EU DisinfoLab

Developing a disinformation detection system and sourcing it live

The case study of the 2018 Italian elections

Authors:

Alaphilippe, Alexandre
Ceccarelli, Chiara
Charlet, Léa
Mycielski, Martin

First published: 17/05/2018 in Brussels

Project funded by Open Society Foundations
Abstract

During the past months, the phenomenon of fake news has drawn significant attention from both researchers and civil society actors. Following the promising results obtained during the 2017 French elections in the identification and sourcing of disinformation content, we tested the implementation of a monitoring system based on an identified ecosystem rather than on a series of keywords.

The purpose of this monitoring process was to give the ability to expose misinformation networks in real time.

As a testing ground for our method we chose the 2018 Italian General Election. During our monitoring we’ve received 171 alerts from the system, with the content recognised as fake, misleading or manipulated a total of 54 times. The Twitter exposure of our findings managed to gather 109,000 impressions.
Introduction

The role that so-called “fake news” played in the public debate during the 2016 US election period has drawn researchers’ attention to this phenomenon, mainly concerning its impact on the democratic electoral process. Naturally, the approaches differed depending on the research field of the various researchers and the entry point to the problem.

The issue was first addressed at the micro level, using numerous examples of disinformation. For instance, the Donbass conflict in Ukraine exemplifies how information warfare can exploit multiple media to spread propaganda in conflict zones, especially through Russian state TV (Khaldarova & Pantti, 2016). The 2016 United States presidential elections showed the importance of social media in fake news dissemination during the campaign period, while also highlighting the difficulty to provide accurate assessment of its impact on the electoral process (Allcott & Gentzkow, 2017). Even in the so-called "swing states" - states that played a deciding role in the result of the elections - where misinformation and fake news was most shared, an impact assessment has been difficult to produce (Howard, Kollanyi, Bradshaw & Neudert, 2018).

Still, the majority of academic work focuses on a macro-level analysis. We can identify three major approaches:

The first approach takes into account the role of echo chambers - questioning the behaviour of the media sphere and their influence in creating partisanship echo chambers. For instance, the role of misinformation content in generating homogenous and polarised echo chambers (Del Vicario et al., 2016), as well as the role of confirmation bias (Quattrociocchi, Scala & Sunstein, 2016), have been demonstrated on Facebook. The influence of recently created alt-right media in the Republican partisanship echo chambers in the US has also been analysed, with Democrat echo chambers relying more on mainstream media and traditional networks like CNN (Faris et al., 2017).

Secondly, we noticed the emergence of research on “computational propaganda”, which relates to artificial amplification of false narratives. The role of automated or semi-automated accounts (bots) in amplifying false narratives has been demonstrated especially during the US elections and the Brexit referendum (Howard & Kollanyi, 2016).

Finally, research on “fake news” detection, combining content analysis with network analysis through the use of semantic tools and machine learning (Conroy, Rubin & Chen, 2015), which highlighted the necessity to have a combined machine/human approach and to fuse techniques to assess veracity of information. Note that instead of the buzz term "fake news", we favour the scientific concept of "disinformation", defined as “content deliberately created with the intent to disrupt electoral processes". (Giglietto, Iannelli, Rossi, & Valeriani, 2016). In our context, we postulated that the content is amplified by social media.

We position ourselves in this third area of research, focusing on identifying the actors and communities susceptible to disinformation and comparing them to the actors and communities actually spreading it.

Indeed, during the 2017 French elections, EU DisinfoLab conducted a pilot project on disinformation spreading on social media. Studying the false information spreading, we have tried to confirm or disprove the theory that there was a Russian influence on the electoral process. To do that, our approach was firstly to identify the disinformation network by analysing the spread of 11 top disinformation pieces during the French election. Next, we
compared it to a “pro-Russian” ecosystem, gathering 5000 most active accounts relaying Sputnik or Russia Today content. Our results showed 75% of correlation between the Russian ecosystem and the ecosystem of people that spread 3 disinformation pieces during the elections. For people that have spread 5 different pieces of disinformation, the correlation went up to 92%, showing high correlation between the two networks.

Building this correlation matrix enabled faster reactivity in a disinformation crisis. The two tweets from Nicolas Vanderbiest during the “Macron Leaks” (alleged email leak from Emmanuel Macron's team combined with a high amount of disinformation) and the “Bahamas rumour” (disinformation that Emmanuel Macron has a secret account in the Bahamas), exposing the network and the correlation, reached 802,564 impressions and 99,135 engagements. This communication’s effectiveness stemmed not from the disinformation having been fact-checked, but having been “sourced”, which, according to our definition, serves “to expose the origin of the information in order to instil in the audience reasonable doubt as to its veracity”.

