Polarization and Fake News: Early Warning of Potential Misinformation Targets

MICHELA DEL VICARIO, IMT School for Advanced Studies Lucca, Italy
WALTER QUATTROCIOCCHI, Ca' Foscari University of Venice, Italy & ISC-CNR Sapienza University of Rome, Italy
ANTONIO SCALA, ISC-CNR Sapienza University of Rome, Italy
FABIANA ZOLLO, Ca' Foscari University of Venice & Center for the Humanities and Social Change, Italy & ISC-CNR Sapienza University of Rome, Italy

Users polarization and confirmation bias play a key role in misinformation spreading on online social media. Our aim is to use this information to determine in advance potential targets for hoaxes and fake news. In this paper, we introduce a framework for promptly identifying polarizing content on social media and, thus, "predicting" future fake news topics. We validate the performances of the proposed methodology on a massive Italian Facebook dataset, showing that we are able to identify topics that are susceptible to misinformation with 77% accuracy. Moreover, such information may be embedded as a new feature in an additional classifier able to recognize fake news with 91% accuracy. The novelty of our approach consists in taking into account a series of characteristics related to users behavior on online social media such as Facebook, making a first, important step towards the mitigation of misinformation phenomena by supporting the identification of potential misinformation targets and thus the design of tailored counter-narratives.

Additional Key Words and Phrases: social media, fake news, misinformation, polarization, classification

ACM Reference Format:
Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and Fake News: Early Warning of Potential Misinformation Targets. 1, 1 (February 2019), 22 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As of the third quarter of 2017, Facebook had 2.07 billion monthly active users [1], leading the rank of most popular social networking sites in the world. In the meantime, Oxford Dictionaries announced "post-truth" as the 2016 international Word of the Year [2]. Defined as an adjective "relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief", the term has been largely used in the context

The authors acknowledge financial support from IMT/Extrapola Srl project and P0000326 project AMOFI (Analysis and Modeling OF social medIa). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Authors' addresses: Michela Del Vicario, IMT School for Advanced Studies Lucca, Piazza S. Ponziano, 6, Lucca, 55100, Italy; Walter Quattrociocchi, Ca' Foscari University of Venice, Via Torino, 155, 30172, Venice, Italy, & ISC-CNR Sapienza University of Rome, Via dei Taurini, 19, 00185, Rome, Italy; Antonio Scala, ISC-CNR Sapienza University of Rome, Via dei Taurini, 19, 00185, Rome, Italy; Fabiana Zollo, Ca' Foscari University of Venice, Via Torino 155, 30172, Venice, & Center for the Humanities and Social Change, Dorsoduro 3246, 30123, Venice, Italy, & ISC-CNR Sapienza University of Rome, Via dei Taurini, 19, 00185, Rome, Italy.

XXXX-XXXX/2019/2-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

of Brexit and Donald Trump's election in the United States and benefited from the rise of social media as news source. Internet changed the process of knowledge production in an unexpected way. The advent of social media and microblogging platforms has revolutionized the way users access content, communicate and get informed. People can access to an unprecedented amount of information –on Facebook more than 3M posts are generated per minute [3]– without the intermediation of journalists or experts, thus actively participating in the diffusion as well as the production of content. Social media have rapidly become the main information source for many of their users: over half (51%) of US users now get news via social media [4]. However, recent studies found that confirmation bias –i.e., the human tendency to acquire information adhering to one's system of beliefs– plays a pivotal role in information cascades [5]. Selective exposure has a crucial role in content diffusion and facilitates the formation of echo chambers –groups of like-minded people who acquire, reinforce and shape their preferred narrative [6–8]. In this scenario, dissenting information usually gets ignored [9], thus the effectiveness of debunking, fact-checking and other similar solutions turns out to be strongly limited.

Since 2013 the World Economic Forum (WEF) has been placing the global danger of massive digital misinformation at the core of other technological and geopolitical risks [10]. Hence, a fundamental scientific challenge is how to support citizens in gathering trustworthy information to participate meaningfully in public debates and societal decision making. However, attention should be paid: since the problem is complex, solutions could prove to be wrong and disastrous. For instance, relying on machine learning algorithms alone to separate the truth from the false is naïf and dangerous, and might have severe consequences.

As far as we know, misinformation spreading on social media is directly related to the increasing polarization and segregation of users [5, 9, 11, 12]. Given the key role of confirmation bias in fostering polarization, our aim is to use the latter as a proxy to determine in advance the targets for hoaxes and fake news. In this paper, we introduce a framework for identifying polarizing content on social media in a timely manner –and, thus, "predicting" future fake news topics. We validate the performances of the proposed methodology on a massive Italian Facebook dataset with more than 300K news from official newspapers and 50K posts from websites disseminating either fake or unsubstantiated information. Our results show that we are able to identify topics susceptible to misinformation with 77% accuracy (0.73 AUC). Our approach may be of great importance to tackle misinformation spreading online, and could represent a key element of a system (*observatory*) to constantly monitor information flow in real time, and issue a warning about topics that require special caution. Moreover, we show that the output of our framework –i.e., whether a topic is susceptible to misinformation– may also be used as a new feature in a classifier able to recognize fake news with 91% accuracy (0.94 AUC).

Despite the goodness of our results, we are aware of the limits of this approach. Indeed, in spite of the great benefits w.r.t. pure misinformation, the identification of disinformation or propaganda has to be tackled with due caution. However, the novelty of our approach consists in also taking into account a series of characteristics related to users behavior on Facebook –e.g., in terms of interactions or sentiment– to derive novel features.

The manuscript is structured as follows: in Section 2 we provide an overview of the related work; in Section 3 we introduce our framework for the early warning, define the new features and present the classification task; in Section 4 we describe a real use-case of the framework on Facebook data; in Section 5 we show how information provided by our framework may be exploited for fake news detection and classification; finally, in Section 6 we discuss the limitations of this work and we draw our conclusions in Section 7.

2 RELATED WORK

A review of previous literature reveals a series of works aiming at detecting misinformation on Twitter, ranging from the identification of suspicious or malicious behavioral patterns by exploiting supervised learning techniques [13, 14], to automated approaches for spotting and debunking misleading content [15–17] or fake images during crisis events [18], to the assessment of the "credibility" of a given set of tweets [19, 20]. A complementary line of research addressed similar issues on other platforms, trying to identify hoax articles on Wikipedia [21], study users' commenting behavior on YouTube and Yahoo! News [22], or detect hoaxes, frauds, and deception in online documents [23]. A large body of work targeted controversy in the political domain [24, 25], and studied controversy detection using social media network structure and content [26].

Users polarization seems to inevitably dominate online debates and discussions [27–29]. Users tend to confine their attention on few sources, often determining a sharp community structure among news outlets [6, 30]. Previous work [31] showed that it is difficult to carry out an automatic classification of misinformation considering only structural properties of content propagation cascades. Thus, taking into account users behavior may prove beneficial to address the problem of misinformation on online social media.

Recent literature reviews [32–35] highlighted the need of designing and developing real-time solutions to deal with false information, and performing early fake news detection. Indeed, researchers already focused on the detection of rumors and rumor sources in networks and social media environments [36–42], however in this work we introduce a framework for a timely identification of possible, *future* fake news topics. To our knowledge, this is the first attempt in that direction, although early warning systems have already been explored for different applications, such as the problem of detecting vandals on Wikipedia [43], or the timely identification of adverse drug reactions [44]. The output information produced by our framework –i.e., whether a topic is susceptible to misinformation– may also be used as a new feature in a classifier aiming at recognizing fake news. The novelty of our contribution is two-fold:

(1) To our knowledge, this is the first work introducing a specific framework for the early warning of possible misinformation targets on social media;
(2) We introduce new features that account for how news are presented to and perceived by users on the social network.

3 EARLY WARNING: A FRAMEWORK

In this paper, we introduce a framework to promptly determine polarizing content and detect potential breeding ground (*early warning*) of either hoaxes, or fake, or unsubstantiated news. The proposed approach is meant for social media platforms such as Facebook and consists in four main phases, as shown in Fig. 1:

1. **Data collection:** First, we identify two categories of news sources: 1) official, and 2) fake –i.e., aiming at disseminating unsubstantiated or fake information. Then, for each category, we collect all data available on the platform under analysis.
2. a) **Topic extraction:** We extract the topics (*entities*) associated to the textual content in the data (e.g. articles). Entity extraction adds semantic knowledge to content to help understand the subject and context of the text that is being analyzed, allowing to identify items such as persons, places, concepts, organizations that are present in the input text. As an example, let us consider the following text: "Dijkstra's algorithmic work plays an important role in many areas of computing science". The entities associated to the text will be `{Edsger W. Dijkstra, Algorithm, Computer Science}` i.e., a person, and two concepts.

