Pierri et al. *EPJ Data Science* (2020) 9:35
https://doi.org/10.1140/epjds/s13688-020-00253-8

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

A multi-layer approach to disinformation detection in US and Italian news spreading on Twitter

Francesco Pierri[1*], Carlo Piccardi[1] and Stefano Ceri[1]

*Correspondence:
francesco.pierri@polimi.it
[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy

Abstract

We tackle the problem of classifying news articles pertaining to *disinformation* vs *mainstream* news by solely inspecting their diffusion mechanisms on Twitter. This approach is inherently simple compared to existing text-based approaches, as it allows to by-pass the multiple levels of complexity which are found in news content (e.g. grammar, syntax, style). As we employ a multi-layer representation of Twitter diffusion networks where each layer describes one single type of interaction (tweet, retweet, mention, etc.), we quantify the advantage of separating the layers with respect to an aggregated approach and assess the impact of each layer on the classification. Experimental results with two large-scale datasets, corresponding to diffusion cascades of news shared respectively in the United States and Italy, show that a simple Logistic Regression model is able to classify disinformation vs mainstream networks with high accuracy (AUROC up to 94%). We also highlight differences in the sharing patterns of the two news domains which appear to be common in the two countries. We believe that our network-based approach provides useful insights which pave the way to the future development of a system to detect misleading and harmful information spreading on social media.

Keywords: Computational social science; Disinformation; Multi-layer networks; Twitter

1 Introduction

In recent years there has been increasing interest in the issue of disinformation spreading on online social media. Global concern over false (or "fake") news as a threat to modern democracies has been frequently raised—ever since 2016 US Presidential elections—in correspondence of events of political relevance, where the proliferation of manipulated and low-credibility content attempts to drive and influence people's opinions [1–4].

Researchers have highlighted several drivers for the diffusion of such malicious phenomenon, which include human factors (confirmation bias [5], naive realism [6]), algorithmic biases (*filter bubble* effect [1]), the presence of deceptive agents on social platforms (bots and trolls [7]) and, lastly, the formation of *echo chambers* [8] where people polarize their opinions as they are insulated from contrary perspectives.

The problem of automatically detecting online disinformation news has been typically formulated as a binary classification task (i.e. credible vs non-credible articles), and tackled with a variety of different techniques, based on traditional machine learning and/or deep learning, which mainly differ in the dataset and the features they employ to perform the classification. We may distinguish three approaches: those built on content-based features, those based on features extracted from the social context, and those which combine both aspects. A few main challenges hinder the task, namely the impossibility to manually verify all news items, the lack of gold-standard datasets and the adversarial setting in which malicious content is created [4, 7, 9].

In this work we follow the direction pointed out in a few recent contributions on the diffusion of disinformation compared to traditional and objective information. These contributions have shown that false news spread faster and deeper than true news [10], and that social bots and echo chambers play an important role in the diffusion of malicious content [7, 8]. Therefore, we focus on the analysis of spreading patterns which naturally arise on social platforms as a consequence of multiple interactions between users, due to the increasing trend in online sharing of news [1].

We propose a classification framework based on a multi-layer formulation of Twitter diffusion networks, extending the results of our previos work [11]. For each article we disentangle different social interactions on Twitter, namely tweets, retweets, mentions, replies and quotes, to accordingly build a diffusion network composed of multiple layers (one for each type of interaction), and we compute structural features separately for each layer. We pick a set of global network properties from the network science toolbox which can be qualitatively explained in terms of social dimensions and allow us to encode different networks with a tuple of features. These include traditional indicators, e.g. network density, number of strong/weak connected components, and diameter, and more elaborated ones such as main K-core number [12] and structural virality [13]. Our work is driven by the following research questions:

- *RQ1*: Does a multi-layer representation of Twitter diffusion networks yield a significant advance in terms of classification accuracy over a conventional single-layer diffusion network?
- *RQ2*: Which of the above features, and which layer, are the most effective in the classification task?

We perform classification experiments with an off-the-shelf Logistic Regression model on two different datasets of mainstream and disinformation news shared on Twitter respectively in the United States and in Italy during 2019. In the former case we also perform multiple disaggregated tests to control for political biases inherent to different news sources, referring to the procedure proposed in [3] to label different outlets. Overall we show that we are able to classify credible vs non-credible diffusion networks (and consequently news articles) with high accuracy (AUROC up to 94%), also when controlling for the political bias of sources (and training only on left-biased or right-biased articles). We observe that the layer of mentions alone conveys useful information for the classification, denoting a different usage of this functionality when sharing news belonging to the two news domains. We also show that the most discriminative features, which are relative to the breadth and depth of the largest cascades in different layers, are the same across the two countries.

As our datasets are collected in different countries, we also investigate whether disinformation can be detected independently from the country where it originates. Cross-country experiments show that our methodology fails to distinguish reliable vs non-reliable news regardless of where it originates from. We argue that this might be due either to the high imbalance of data or to the class discrepancies which are country specific. It emerges that a classifier based on our methodology should be trained in a country-wise fashion.

The outline of this paper is the following: we first provide a description of related literature; next, we describe our methodology acknowledging its intrinsic limitations; then, we provide experimental results and finally we draw conclusions and future directions.

2 Related literature

A deep learning framework for detecting fake news cascades is proposed in [14], where the authors refer to [10] in order to collect Twitter cascades pertaining to verified false and true rumors. They employ *geometric* deep learning, a novel paradigm for graph-based structures, to classify cascades based on four categories of features, such as user profile, user activity, network and spreading, and content. They also observe that a few hours of propagation are sufficient to distinguish false news from true news with high accuracy.

Diffusion cascades on Weibo and Twitter are analyzed in [15], where the authors focus on highlighting different topological properties, such as the number of *hops* from the source or the heterogeneity of the network, to show that fake news shape diffusion networks which are highly different from credible news, even at early stages of propagation.

In this work, we consider the results of [11] as our baseline. In that paper, we classified US news articles leveraging Twitter diffusion networks. Here we further build on that approach by exploiting a disaggregated multi-layer representation of Twitter social interactions, in order to take advantage of a larger and more specific set of features. The methodology used in [11] has several analogies with [16], where authors successfully detect Twitter *astroturfing* content, i.e. political campaigns disguised as spontaneous grassroots, with a machine learning framework based on network features.

