharing of fake news, misinformation concerns that a set of fake news is not disseminated with the intention of misleading recipients. Many times citizens don't verify the news they come across, but they tend to believe and/or share information with which they tend to agree. In fact, previous research has shown that people are more likely to accept authentic information that confirms and corroborates their pre-existing certainties (Del Vicario et al., 2016). The automatisms of merging the information generated on the network on the one hand and the confirmation bias[2] on the other, contribute to the polarization of positions. The role of the social sciences for the study and understanding of this phenomenon

[1] The paper is the result of a common work especially in its introductory part. However, Miriam Di Lisio edited paragraphs 1, 2, 4. Domenico Trezza edited paragraphs 3, 5, 6.

[2] A tendency to privilege information that confirms our opinions.

appears undoubtedly essential. Through previous research, the analysis of the content of the analyzed false news led to the identification of precise structural and content aspects of the aforementioned. It has been noted that there has been an evolution over the years regarding the actual construction of fake news. Just think that a few years ago fake contents were more easily recognizable for the elements that made up the corpus: the news was rich in information and showed images, links, and videos on the topic addressed. Over time, however, the way to compose fake news has gradually improved, trying to disguise the author's fallacious purposes, creating a scientific basis to give a certain credibility to the article in the eyes of the reader. There is an increasing need for short, intuitive news, which adapts to the size of the screens of the devices of the new millennium (Pira, Altinier, 2018, p. 27).

The history of fake news denotes the existence of particular predominant categories: the pseudoscientific one, which groups together all those news that exploit the scientific value to disseminate information and this make it credible; that conspiracy theorist who collects those news that claim that conspiracies or plots are hidden behind the most distinct events; that relating to pseudo-medicine/nutrition which refers to all those sites that are not based on any scientific principle, but offer cures, treatments etc; that relating to pseudo-journalism / politics refers to sites that use the journalistic idiom to disseminate articles that evoke ideas and opinions already consolidated in the minds of readers; that terrorist weather that concerns all those sites that divulge disinformation about the weather conditions; the pseudo-satirical one that refers to all those sites that use satire as a facade to propagate viral hoaxes, dedicated to that public that does not know what a disclaimer[3] is (Coltelli, 2018)[4]. With time, the detection/classification of fake news is gradually becoming of fundamental importance for the community in order to defend, in particular, the less erudite people. Over the years various machine learning techniques have been proposed, which represents the area of greatest impact of Artificial Intelligence, due to the ability of algorithms to perceive patterns and rules, which exceeds the human cognitive one (Marmo, 2020). Many scholars have tried to work in such a way as to automatically recognize fake news, trying to detect them not only efficiently but, above all, demonstrable (Zhou, Zafarani, 2020). The opportunity to use automatic classification techniques allows, where possible, to identify fake news based on semantic affinities: a semantic search algorithm determines the meaning of a text starting from the relationships between the lemmas of a corpus or a sentence; co-occurrences make it possible to recall the entity and category of the relevance of the topic. The realization of algorithms, however, takes time, especially in the field of machine learning: there is the need to do a lot of tests to find a suitable model and to optimize the parameters. What has changed with the coronavirus? Are these still the predominant categories in the disinformative scenery? In this scenario, digital communication and the context of social media appear to be decisive in the erroneous, fallacious, or illusory disclosure of information. The epidemic has triggered in the social networks they need to contain the spread of news from unreliable sources, however, the production and sharing of fake news has not stopped (Sala, Scaglioni, 2020). This work aims to answer these questions and to focus its work on the automatic classification of texts, applied to a corpus of about 230 thousand tweets obtained in the week from 9 to 15 March 2020. Working on

[3] A disclaimer is typically a statement intended to define or outline the extent, rights, and obligations between two or more parties involved in a legally recognized relationship (Wikipedia).

