Classification of unlabeled online media

Sakthi Kumar Arul Prakash

Carnegie Mellon University

Conrad Tucker (✉ conradt@andrew.cmu.edu)

Carnegie Mellon University

Research Article

Keywords: entropy, media likes, fake vs. authentic, humans or machine learning algorithms

Posted Date: December 2nd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-107002/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Scientific Reports on March 25th, 2021. See the published version at <https://doi.org/10.1038/s41598-021-85608-5>.

Classification of unlabeled online media

Sakthi Kumar Arul Prakash¹ and Conrad Tucker^{1,2,*}

¹Department of Mechanical Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA

²Department of Machine Learning, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA

*conradt@andrew.cmu.edu

ABSTRACT

This work investigates the ability to classify misinformation in online social media networks in a manner that avoids the need for ground truth labels. Rather than approach the classification problem as a task for humans or machine learning algorithms, this work leverages user-user and user-media (i.e., media likes) interactions to infer the type of information (fake vs. authentic) being spread, without needing to know the actual details of the information itself. To study the inception and evolution of user-user and user-media interactions over time, we create an experimental platform that mimics the functionality of real world social media networks. We develop a graphical model that considers the evolution of this network topology to model the uncertainty (entropy) propagation when fake and authentic media disseminates across the network. The creation of a real-world social media network enables a wide range of hypotheses to be tested pertaining to users, their interactions with other users, and with media content. The discovery that the entropy of user-user, and user-media interactions approximates fake and authentic media likes, enables us to classify fake media in an unsupervised learning manner.

Introduction

The idea that people may not be suitable to assess the authenticity of information without the aid of additional tools is widely explored by the forensics community^{1,2}. Information sharing in social media is highly dependent on data modality, the behavior of fellow networkers, the habit of using online services, and the widespread presence of opinion leaders in posts and comments^{3,4}. In some cases, information is doctored using *deep fakes*⁵, or manipulated by spreading misinformation relating to health outbreaks such as COVID-19^{6,7}. Beyond affecting the conduit of social media networks, much of algorithmic fake news detection^{2,8} or recommender systems⁹ rely on data type identification and data labeling. Such information is typically manually compiled by content moderators in social media companies who may undergo post-traumatic stress disorder as a result of viewing disturbing media^{10,11}.

In this study, we consider i) work that seeks to understand why misinformation spreads and ii) work that seeks to prevent the spread of misinformation through classification. Current literature has improved our understanding of truth and falsity diffusion through quantitative analyses on the spread of rumors such as the discovery of the Higgs boson¹², and the spread of anti-vaccine campaigns that led to decreases in vaccination rates against measles^{13,14}. Other studies have investigated modeling rumor diffusion¹⁵, performing credibility evaluation^{16,17} and intervening to curtail the spread of rumors¹⁸. Similar analyses have been conducted in social media networks such as Facebook, where Del Vicario et al.¹⁹ analyze the diffusion of science and conspiracy-theory stories, and in Twitter where Bovet and Kakse²⁰ study the influence of fake news during the 2016 presidential election, and Vosoughi et al.²¹ study the characteristics of fact-checked rumors.

The spread of information, authentic or fake, is however significantly impacted by the credibility of the users in the network and their connections²². In a social media network such as Twitter, users form opinions and beliefs²³ based on outside influences^{20,24}, and use these opinions to test if information adheres to or opposes their beliefs²⁵. Though the most direct measure of user credibility involves asking the other users²⁶, such a measure is often subjective. Additionally, a subjective measure does not apply to sources or users previously unknown to a user²⁷. However, such occasions are likely to occur during times of disasters such as COVID-19⁷, or during measles outbreaks^{13,14} where authentic information was provided by those physically attending the event²⁷. Hence, we are interested in understanding how fake information spreads in a controlled social media network where users respond to unknown users with the help of an objective credibility measure. In this study, we introduce a credibility measure and show how it is proportionally related to a user's interaction with other users and media.

Node-edge homogeneity and the homogeneous sharing paths created by social media users during the evolution of a social media network, play a critical role in the spread of fake information^{19,28}. When presented with uncertainty, human behaviour in social media networks tends to react differently, employing unconventional social heuristics²⁹. The use of information-theoretic measures such as entropy, avoids making assumptions about human behavior, thus allowing statistical characterization of

entropy³⁰ in social media networks. We develop a metric which captures user interactions in an online social media network such that entropy within interactions captures the spread of authentic and fake media.