3 Methodology

3.1 Disinformation and mainstream news

In this work we formulate our classification problem as follows: given two classes of news articles, respectively D (*disinformation*) and M (*mainstream*), a set of news articles A_i and associated class labels $C_i \in \{D, M\}$, and, for each article A_i, a set of tweets $\Pi_i = \{T_i^1, T_i^2, \ldots\}$ each containing an Uniform Resource Locator (URL) pointing explicitly to article A_i, predict the class C_i of each article A_i.

There is huge debate and controversy on a proper taxonomy of malicious and deceptive information [2–4, 11, 17–19]. In this work we prefer the term *disinformation* to the more specific *fake news* to refer to a variety of misleading and harmful information. Therefore, we follow a *source-based* approach, a consolidated strategy also adopted by [2, 3, 7, 18], in order to obtain relevant data for our analysis. We collected:

1. Disinformation articles, published by websites which are known for producing low-credibility content, false and misleading news reports as well as extreme propaganda and hoaxes, and flagged as such by reputable journalists and fact-checkers;

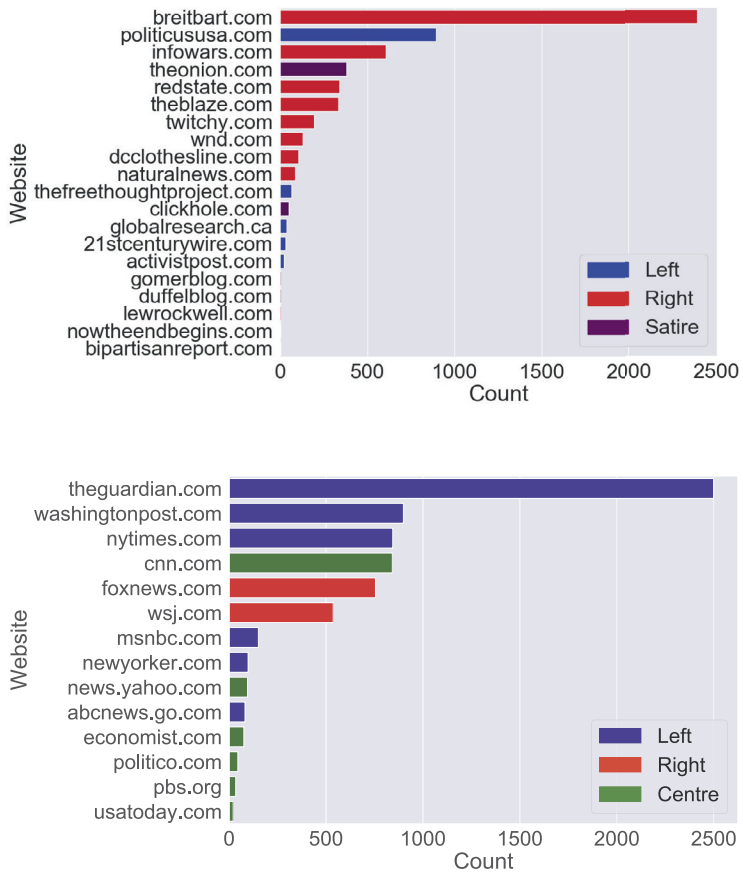

Figure 1 Distribution of the number of articles per source for US (**a**) disinformation and (**b**) mainstream news. Colors indicate the political bias label of each source

2. Mainstream news, referring to traditional news outlets which deliver factual and credible information.

3.2 US dataset

We collected tweets associated to a dozen US *mainstream* news websites, i.e. most trusted sources described in [20], with the Streaming API, and we referred to Hoaxy API [18] for what concerns tweets containing links to 100+ US *disinformation* outlets. We filtered out articles associated to less than 50 tweets to reduce noisy observations. The resulting dataset contains overall ∼1.7 million tweets for mainstream news, collected in a period of three weeks (February 25th, 2019–March 18th, 2019), which are associated to 6978 news articles, and ∼1.6 million tweets for disinformation, collected in a period of three months (January 1st, 2019–March 18th, 2019) for the sake of balance of the two classes, which hold 5775 distinct articles. Diffusion censoring effects [13] were correctly taken into account in both collection procedures. We provide in Fig. 1 the distribution of articles by source and political bias for both news domains.

As it is reported that conservatives and liberals exhibit different behaviors on online social platforms [21–23], we further assigned a political bias label to different US outlets (and therefore news articles) following the procedure described in [3]. In order to assess the robustness of our method, we performed classification experiments by training only

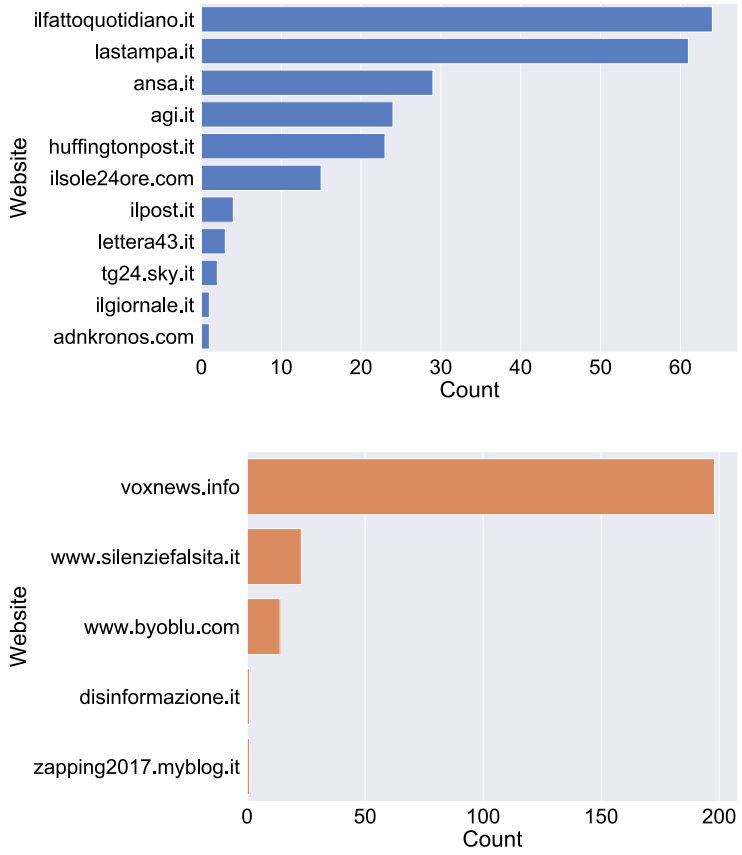

Figure 2 Distribution of the number of articles per source for Italian (**a**) mainstream and (**b**) disinformation news

on left-biased (or right-biased) outlets of both disinformation and mainstream domains and testing on the entire set of sources. As additional test, we excluded particular sources that outweigh the others in terms of samples to avoid over-fitting.