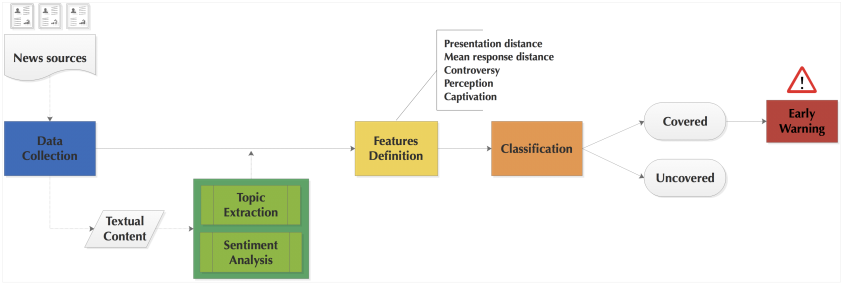

Fig. 1. Flowchart of the proposed framework.

b) **Sentiment analysis:** We associate news and users' comments in the dataset to a sentiment score ranging from *totally negative* to *absolutely positive*.

3. **Features definition:** We now use the information collected in the previous steps to derive a series of features that take into account how information is presented and perceived on the platform.
4. **Classification:** Finally, we perform the classification task using different state-of-the-art machine learning algorithms and comparing their results. Once detected the best algorithms (and the related feature sets), we are ready to classify entities and, thus, identify potential targets for fake news i.e., what we define *early warning*.

3.1 Features

In the following, we define our features for Facebook. A post is a content (e.g., text, external link, video, photo(s)) that has been shared on the platform by a user or a page. Let e be an entity –i.e., one of the main items/topics in a post. We say a user to be engaged in entity e if she/he left more than 95% of her/his comments on posts containing e. Notice that, since a post may contain several entities, then a user may be engaged with more than one entity at the same time. For example, we say Alice to be engaged with "computer science" and "philosophy" if she left more than 95% of her comments to posts that deal with the entities `Computer Science` and `Philosophy`.

We define the following features:

(1) The *presentation distance* $d_p(e)$ i.e., the absolute difference between the maximum and the minimum value of the sentiment score of all posts containing entity e[1].
(2) The *mean response distance* $d_r(e)$ i.e., the absolute difference between the mean sentiment score on the posts containing the entity and the mean sentiment score on their comments.[2]
(3) The *controversy* of the entity:

$$V(e) = \begin{cases} 0, & \text{if } d_p(e) < \delta_p \\ 1, & \text{if } d_p(e) \geq \delta_p \end{cases}$$

where δ_p is a specific threshold dependent on the data;
(4) The *perception* of the entity $P(e)$ as:

$$P(e) = \begin{cases} 0, & \text{if } d_r(e) < \delta_r \\ 1, & \text{if } d_r(e) \geq \delta_r \end{cases}$$

where δ_r is a specific threshold dependent on the data.

[1]The sentiment is computed over the textual content associated to the post containing the entity.
[2]We also consider minimum, maximum, and the standard deviation for this measure.

(5) The *captivation* of the entity $\kappa(e)$:

$$\kappa(e) = \begin{cases} 0, \text{ if } u_e < \rho_e \\ 1, \text{ if } u_e \geq \rho_e \end{cases} \quad \rho_e \in [0, 1]$$

where u_e is the fraction of users engaged in entity e and ρ_e is a threshold dependent on the data.

The controversy of an entity is directly related to how the entity is communicated to users. The greater the presentation distance, the greater is the difference between the sentiment of the posts involving such an entity. Following previous works [7, 45], the idea behind this metric is that a greater distance may be a valid indicator of the controversy of an entity, intended as the difference of opinion w.r.t. a certain topic. Similarly, the perception of an entity is defined on the basis of the sentiment that posts containing such an entity elicit in the reader. The greater the mean response distance, the greater is the difference between how the topic is presented and how it is instead perceived by users involved in the discussion around it. Finally, the captivation of an entity gives an idea of how much attention the entity receives in terms of users interaction. Summarizing, we say an entity e:

- to be controversial if $V(e) = 1$, uncontroversial otherwise;
- to arouse controversial response if $P(e) = 1$, and vice versa;
- to be captivating if $\kappa(e) = 1$, non-captivating otherwise.

Finally, let E be the set of all entities. We say entities to be *covered* if they appear in both categories of news sources –official and fake. We define such a set as $C \in E$.

To compute the features $V(e)$, $P(e)$, and $\kappa(e)$ we need to find the thresholds δ_p, δ_r, and ρ_e, which are clearly dependent on the data –and, thus, on the specific platform under analysis. To this aim, we need to define the following pairs:

- $(E_{\delta_p}, C_{\delta_p})$, where E_{δ_p} (respectively, C_{δ_p}) is the number of all (respectively, covered) entities in E for which $d_p(e) \geq \delta_p$;
- $(E_{\delta_r}, C_{\delta_r})$, where E_{δ_r} (respectively, C_{δ_r}) is the number of all (respectively, covered) entities in E for which $d_r(e) \geq \delta_r$;
- (E_{ρ_e}, C_{ρ_e}), where E_{ρ_e} (respectively, C_{ρ_e}) is the number of all (respectively, covered) entities in E for which $u_e \geq \rho_e$.

In the following, we will see that a deep analysis of such metrics allows to define the proper thresholds for one's own data. In Section 4 we discuss a real use-case of our framework and show how to select thresholds and thus define the above listed features for Facebook.

3.2 Classification

To identify topics that are potential targets for fake news, we compare the performance of several state-of-the-art classification algorithms, thus select the best ones and extract a set of features capable of ensuring a noteworthy level of accuracy. To this aim, we rely on the Python scikit-Learn package [46] and, on the basis of the most recent literature [47–54], we consider the following classifiers: Linear Regression (LIN) [55], Logistic Regression (LOG) [56], Support Vector Machine (SVM) [57] through support vector classification, K-Nearest Neighbors (KNN) [58], and Neural Network Models (NN) [59] through the Multi-layer Perceptron L-BFGS algorithm, and Decision Trees (DT) [60].

To validate the results, we split the data into training (60%) and test sets (40%) and make use of the following metrics:

— *Accuracy*, that is the fraction of correctly classified examples (both true positives T_p and true negatives T_n) among the total number of cases (N) examined:

$$Accuracy = \frac{T_p + T_n}{N}$$

— *Precision*, that is the fraction of true positives (T_p) over the number of true positives plus the number of false positives (F_p):

$$Precision = \frac{T_p}{T_p + F_p}$$

— *Recall*, that is the fraction of true positives (T_p) over the number of true positives plus the number of false negatives (F_n):

$$Recall = \frac{T_p}{T_p + F_n}$$

— *F_1-score*, that is the harmonic mean of *Precision* and *Recall* [61]:

$$F_1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

— Finally, the *False Positive* (*FP*) *Rate* (or *Inverse Recall*), that is the fraction of false positives (F_p) over the number of false positives plus the number of true negatives (T_n):

$$FP\ Rate = \frac{F_p}{F_p + T_n}$$

To evaluate and compare the classifiers we measure the accuracy of the predicted values through the *Area Under the ROC Curve* (AUC), where the *Receiver Operating Characteristic* (ROC) is the curve that plots the *Recall* against the *FP Rate* at various thresholds settings [62].

4 A REAL USE-CASE: FACEBOOK

In this section, we describe a real use-case of the proposed framework on Facebook. Our final aim is to identify the topics that are most likely to become a target for future fake news.

4.1 Data collection

We identify two main categories of Facebook pages associated to:

(1) Italian official newspapers (*official*);
(2) Italian websites that disseminate either hoaxes or unsubstantiated information or fake news (*fake*).

To produce our dataset, for set *(1)* we followed the exhaustive list provided by ADS[3] [63], while for set *(2)* we relied on the lists provided by very active Italian debunking sites [64, 65]. To validate the lists, all the pages have then been manually checked by looking at their self-description and the type of promoted content. For each page, we downloaded all the posts in the period 31.07–12.12 2016, as well as all their likes and comments. The exact breakdown of the dataset is provided in Table 1. The entire data collection process was performed exclusively by means of the Facebook Graph API [66], which is publicly available and can be used through one's personal Facebook user account. We used only public available data (users with privacy restrictions are not included in our dataset). Data was downloaded from Facebook pages that are public entities. When allowed by users' privacy specifications, we could have accessed public personal information. However, in

[3] ADS is an association for the verification of newspaper circulation in Italy. Their website provides an exhaustive list of Italian newspaper supplying documentation of their geographical and periodical diffusion.

Table 1. Breakdown of the dataset (Facebook).

	Official	Fake
Pages	58	17
Posts	333, 547	51, 535
Likes	74, 822, 459	1, 568, 379
Comments	10, 160, 830	505, 821
Shares	31, 060, 302	2, 730, 476

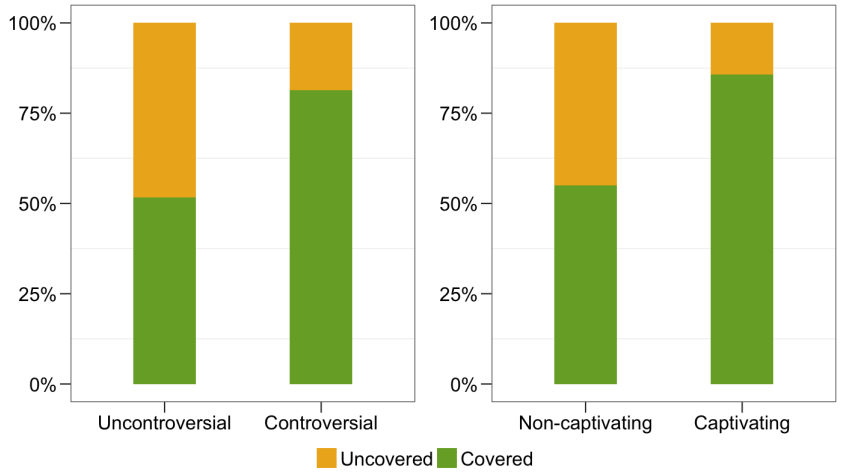

Fig. 2. Fraction of entities found in fake news (covered), for both controversial/uncontroversial and captivating/non-captivating entities.

our study we used fully anonymized and aggregated data. We abided by the terms, conditions, and privacy policies of Facebook.