[4] The Black List | Butac - Bufale Un Tanto Al Chilo

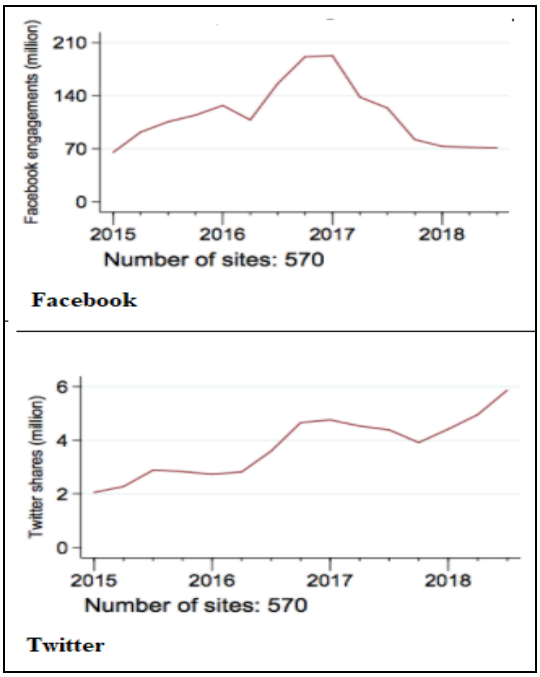
tweets means dealing with a type of content different from the usual, a form of short writing (Chiusaroli, Zanzotto, 2012) with brevity, synthesis, the reduction into significantly reduced elements, a type of text peculiar to the contemporary criterion of online communication. This methodology is presented as a necessary strategy for the realization of a mapping of bad information about covid-19. The aforementioned strategy has made it possible to shed light on the initial questions and to find lexical and syntactic tendencies within this type of texts, identifying the polarity of the tweets concerning three properties originally identified: the author (user), the source and the theme; moreover, it was possible to study how these trends affect the veracity or otherwise of news. The work is articulated into six paragraphs: the first and second introduce the study background. The first concerning social media strategies, with a focus on Twitter and Facebook, to reduce the risk of the virality of bad information and exploring the usefulness - with some examples - of machine learning techniques to identify fake contents. The second doing a review on the phenomenon of fake news in relation to the coronavirus emergency. The third paragraph is about methodology. The research questions and the technique used, supervisioned modeling, are defined. The fourth and fifth paragraph are about analysis. The fourth building training with the manual classification of tweets and exploring the most frequent fake issues. The fifth paragraph concerns the start of the model. The last section will discuss model results and emerging perspectives for supervised fake news analysis.

## 1. Fake news, social media and machine learning. Related works

Twitter is the context of our analysis. It has some important advantages for those who search with the textual contents of social networks: popularity (it has more than 300 million users worldwide), the few privacy constraints, and the tendency to standardize content (280 characters). This entails the possibility for a researcher to create easily huge dataset (accessibility to the API is rather fast). How does the platform interact with the problem of fake news, which is our object of investigation? Twitter, like all other social platforms, is a digital environment in which fake content can become viral in a short time and this can have disastrous consequences for the global community, especially because it can affect people's behavior. Proof of how dangerous the virality of fake content comes with the US presidential elections in 2016, during which there was a significant increase in fake content which, according to what political analysts and data scientists say, have influenced the electoral behavior of Americans (Allcott et al. 2018). This episode represents a strong alarm bell and has encouraged the owners and managers of large social networks to put a stop to the phenomenon, with different strategies and outcomes.

Allcott et al. (ibid.) studied the virality of fake content verified on Facebook and Twitter, over the period 2015-2018. The respective trends (fig. 1) shows how, based on very different numbers of engagement (significantly higher for Facebook), it is plausible that after 2017 Facebook began to implement effective algorithms to reduce the phenomenon of fake news (the trend of the engagement of fake pages is down), while this still seems not to happen for Twitter, with the number of fake shares that is even on the rise.