3.3 Italian dataset

For what concerns the Italian scenario we first collected tweets with the Streaming API in a 3-week period (April 19th, 2019–May 5th, 2019), filtering those containing URLs pointing to Italian official newspapers websites as described in [24, 25]; these correspond to the list provided by the association for the verification of newspaper circulation in Italy (Accertamenti Diffusione Stampa).[a] We instead referred to the dataset provided by [26] to obtain a set of tweets, collected continuously since January 2019 using the same Twitter endpoint, which contain URLs to 60+ Italian disinformation websites (the list is available in the Additional file 1); these were obtained by using black-lists from Italian fact-checking websites and agencies (PagellaPolitica.it, Bufale.net and Butac.it). In order to get balanced classes, we retained data collected in a longer period w.r.t to mainstream news (April 5th, 2019–May 5th, 2019). In both cases we filtered out articles with less than 50 tweets; overall this dataset contains ∼160k mainstream tweets, corresponding to 227 news articles, and ∼100k disinformation tweets, corresponding to 237 news articles. We provide in Fig. 2 the distribution of articles according to distinct sources for both news domains. As in the US

Table 1 Breakdown of US and IT datasets in terms of different Twitter interactions

Country	Class	Mentions	Replies	Retweets	Quotes	Tweets
United States	Mainstream	87,183	30,745	1,482,261	29,365	409,544
	Disinformation	123,047	22,599	1,207,243	94,027	220,891
Italy	Mainstream	1578	473	18,794	1378	4832
	Disinformation	929	186	35,323	3192	5302

dataset, we took into account censoring effects [13] by excluding tweets published before (*left-censoring*) or after two weeks (*right-censoring*) from the beginning of the collection process.

The different volumes of news shared on Twitter in the two countries are due both to the different population size of US and Italy (320 vs 60 million) but also to the different usage of Twitter platform (and social media in general) for news consumption [27]. Both datasets analyzed in this work are available from the authors on request.

3.4 Breakdown of Twitter interactions

We disentangle different social interactions on Twitter according to five categories:

Mention (M): Including in a tweet another account's Twitter user name, preceded by the "@" symbol;

Reply (R): Responding to another account's tweet;

Retweet (RT): Re-posting a tweet;

Quote (Q): Retweeting with the addition of a comment;

Tweet (T): Posting a tweet containing an article URL.

We show in Table 1 the breakdown of our datasets for what concerns cardinalities of different Twitter interactions across news domains. We notice that news sharing mostly involves retweeting and tweets in both countries and for both classes of news articles.

For what concerns different Twitter actions, users primarily interact with each other using retweets and mentions [22]. The former are the main engagement activity and act as a form of endorsement, allowing users to rebroadcast content generated by other users [28]. Besides, when node B retweets node A we have an implicit confirmation that information from A appeared in B's Twitter feed [16]. Quotes are simply a special case of retweets with comments. Mentions usually include personal conversations as they allow someone to address a specific user or to refer to an individual in the third person; in the first case they are located at the beginning of a tweet and they are known as replies, otherwise they are put in the body of a tweet [22]. The network of mentions is usually seen as a stronger version of interactions between Twitter users, compared to the traditional graph of follower/following relationships [29].

3.5 Building diffusion networks

Using the notation described in [30] we employ a multi-layer representation for Twitter diffusion networks. Sociologists have indeed recognized decades ago that it is crucial to study social systems by constructing multiple social networks where different types of ties among the same individuals are used [31]. Therefore, for each news article we build a multi-layer diffusion network composed of four different layers, one for each type of social interaction on Twitter platform, namely retweet (RT), reply (R), quote (Q) and mention (M), as shown in Fig. 3. These networks are not necessarily *node-aligned*, i.e. users might be missing in some layers. We do not insert "dummy" nodes to represent users not active in

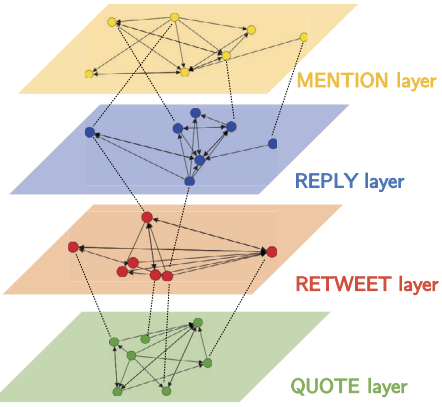

Figure 3 A visualization of a Twitter multi-layer diffusion network with four layers

a given layer as it would have severe impact on the global network properties (e.g. number of weakly connected components). Alternatively one may look at each multi-layer diffusion network as an ensemble of individual graphs [30]; since global network properties are computed separately for each layer, they are not affected by the presence of *inter-layer* edges, which nonetheless allow the diffusion of information across layers.

In our multi-layer representation, each layer is a directed graph where we add edges and nodes for each tweet of the layer type. While the direction of information flow—thus the edge direction—is unambiguous for some layers, e.g. RT, the same is not true for others. Here we follow the conventional approach described e.g. in [7, 16, 19, 22] to define the direction of edges. For the RT layer: whenever user a retweets account b we first add nodes a and b if not already present in the RT layer, then we build an edge that goes from b to a if it does not exist. Similarly for the other layers: for the R layer edges go from user a (who replies) to user b, for the Q layer edges go from user b (who is quoted by) to user a and for the M layer edges go from user a (who mentions) to user b. Note that, by construction, our layers do not include isolated nodes; they correspond to "isolated tweets", i.e. tweets which have not originated any interactions with other users. However, they are present in our dataset, and their number is exploited for classification, as described below.

3.6 Global network properties

We used a set of global network indicators which encode each network layer by a tuple of features. Then we simply concatenated tuples as to represent each multi-layer network with a single feature vector. We used the following global network properties:

1. *Number of Strongly Connected Components (SCC)*: a Strongly Connected Component of a directed graph is a maximal (sub)graph where for each pair of vertices u, v there is a path in each direction ($u \to v$, $v \to u$).
2. *Size of the Largest Strongly Connected Component (LSCC)*: the number of nodes in the largest strongly connected component of a given graph.
3. *Number of Weakly Connected Components (WCC)*: a Weakly Connected Component of a directed graph is a maximal (sub)graph where for each pair of vertices (u, v) there is a path $u \leftrightarrow v$ ignoring edge directions.
4. *Size of the Largest Weakly Connected Component (LWCC)*: the number of nodes in the largest weakly connected component of a given graph.