4.2 Data Insights

Before going into the details of the framework application, it is worth examining how the features presented in section 3.1 may contribute to our aim. In this direction, some valuable insights can be drawn directly from the data.

Fig. 2 shows the fraction of controversial/captivating entities also found in fake news i.e., what we define as *covered*. We may notice that both controversial and captivating entities are much more present in fake news than their counterpart, thus highlighting the potential of such features in identifying topics that are likely to be subject to misinformation.

Looking at sentiment-based features, Fig. 3 shows the mean number of likes and comments received by entities for increasing levels of presentation distance. We may observe a positive relationship between the entity's presentation distance and the attention received on Facebook in terms of users activity. Indeed, entities that are presented in a different way by official and fake news sources –i.e. with higher values of δ_p– get on average a higher number of likes and comments and, hence, a higher attention.

Moving to users' response, Fig. 4 shows a series of violin plots representing the estimated probability density function of the mean response distance. The measure is computed for the

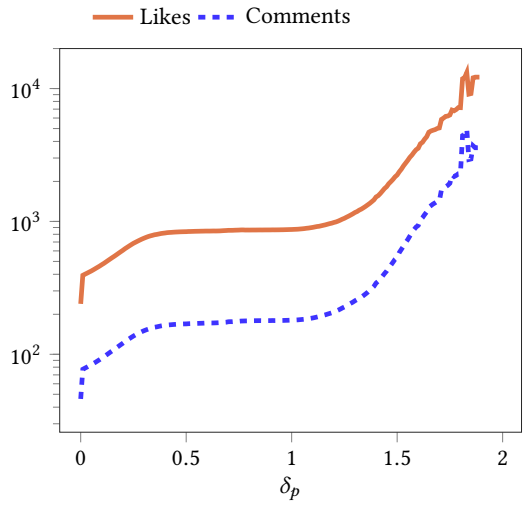

Fig. 3. Mean number of likes (solid orange) and comments (dashed blue) for entities whose presentation distance $d_p(e) \geq \delta_p$.

two classes of entities –controversial and uncontroversial– and for both covered and uncovered entities. We may notice some significant differences: distributions of controversial entities show a main peak around 0.4, which is broader in the case of covered entities, whereas distributions of uncontroversial entities are centered around smaller values and present two peaks, which are of similar size in the case of covered entities. This evidence suggests that for controversial entities users' response is usually divergent from the post presentation (mean response distance is near to 0.4), while uncontroversial ones show mixed responses, that can be either similar to (mean response distance near to 0) or slightly divergent from (mean response distance near to 0.25) how the news is presented. This may also indicate that there is a higher probability of divergent response for covered entities.

Finally, we focus on the temporal patterns associated to the appearance of a certain topic (entity) on fake news. More specifically, we compute the temporal difference, in term of hours, between the first appearance of an entity on official information and its consequent first appearance on fake news. Considering all entities appearing on official news in our dataset, about 50% is present in fake news as well, and half of these entities appear on fake news only after they do on official information. Fig. 5 shows the Probability Density Function (PDF) of the temporal distance for the following categories of entities: all, controversial (i.e. $V(e) = 1$), arousing controversial response (i.e. $P(e) = 1$), and captivating (i.e. $k(e) = 1$). We may notice a main peak at the first three hours, for all categories, followed by a smaller peak between 20 and 22. Also, we may observe that the emergence of an entity on fake news is confined to about 25 hours after its first appearance on official news.

Summarizing, such insights are interesting and in our opinion motivate the choice of the newly introduced features. Indeed, we have observed that controversial and captivating entities are more likely to be found in fake news, and that covered entities show a higher probability of divergent response. Also, presentation distance proves to be a good indicator of the attention received by the topic in terms of likes and comments. Finally, the idea of an early warning strategy to counteract

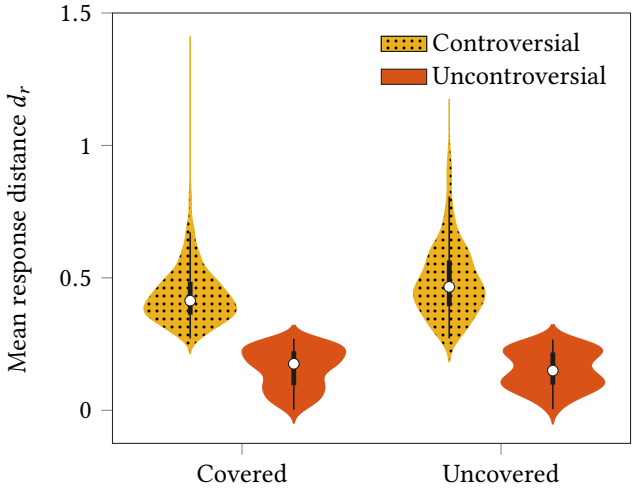

Fig. 4. Estimated probability density function of the mean response distance for controversial (dotted yellow) and uncontroversial (orange) entities by covered and uncovered entities.

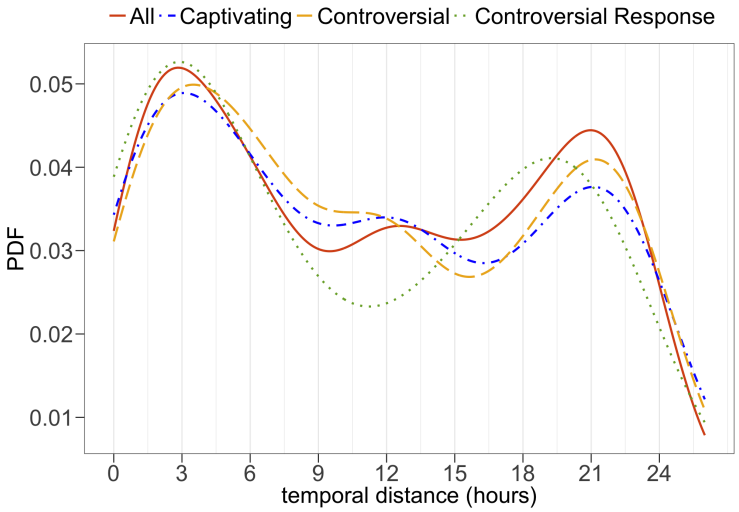

Fig. 5. Probability distribution function (PDF) of the temporal distance (in hours) between the first appearance of an entity on official news and its consequent first appearance on fake news, for the following classes of entities: all, controversial, arousing controversial response, and captivating.

to possible fake news is all the more justified by the fact that most topics become subject to misinformation within a day from their first appearance on official news.

4.3 The Framework in Action

4.3.1 Topic extraction and sentiment analysis. To perform topic extraction and sentiment analysis, we rely on Dandelion API [67], that is particularly suited for the Italian language and gets good performances on short texts as well [68]. By means of the Dandelion API service, we extract the

main entities and the sentiment score associated to each post of our dataset, whether it has a textual description or a link to an external document. Entities represent the main items in the text that, according to the service specifications, could fall in one of six categories: person, works, organizations, places, events, or concepts. Thus, for each post we get a list of entities and their related confidence level, and a sentiment score ranging from −1.0 (totally negative) to 1.0 (absolutely positive). During the analysis, we only considered entities with a confidence level greater than or equal to 0.6, hereafter referred as sample E_1. Moreover, we selected all entities with a confidence level greater than or equal to 0.9 and occurring in at least 100 posts. For these entities, hereafter referred as sample E_2, we selected all the posts where they appeared and run Dandelion API to extract the sentiment score associated to their comments. Details of both samples are shown in Table 2.

Table 2. Entities Samples.