Following these promising results, we wanted to test the reversal of the methodology. If we were able to find a correlation between disinformation spreading and the Russian ecosystem, would it work to firstly monitor the Russian ecosystem in order to detect disinformation in real time? Furthermore, we wanted to test if this monitoring technique would contribute to raising awareness amongst the population about the potential issue.

Methodology

In order to verify our hypotheses, we conducted the same methodology, which consists of five practical steps:

a) Triangulation

First, we collected a series of accounts representing the ecosystem that we wished to identify. In this case, we selected four accounts spreading pro-Russian narratives in Italy:

- Russia Beyond (@rbth_it), a Russian lifestyle magazine;
- L'antiDiplomatico (@Lantidiplomatico): a highly-followed account pushing pro-Russian content;
- The Russian Embassy in Italy (@Rusembitaly);
- The Italian-language account of Russian media outlet Sputnik (@sputnik_italia).

We then identified recurring accounts in their audience. These recurring accounts form the first level of identification, which by design is not sufficient as it contains too many non-relevant accounts.

1 See Appendix 1.
The following step was to analyse how these accounts were following each other in order to identify which Twitter accounts are the most followed by other individuals in the same ecosystem; the triangulation was then launched on the 25 most followed accounts. Again, we analysed how they were following each other and used the Social Network Analysis (SNA) method to describe the ecosystem we were monitoring:

Using the Louvain algorithm (Blondel, Guillaume, Lambiotte & Lefebvre, 2008), we identified three main communities:
• An anti-Euro community (in green)
• The main M5S community (in red)
• An extreme-right community (in pink/purple)

We made a comparison of communities found with different community detection algorithms:

<table>
<thead>
<tr>
<th></th>
<th>FastGreedy*</th>
<th>MultiLevel*</th>
<th>WalkTrap</th>
<th>InfoMap</th>
<th>LabelProp.**</th>
<th>DisinfoLab/Louvain</th>
</tr>
</thead>
<tbody>
<tr>
<td>FastGreedy</td>
<td>1</td>
<td>0.8677116</td>
<td>0.6463084</td>
<td>0.4935171</td>
<td>0.4708884</td>
<td>0.8502247</td>
</tr>
<tr>
<td>MultiLevel</td>
<td>-</td>
<td>1</td>
<td>0.6618715</td>
<td>0.4805499</td>
<td>0.4430403</td>
<td>0.8903118</td>
</tr>
<tr>
<td>WalkTrap</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.6370240</td>
<td>0.2686518</td>
<td>0.6709527</td>
</tr>
<tr>
<td>InfoMap</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.5739915</td>
<td>0.4723503</td>
</tr>
<tr>
<td>LabelProp.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.3225681</td>
</tr>
<tr>
<td>DisinfoLab/Louvain</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Random***</td>
<td>0.5527120</td>
<td>0.5666234</td>
<td>0.6540370</td>
<td>0.5053527</td>
<td>0.2881368</td>
<td>0.588742</td>
</tr>
</tbody>
</table>

Table 1: Comparison of communities found with different community detection algorithms, the partition given by the Louvain algorithm and a random partition.

The comparison score is the Rand similarity index where 1 is an exact match and 0 is no match at all:
* Modularity-based community detection algorithms
** Mean similarity value of over 100 iterations
*** Random partition was done assuming 4 large communities were defined (based on the communities found by EU DisinfoLab)

b) Monitoring

The second step was to monitor this ecosystem. Our aim was to test if, by closely monitoring it, we could detect any kind of disinformation operations; also, we wanted to verify the validity of account-based monitoring over semantic monitoring. Whenever our system detected, through the use of Twitter’s API, that the same content was shared at least 3 times within the ecosystem, we received an alert in a dedicated Slack channel. The monitoring process for this study was conducted over the span of ten weeks, from December 15th until the election day, March 4th.

c) Adapting the monitoring

A monitoring process is never static, it has to change and be adapted to the findings and the potential new leads for the research. This explains why the scope of monitoring needed to be refined in order to reduce potential noise, such as unrelated accounts and tweets. Furthermore, to expand our monitoring capabilities, we also relied on our network and partners in order to receive alerts of suspicious pieces of news shared outside of the monitored network. All disinformation operations detected this way would then also be sourced, and the obtained results compiled with the final findings.
d) Managing the alerts

The detected content was then classified according to the following criteria:

- Butac Classification and Black List, which aims to classify a large number of Italian websites according to various categories/issues;
- First Draft classification, as defined by Claire Wardle (2017), which aims at segregating the various types of misinformation;
- Issues addressed, from a list created by our Project Officer;
- Timestamp of the content;
- Source name of the content;
- Source website of the content.