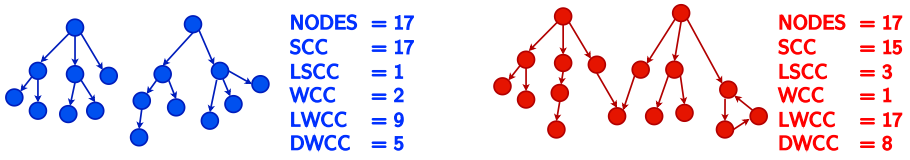

Figure 4 Two illustrative examples of diffusion layers. Left: The same news spreads, in a pure top-down broadcast manner, along two distinct cascades. Thus, SCC equals the number of nodes, since each strongly connected component is a single node, while WCC is the number of distinct cascades. Right: The two cascades merge in a common node (thus WCC = 1) and, additionally, mono-directionality is broken by a loop (thus SCC is less than the number of nodes)

5. *Diameter of the Largest Weakly Connected Component (DWCC)*: the largest distance (number of edges of the shortest path) between two nodes in the (undirected version of) largest weakly connected component of a graph.
6. *Average Clustering Coefficient (CC)*: the average of the local clustering coefficients of all nodes in a graph; the local clustering coefficient of a node quantifies how close its neighbourhood is to being a complete graph (or a clique). It is computed according to [32].
7. *Main K-core Number (KC)*: a K-core [12] of a graph is a maximal sub-graph that contains nodes of internal degree K or more; the main K-core number is the highest value of K (in directed graphs the total degree is considered).
8. *Density (d)*: the density for directed graphs is $d = \frac{|E|}{|V||V-1|}$, where $|E|$ is the number of edges and $|V|$ is the number of vertices in the graph; the density equals 0 for a graph without edges and 1 for a complete graph.
9. *Structural virality of the largest weakly connected component (SV)*: this measure is defined in [13] as the average distance between all pairs of nodes in a cascade tree or, equivalently, as the average depth of nodes, averaged over all nodes in turn acting as a root; for $|V| > 1$ vertices, $\mathrm{SV} = \frac{1}{|V||V-1|} \sum_i \sum_j d_{ij}$ where d_{ij} denotes the length of the shortest path between nodes i and j. This is equivalent to compute the Wiener's index [33] of the graph and multiply it by a factor $\frac{1}{|V||V-1|}$. In our case we computed it for the undirected equivalent graph of the largest weakly connected component, setting it to 0 whenever $|V| = 1$.

We used `networkx` Python package [34] to compute all features. Whenever a layer is empty, we simply set to 0 all its features. In addition to computing the above nine features for each layer, we added two indicators for encoding information about isolated tweets, namely the number T of isolated tweets (containing URLs to a given news article) and the number U of unique users authoring those tweets. Therefore, a diffusion network for a given article is represented by a vector with $9 \cdot 4 + 2 = 38$ entries.

3.7 Interpretation of network features and layers

The aforementioned network properties can be qualitatively explained in terms of social footprints as follows (see the illustrative examples in Fig. 4): in this specific class of networks, *SCC* correlates with the size (i.e. number of nodes) of the diffusion layer, as the propagation of news occurs in a broadcast manner in most cases, i.e. re-tweets dominate on other interactions, while *LSCC* allows to distinguish cases where such mono-directionality is somehow broken. *WCC* equals (approximately) the number of distinct diffusion cascades pertaining to each news article, with exceptions corresponding to those

cases where some cascades merge together via Twitter interactions such as mentions, quotes and replies, and accordingly *LWCC* and *DWCC* equals the size and the depth of the largest cascade. *CC* corresponds to the level of connectedness of neighboring users in a given diffusion network whereas *KC* identifies the set of most influential users in a network [19]. Finally, *d* describes the proportions of potential connections between users which are actually activated and *SV* indicates whether a news item has gained popularity with a single and large broadcast or in a more viral fashion through multiple generations [13].

3.8 Limitations

As mentioned beforehand, we use a coarse approach to label articles at the source level relying on a huge corpus of literature on the subject. We believe that this is currently the most reliable classification approach, although it entails obvious limitations, as disinformation outlets may also publish true stories and likewise misinformation is sometimes reported on mainstream media [4]. Also, given the choice of news sources, we cannot test whether our methodology is able to classify disinformation vs factual but not mainstream news which are published on niche, non-disinformation outlets [11].

Another crucial aspect in our approach is the capability to fully capturing sharing cascades on Twitter associated to news articles. It has been reported [35] that the Twitter streaming endpoint filters out tweets matching a given query if they exceed 1% of the global daily volume[b] of shared tweets, which nowadays is approximately $5 \cdot 10^8$; however, as we always collected less than 10^6 tweets per day, we did not incur in this issue and we thus gathered 100% of tweets matching our query.

We built Twitter diffusion networks using an approach widely adopted in the literature [3, 7, 19]. We remark that there is an unavoidable limitation in Twitter Streaming API, which does not allow to retrieve *true* re-tweeting cascades because re-tweets always point to the original source and not to intermediate re-tweeting users [10, 13]; thus we adopt the only viable approach based on Twitter's public availability of data. However, by disentangling different interactions with multiple layers we potentially reduce the impact of this limitation on the global network properties compared to the approach used in our baseline.

Finally, a limitation of the present work is the lack of a direct comparison of our methodology with other techniques, an exercise which boils down to assessing several classification metrics on the same dataset(s). As thoroughly discussed in [36], the problem of reliably comparing fake-news classifiers is open and faces many types of challenges that go out the scope of this work. We just mention that the performance of our classification framework is quantitatively comparable (in terms of AUROC value) to that of state-of-the-art deep learning models for fake news detection [14, 37]. However, this result is only indicative, because obtained on different datasets and, in one case [37], with different focus of the classification task.

4 Experiments

4.1 Setup

We performed classification experiments using a basic off-the-shelf classifier, namely Logistic Regression (LR) with L2 penalty; this also allows us to compare results with our baseline [11]. We applied a standardization of the features and we used the default configuration for parameters as described in `scikit-learn` package [38]. We also tested other

Table 2 Composition of the US dataset according to domain (mainstream vs disinformation), size class (number of unique users who interact with a given news) and political bias

	No. mainstream			No. disinformation		
Size class	Left	Right	Tot.	Left	Right	Tot.
[0, 100)	774	2746	4177	379	2086	2640
[100, 1000)	1712	464	2605	654	1946	2900
[1000, +∞)	115	54	196	19	162	235
[0, +∞)	4573	1292	6978	1052	4194	5575

Table 3 Composition of the Italian dataset according to domain (mainstream vs disinformation) and size class (number of unique users who interact with a given news)

Size class	No. mainstream	No. disinformation
[0, 100)	165	79
[100, 1000)	61	158
[0, +∞)	226	237

classifiers (such as K-Nearest Neighbors, Support Vector Machines and Random Forest) but we omit results as they give comparable performance. We remark that our goal is to show that a very simple machine learning framework, with no parameter tuning and optimization, allows for accurate results with our network-based approach.