*Figure 1. Fake trends in Facebook and Twitter engagement*




*Source: Allcott et al. 2018*

Twitter's concern as a fake-related environment seems fueled by research by Vosoughi and collaborators (2018), who found that fake content spread on Twitter is 70% more likely to be retweeted than real content. Twitter, like other social networks, have adopted strategies to reduce the phenomenon. These strategies are often linked to the use of machine learning techniques that train the algorithm to recognize user face (bots) or suspicious content and therefore new nip in the bud possible sources of bad information. But can this be enough? Although Twitter has claimed to have achieved some goals in the fight against the fake world (for example their supervised algorithms contributed to the elimination of + 214% of bot accounts compared to the previous year), many studies on the virality of bad information on Twitter suggest that the problem is still very relevant (Castillo et al. 2011). There are not a few jobs that aim to identify ML models to automatically detect suspicious content on social networks. Castillo and other collaborators, for example, were among the first to build a model based on certain aspects of the tweets that best discriminated against their credibility. Much of this type of analysis uses text as its main feature, using the frequencies and type of words present in the tweets. Other researches, which also achieved good results in terms of model accuracy, took into consideration the characteristics of the users (for example, the time of registration to the platform, the nickname, and the network of contacts). Although the models developed have achieved satisfactory results, the uncertain definition of fake news however reminds us that the task of reducing the phenomenon of viral disinformation cannot be completely entrusted to machine learning. In this work, we try to explore some machine learning techniques in a context in which the researcher has full awareness of the processes.


Culture e Studi del Sociale-CuSSoc, 2021, 6(1), Special, pp. 143-156
ISSN: 2531-3975

## 2. The problem of fake news during the Covid-19 outbreak. New disinformation classes emerging

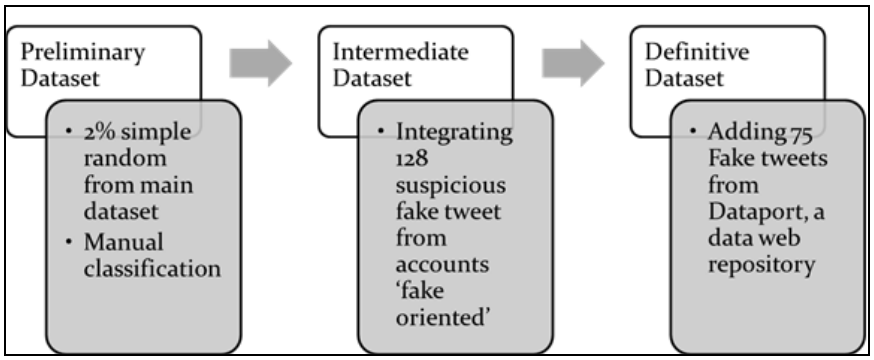
In a difficult historical period like the one the world is going through since December 2019, fear, anguish and lack of knowledge of the virus and the disease have led citizens to produce, share and nevertheless, to believe in unverified news, generating an unprecedented flow of disinformation[5]. The WHO (World Health Organization) has not only announced the health dangers caused by the coronavirus, but has also defined the moment as highly infodemic, due to the amount of information, true and, above all, false, circulating on the net about this topic (Pulido et al, 2020). According to an English study (Julii Brainard, 2019), funded by the National Institute for Health Research and presented in late February by East Anglia University, "disinformation on health can intensify outbreaks of infectious diseases".As of December 2019, news about the virus, both true and false, began to populate the web (Orso et al, 2020). The AGCOM (Italian Authority for Communications Guarantees), following the analysis of the textual content of all the disinformation articles that it managed to detect on the coronavirus, highlights the emergence of some dominant narratives on the epidemic, such as risks, conspiracy theories and the news, centered on a disclosure built on the repeated use of terms aimed at leveraging negative emotions. In particular, it shows the fact-checks[6] of the top 10 fake news reports on the coronavirus epidemiological emergency. Among the most viral fake news Covid-related in the Net: those on the remedies to wipe out the virus by drinking water every 15 minutes or taking vitamin C daily, or those that denounce the use of ibuprofen because it would accelerate the outflow of the disease , or even those relating to the reduction of salaries by Italian political offices to deal with the country's economic emergency. Therefore fake news could have a terribly negative effect, mainly in a critical phase such as the one faced by Italy starting from the end of February 2020. The amount of fallacious data disclosed in parallel with the spread of the virus which has influenced communication on collective health, prompted us to choose fake news relating to the pandemic as the object of study (Brennen et al, 2020). It is not easy to define the concept of fake news. It includes several meanings, from disinformation disseminated for specific purposes (for example, political purposes) to that unknowingly spread or made viral in good faith. On the other hand, if two people with different ideas and opinions were asked to define fake news, they would most likely give two completely different answers, based, in fact, on their beliefs (political, values, etc.). This is to say that studying this phenomenon can be very complex. Of course, blatantly false information exists and is easily identifiable, that is, linked to non-existent people, things, or facts. But there is a gray area that is often not easily refutable because it is necessary to contextualize it (for example, a single excerpt that may have opposite meanings with respect to the textual context) or because its truthfulness cannot be established with certainty. The uncertain delimitation of the meaning of fake news suggests that doing research on this issue is quite complex, but at the same time it is useful given the growing problem of the virality of bad information. The application of machine learning techniques could represent a very effective way for this type of analysis because they allow you to use classification algorithms trained on specific contexts and then applied to large corpora.