	E_1		E_2	
	Official	Fake	Official	Fake
Entities	82, 589	19, 651	1, 170	763
Posts	121, 833	5, 995	16, 098	8, 234
Comments	6, 022, 299	135, 988	1, 241, 703	171, 062

4.3.2 Features. Following the features presented in Section 3.1, it is straightforward to compute the presentation distance $d_p(e)$ and the mean response distance $d_r(e)$ for each entity of our samples E_1 and E_2. As anticipated in Section 3.1, to calculate controversy, perception and captivation, we first need to find the proper thresholds for our data. To develop an intuition on how to determine such quantities, in the following we will analyze the behavior of the number of covered entities as a function of the thresholds δ_p, δ_r, and ρ_e. In Fig. 6 we show for both samples E_1 (left) and E_2 (right) the number of covered entities C_{δ_p} with presentation distance $\geq \delta_p$ normalized with respect to the total number of covered entities C. We observe that C_{δ_p}/C presents a plateau between two regions of monotonic decrease with respect to δ_p. This behavior indicates that entities are clearly separated in two sets and that all the entities with $d_p(e) \geq \delta_p$ are indeed controversial. Consequently, we may take the inflection point corresponding to the second change in the curve concavity as our threshold $\delta_p(e)$, since it accounts for the majority of covered entities. To do that, we fit our data to polynomial functions and compute all inflection points. We get the following thresholds for $V(e)$: $\delta_p(e) = 1.1$ for sample E_1 and $\delta_p(e) = 0.98$ for sample E_2. Notice that the height of the plateau corresponds to the size of controversial news (i.e. $d_p(e) > \delta_p$): hence, for E_1 and E_2 we have that respectively $\sim 55\%$ and $\sim 40\%$ of the covered entities are controversial.

In Fig. 6 we also show the ratio $C_{\delta_p}/E_{\delta_p}$, that measures the correlation between covered entities C and all entities E by varying presentation distances. Notice that the fact that $C_{\delta_p}/E_{\delta_p} \to 1$ when δ_p grows indicates that the all highly controversial entities are also covered entities. Like C_{δ_p}/C, also $C_{\delta_p}/E_{\delta_p}$ shows a plateau in the same region of δ_p's. We observe that the main difference between the two samples E_1 and E_2 is in the initial value of $C_{\delta_p}/E_{\delta_p}$ and the height of the plateaus. For sample E_2, the total number $C_{\delta_p=0}$ of covered entities is about $\sim 60\%$ of the total number of entities $E_{\delta_p=0}$, while for E_1 it is only $\sim 20\%$. On the same footing, we have that the different height of the plateaus indicates that while for set E_1 45% of controversial news are also covered (i.e. they belong both to the official and fake categories), for set $E2$ this ratio becomes $\sim 80\%$.

An analogous approach may be used to find the thresholds δ_r and ρ_e for features perception $P(e)$ and captivation $\kappa(e)$, respectively. Notice that such features are only applicable to sample E_2, for which the sentiment score is available for comments too.

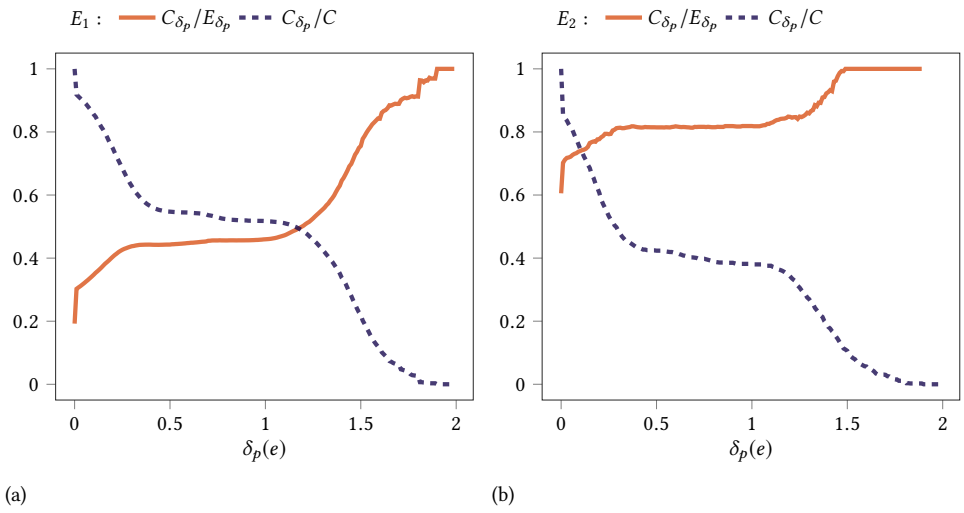

Fig. 6. We show in dashed blue C_{δ_p}/C, the fraction of covered entities with presentation distance greater than or equal to δ_p, and in solid orange $C_{\delta_p}/E_{\delta_p}$, the ratio of covered entities w.r.t. all entities with presentation distance greater than or equal to δ_p, both for sample $E1$ (a) and $E2$ (b). The plateaus in the curves indicate that entities are well separated into a set on uncontroversial (low δ_p) and a set of controversial entities (high δ_p).

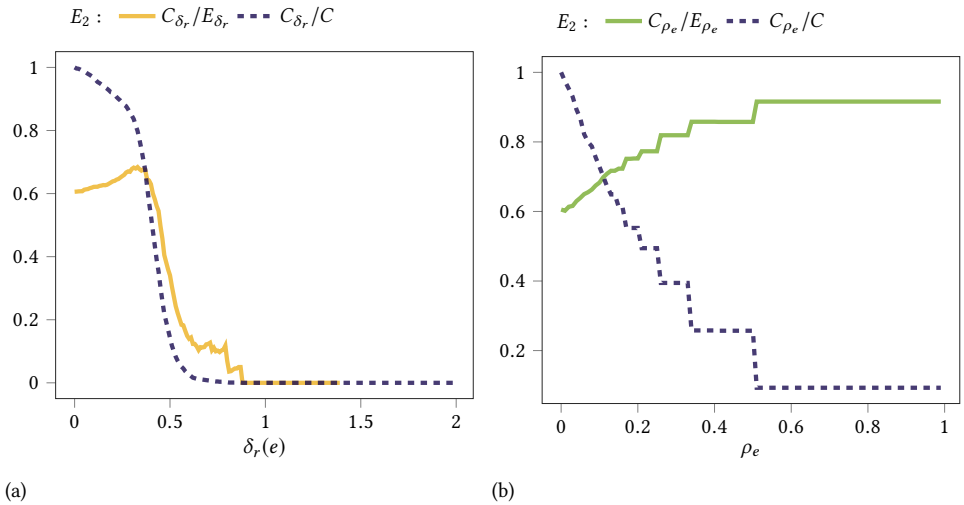

Fig. 7. Panel (a): C_{δ_r}/C (dashed blue) and $C_{\delta_r}/E_{\delta_r}$ (solid yellow). The monotonic decrease of the ratio among covered entities and entities with response distance $\geq \delta_r$ indicates that covered entities do not generate much debates and critiques from their audience. Panel (b): C_{ρ_e}/C (dashed blue) and C_{ρ_e}/E_{ρ_e} (solid green). The monotonic increase of the ratio among covered entities and entities with user share $\geq \rho_e$ indicates that most viral posts can reach up to $\sim$ 90% of the basin of possible users.

In Fig. 7 (a) we show the fraction C_{δ_r}/C of covered entities with response distance d_r greater than or equal to δ_r and the ratio $C_{\delta_r}/E_{\delta_r}$ among covered entities with $d_r \geq \delta_r$ and entities with

the same $d_r \geq \delta_r$. By definition, C_{δ_r}/C is a decreasing function of δ_r but this time it is not present an evident plateau that helps us setting a clearcut division into two distinct subsets. Also, unlike Fig. 6, the quantity $C_{\delta_r}/E_{\delta_r}$ is monotonically decreasing. Such a behavior indicates that the higher the presentation distance, the lower the probability that such post is covered, i.e. there is a higher probability to have covered entities – and hence a higher probability of dealing with fake news – when the users' response is consonant with the presentation given by news sources. The quantity $C_{\delta_r}/E_{\delta_r}$ starts from an initial value of ~ 0.6 that is followed by a sudden downfall.

In Fig. 7 (b) we show the fraction C_{ρ_e}/C of covered entities engaging a fraction of users $\geq \rho_e$ and the ratio C_{ρ_e}/E_{ρ_e} among covered entities and entities engaging a fraction of users $\geq \rho_e$. We observe that again C_{ρ_e}/C does not show a clearcut plateau. The same lack of a plateau happens for C_{ρ_e}/E_{ρ_e}, that this time is an increasing quantity. The results of Fig. 7 indicate that the higher the share of engaged users, the higher the probability that a post is debated among the two communities: in fact, we can read from the curve ($\rho_e \longrightarrow 1$) that $\sim 90\%$ of the most viral posts is covered. Notice that, since we observe a monotonic behavior, the inflection point is just a proxy of the separation point among two subsets fo entities with high and low values of the analyzed threshold parameter. By fitting our data to polynomial functions and computing all inflection points, we get the thresholds $\delta_r(e) = 0.27$ and $\rho_e = 0.42$.[4]

To recap, our features will be thus defined:

$$V(e)_{E_1} = \begin{cases} 0, & \text{if } d_p(e) < 1.1 \\ 1, & \text{if } d_p(e) \geq 1.1 \end{cases} \qquad V(e)_{E_2} = \begin{cases} 0, & \text{if } d_p(e) < 0.98 \\ 1, & \text{if } d_p(e) \geq 0.98 \end{cases}$$

$$P(e)_{E_2} = \begin{cases} 0, & \text{if } d_r(e) < 0.27 \\ 1, & \text{if } d_r(e) \geq 0.27 \end{cases} \qquad \kappa(e)_{E_2} = \begin{cases} 0, & \text{if } u_e < 0.42 \\ 1, & \text{if } u_e \geq 0.42 \end{cases}$$

4.3.3 Classification. Now that we have determined all the necessary thresholds and thus defined our features, we are ready for the last step of our framework i.e., the classification task. As an example, let us consider the following news:

Official: "This year the flu will come earlier and be worse, 6 million (people) bedridden" posted by *La Stampa* on 08/10/2016[5].