We used the following evaluation metrics to assess the performance of different classifiers (TP = true positives, FP = false positives, FN = false negatives):

1. *Precision* = $\frac{TP}{TP+FP}$, the ability of a classifier not to label as positive a negative sample.
2. *Recall* = $\frac{TP}{TP+FN}$, the ability of a classifier to retrieve all positive samples.
3. *F1-score* = $2\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$, the harmonic average of Precision and Recall.
4. *Area Under the Receiver Operating Characteristic curve (AUROC)*: the Receiver Operating Characteristic (ROC) curve [39], which plots the TP rate versus the FP rate, shows the ability of a classifier to discriminate positive samples from negative ones as its threshold is varied; the AUROC value is in the range $[0, 1]$, with the random baseline classifier holding AUROC = 0.5 and the ideal perfect classifier AUROC = 1; thus larger AUROC values (and steeper ROCs) correspond to better classifiers.

In particular we computed so-called *macro* average—simple unweighted mean—of these metrics evaluated considering both labels (*disinformation* and *mainstream*). We employed stratified shuffle split cross validation (with 10 folds) to evaluate performance.

Finally, we partitioned networks according to the total number of unique users involved in the sharing, i.e. the number of nodes in the aggregated network represented with a single-layer representation considering together all layers and also isolated tweets. A breakdown of both datasets according to size class (and political biases for the US scenario) is provided in Table 2 and Table 3.

4.2 Classification performance

In Table 4 we first provide classification performance on the US dataset for the LR classifier evaluated on the size classes described in Table 2. We can observe that in all instances our methodology performs much better than a random classifier (50% AUROC), with AUROC values above 85% in all cases.

For what concerns political biases, as the classes of mainstream and disinformation networks are not balanced (e.g., 1292 mainstream and 4149 disinformation networks with

Table 4 Performance of the LR classifier (using a multi-layer approach) evaluated on different size classes on both the US (top rows) and the Italian (bottom rows) dataset

Size class	AUROC	Precision	Recall	F1-score
(US) [0, 100)	0.87 ± 0.01	0.79 ± 0.01	0.77 ± 0.01	0.78 ± 0.01
(US) [100, 1000)	0.93 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01
(US) [1000, $+\infty$)	0.94 ± 0.02	0.86 ± 0.05	0.86 ± 0.05	0.86 ± 0.05
(US) [0, $+\infty$)	0.88 ± 0.01	0.81 ± 0.01	0.80 ± 0.01	0.80 ± 0.01
(IT) [0, 100)	0.89 ± 0.06	0.81 ± 0.11	0.82 ± 0.11	0.81 ± 0.11
(IT) [100, 1000)	0.86 ± 0.07	0.83 ± 0.08	0.78 ± 0.06	0.80 ± 0.06
(IT) [0, $+\infty$)	0.90 ± 0.02	0.81 ± 0.05	0.81 ± 0.05	0.81 ± 0.05

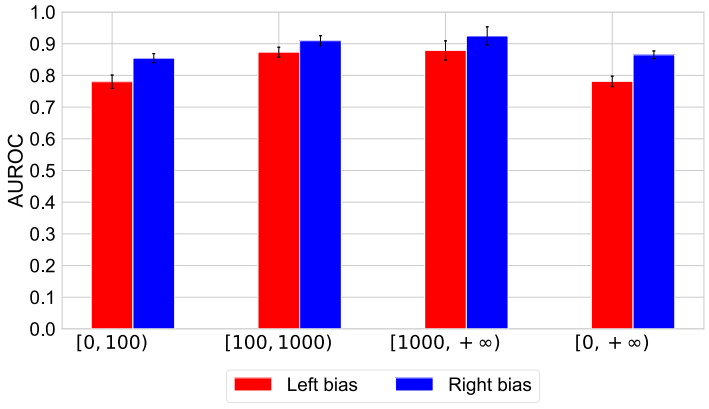

Figure 5 AUROC values for the Balanced Random Forest classifier trained on left-biased (red) and right-biased (blue) news articles in the US dataset, and tested on the entire dataset. Error bars indicate the standard deviation of AUROC values over different folds of the cross validation

right bias) we employ a Balanced Random Forest with default parameters (as provided in `imblearn` Python package [40]). In order to test the robustness of our methodology, we trained only on left-biased networks or right-biased networks and tested on the entire set of sources (relative to the US dataset); we provide a comparison of AUROC values for both biases in Fig. 5. We can notice that our multi-layer approach still entails significant results, thus showing that it can accurately distinguish mainstream news from disinformation regardless of the political bias. We further corroborated this result with additional classification experiments, that yield similar performance, in which we excluded from the training/test set two specific sources (one at a time and both at the same time) that outweigh the others in terms of data samples—respectively "breitbart.com" for right-biased sources and "politicususa.com" for left-biased ones [36].

We performed classification experiments on the Italian dataset using the LR classifier and different size classes (notice that [1000, $+\infty$) is empty for the Italian dataset); we show results for different evaluation metrics in Table 4. We can see that despite the limited number of samples (one order of magnitude smaller than the US dataset) the performance is overall in accordance with the US scenario.