[5] Disinformation means the intentional dissemination of incorrect or distorted news or information to influence someone's actions and choices [http://www.treccani.it/vocabolario/disinformazione/].
[6] https://www.agcom.it/factcheckcovid19

## 3. Research questions, data and methods

Our research questions are relates to phenomon of the fake news and the methodological implications in the scientific research. What are the categories of fake news and with what intensity did they circulate during the first week of the emergency? What are implications on our data and on the fake categories about using supervisioned model? The work seeks to respond by applying supervised analysis techniques to the text, through a process of constructing reasoned training on specific emergency issues fakes. A strategy that involves the construction of the algorithm starting from the researcher's knowledge and reasoned choices. The work concerns the analysis of fake news circulated on Twitter during the first week of the Covid emergency, March 9-15 with the start of the Italian lockdown. As a result, a large communication Covid-related has developed on the social network. The tweet retrieval was carried out by the basic Twitter API 1.1, through the 'rtweet' R package, using four key extraction: #coronavirus, #coronavirusitalia, #covid, #Covid-19, that are the main hashtags of the emergency topic trends in the week considered. The main dataset consists of 238829 tweets. The analysis of the tweets was carried out through a manual classification from a sample of tweets and following an automatic classification through a machine learning procedure (regression logistic) using R software, and 'tm'package for the text classification. The difficulty and challenge of this work was trying to build an effective training dataset to form and test the supervisioned model. For this purpose, three steps were developed: 1. Preliminary (4522 tweets): extraction of 2% of the tweets from the main dataset by simple random sampling to analyze the content according to the fake related/no fake dichotomy. 2.Intermediate (4650 tw.): integrating 128 suspicious fake tweet from accounts 'fake oriented'. 3. Definitive (4722 tw.): adding 75 Fake tweets from a data web repository

*Figure 2. Steps of dataset training creation*




*Source: Our elaboration*

The first step involved the random extraction of 4522 tweets (2% of the main dataset). This share needs to balance between the manual classification of the tweets and a good quota forming the model. The only criterion envisaged for the extraction was the selection among only the no verified accounts[7], because it is

[7] Twitter distinguishes between verified accounts (with a blue check) and non-verified accounts. Verified Twitter accounts are those belonging to profiles of public or private bodies, foundations, companies, or persons of particular importance, recognized by Twitter.

plausible that they are more "fake sharing - oriented". The second step involved the analysis of 4522 tweets and the attribution of each tweet to the fake / no fake dichotomy. We have classified as fake the contents that took up themes that are notoriously false or tend to be suspect, including those that suggested sharing in good faith (misinformation). In order to orient in the classification, our references have been the fake issues detected by two official sources: Butac.it ('Speciale Coronavirus' section) and Health Ministry website ('Attentiallebufale' section), which deals with Covid monitoring disinformation. As a result, our results may be weighted toward diffusion of misinformation that Butac and the Italian Healthy Minister is aware of, and may not fully capture trends in misinformation that they are not aware of. It is difficult to assess how large this latter group might be. Our study object almost certainly includes the most important issues of false stories on the first period of the Italia pandemic. Since this first phase, 54 'suspicious' tweets were identified (1.2% of the total[8]).