Fake: "THE FLU VACCINE? NOBODY TELLS YOU THAT IT CONTAINS 25K TIMES MORE MERCURY THAN THAT ALLOWED IN DRINKING WATER" posted by *La Gazzetta della Sera* on 17/11/2016[6].

Both news are about influenza vaccine, indeed the entity "Flu Vaccine"[7] is common to both Facebook posts. In our data, this entity is thus said to be *covered* (i.e., it appears both in the official and fake news). Moreover, it is controversial i.e., $C(e) = 1$. We will see that this information may prove helpful to timely identify topics that are potential targets for fake news. Starting from the previously mentioned features, our aim is hence to assess if an entity found on a post is likely to be soon found on fake news as well.

[4]Notice that, given the stepwise behavior of the curves in Fig. 7, the location of the inflection point depends on the degree if the polynomial. While for high degree polinomial several inflection point will be present, using low degree polynomials has a smoothing effect on the data; in particular, the inflection point chosen correspond to separating the smallest set of most disputed entities from the rest.

[5]Original text (in Italian): "Quest'anno l'influenza arriva prima e sarà più cattiva, a letto in 6 milioni".

[6]Original text (in Italian): "IL VACCINO INFLUENZALE? NESSUNO TI DICE CHE CONTIENE 25MILA VOLTE PIU' MERCURIO DI QUELLO CONSENTITO NELL'ACQUA POTABILE".

[7]Original text (in Italian): "Vaccino Antinfluenzale".

Table 3. **Early Warning.** Features.

ID	Feature name	N. of features	Sample
1	*Occurrences*	1	E_1, E_2
2	*Min/Max/Mean/Std sentiment score on posts*	4	E_1, E_2
3	*Presentation distance*	1	E_1, E_2
4	*Number of negative posts*	1	E_1, E_2
5	*Controversy*	1	E_1, E_2
6	*Min/Max/Mean/Std sentiment score on comments*	4	E_2
7	*Min/Max/Mean/Std response distance*	4	E_2
8	*Comments count*	1	E_2
9	*Number of negative comments*	1	E_2
10	*Perception*	1	E_2
11	*Captivation*	1	E_2

Table 3 provides a list of all the features employed in our classifiers. Notice that the main difference consists in the fact that sample E_2 also benefits from features involving the sentiment score of users comments (features 6–11).

As illustrated in Section 3.2, we compare the results of different classifiers: Linear Regression (LIN), Logistic Regression (LOG), Support Vector Machine (SVM) with linear kernel, K-Nearest Neighbors (KNN) with $K = 5$, Neural Network Models (NN) through the Multi-layer Perceptron L-BFGS algorithm, and Decision Trees (DT) with Gini Index. Given the asymmetry of our dataset in favor of official news sources, we re-sample the data at each step in order to get two balanced groups and avoid bias. Although an undersampling drawback is that potentially useful information is neglected, such a choice is a popular procedure when the number of observations belonging to one class is significantly lower than those belonging to the other class, and is supported by the literature [69–71].

For the sake of simplicity, in Table 4 we report the classification results only for the four best performing algorithms. We may notice an appreciable high accuracy –especially for the case of data sample E_1– and observe that all algorithms are able to accurately recognize uncovered topics, however their ability decreases in the case of covered ones. Moreover, we notice a significantly low FP Rate for the case of covered entities achieved by both LOG and NN, meaning that even though they are more difficult to detect, there is also a smaller probability to falsely label a covered entity as not, which takes on particular importance given the scope of this work. Indeed, such a result could be eventually implemented to support newsrooms by raising warnings on *sensitive* topics. By its very nature, such a system would be both *i)* time-constrained (there is little time to counteract), and *ii)* resource-constrained (human resources –e.g., journalists, communicators– are generally scarce w.r.t. information load). Thus, it is for us more important to correctly classifying covered entities (high precision) than getting them all (high recall). However, future work will consider a refinement of the classification models to increase the recall and improve the results.

Once detected the two best algorithms –i.e., LOG and NN– we use them to classify entities again. Specifically, we take the whole samples –E_1 and E_2– and we make predictions about the potentiality of each entity to become object of fake news by using either LOG or NN. We then keep the two predicted values for each entity. Looking at the AUC score, we can determine the features performance for our classifier. We use the forward stepwise features selection where, starting from an empty set of features, we iteratively add the best performing one among the unselected, when

Table 4. **Early Warning: Classification Results.** We report the performances for the four best performing algorithms (LOG, SVM, KNN, NN). The first reported value refers to E_2, while that for E_1 is in parenthesis. Values in bold denote the two best algorithms. W. Avg. denotes the weighted average across the two classes.

	AUC	Accuracy	M. Abs. Err.		Precision	Recall	FP Rate	F1-score
LOG	**0.73 (0.76)**	**0.77 (0.79)**	**0.23 (0.21)**	*Uncovered*	0.74(0.77)	0.95(0.92)	0.50(0.42)	0.83(0.84)
				Covered	0.87(0.84)	0.50(0.59)	0.05(0.08)	0.63(0.69)
				W. Avg.	0.79(0.80)	0.77(0.79)	0.28(0.25)	0.75(0.78)
SVM	0.68 (0.74)	0.71(0.80)	0.29(0.20)	*Uncovered*	0.75(0.81)	0.77(0.89)	0.38(0.32)	0.76(0.85)
				Covered	0.64(0.80)	0.62(0.67)	0.32(0.11)	0.63(0.73)
				W. Avg.	0.71(0.80)	0.71(0.80)	0.35(0.21)	0.71(0.80)
KNN	0.60 (0.75)	0.67(0.78)	0.33(0.22)	*Uncovered*	0.71(0.80)	0.77(0.84)	0.58(0.31)	0.74(0.82)
				Covered	0.60(0.74)	0.53(0.69)	0.23(0.16)	0.56(0.71)
				W. Avg.	0.66(0.78)	0.67(0.78)	0.35(0.23)	0.67(0.78)
NN	**0.68 (0.77)**	**0.72 (0.80)**	**0.28 (0.20)**	*Uncovered*	0.71(0.79)	0.92(0.92)	0.57(0.37)	0.80(0.85)
				Covered	0.77(0.84)	0.42(0.63)	0.08(0.08)	0.55(0.72)
				W. Avg.	0.73(0.81)	0.72(0.80)	0.32(0.22)	0.70(0.80)

Table 5. **Early Warning.** Features performance.

	LOG		NN	
	E_1	E_2	E_1	E_2
1	**Presentation distance**	Occurrences	**Presentation distance**	**Presentation distance**
2	**Captivation**	**Std response distance**	**Captivation**	Occurrences
3	Mean post sent. score	**Min response distance**	Min post sent. score	**Captivation**
4	**Controversy**	Std comm. sent. score	Max post sent. score	Mean comm. sent. score
5	Min post sent. score	Min comm. sent. score	Std post sent. score	**Mean response distance**
6	Max post sent. score	**Captivation**	Number negative posts	**Perception**
7	Number negative posts	Mean comm. sent. score	**Controversy**	**Max response distance**

tested together with the best ones selected so far. In other words, we begin by finding the best single feature, and gradually add the feature that improves performance most. The process ends when the AUC score remains unchanged. Table 5 reports the best features, for E_1 and E_2, for both algorithms LOG and NN. The newly introduced measures –i.e., presentation distance, response distance, controversy, perception, and captivation– are denoted in bold. We may observe that the presentation distance is the best performing feature in all cases, with the exception of the logistic regression on E_2, where instead we find the response distance to be among the first three.

5 FAKE NEWS DETECTION

As we have seen, we are now able to issue a warning about topics that require special caution. This could be a crucial element of a broader system aiming at monitoring information flow constantly and in real time. As a further step, we want to exploit the output of our framework –i.e., whether a topic would appear in fake news or not– to build a new feature set for a classification task with the aim of distinguishing fake news from reliable information. To this end, we show a possible application to Facebook data.

5.1 Instantiation to Italian News on Facebook

Let us consider two samples, P_1 and P_2, that are built from E_1 and E_2 (see Section 4.3.1) by taking all the news –i.e. the posts– containing such entities. More specifically, let $\mathcal{P}$ be the set of all posts

in our dataset. We define: $P_i = \{p \in \mathcal{P} : \exists e \in E_p | e \in E_i\}$ for $i \in \{1, 2\}$ where E_p represents the set of all entities contained in post p. As before, the main difference between the samples is the availability of sentiment scores for comments in the second one.