As shown in Table 5, we obtain results which are much better than our baseline in all size classes:

- In the US dataset our multi-layer methodology performs much better in all size classes except for large networks ([1000, $+\infty$) size class), reaching up to 13% improvement on smaller networks ([0, 100) size class);

Table 5 Comparison of performance of our multi-layer approach vs the baseline (single-layer). We show AUROC values for the LR classifier evaluated on different size classes of both US and IT datasets

Size class	Single-layer	Multi-layer
(US) [0, 100)	0.74 ± 0.02	0.87 ± 0.01
(US) [100, 1000)	0.85 ± 0.02	0.93 ± 0.01
(US) [1000, +∞)	0.93 ± 0.03	0.94 ± 0.02
(US) [0, +∞)	0.78 ± 0.02	0.88 ± 0.01
(IT) [0, 100)	0.77 ± 0.08	0.89 ± 0.06
(IT) [100, 1000)	0.66 ± 0.14	0.86 ± 0.07
(IT) [0, +∞)	0.74 ± 0.12	0.90 ± 0.02

Table 6 Different evaluations metrics for LR classifier evaluated on different size classes of the US dataset and trained using features separately for each layer. Best scores for each row are written in bold

Size class	Metric	Quotes	Retweets	Mentions	Replies
[0, 100)	AUROC	**0.75 ± 0.02**	0.63 ± 0.02	**0.75 ± 0.02**	0.61 ± 0.02
	Precision	**0.71 ± 0.02**	0.59 ± 0.02	0.70 ± 0.02	0.60 ± 0.04
	Recall	0.66 ± 0.01	0.55 ± 0.01	**0.67 ± 0.01**	0.54 ± 0.02
	F1-score	0.66 ± 0.02	0.53 ± 0.02	**0.68 ± 0.02**	0.50 ± 0.06
[100, 1000)	AUROC	**0.81 ± 0.02**	0.63 ± 0.02	**0.81 ± 0.02**	0.65 ± 0.03
	Precision	0.73 ± 0.02	0.61 ± 0.02	**0.75 ± 0.02**	0.65 ± 0.02
	Recall	0.73 ± 0.02	0.60 ± 0.02	**0.75 ± 0.02**	0.62 ± 0.02
	F1-score	0.73 ± 0.02	0.60 ± 0.02	**0.75 ± 0.02**	0.60 ± 0.02
[1000, +∞)	AUROC	**0.85 ± 0.08**	0.62 ± 0.08	0.84 ± 0.04	0.66 ± 0.06
	Precision	**0.80 ± 0.08**	0.61 ± 0.08	0.75 ± 0.06	0.61 ± 0.10
	Recall	**0.80 ± 0.08**	0.60 ± 0.07	0.75 ± 0.06	0.59 ± 0.07
	F1-score	**0.79 ± 0.08**	0.59 ± 0.08	0.75 ± 0.06	0.58 ± 0.09
[0, +∞)	AUROC	0.76 ± 0.01	0.62 ± 0.01	**0.77 ± 0.01**	0.59 ± 0.04
	Precision	0.70 ± 0.01	0.58 ± 0.01	**0.73 ± 0.01**	0.59 ± 0.05
	Recall	0.69 ± 0.01	0.56 ± 0.01	**0.71 ± 0.01**	0.55 ± 0.03
	F1-score	0.69 ± 0.01	0.53 ± 0.01	**0.71 ± 0.01**	0.52 ± 0.05

- In the IT dataset our multi-layer methodology outperforms the baseline in all size classes, with the maximum performance gain (20%) on medium networks ([100, 1000) size class); the baseline generally reaches worst performance compared to the US scenario.

4.3 Layer importance analysis

In order to understand the impact of each layer on the performance of classifiers, we performed additional experiments considering separately each layer (we ignored T and U features relative to isolated tweets).

In Table 6 we show metrics for each layer and all size classes, computed with a 10-fold stratified shuffle split cross validation, evaluated on the US dataset; in Fig. 6 we show AUROC values for each layer compared with the general multi-layer approach. We can notice that both Q and M layers alone capture adequately most of discrepancies of the two distinct news domains in the United States as they obtain good results with AUROC values in the range 75%–86%; these are comparable with those of the multi-layer approach which, nevertheless, outperforms them across all size classes.

In the Italian dataset we observe that the M layer obtains comparable performance w.r.t the multi-layer approach for what concerns small networks and the dataset altogether, whereas the RT layer performs better on large networks (see Table 7 and Fig. 7). We also notice higher values in standard deviations of performance metrics which are likely due to the limited size of the training/test data.

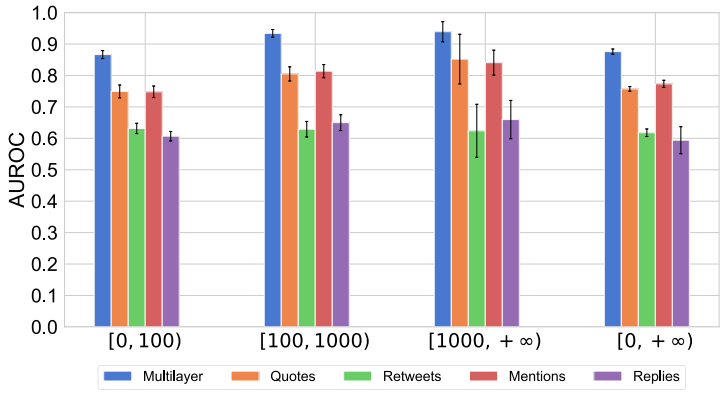

Figure 6 AUROC values for the LR classifier (evaluated on different size classes of the US dataset) trained using different layers separately and together (our multi-layer approach). Error bars indicate the standard deviation of AUROC values over different folds of the cross validation

Table 7 Different evaluations metrics for LR classifier evaluated on different size classes of the IT dataset and trained using features separately for each layer. Best scores for each row are written in bold

Size class	Metric	Quotes	Retweets	Mentions	Replies
[0, 100)	AUROC	0.49 ± 0.12	0.73 ± 0.08	**0.74 ± 0.06**	0.49 ± 0.09
	Precision	0.34 ± 0.00	**0.61 ± 0.15**	0.58 ± 0.08	0.34 ± 0.00
	Recall	0.50 ± 0.00	**0.63 ± 0.13**	0.57 ± 0.07	0.50 ± 0.00
	F1-score	0.40 ± 0.00	**0.61 ± 0.13**	0.57 ± 0.07	0.40 ± 0.00
[100, 1000)	AUROC	0.64 ± 0.10	**0.80 ± 0.07**	0.62 ± 0.11	0.51 ± 0.07
	Precision	0.59 ± 0.18	**0.77 ± 0.13**	0.74 ± 0.15	0.66 ± 0.18
	Recall	0.56 ± 0.08	**0.67 ± 0.10**	0.64 ± 0.08	0.56 ± 0.07
	F1-score	0.55 ± 0.11	**0.67 ± 0.11**	0.65 ± 0.10	0.56 ± 0.08
[0, +∞)	AUROC	0.72 ± 0.08	0.72 ± 0.06	**0.82 ± 0.07**	0.51 ± 0.05
	Precision	0.66 ± 0.09	0.75 ± 0.06	**0.76 ± 0.06**	0.53 ± 0.06
	Recall	0.66 ± 0.09	0.70 ± 0.04	**0.75 ± 0.06**	0.51 ± 0.03
	F1-score	0.66 ± 0.09	0.70 ± 0.04	**0.75 ± 0.06**	0.47 ± 0.04

4.4 Feature importance analysis and cross-country experiments

We further investigated the importance of each feature by performing a χ^2 test, with 10-fold stratified shuffle split cross validation, considering the entire range of network sizes $[0, +\infty)$. We show the Top-5 most discriminative features for each country in Table 8.