Data-driven supervised methods benefit from superior classification accuracy due to the availability of labels but may fail to generalize to unseen data and often rely on manual human labeling. Alternatively, unsupervised methods do not rely on labels to classify data and address the trade-off in accuracy with the help of additional features, feature representation techniques, or complex models. Though supervised approaches such as neural networks, learn clusters from encoded media representations, they require fine-tuning models as new media is created and shared in social media networks³¹. Furthermore these methods may be constrained by storage limitations based on media quality. Although one can argue that media can be stored in a compressed format as a work around to storage limitation, it has been reported by Rossler et al.³² that media compression of raw data leads to a loss in classification accuracy as much as 19%. Thus, there is a necessity to shift towards unsupervised approaches for content discovery or fake media discovery^{33–35}.

The paper is presented as follows. First, we annotate the controlled experiment conducted to collect real-time social network data. We then define media node-edge homogeneity, followed by a comprehensive analysis of how entropy explains the spread of fake information. Finally, we propose an unsupervised model that uses entropy to select users to classify information as fake or authentic.

Methods

Data Collection

We conduct a real-life social experiment, wherein participants are recruited from the crowd-sourcing platform *Amazon MTurk* (AMT) to participate in the study. The study was conducted with the approval of and in accordance with the relevant guidelines and regulations of The Pennsylvania State University’s *Institutional Review Board* (IRB). Prior to conduct of the study, an informed consent was obtained from all participants for participation in the study. Only human subjects of age 18 or above were included in the study. We created a mock online social media platform that is similar to Twitter and follows their directional model of interactions. We restrict users from posting their own media and from unfollowing a followed user to conduct a controlled real-time, real-world social experiment. This web application is released for public use to replicate or conduct similar online social experiments. The focus of the data collection is to organically acquire the interactions between unknown users rather than from known users within the network in order to gain a fundamental understanding of how fake information can be classified using random user opinions. We populate the social network with *authentic* and *fake* videos from the *FaceForensics* data set³² and verified fake YouTube videos. The users were monetarily incentivized to spend at least 45 minutes per Human Intelligence Task (HIT). A user was allowed to participate in any number of HITs over the period of study, ensuring that the same user population could log into the network and continue accruing credibility and followers. We chose to use 40 random videos, with 20 being authentic and the other 20 fake. The total number of enrolled users in the study was 620, which falls within the range of participants recruited for other social science experiments that utilize AMT^{36,37}.

To simulate a real-world scenario in Twitter where users follow users or are followed by users thereby forming explicit connections³⁸, we introduce the notion of user credibility. According to Resnick et al.³⁹, reputation mechanisms provide an incentive for honest behavior and help users make decisions involving trust. In this study, we develop an approach to compute user credibility using trust modeling. Specifically, we use the Beta Reputation System (BRS)⁴⁰ used in e-commerce to calculate the credibility of users. We select this method as it is grounded in probability theory and uses expectations on probability distributions to derive the predicted benefit of trust to help users form explicit connections in a participatory media setting. All users start with a baseline credibility score in order to avoid biases. We use a Bayesian credibility update rule (for further details see [Supplementary Information, section 2](#)) to update the credibility of users based on the number of authentic and fake media liked. Liking authentic media increases the credibility score while liking fake media decreases the credibility score. We also provide a monetary incentive to the top three users who earn the highest credibility score and get the highest user following by the end of the two-day experiment (for further details see [Supplementary Information, section 1](#)).

Preliminaries and Definitions

The social network under study is a directed network having links specifying the followers (in-degree) and the following (out-degree). A social network $\mathcal{G}(t) := \{\mathcal{V}, \mathcal{E}\}$ is an order 2 Tensor such that $\mathcal{G} \in \mathbb{R}^{N \times N}$ where N represents the number of users or $\mathcal{V}$ such that $\mathcal{V} \in \mathbb{R}^N$, $\mathcal{E} := \{(i, j) | i, j \in \mathcal{V}, i \neq j, \mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}\}$ is the set of all pairs of distinct nodes, called edges and $t \in T$ denotes timestamp index when all network information is saved in the database. Let the degree matrix $\mathbf{D}$ of $\mathcal{G} = \text{diag}(d(1), \dots, d(N))$ and $\mathbf{A}$ denote the adjacency matrix such that $A_{ij} = 1$ if and only if i follows j in case of out-degree network ($\mathcal{G}_{out}$) or j is followed by i in case of in-degree network ($\mathcal{G}_{in}$). The media liked by node i is denoted as $\mathbf{Z}_i$ which is a one-hot vector, where $\mathbf{Z}_i = [Z_1, Z_2, \dots, Z_L]$ and Z_l denotes the user i liking category l such that $Z_l = 0$ indicates no like and $Z_l = 1$ indicates media like.