*Table 1. Fake issue detected by the Butac site*

| *Anti-europeism* | *Pseudoscience* | *Alarmism* | *Denial* | *Anti-immigration* |
|---|---|---|---|---|
| - American soldiers and EU Subjection | - Vaccine developed in Australia<br>- Vitamin C and Covid<br>- Covid supplements | - I have my brother-in-law who is a doctor…<br>- Detention of the good<br>- BSL 4 Biocontainment<br>- Young people hospitalized in Como<br>- Underestimated deaths | - Sgarbi and denialbehaviour | - Uncontrolledinfectedimmigrants |

*Source: Butac.it and Italian Healthy Minister website*

The second step involved the analysis of the network of fake-oriented users from main dataset (tab.2). This strategy has allowed us to found new 128 fake tweets and to integrate them in the training dataset. The third phase has tried to increase training dataset integrating 75 fake sentences on Covid, founded in the same period, from a Tweet database of IEEEDataport, a web repository (Lamsal 2020). Therefore, the definitive dataset consists of 4727 tweets, with 257 fake-related. The analysis section involves the application of the supervised logistic regression model, which operates on the text and classifies it automatically.

[8] Being a random sample, it is plausible that this share does not differ significantly from the real parameter.

*Table 2. Account nickname, fake tweets and topic fake*

| *Account nickname* | *Fake tweets* | *Topic fake* |
|---|---|---|
| santini1965 | 17 | Virus denial (7) – Cospiracy (7) |
| GianvitArmenise | 11 | Cospiracy (6) |
| Esticatzhi | 8 | Antieuropeism (6) |
| MauLazio29 | 8 | Cospiracy (7) |
| Saul95153757 | 7 | Pseudoscience (7) |
| euright9 | 6 | Cospiracy (4) |
| IeeeMatteo | 6 | Virus denial (5) |
| paolopasquale | 6 | Cospiracy (5) |
| ShootersykEku | 6 | Pseudoscience (4) |

*Source: Our elaboration*

## 4. Before testing the model. Exploring the fake-issues

The tweets fake content - related are about 4% of the sample.As we can see in table 3, engagement is different between fake and no fake tweets. Fake tweets on average have less followers, but, as expected, they have higher virality values than others. They are more 'retweeted' (7.8 vs 2.5) and receive more likes (15.8 vs 7.6). Before testing the model, we have explored the most common fake issues and their engagement (follower, retweet, favorite count). We have found and classified 8 issues: Conspiracy, Pseudoscience, Virus denial, Antieuropeism, News not verified, Anti immigration and Anti China. Table 3 higlights that most of the fake tweets (65) of our sample are linked to conspiracy themes (for example, the virus and 5g, Bill Gates' vaccine, etc). A marginal share of fake tweets refers to Covid alarmism and the attitude against China. However, it is noted that the first issues are more common but also less viral, because they have low engagement values. Instead, the latest issues are less widespread but more oriented towards virality. As we expected, the tweets related to news not verified have very low engagement values: they are short news, not easily classifiable, linked to particular events and that did not circulate enough on the web. This first part of the analysis offers us an insight into tweet fake related. At this point, how can an automatic classification help us with large amounts of textual data? The authors tried to explore this question by applying a supervised model.

*Table 3. Type, Topic and Engagement of tweets*

| | *nr* | *%* | avg nr of Followers | Retweets | Favorite |
|---|---|---|---|---|---|
| Tweets no fake-related | 4468 | 96% | 4450 | 2,5 | 7,6 |
| Tweets fake-related | 182 | 4% | 2867,3 | 7,8 | 15,8 |
| Cospiracy (5g, Bill Gates, biologic war) | 65 | *36%* | 1429 | 2,7 | 6,1 |
| Pseudoscience (vitamine C, Panzironi method, religion or magic solutions) | 29 | *16%* | 1608 | 1,6 | 3,1 |
| Virus denial | 24 | *13%* | 2584 | 38,8 | 69,8 |
| Antieuropeism / pro China (enemy Europe, China friend) | 21 | *12%* | 2480 | 7,1 | 13,1 |
| News not verified (suspicious events) | 18 | *10%* | 166 | 0,3 | 0,3 |
| Anti immigration (infected immigrants, uncontrolled arrival of immigrants..) | 12 | *7%* | 3615 | 4,8 | 9,7 |
| Alarmism (doctors, nurses and staff predicting bad events through viral audio messages) | 10 | *5%* | 3162 | 0,8 | 2,0 |
| Anti China (China guilty) | 3 | *2%* | 2969 | 17,8 | 42,2 |
| | | | | high | low |