We find exactly the same set of features (with different rankings in the Top-3) in both countries; these correspond to two global network properties—LWCC, which indicates the size of the largest cascade in the layer, and SCC, which correlates with the size of the layer ($\rho \approx 0.99$, with $p \approx 0$ in all cases)—associated to the same set of layers (Q, RT and M).

We further performed a χ^2 test to highlight the most discriminative features in the M layer of both countries, which performed equally well in the classification task as previously highlighted; also in this case we focused on the entire range of network sizes $[0, +\infty)$. Interestingly, we discovered exactly the same set of Top-3 features in both countries, namely LWCC, SCC and DWCC (which indicates the depth of the largest cascade in the layer). We performed a Kolmogorov–Smirnov two-sample test to assess whether distributions of these features are statistically equivalent across the two news domains; the hypothesis was rejected in all cases at $\alpha = 0.05$.

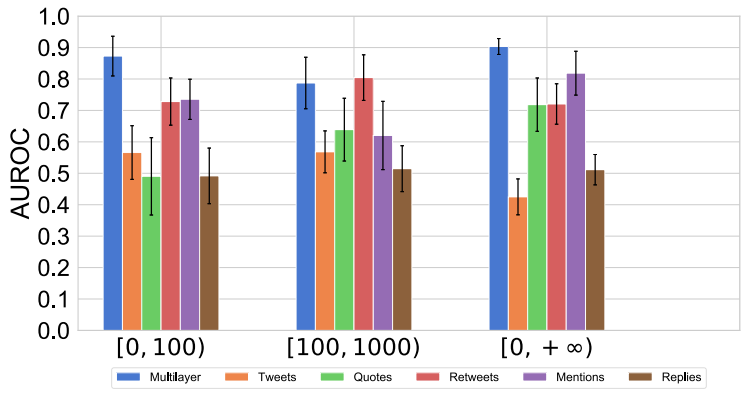

Figure 7 AUROC values for the LR classifier (evaluated on different size classes of the IT dataset) trained using different layers separately and together (our multi-layer approach). Error bars indicate the standard deviation of AUROC values over different folds of the cross validation

Table 8 Top-5 most discriminative features according to χ^2 test evaluated on both US and IT datasets (considering networks in the $[0, +\infty)$ size class)

Rank	US	IT
#1	SCC (Quotes)	LWCC (Retweets)
#2	LWCC (Retweets)	SCC (Retweets)
#3	SCC (Retweets)	SCC (Quotes)
#4	LWCC (Quotes)	LWCC (Quotes)
#5	LWCC (Mentions)	LWCC (Mentions)

The similarities evidenced so far in both countries—i.e., classification performance of single layers and features importance—might suggest that the two news domains exhibit discrepancies which are geographic-independent. We further investigated this hypothesis by testing the performance of both LR and Balanced Random Forest classifiers in several cross-country settings, e.g. training on the US dataset and testing on the Italian (and viceversa), performing feature normalization either over the entire data or separately for training and test sets, to investigate whether we can classify disinformation vs mainstream news regardless of the country where they originate. Interestingly, performance is in all cases worse or equal than those of a random classifier (AUROC = 50%); this might be due either to the high imbalance of data across the two countries, or most likely suggests that sharing patterns of the two news domains exhibit coupled dissimilarities which are very country specific.

5 Conclusions

In this work we tackled the problem of the automatic classification of news articles in two domains, namely *mainstream* and *disinformation* news, with a language-independent approach which is based solely on the diffusion of news items on Twitter social platform. We disentangled different types of interactions on Twitter to accordingly build a multi-layer representation of news diffusion networks, and we computed a set of global network properties—separately for each layer—in order to encode each network with a tuple of features. Our goal was to investigate whether a multi-layer representation performs better than an aggregated one [11], and to understand which of the features, observed at given layers, are most effective in the classification task.

Experiments with an off-the-shelf classifier such as Logistic Regression on datasets pertaining to two different media landscapes (US and Italy) yield very accurate classification results (AUROC up to 94%), also when controlling for the different political bias of news sources, which are far better than our baseline [11] with improvements up to 20%. Classification performance using individual layers shows that the layer of mentions alone entails better performance w.r.t. other layers in both countries, pointing in both cases to a peculiar usage of this type of Twitter interaction across the two domains.

We also highlighted the most discriminative features across different layers in both countries; we noticed the exact same set of features, suggesting, at first glance, that differences between the two news domains might be country-independent and rather due only to the typology of content shared. However, the two news domains exhibit coupled dissimilarities in sharing patterns which appear to be very country specific, and our methodology fails to detect disinformation regardless of where it originates.

Overall, our results prove that the topological features of multi-layer diffusion networks might be effectively exploited to detect online disinformation. Notice that we do not deny the presence of deceptive efforts to orchestrate the regular spread of information on social media via content amplification and manipulation [41, 42]. On the contrary, we postulate that such hidden forces might play to accentuate the discrepancies between the diffusion patterns of disinformation and mainstream news (and thus to make our methodology effective).

In the future we aim to further investigate two main directions: (1) employ temporal networks to represent news diffusion and apply classification techniques (e.g. recurrent neural networks) that take into account the sequential aspect of data; (2) leverage our network-based features in addition to state-of-the-art text-based approaches for "fake news" detection in order to deliver a real-world system to detect misleading and harmful information spreading on social media.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-020-00253-8.

Additional file 1. Supplementary information (ZIP 1 kB)

Acknowledgements
F.P. and S.C. are supported by the PRIN grant HOPE (FP6, Italian Ministry of Education). S.C. is partially supported by ERC Advanced Grant 693174. F.P. is grateful to Manlio De Domenico and Riccardo Gallotti for insightful discussions on multi-layer networks.

Funding
Not applicable.

Availability of data and materials
Datasets analyzed in this work are available from the corresponding author on request. In particular, we can only release the list of IDs which can be used to retrieve tweets in agreement with Twitter terms of service.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
FP designed and performed the experiments. CP and SC supervised the experiments. All authors wrote and reviewed the manuscript.