If the collected piece of information was classified as disinformation or seemed otherwise suspicious, we proceeded to a sourcing analysis.

e) Sourcing the accurate alerts

We then progressed to our Social Network Analysis (SNA) framework, analysing interactions on the social platform in order to understand the source of the disinformation and how the narrative evolved online. Our analysis focused on the emergence of the disinformation piece and on the accounts that contributed to the propagation of the information. To identify and visualise conversations, ecosystems and their influencers, we use our Social Network Analysis (SNA) methodology to analyse the relationships between social actors and to represent complex social systems (Lazega, 1998).

We use three types of processes to analyse social network conversations:

a) Individual Indicators for every node/account
b) Community identification and labelling, through the use of modularity algorithms (Newman & Girvan, 2004) and the so-called “Louvain method” (Blondel, Guillaume, Lambiotte & Lefebvre, 2008).

c) Spatialisation and visualisation using Force Atlas

Results

a) Monitoring

Between December 17th, 2017 and March 4th, 2018, we analysed 171 pieces of news from various media outlets caught by our monitoring system². Most of these articles were from mainstream media outlets such as La Repubblica, Il Fatto Quotidiano, La Stampa and several others, but some were from more recently created outlets, manifesting a strong pro-Russian, anti-European point of view, such as Riscatto Nazionale, Primato Nazionale or Scenari Economici.

The content of these articles was rarely completely fake, but we noticed a certain tendency for misinformation and biased content coming from some specific websites (such as ImolaOggi, Il Populista, Scenari Economici, Dagospia and Riscatto Nazionale).

Using First Draft News’ categorisation of disinformation and misinformation, the news spotted can be classified as follows:

![First Draft - Nuzzel Classification](image)

Out of 171 news stories spotted, 54 were recognised as fake, misleading or manipulated to push a specific message, which was encased either in a Eurosceptic, anti-European discourse, or anti-migrant. Still, most of the news we spotted was either real, meaning that the news was reported as accurate and neutral (100), or an expression of personal views on political, economic or European affairs (18)³.

---

² See appendix 2 for a comprehensive list
³ See appendix 2 for a comprehensive list of all news sourced and the narratives behind them
By looking at the various narratives detected we can determine that most of the misleading content we spotted concerned directly the Italian elections (5 articles), the Italian economy (4 articles) or alleged acts of violence committed by immigrants (4 articles). In the false connections category, the elections narrative is the most present as well (5 articles), followed by a more economy-oriented narrative (2 articles). The only article included in the false context category is structured around an anti-EU narrative; this same narrative is also the most occurring in the manipulated content category (3 articles), followed by anti-migrant narratives (2 articles).

![Misinformation content](image)

**Figure 5: Classification of problematic content in terms of narratives**

If we look at how the narratives evolved from a chronological point of view, we can see that at the start of this project, the content of most articles revealed an anti-European sentiment and anti-migrant concerns, coupled with a strong interest in Italian politics and the electoral campaign.

Starting from Week 4, the presence of anti-migrant discourse in the news started to become more consistent and evolved towards a narrative focused on acts of violence conducted by migrants, a narrative that reached its peak during the last weeks of the analysis.

Furthermore, starting from Week 7, we also notice the rise of a new type of narrative, which focuses on the clash between fascist and anti-fascist movements, which is explainable by two major events that happened in Italy in the same period: the Macerata shooting and the death of 18 year old Pamela Mastropietro, which will both be explained in the next pages. Once these two events occurred, we notice a renewed interest in immigrant violence and anti-migrant narratives, which are only shadowed during the very last week of analysis by a stronger interest in the Italian elections as a whole, with several articles concerning the different candidates and the possible outcomes of the elections.

---

4 See appendix 3 for the full narrative distribution

5 See appendix 4 for the chronology of the articles
The evolution of the prevailing narrative within the content received through alerts is worth pointing out, as it seems strongly influenced by certain key events we noticed.

In the first period, from December 2017 to the beginning of February 2018, the content was mostly characterized by economic and political news that depicted the European Union and the Euro as a threat for stability and growth. In this period, almost 80% of the news spotted via our monitoring concerned either the failures of the European Union in terms of stability, economy and governance (48%), or what this entails for the Italian economy (52%).