*Source: Our elaboration*

## 5. The construction of the supervised model from the textual analysis to the training and testing model

The supervised model used is that of logistic regression, applied on definitive dataset of 4725 tweets. The purpose of this classification algorithm will be to identify the line that best manages to separate the two classes (in our case fake or not fake) in the space of characteristics, that are the text of the tweet. The text of the tweet has been processed and transformed in a corpus. The text has been pre-processed with R software, involved four steps: normalization to omologate words written in uppercase, and to strip whitespace; remove italianstopwords, to remove all 'empty' words such as prepositions, conjunctions, articles, etc; stem document or lemmatization to aggregate words that belong to the same root; remove punctuation to avoid getting punctuation as single textual form. Pre-processed the corpus, the frequency of each single term was extracted by converting the corpus into a documents-terms matrix, in which frequency is occurrence of the term within the tweets (tfweight). The vocabulary of tab.4, reduced by removing terms with relevant sparsity, shows the most frequent words. They are related to the hashtags #coronavirus, #iorestoacasa, #covid19, and #coronavirusitalia and to common words such as emergency, government, virus, now that better define the context of the situation and the urgency to act.

*Table 4. Term frequency*

| *Term* | *Freq* |
|---|---|
| coronavirus | 2999 |
| iorestoacasa | 1095 |
| covid19 | 1092 |
| coronavirusitalia | 531 |
| italia | 417 |
| casa | 356 |
| covid2019 | 269 |
| fare | 241 |
| covid19italia | 227 |
| solo | 227 |
| restiamoacasa | 202 |
| emergenza | 193 |
| ora | 184 |
| cosa | 181 |
| prima | 172 |
| oggi | 168 |
| governo | 152 |
| virus | 150 |
| bene | 147 |
| quarantena | 142 |
| pandemia | 139 |

*Source: Our elaboration*

The processed text allows us to test the supervised regression model (tab.5). The procedure involves taking the definitive dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset. As suggested by literature (Caruana, Niculescu-Mizil, 2006), there is no optimal split percentage, but common split percentages include three chances: 1)Train: 80%, Test: 20%; 2) Train: 67%, Test: 33%; 3) Train: 50%, Test: 50%. It could depend from the size of the sample. In our case, given the small sample, we have opted for 80% and 20% solution. The training was started with a logit regression, in which the dichotomous condition of fake or not of the tweet (labeled variable) is the dependent variable, while the features of the model are the textual forms that so have the faculty of predictors. Logistic regression models the probability of one class or another. In fact, we are modeling the probability that a tweet belongs to fake content-related class or not. Model has been tested on subset 'test', that is not labeled dataset, creating as an output a predicting variable.

*Table 5. Setting, starting and output of the Supervisioned model*

| *1.Setting train and test set* |
|---|
| Train = 3750 tw. (80%) |
| Test = 975 tw. (20%) |

| *2. Starting model* |
|---|
| Logistic regression model |
| Mod →Glm(Fake content ~ text , Train, family = binomial(logit))<br>Predicting function →Pred Fake = predict (mod, Test) |