Endnotes
[a] http://www.adsnotizie.it. Accessed: April 18th, 2019.
[b] https://www.internetlivestats.com/twitter-statistics/

Publisher's Note

Received: 30 April 2020 Accepted: 12 November 2020 Published online: 23 November 2020

References

1. Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31(2):211–236
2. Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Fake news on Twitter during the 2016 U.S. presidential election. Science 363(6425):374–378
3. Bovet A, Makse HA (2019) Influence of fake news in Twitter during the 2016 US presidential election. Nat Commun 10(1):7
4. Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, Schudson M, Sloman SA, Sunstein CR, Thorson EA, Watts DJ, Zittrain JL (2018) The science of fake news. Science 359(6380):1094–1096
5. Nickerson RS (1998) Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol 2(2):175
6. Reed ES, Turiel E, Brown T (2013) Naive realism in everyday life: implications for social conflict and misunderstanding. Values Knowl 113–146
7. Shao C, Ciampaglia GL, Varol O, Yang K-C, Flammini A, Menczer F (2018) The spread of low-credibility content by social bots. Nat Commun 9(1):4787
8. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W (2016) The spreading of misinformation online. Proc Natl Acad Sci 113(3):554–559
9. Pierri F, Ceri S (2019) False news on social media: a data-driven survey. ACM Sigmod Rec 48(2)
10. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151
11. Pierri F, Piccardi C, Ceri S (2020) Topology comparison of Twitter diffusion networks effectively reveals misleading news. Sci Rep 10:1372
12. Batagelj V, Zaversnik M (2003) An O(m) algorithm for cores decomposition of networks. arXiv preprint. arXiv:cs/0310049
13. Goel S, Anderson A, Hofman J, Watts DJ (2015) The structural virality of online diffusion. Manag Sci 62(1):180–196
14. Monti F, Frasca F, Eynard D, Mannion D, Bronstein MM (2019) Fake news detection on social media using geometric deep learning. arXiv preprint. arXiv:1902.06673
15. Zhao Z, Zhao J, Sano Y, Levy O, Takayasu H, Takayasu M, Li D, Wu J, Havlin S (2020) Fake news propagates differently from real news even at early stages of spreading. EPJ Data Sci 9(1):7
16. Ratkiewicz J, Conover M, Meiss M, Gonçalves B, Patil S, Flammini A, Menczer F (2011) Detecting and tracking political abuse in social media. In: ICWSM 2011, 249
17. Davis CA, Varol O, Ferrara E, Flammini A, Menczer F (2016) Botornot: a system to evaluate social bots. In: Proceedings of the 25th international conference companion on world wide web, pp 273–274. International World Wide Web Conferences Steering Committee
18. Shao C, Ciampaglia GL, Flammini A, Menczer F (2016) Hoaxy: a platform for tracking online misinformation. In: Proceedings of the 25th international conference companion on world wide web. WWW'16 companion. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp 745–750.
19. Shao C, Hui P-M, Wang L, Jiang X, Flammini A, Menczer F, Ciampaglia GL (2018) Anatomy of an online misinformation network. PLoS ONE 13(4):1–23
20. Mitchell A, Gottfried J, Kiley J, Matsa KE (2014) Political polarization & media habits. Pew Res Centr 21
21. Barberá P, Jost JT, Nagler J, Tucker JA, Bonneau R (2015) Tweeting from left to right: is online political communication more than an echo chamber? Psychol Sci 26(10):1531–1542
22. Conover MD, Gonçalves B, Flammini A, Menczer F (2012) Partisan asymmetries in online political activity. EPJ Data Sci 1(1):6
23. Bovet A, Morone F, Makse HA (2018) Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. Sci Rep 8(1):8673
24. Vicario MD, Quattrociocchi W, Scala A, Zollo F (2019) Polarization and fake news: early warning of potential misinformation targets. ACM Trans Web 13(2):10
25. Pierri F (2020) The diffusion of mainstream and disinformation news on Twitter: the case of Italy and France. In: Companion proceedings of the web conference 2020, pp 617–622
26. Pierri F, Artoni A, Ceri S (2020) Investigating Italian disinformation spreading on Twitter in the context of 2019 European elections. PLoS ONE 15(1):e0227821
27. Nielsen RK, Newman N, Fletcher R, Kalogeropoulos A (2019) Reuters institute digital news report 2019. Report of the Reuters Institute for the Study of Journalism
28. Boyd D, Golder S, Lotan G (2010) Tweet, tweet, retweet: conversational aspects of retweeting on Twitter. In: 2010 43rd Hawaii international conference on system sciences. IEEE Press, New York, pp 1–10
29. Grabowicz PA, Ramasco JJ, Moro E, Pujol JM, Eguiluz VM (2012) Social features of online networks: the strength of intermediary ties in online social media. PLoS ONE 7(1):e29358
30. Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA (2014) Multilayer networks. J Complex Netw 2(3):203–271
31. Wasserman S, Faust K (1994) Social network analysis. Methods Appl 8
32. Saramäki J, Kivelä M, Onnela J-P, Kaski K, Kertesz J (2007) Generalizations of the clustering coefficient to weighted complex networks. Phys Rev E 75(2):027105
33. Wiener H (1947) Structural determination of paraffin boiling points. J Am Chem Soc 69(1):17–20
34. Hagberg A, Swart P, Chult D (2008) Exploring network structure, dynamics, and function using networkX. Technical report, Los Alamos National Lab. (LANL), Los Alamos, NM (United States)
35. Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from Twitter's streaming api with Twitter's firehose. In: Seventh international AAAI conference on weblogs and social media

36. Bozarth L, Budak C (2020) Toward a better performance evaluation framework for fake news classification. In: Proceedings of the international AAAI conference on web and social media, vol 14, pp 60–71
37. Zellers R, Holtzman A, Rashkin H, Bisk Y, Farhadi A, Roesner F, Choi Y (2019) Defending against neural fake news. In: Advances in neural information processing systems, pp 9051–9062
38. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al(2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
39. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27(8):861–874
40. Lemaître G, Nogueira F, Aridas CK (2017) Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 18(17):1–5
41. Stewart LG, Arif A, Starbird K (2018) Examining trolls and polarization with a retweet network. In: Proceedings ACM WSDM, workshop on misinformation and misbehavior mining on the web
42. Badawy A, Ferrara E, Lerman K (2018) Analyzing the digital traces of political manipulation: the 2016 Russian interference Twitter campaign. In: 2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE Press, New York, pp 258–265