We apply our framework for the early warning to each sample E_1 and E_2, and consider their respective parameters and best classifiers, obtaining two separate sets of new features. In specifying such features, we account for the total number of predicted covered entities and their rate, getting a total of four features, two for any of the two adopted algorithms. Under this perspective, the particularly low FP Rate of covered entities results to be especially suited to our aim. Indeed, we may retain less information, but with a higher level of certainty. As shown in Section 4.3.3, to assess the potentiality of a topic to become object of fake news, our framework provides two predicted values for each entity, one for each classifier –LOG and NN. We now use these values to build two new features, that will be used in a new classifier –along with other features– with the aim of detecting fake news. Specifically, we may define five categories of features:

(1) *Structural features* i.e., related to news structure and diffusion;
(2) *Semantic features* i.e., related to the textual contents of the news;
(3) *User-based features* i.e., related to users' characteristics in terms of engagement and polarization;
(4) *Sentiment-based features*, which refer to both the way in which news are presented and perceived by users;
(5) *Predicted features*, obtained from our framework.

The complete list of features used for both samples is reported in Table 6. To assess the relevance and the full predictive extent of each features category, we perform two separate five-step experiments. In the first one, experiment A, we test the performance of the introduced state-of-the-art binary classifiers by first considering structural features alone and then adding only one of the features' category at each subsequent step, for both samples P_1 and P_2. While in experiment B we consider again only structural features in the first step and then we sequentially add one of the features' category at each step. Hence for each algorithm and each experiment we get ten different trained classifiers, five for P_1 and five for P_2, where the considered categories for experiment A are: 1) structural (ST), 2) structural and semantic (ST+S), 3) structural and user-based (ST+UB), 4) structural and sentiment-based (ST+SB), 5) structural and predicted features (ST+P); while for experiment B they are: 1) structural (ST), 2) structural and semantic (ST+S), 3) structural, semantic, and user-based (ST+S+UB), 4) structural, semantic, user-based, and sentiment-based (ST+S+UB+SB), 5) structural, semantic, user-based, sentiment-based, and predicted features (ST+S+UB+SB+P). Again, we apply different classifiers: Linear Regression (LIN), Logistic Regression (LOG), Support Vector Machine (SVM) with linear kernel, K-Nearest Neighbors (KNN) with $K = 5$, Neural Network Models (NN) through the Multi-layer Perceptron L-BFGS algorithm, and Decision Trees (DT) with Gini Index. Given the asymmetry of our dataset in favor of official news sources, we re-sample the data at each step in order to get two balanced groups. We compare the results by measuring the accuracy of the predicted values through the AUC score. In Fig. 8 we report the AUC values for the two five-steps experiments. We only focus on the results of the four best performing algorithms i.e., LIN, LOG, KNN, and DT. Experiment A, shows that semantic features are the ones bringing the highest improvement w.r.t. the baseline on structural features. However, we also observe remarkable improvements for the predicted features, regardless of their small number. When looking at experiment B, we observe a significant increment in the AUC during the last step (ST+S+UB+SB+P) with respect to previous ones[8], and this is especially evident for logistic regression and decision trees on P_2. Moreover, from

[8]The only exceptions are for k-nearest neighbors on P_1, where we observe the highest accuracy on step 2, and for decision trees on P_1, where instead we observe a decrement in the accuracy in the last step.

Fig. 8, we can see that logistic regression is the best performing algorithm and that it achieves especially high results on P_2. We should also notice the relative predictive power of structural features, as a matter of fact the introduction of semantic features brings, in all cases, the largest jump in the accuracy.

Table 7 reports classification results for the four top-ranked classifiers –LIN, LOG, KNN, and DT– on our two samples P_1 and P_2, considering all defined features categories. We notice an overall very high level of accuracy, where the best score is 0.91 and it is achieved by logistic regression on P_2, with respective precision rates in the detection of *fake* and *not fake* equal to 0.88 and 0.94. Our classifiers are generally more accurate in the detection of not fake information, however both false positive rates for fake and not fake are significantly low (especially in the LOG case), with a slightly smaller probability of falsely labeling a not fake as fake.

To evaluate the features performance, we compute their AUC score and then employ the forward stepwise features selection (see Section 4.3.3). As before, the process stops when the AUC score remains unchanged. Table 8 reports features performance for both samples P_1 and P_2. We may note that predicted features show good results in both cases, however structural and semantic features are the most represented. We may deduce that the newly introduced sentiment-base and predicted features are extremely relevant for the purpose of fake news identification. Moreover, the potential influential character of the commenters, embodied in the average number of comments to the commenters, is either the first or the second best performing features, underling the often neglected primary role of intermediate nodes in the diffusion of fake news.

6 LIMITATIONS

Despite the goodness of our results, we are aware that our work present some limitations. First, for the sake of simplicity we assume that all fake news, hoaxes and unsubstantiated information come from unofficial news sources. This is not necessarily true, especially when dealing with disinformation and propaganda, and not mere misinformation. On the other hand, not all news and information published by unofficial news sources are necessarily false. However, all unofficial sources in our dataset have been extensively reported as untrustworthy/unreliable by well-known Italian debunking sites [64, 65].

Second, although we have showed that controversial and captivating entities are more vulnerable to fake news, there is no guarantee that only such topics will be targeted by fake news. Moreover, a thorough evaluation of how efficient the current framework is when applied in a real time- and resource-constrained system is currently lacking, and needs to be addressed in future works.

Finally, a comparison with other datasets and/or social media platforms is desirable and would allow to prove the generality of the proposed framework. Unfortunately, despite the availability of public resources for fake news investigation [32, 72, 73], such datasets would not be suitable to the task, since related to websites (e.g., BuzzFeed or PoliFact) that do not have the same characteristics of social media. Indeed, the lack of information about users' activity would not make it possible to compute the features as defined in Section 3.1. However, our results on Facebook are promising and may pave the way to the design of new tools for the mitigation of misinformation spreading on the platform. In this direction, future work will be devoted to test and extend the proposed approach to other datasets and social media.

Table 6. **Features adopted in the classification task.** We count a total of 52 features for sample P_2 and 44 for sample P_1.

Class	Feature name	Posts	Comments	N. of features	Sample
STRUCTURAL	Number of likes/comments/shares	x	-	3	P_1, P_2
	Number of likes/comments on comments	-	x	2	P_1, P_2
	Average likes/comments on comments	-	x	2	P_1, P_2
SEMANTIC	Number of characters	x	x	2	P_1, P_2
	Number of words	x	x	2	P_1, P_2
	Number of sentences	x	x	2	P_1, P_2
	Number of capital letters	x	x	2	P_1, P_2
	Number of punctuation signs	x	x	2	P_1, P_2
	Average word length[a]	x	x	2	P_1, P_2
	Average sentence length[a]	x	x	2	P_1, P_2
	Punctuation rate[b]	x	x	2	P_1, P_2
	Capital letters rate[b]	x	x	2	P_1, P_2
USER-BASED	Av./Std comments to commenters	-	-	2	P_1, P_2
	Av./Std likes to commenters	-	-	2	P_1, P_2
	Mean std likes/comments to commenters	-	-	2	P_1, P_2
	Av./Std comments per user	-	-	2	P_1, P_2
	Av./Std pages per user	-	-	2	P_1, P_2
	Total engaged users[c]	-	-	1	P_1, P_2
	Rate of engaged users[c]	-	-	1	P_1, P_2
SENTIMENT-BASED	Sentiment score	x	-	1	P_1, P_2
	Av./Std comments' sentiment score	-	x	2	P_2
	Rate positive/negative comments	-	-	2	P_2
	Number of positive over negative comments	x	-	1	P_2
	Mean/Std presentation distance	x	-	2	P_1, P_2
	Number/Rate of captivating entities[d]	x	-	2	P_2
	Av. response distance	-	-	1	P_2
PREDICTED	Numb. of pred. D entities[e] (LOG, NN)	-	-	2	P_1, P_2
	Rate of pred. D entities[e] (LOG, NN)	-	-	2	P_1, P_2

[a] Average length computed w.r.t. both posts and comments.
[b] Over total number of characters.
[c] Users engaged with any of the entities detected in the post.
[d] Entities for which $\kappa(e) = 1$ (see Section 3.1 for details).
[e] Our framework for the early warning allows to classify an entity as covered or not. Here we consider covered entities predicted through logistic regression (LOG) and neural networks (NN), since they proved to be the best performing algorithms (see Section 4.3.3 for details).

7 CONCLUSIONS AND FUTURE WORKS

In this article, we presented a framework for a timely identification of polarizing content that enables to 1)"predict" future fake news topics on social media, and 2) build a classifier for fake news detection. We validated the performances of our methodology on a massive dataset of official news and hoaxes on Facebook. Our analysis shows that a deep understanding of users' behavior and polarization is crucial when dealing with the problem of misinformation. To our knowledge, this is the first attempt towards the early detection of possible future topics for fake news, still not without limitations –mainly due to the fact that fake or unsubstantiated information is often diffused even by official newspapers. When dealing with a complex issue such as massive digital misinformation, special caution is required. Our framework is in its early stage and there is certainly room for improvements, especially in the case of an implementation in a real-time "observatory". Future works will be devoted to refine the "prediction" task, to assess the relevance of the early warnings, and to identify what kind of entities are more prone to misinformation. However, our results are promising and bode well for a system enabled for monitoring information flow in real time and issuing a warning about delicate topics. In this direction, our approach could represent a pivotal step towards the mitigation of misinformation phenomena on Facebook.