Starting from the beginning of February, the narrative clearly shifted towards an even more controversial one: racism and illegal immigration. This major shift in terms of narrative coincided with the so-called “Macerata incident” - the death of Pamela Mastropietro, a young girl from the Marche region, who was supposedly killed by a Nigerian man. Along with the shooting and wounding of 9 black persons, committed by Luca Traini, a former Northern League candidate in the local elections in Macerata, it triggered the spread of a nationalistic, anti-migrant narrative, leading to a demonization of leftist movements, often presented as “the true evil” and “the violent ones”.

From this moment onward, most of the content reported showed a clear focus on an anti-migrant narrative: most of the news spotted between February and March concerned either racist violence and verbal attacks, or the clash between fascist and anti-fascist movements. Luca Traini’s salutation after his attack on black people was identified as a fascist salute, which contributed to the rise of clashes between fascist and anti-fascist groups. These clashes quickly spread through the streets in Italy, resulting in an overall rise of violence and aggressive behaviour.

b) Sourcing

Having developed a network of journalists, experts and fact-checkers who supported us throughout this project as well as having the Slack channel and conducting manual monitoring...
of the news, we were able to be promptly alerted about any major piece of fake news appearing on Twitter. Eventually, through our community detection (calls and collaborations with local fact-checkers and researchers), we fully sourced a total of 178 problematic content over a 10 weeks period, which represents around 2,5 false stories per day. Seven of the news we spotted were also sourced and mapped in real time.

In order to demonstrate how our methodology works in practice we will now present two examples of these real time-sourced news. The results of our research and analysis were shared on EU DisinfoLab’s Twitter accounts, one of them being specifically created for the Italian audience in order to gain more visibility for the findings.

1. February 13th, 2018 – The Trenitalia case

On February 13th, we were alerted about a new suspicious piece of news, which had already been debunked as fake. In the morning of February 12th, a man named Luca Caruso shared a Facebook post concerning a migrant who was supposedly traveling on a Frecciarossa train without having a valid ticket.

After crawling data on Twitter through the use of keywords and Boolean indicators, we came to realise that the news was fake, and that a whole debunking echochamber had been quickly activated: a whole network of fact-checkers, journalists and bloggers started sharing the debunked news of the evening of February 12th.

The post, which was published in the early morning hours, reached more than 120.000 reactions and 80.000 shares within a few hours. It created a massive uproar on Facebook, and triggered a quick response from the Italian train company Trenitalia, owner of Frecciarossa trains. In their statement, Trenitalia exposed the facts behind what truly happened: the man was sitting on the wrong seat and his knowledge of Italian was not sufficient for him to properly communicate with the officer. The officer saw that the scene was attracting a lot of attention from the other passengers, so he took him out to the corridor to figure out what was happening.

2. March 4th, 2018 - The Sicilian ballots case

We kept our monitoring system active even during the electoral weekend, to be sure we do not miss any last-minute attempt at a disinformation operation aimed at interfering with the voting process. This is when we spotted a new disinformation operation, which was published right before the opening of voting stations.

According to Il Fatto, an Italian news platform whose name resembles the better known media outlet Il Fatto Quotidiano, more than 500.000 ballot sheets had supposedly been discovered in Sicily on the day of the elections as part of an operation to shift the vote in favour or Democratic Party, the logo of which was supposedly visible on the ballot sheets (Sicilia: Trovate 500 mila schede precompilate, 2018).

As in the previous example, we analysed all tweets related to this topic, which we located through selected keywords, and again discovered this news was fake. The news had been quickly debunked once again, from fact-checkers and journalists who investigated the news and then spread their findings on Twitter.
This turned out to be the most shared fake news of the whole election period: at 6:35am of March 4th, just on Facebook, the story counted 26037 reactions, 21221 comments and 18067 shares.

Discussion on the results

First of all, there are several differences between the Italian case and the French case. In the Italian context, the very few disinformation operations we witnessed were mainly spread by individuals who were somehow directly linked to the political forces involved in the elections; hence, it was not possible to provide clear evidence of a foreign actor meddling with the elections in favour of one party or the other.

Yet, thanks also to our network of experts and journalists, we managed to set up an alert system that turned out to be quite effective: in fact, a third of the alerts we received led us to content that was either misinformation or disinformation.

However, we also encountered a series of challenges which should not be disregarded.

One of the issues we faced was that the majority of the content we spotted through alerts, our network of experts and our own monitoring process, was not pure disinformation. Many of the articles we found can be rather labelled as misinformation, with the content being mostly accurate apart from one misleading aspect, like figures, timing, or phrasing. The fact that most of the content we detected was of the misleading kind suggests that in Italy, when aiming to spread disinformation through the news, it is mostly achieved through publishing content written and presented in a manner leading the audience to interpret it in a specific, predefined way.