3. Output table

| text | obs | pred |
|---|---|---|
| #Gravidanza e post-parto. Come comportarsi di fronte alla ... | no fake | no fake |
| Voi ce l'avete la benedizione a domicilio? <U+2728><U+00... | no fake | no fake |
| Misure urgenti per prevenzione e gestione Coronavirus, nuo... | no fake | no fake |
| Se anche dovessi sfuggire al coronavirus, il colesterolo non ... | no fake | no fake |
| Nel caso ti stia chiedendo quanto i media controllino le pers... | fake | fake |
| <U+0001F5D3><U+FE0F>Il #Coronavirus sposta anche le d... | no fake | no fake |
| @micheleemiliano visto il decreto che costringe tutti in cas... | no fake | no fake |
| 12. I dati ISS confermano che pochi dei morti sono deceduti... | fake | no fake |
| Coronavirus diretta. Appello del governo: «Non viaggiate ne... | no fake | no fake |
| Ora il nostro giornalista in diretta al @MediasetTgcom24 pe... | no fake | no fake |

*Source: Our elaboration*

The assessment of the model has been carried out over three parameters: accuracy, as the the percentage of predicted correct on the total, precision, as the predicted correct on each observed class, and recall, which calculates the percentage of predicted for each class.The table 7 summarizing the results of the 2x2 crosstable between predicted and observed values (tab.6) shows us a good fit of the model for the prediction of the no-fake, superior to the good quota of 90%, for accuracy, precision and recall. Instead, several problems are detected for the prediction of the fakes, which reaches only 25%. It is likely that this is related to many factors: for example, the small size of the sample, so it will have to be expanded. We believe that one of the problematic factors is probably the ambiguity of the concept of fake news. As previously seen, in fact, that there are some categories of fake news that are deeply ambiguous. This is the case of not verified news. In fact, exploring the tested tweets, we observe that only 1 out of 11 tweets of that category was predicted correctly.

*Table 6. Predicted / Observed Crosstable output (uncorrected predicteds are highlighted in grey)*

| | | *Observed* | |
|---|---|---|---|
| | | No Fake | Fake |
| *Predicted* | No Fake | 869 | 46 |
| | Fake | 45 | 15 |

*Source: Our elaboration*

*Table 7. Accuracy, Precision and Recall output*

| | *Precision* | | *Recall* | |
|---|---|---|---|---|
| *Accuracy* | No Fake | Fake | No Fake | Fake |
| 91% | 95% | 25% | 94,9% | 25% |

*Source: Our elaboration*

## 6. Perspectives to improve the algorithm and better track the pandemic disinformation. Strengths and weaknesses

The Covid-19 pandemic has caused an unprecedented health emergency. Not only that. It was and still is an event that turned our lives upside down, and as result it also became a huge communication event. We believe that, looking at the production numbers of tweets related to the emergency, especially in the worst days of the epidemic, the media coverage and buzz on social media were almost total. A true infodemic, as an uncontrolled flow of information about the virus. Of course, it is not just about good information. The phenomenon of fake news, which in recent years has reached such alarming levels that it has also affected public opinion, has found fertile ground in such a complex emergency phase. Starting from the analysis of the concept of fake news and how the phenomenon has evolved in recent years, our research assumes that the study of bad information circulating on social platforms should not leave out of consideration the analysis strategies involving digital data, or large text corpora. Especially if the context of the research is that of an infodemic, in which the virality of information becomes maximum. This is the reason why analysis strategies, based on automatic text classification, are becoming more and more popular in the field of content analysis. Our work aimed to test the application of a machine learning model, based on supervised logistic regression, on a sample of tweets from the first week of emergence. To train the algorithm, it was necessary to create a base of labeled tweets. The task was not easy. In fact, we realized that the number of suspicious tweets, at least for the week under review, was not sufficient to create a solid base for the algorithm. Although this was a critical problem for the methodological process, it also represented initial evidence, useful for answering the first question: during the onset of the pandemic, the circulation of fake content was relatively small. Exploring our second question, we observe how these are associated with dimensions already experienced by the 'dialectic' of disinformation, such as conspiracy (the virus caused by strong powers), pseudoscience (virus that cures itself with non-scientific approaches), denialist attitudes (the virus does not exist). As if that were not enough, there is also room for the virus-immigrant association. We observed how the sources of these tweets are often accounts fake-oriented. For example, one account alone tweeted (not counting retweets) up to 17 suspicious