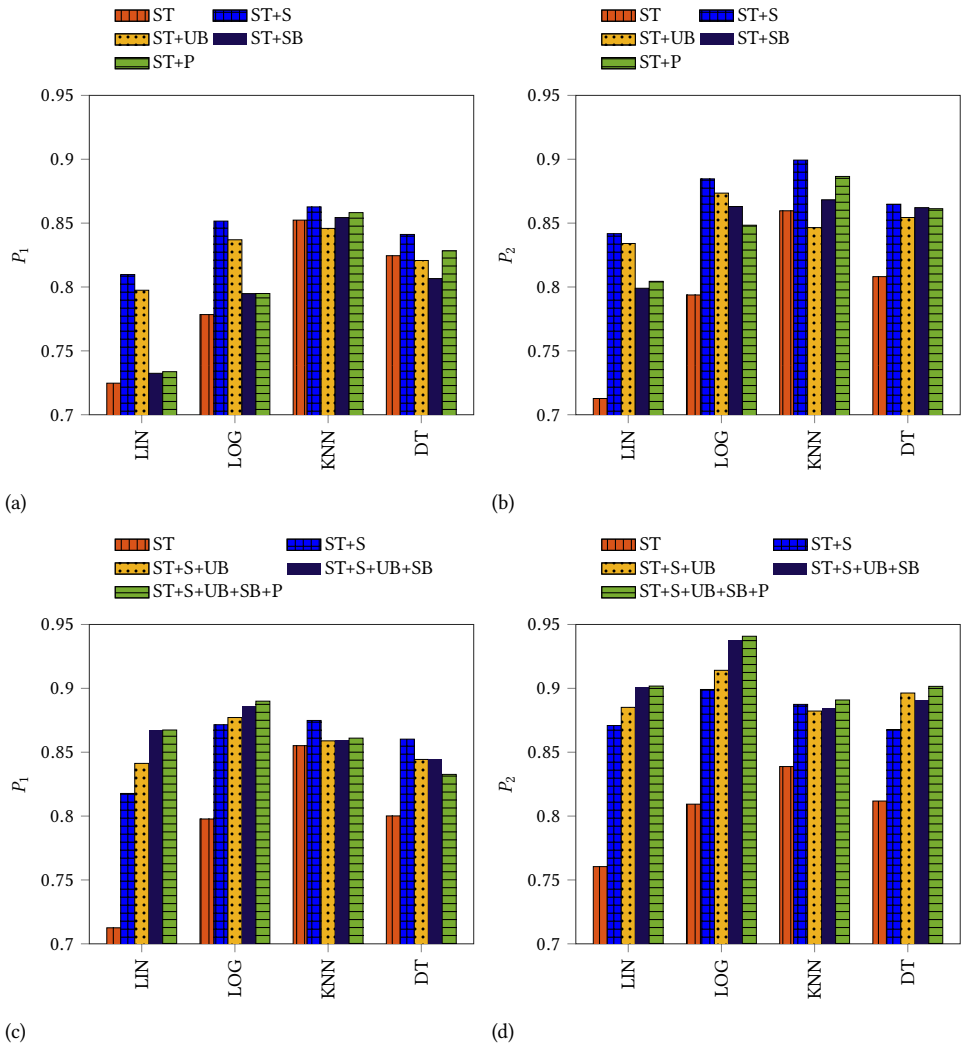

Fig. 8. AUC values for experiment A on (a) P_1 and (b) P_2 and experiment B on (c) P_1 and (d) P_2 for the following categories of features: structural (ST), semantic (S), user-based (UB), sentiment-based (SB), and predicted (P).

REFERENCES

[1] Statista, "Number of monthly active facebook users worldwide as of 3rd quarter 2017 (in millions)," Website, 2018. [Online]. Available: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

[2] O. Dictionaries, "Oxford dictionaries word of the year 2016 is...post-truth," Website, 2017. [Online]. Available: https://www.oxforddictionaries.com/press/news/2016/12/11/WOTY-16

[3] R. Allen, "What happens online in 60 seconds?" Website, 2017. [Online]. Available: https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/

[4] N. Newman, R. Fletcher, A. Kalogeropoulos, D. A. Levy, and R. K. Nielsen, "Reuters digital news report," 2017.

[5] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "The spreading of misinformation online," *Proceedings of the National Academy of Sciences*, vol. 113, no. 3, pp. 554–559, 2016.

Table 7. **Classification results.** We report the performances for the 4 best performing algorithm (LIN, LOG, KNN, DT). The first reported values refer to P_2, as in that case we have more features for the classification and we also get better results, while those for P_1 are in parentheses. W. Avg. denotes the weighted average across the two classes.

	AUC	Accuracy	M. Abs. Err.		Precision	Recall	FP Rate	F1-score
LIN	0.90 (0.87)	0.90(0.84)	0.10(0.16)	*Not Fake*	0.91(0.85)	0.90(0.84)	0.11(0.15)	0.91(0.85)
				Fake	0.88(0.84)	0.90(0.85)	0.10(0.16)	0.89(0.84)
				W. Avg.	0.90(0.84)	0.90(0.84)	0.11(0.16)	0.90(0.84)
LOG	**0.94 (0.89)**	**0.91 (0.88)**	**0.09 (0.12)**	*Not Fake*	0.94(0.90)	0.90(0.87)	0.07(0.10)	0.92(0.88)
				Fake	0.88(0.87)	0.93(0.90)	0.10(0.13)	0.90(0.88)
				W. Avg.	0.91(0.88)	0.91(0.88)	0.08(0.12)	0.91(0.88)
KNN	0.89 (0.86)	0.87(0.82)	0.13(0.18)	*Not Fake*	0.90(0.82)	0.86(0.82)	0.11(0.18)	0.88(0.82)
				Fake	0.84(0.81)	0.89(0.82)	0.14(0.18)	0.87(0.82)
				W. Avg.	0.87(0.82)	0.87(0.82)	0.13(0.18)	0.87(0.82)
DT	0.90 (0.83)	0.89(0.85)	0.11(0.15)	*Not Fake*	0.92(0.86)	0.86(0.83)	0.09(0.14)	0.89(0.84)
				Fake	0.85(0.83)	0.91(0.86)	0.14(0.17)	0.88(0.85)
				W. Avg.	0.89(0.85)	0.89(0.84)	0.12(0.16)	0.89(0.84)

Table 8. **Features performance for post classification.** For each sample P_1 and P_2, we report the feature and its respective category. Labels (p) and (c) indicate if the feature is computed w.r.t. either posts or comments.

	P_1		P_2	
	Feature	*Cat.*	*Feature*	*Cat.*
1	**Av. numb. of comments to comm.ers**	UB	Number of words (p)	S
2	**Numb. of predicted covered entities (NN)**	P	**Av. numb. of comments to comm.ers**	UB
3	Number of words (p)	S	Number of likes (p)	ST
4	Number of likes (p)	ST	Number of shares (p)	ST
5	Number of shares (p)	ST	Capital letters rate (p)	S
6	**Std numb. of likes to comm.ers**	UB	Number of comments (c)	ST
7	Number of capital letters (p)	S	Number of comments (p)	ST
8	Number of punctuation signs (p)	S	Av. comments per user	UB
9	Std numb. of comments per user	UB	Number of sentences	S
10	Std sentiment score (c)	SB	Number of punctuation signs	S
11	Number of characters (c, p)	S	**Numb. of predicted covered entities (NN)**	P
12	Av. sentence length (p)	S	**Mean presentation distance**	SB
13	**Mean presentation distance**	SB	Av. numb. of comments to comm.	ST
14	Number of comments (c)	ST	Rate of polarized users	UB
15	**Rate of predicted covered entities (LOG)**	P	**Std presentation distance**	SB
16	**Rate of predicted covered entities (NN)**	P	Av. pages per user	UB

[6] A. L. Schmidt, F. Zollo, M. Del Vicario, A. Bessi, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "Anatomy of news consumption on facebook," *Proceedings of the National Academy of Sciences*, vol. 114, no. 12, 2017.

[7] M. Del Vicario, F. Zollo, G. Caldarelli, A. Scala, and W. Quattrociocchi, "Mapping social dynamics on Facebook: The Brexit debate," *Social Networks*, vol. 50, no. Supplement C, pp. 6 – 16, 2017.

[8] A. L. Schmidt, F. Zollo, A. Scala, C. Betsch, and W. Quattrociocchi, "Polarization of the vaccination debate on facebook," *Vaccine*, vol. 36, no. 25, pp. 3606 – 3612, 2018.

[9] F. Zollo, A. Bessi, M. Del Vicario, A. Scala, G. Caldarelli, L. Shekhtman, S. Havlin, and W. Quattrociocchi, "Debunking in a world of tribes," *PLOS ONE*, vol. 12, no. 7, pp. 1–27, 07 2017.

[10] W. L. Howell, "Digital wildfires in a hyperconnected world," World Economic Forum, Tech. Rep. Global Risks, 2013.

[11] F. Zollo and W. Quattrociocchi, "Social dynamics in the age of credulity: the misinformation risk and its fallout," in *Digital Dominance. The Power of Google, Amazon, Facebook, and Apple*, M. Moore and D. Tambini, Eds. Oxford: Oxford University Press, 2018 (In Press).