It is also worth mentioning that in Italy the preferred platform used for conducting disinformation and misinformation operations is Facebook, which is also considered a trustworthy source of information for the Italian public (Russo, 2015). Unfortunately, with no possibility to run social network analysis on the platform due to lack of data access, we were not able to track how disinformation spread or which communities were involved. Looking back at the Trenitalia story, for example, it would have been interesting to understand which networks on Facebook propagated this false information and if this spread was organic or driven by advertisement.

It is also important to note that many of the disinformation operations we investigated were received through alerts from our partner network, showing our monitoring potential needs to be combined with strong local partnerships.

This leads us to question the capability of our methodology to adapt to countries/zones where Twitter has less penetration within the population. Twitter was used less as a channel to spread disinformation and more as a debunking echo chamber: in fact, all the news we investigated, which had a significant amount of visibility on Twitter, were mostly traceable because of the activity of debunkers on the platform. Unfortunately, the same cannot be said of Facebook, where a lot of pieces of fake information was shared by thousands of people, yet they could not be traced back to their original network or propagators (Università degli studi di Urbino Carlo B. 2018). This tendency can be explained in two ways: firstly, the use of Twitter in Italy is less distributed than in other countries (Leading countries based on number of Twitter users as of April 2018, 2018), and secondly, because of the accessibility of Twitter
data, disinformation agents tend to use platforms where there are almost no possibilities of being exposed, Facebook being one of these platforms.

Our local partners and experts had a key role not only in warning us whenever they spotted a piece of disinformative content but also in sharing our findings and analyses once we published them on Twitter. This shows that however accurate our methodology is, it still requires cooperation with a solid partner network. Successfully assembling a group of experts, located both in Brussels and in Italy, had a positive impact on the way this study was managed and delivered, which demonstrates how projects like this can be conducted if there is a strong collaborative effort between researchers and experts. This objective can only be fulfilled by continuing and strengthening our outreach efforts, through collaborations and discussions with our community.

The success of the project was largely dependent on reaching a wide audience among Italian media and relevant stakeholders. During the two months of our monitoring process, we were able to attract close to 300 followers on our dedicated Italian Twitter account and, according to Twitter analytics, our tweets earned 109000 impressions between January and the end of March. Furthermore, the most popular content we produced included exposing disinformation in real time with visual aids like maps, showing the community’s need for real time analysis with results presented in a clear, graphic form.

Conclusions

Disinformation cannot always be detected and identified. It evolves, it does not repeatedly appear in the same form and the patterns it can follow depend on numerous criteria, such as geographic location, political landscape and culture, as well as experience, resources and tools available to disinformation agents. For instance, after being banned from Twitter in 2014, many ISIS terrorists abandoned both Twitter and Facebook to then regroup on Telegram (Robins-Early, 2017) because its encryption capabilities made it possible for them to conceal their conversations and network.

Still, Facebook’s refusal to share data with researchers and experts is one of the leading obstacles in successfully understanding and countering disinformation and misinformation, as most of it is gathered and shared on that platform.

Despite this obstacle we nevertheless managed to find a way to monitor disinformation spreading in an electoral context. But to have meaningful impact our method - apart from requiring data access from platforms - needs to be complemented with a strong network of local journalists, consultants and experts.

Looking to the future and exploring optimal areas of research, we believe that in order to pierce filter bubbles, social researchers should focus on studying the possible ways and efficiency of using semi-automated or automated tools to retarget verified content to networks spreading disinformation. It would support media literacy and other awareness-raising initiatives, while encouraging responsibility from individuals and organisations facing misleading or disinformative content.
Figure list

Figure 1: The triangulation process .................................................................. 4
Figure 2: The susceptible communities ................................................................. 4
Figure 3: Hybrid approach to detect disinformation networks ................................. 6
Figure 4: Classification of our monitoring findings according to First Draft’s criteria .... 7
Figure 5: Classification of problematic content in terms of narratives ..................... 8
Figure 6: Chronological evolution of the principal narratives during our monitoring project.... 9

Table list

Table 1: Comparison of communities found with different community detection algorithms, the partition given by the Louvain algorithm and a random partition. .............................................. 5

Appendices

Appendix 1: Audience and engagement obtained with disinformation sourcing
Appendix 2: First Draft classification of Nuzzel findings
Appendix 3: Misinformation content narratives distribution
Appendix 4: Chronology
Reference List


Leading countries based on number of Twitter users as of April 2018 (in millions), Statista, Retrieved from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/