tweets within a few days. However, this list of accounts was also a methodological tool, as tracking them in the dataset was useful for us to find another group of fake tweets. The supervisioned model did not return satisfactory results: the good accuracy value could be due to the high number of non-fake tweets that the model detected (in fact, the accuracy of the fake is very low). Outcomes suggests that in order to increase the performance of the algorithm it is necessary to increase the number of the training base. This is not an easy challenge because, as we have seen, the base of fakes related to covid is not large, and this in our opinion can be solved through the shared effort of the scientific community to make available databases of fakes to better train algorithm and allow more reliable machine learning analyses. On the other hand, the misinformation concept is extremely complex due to its ambiguity, especially in unverified simply news, and this could be a problem when applying supervised techniques.There is a need for social networks, where much of the misinformation often circulates, to take this ambiguity into account. We believe it is important that within the framework of these predictive models, the researcher or data analyst has sufficient control over the entire analysis process. Tracking misinformation content on the web is not just possible but needed if we consider the large flow of textual data that defines our daily backgrounds.

**References**


Allcott, H., Gentzkow, M.,&Yu, C., (2018). Trends in the diffusion of misinformation on social media, in *ResearchGate*, april 2018.

Brainard, J.,& Hunter, P., (2019). Misinformation making a disease outbreak worse: outcomes compared for influenza, monkeypox, and norovirus, in *Sage Journal*, 12 Novembre 2019 [https://journals.sagepub.com/doi/pdf/10.1177/0037549719885021].

Brennen, J. S., Simon, F. M., &Nielsen, R. K., (2020). Beyond (Mis)Representation: Visuals in COVID-19 Misinformation, in *The International Journal of Press/Politics*, October 2020 [https://journals.sagepub.com/doi/pdf/10.1177/1940161220964780]

Caruana, R., &Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168).

Castillo, C., Mendoza, M., &Poblete, B., (2011). Information credibility on Twitter, in *ResearchGate*, january 2011.

Chiusaroli, F., & Zanzotto, F. M. (a cura di), (2012). *Scritture brevi di oggi. Quaderni di Linguistica Zero*, 1, Napoli, Università degli studi di Napoli "L'Orientale", ISBN: 978-88-6719-017-1.

Coltelli, M. (2018). The Black List | Butac - Bufale Un Tanto Al Chilo.

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli,G., & Quattrociocchi, W., (2016). EchoChambers: EmotionalContagion and Group Polarization on Facebook, in *Scientific Reports 6*, articlenumber: 37825.

Lamsal, R., (2020). Coronavirus (Covid-19) tweets dataset, in *IEEE DataPort* [https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset].

Marmo, R., (2020). *Algoritmi per l'intelligenza artificiale. Progettazione dell'algoritmo, dati e machine learning, neural network, deeplearning*, Hoepli, Milano.

Orso, D., Federici, N., Copetti, R., &Vetrugno, L., (2020). Infodemic and the spread of fake news in the COVID-19-era, in *European Journal of Emergency Medicine*, april 2020.

Pira, F., &Altinier, A., (2018), *Giornalismi. La difficile convivenza con fake news e misinformation*, libreriauniversitaria.it, Cassino (FR).

Pulido, C. M., Villarejo-Carballido, B., Redondo-Sama, G., &Gòmez, A., (2020). COVID-19 infodemic: More retweets for science-based information on coronavirus than for false information, in *International Sociology* [https://www.researchgate.net/publication/340658737_COVID19_infodemic_More_retweets_for_science-based_information_on_coronavirus_than_for_false_information].

Quattrociocchi, W., & Vicini, A., (2016). *Misinformation. Guida alla società· dell'informazione e della credulità*, Franco Angeli, Milano.

Sala, M., & Scaglioni, M., (2020). *L'altro virus, comunicazione e disinformazione al tempo del covid-19*, Vita e Pensiero, Milano.

Vosoughi, S., Roy, D., & Aral, S., (2018). The spread of true and false news online, in *ResearchGate*, march 2018 [https://www.researchgate.net/publication/323649207_The_spread_of_true_and_false_news_online].

Zhou, X., & Zafarani, R., (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities, in *ACM Computing Surveys*, september 2020, Article n. 109.