[12] ——, "Misinformation spreading on facebook," in *Complex Spreading Phenomena in Social Systems*, S. Lehmann and Y.-Y. Ahn, Eds. Springer Nature, 2018.

[13] S. Antoniadis, I. Litou, and V. Kalogeraki, "A model for identifying misinformation in online social networks," in *On the Move to Meaningful Internet Systems: OTM 2015 Conferences*, C. Debruyne, H. Panetto, R. Meersman, T. Dillon, G. Weichhart, Y. An, and C. A. Ardagna, Eds. Cham: Springer International Publishing, 2015, pp. 473–482.

[14] M. Rajdev and K. Lee, "Fake and spam messages: Detecting misinformation during natural disasters on social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on*, vol. 1. IEEE, 2015, pp. 17–20.

[15] C. Boididou, S. Papadopoulos, L. Apostolidis, and Y. Kompatsiaris, "Learning to detect misleading content on twitter," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 2017, pp. 278–286.

[16] C. Boididou, S. E. Middleton, Z. Jin, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, and Y. Kompatsiaris, "Verifying information with multimedia content on twitter," *Multimedia Tools and Applications*, pp. 1–27, 2017.

[17] A.-M. Popescu and M. Pennacchiotti, "Detecting controversial events from twitter," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 1873–1876.

[18] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 729–736.

[19] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 675–684.

[20] C. Buntain and J. Golbeck, "Automatically identifying fake news in popular twitter threads," in *Smart Cloud (SmartCloud), 2017 IEEE International Conference on*. IEEE, 2017, pp. 208–215.

[21] S. Kumar, R. West, and J. Leskovec, "Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 591–602.

[22] S. Siersdorfer, S. Chelaru, J. S. Pedro, I. S. Altingovde, and W. Nejdl, "Analyzing and mining comments and comment ratings on the social web," *ACM Transactions on the Web (TWEB)*, vol. 8, no. 3, p. 17, 2014.

[23] S. Afroz, M. Brennan, and R. Greenstadt, "Detecting hoaxes, frauds, and deception in writing style online," in *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, 2012, pp. 461–475.

[24] Y. Mejova, A. X. Zhang, N. Diakopoulos, and C. Castillo, "Controversy and sentiment in online news," *arXiv preprint arXiv:1409.8152*, 2014.

[25] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*. ACM, 2005, pp. 36–43.

[26] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Quantifying controversy in social media," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 2016, pp. 33–42.

[27] A. Guess, B. Nyhan, and J. Reifler, "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign," 2018.

[28] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg, "Structural diversity in social contagion," *Proceedings of the National Academy of Sciences*, vol. 109, no. 16, pp. 5962–5966, 2012.

[29] P. H. C. Guerra, W. Meira Jr, C. Cardie, and R. Kleinberg, "A measure of polarization on social media networks based on community boundaries." in *ICWSM*, 2013.

[30] A. L. Schmidt, F. Zollo, A. Scala, and W. Quattrociocchi, "Polarization rank: A study on european news consumption on facebook," *arXiv preprint arXiv:1805.08030*, 2018.

[31] M. Conti, D. Lain, R. Lazzeretti, G. Lovisotto, and W. Quattrociocchi, "It's always april fools' day! on the difficulty of social network misinformation classification via propagation features," *arXiv preprint arXiv:1701.04221*, 2017.

[32] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[33] S. Zannettou, M. Sirivianos, J. Blackburn, and N. Kourtellis, "The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans," *arXiv preprint arXiv:1804.03461*, 2018.

[34] S. Kumar and N. Shah, "False information on web and social media: A survey," *arXiv preprint arXiv:1804.08559*, 2018.

[35] S. Kumar, M. Jiang, T. Jung, R. J. Luo, and J. Leskovec, "Mis2: Misinformation and misbehavior mining on the web," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 799–800.

[36] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, and M. Gomez-Rodriguez, "Leveraging the crowd to detect and reduce the spread of fake news and misinformation," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 324–332.

[37] N. Karamchandani and M. Franceschetti, "Rumor source detection under probabilistic sampling," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 2184–2188.
[38] S. Kwon, M. Cha, and K. Jung, "Rumor detection over varying time windows," *PloS one*, vol. 12, no. 1, p. e0168344, 2017.
[39] Z. Wang, W. Dong, W. Zhang, and C. W. Tan, "Rumor source detection with multiple observations: Fundamental limits and algorithms," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 42, no. 1. ACM, 2014, pp. 1–13.
[40] S. Spencer and R. Srikant, "Maximum likelihood rumor source detection in a star network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2199–2203.
[41] Y. Liu and S. Xu, "Detecting rumors through modeling information propagation networks in a social media environment," *IEEE Transactions on Computational Social Systems*, vol. 3, no. 2, pp. 46–62, 2016.
[42] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 1395–1405.
[43] S. Kumar, F. Spezzano, and V. Subrahmanian, "Vews: A wikipedia vandal early warning system," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2015, pp. 607–616.
[44] M. Yang, M. Kiang, and W. Shang, "Filtering big data from social media–building an early warning system for adverse drug reactions," *Journal of biomedical informatics*, vol. 54, pp. 230–240, 2015.
[45] M. Del Vicario, S. Gaito, W. Quattrociocchi, M. Zignani, and F. Zollo, "News consumption during the Italian Referendum: A cross-platform analysis on Facebook and Twitter," in *Proceedings of the 4th IEEE International Conference on Data Science and Advanced Analytics*. IEEE, 2017.
[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[47] I. Alsmadi and G. K. Hoon, "Term weighting scheme for short-text classification: Twitter corpuses," *Neural Computing and Applications*, pp. 1–13, 2018.
[48] S. A. Özel, E. Saraç, S. Akdemir, and H. Aksu, "Detection of cyberbullying on social media messages in turkish," in *International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2017, pp. 366–370.
[49] A. Khatua and A. Khatua, "Cricket world cup 2015: Predicting user's orientation through mix tweets on twitter platform," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 2017, pp. 948–951.
[50] D. Antonakaki, I. Polakis, E. Athanasopoulos, S. Ioannidis, and P. Fragopoulou, "Exploiting abused trending topics to identify spam campaigns in twitter," *Social Network Analysis and Mining*, vol. 6, no. 1, p. 48, 2016.
[51] J. Hemsley, S. Tanupabrungsun, and B. Semaan, "Call to retweet: Negotiated diffusion of strategic political messages," in *Proceedings of the 8th International Conference on Social Media & Society*. ACM, 2017, p. 9.
[52] W. van Zoonen and G. Toni, "Social media research: The application of supervised machine learning in organizational communication research." *Computers in Human Behavior*, vol. 63, pp. 132–141, 2016.
[53] S. Vosoughi and D. Roy, "Tweet acts: A speech act classifier for twitter." in *ICWSM*, 2016, pp. 711–715.
[54] C.-C. Chang, S.-I. Chiu, and K.-W. Hsu, "Predicting political affiliation of posts on facebook," in *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*. ACM, 2017, p. 57.
[55] R. M. Rifkin and R. A. Lippert, "Notes on regularized least squares," 2007.
[56] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 1-2, pp. 83–112, 2017.
[57] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
[58] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.
[59] D. E. Rumelhart, G. E. Hinton, R. J. Williams *et al.*, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, p. 1, 1988.
[60] L. Breiman, *Classification and regression trees*. Routledge, 2017.
[61] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
[62] C. E. Metz, "Basic principles of roc analysis," in *Seminars in nuclear medicine*, vol. 8, no. 4. Elsevier, 1978, pp. 283–298.
[63] ADS, "Elenchi testate," Website, 2016. [Online]. Available: http://www.adsnotizie.it/_testate.asp
[64] Bufale.net, "The black list: la lista nera del web," Website, 2016, last checked: 27.11.2017. [Online]. Available: http://www.adsnotizie.it/_testate.asp
[65] BUTAC, "The black list," Website, 2016, last checked: 27.11.2017. [Online]. Available: http://www.butac.it/the-black-list/
[66] Facebook, "Using the Graph API," Website, 8 2013, last checked: 27.11.2017. [Online]. Available: https://developers.facebook.com/docs/graph-api/using-graph-api/
[67] SpazioDati, "Dandelion API," Website, 8 2017. [Online]. Available: https://dandelion.eu/docs/

[68] R. F. Canales and E. C. Murillo, "Evaluation of entity recognition algorithms in short texts," *CLEI ELECTRONIC JOURNAL*, vol. 20, no. 1, 2017.
[69] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, 2009.
[70] C. Drummond, R. C. Holte *et al.*, "C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," in *Workshop on learning from imbalanced datasets II*, vol. 11. Citeseer, 2003, pp. 1–8.
[71] N. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets. sigkdd explor newsl 6: 1–6," 2004.
[72] K. Shu, S. Wang, and H. Liu, "Exploiting tri-relationship for fake news detection," *arXiv preprint arXiv:1712.07709*, 2017.
[73] K. Demidova, "Getting real about fake news," GitHub, 2016. [Online]. Available: https://github.com/demidovakatya/competitions/blob/master/fake-news/README.md