Relating entropy to user credibility

We assume that users recruited in our social network are not known to each other and by showing every user the credibility of all users, we assume that users tend to follow or be followed by other users on the basis of credibility. We observe a positive linear correlation between the number of links connected to a user (in-degree) and credibility, consistent with findings from other works²⁷ (for further details see [Supplementary Information, section 2](#)). However, in reality, such a credibility measure is often network-dependent, or may not be readily available. Hence, we use links and media likes to approximate user credibility. We define a similarity metric, *Media Node-Edge Homogeneity*, which computes similarity as a function of common media liked by any two users, which have an edge connection at time t . We denote the media node-edge homogeneity (MNEH) metric as M_{ij}^t , where,

$$M_{ij}^t = \frac{|\mathbf{Z}_i^t \cap \mathbf{Z}_j^t|}{|\mathbf{Z}_i^t \cup \mathbf{Z}_j^t|} \quad (1)$$

In Eqn. 1, M_{ij}^t between any two users i and j at time (t) , is bounded between the interval $[0, 1]$, where, $0 \leq M_{ij} \leq 1$ expresses the magnitude of similarity in media *likes* between any two users i and j . The node-edge homogeneity interval $[0, 1]$ is discretized with a resolution of 0.01. The MNEH matrix ($\mathbf{M}^t$) is a sum of matrices as shown in Eqn. 2 with varying γ , such that $\tilde{\gamma} = \gamma/100$, $M_{ij,\gamma}^t = 1$ if and only if the similarity in media liked by users i and j is $\gamma\%$.

$$\mathbf{M}^t = \sum_{\gamma=1}^{100} \tilde{\gamma} \mathbf{M}_{\gamma}^t \quad (2)$$

Given $\mathbf{A}^t$ and $\mathbf{M}^t$ from the in-degree and out-degree networks, we assess the Directed Acyclic Graph (DAG) connecting i , j and $\mathbf{M}^t$. We assume liking media and forming links with other users to be conditionally independent events since $j = 1$ is observed thereby breaking the dependence between the events. Hence, we compute the probability of a link between user i and j for each instant in time t as shown in Eqn. 3.

$$\begin{aligned} p(j=1|\mathbf{M}_{\tilde{\gamma}}^t, i) &= p(j=1|i) \cdot p(j=1|\mathbf{M}_{\tilde{\gamma}}^t) \\ &= (\mathbf{A}\mathbf{D}^{-1})_{ij}^t \cdot \frac{p(j=1, \mathbf{M}_{\tilde{\gamma}}^t)}{p(\mathbf{M}_{\tilde{\gamma}}^t)} \\ &= (\mathbf{A}\mathbf{D}^{-1})_{ij}^t \cdot (\mathbf{M}_{\tilde{\gamma}} \mathbf{D}_{\mathbf{M}_{\tilde{\gamma}}}^{-1})_{ij}^t \\ &= ((\mathbf{A} \circ \mathbf{M}_{\tilde{\gamma}}) \mathbf{D}_{\mathbf{A}\mathbf{M}_{\tilde{\gamma}}}^{-1})_{ij}^t \\ &= P_{ij}^t \cdot P_{ij,\tilde{\gamma}}^t \end{aligned} \quad (3)$$

where $p(j=1|i) = P_{ij}^t$ is a conditional probability which represents the probability of a link formed with user j given user i at time t such that P_{ij}^t is an element of the transition matrix ($\mathbf{P}^t$) which is algebraically computed as $(\mathbf{A}\mathbf{D}^{-1})^t$, and $\frac{p(j=1, \mathbf{M}_{\tilde{\gamma}}^t)}{p(\mathbf{M}_{\tilde{\gamma}}^t)} = P_{ij,\tilde{\gamma}}^t$ is computed as follows,

$$\begin{aligned} \frac{p(j=1, \mathbf{M}_{\tilde{\gamma}}^t)}{p(\mathbf{M}_{\tilde{\gamma}}^t)} &= \frac{1 \{M_{ij,\tilde{\gamma}}^t = 1\}}{\sum_{j \in N} 1 \{M_{ij,\tilde{\gamma}}^t = 1\}} \\ &= (\mathbf{M}_{\tilde{\gamma}} \mathbf{D}_{\mathbf{M}_{\tilde{\gamma}}}^{-1})_{ij}^t \\ &= P_{ij,\tilde{\gamma}}^t \end{aligned} \quad (4)$$

where $\mathbf{D}_{\mathbf{A}\mathbf{M}_{\tilde{\gamma}}}$ is a diagonal matrix such that $D_{\mathbf{A}\mathbf{M}_{\tilde{\gamma}},ii} = \sum_{j \in N} (\mathbf{A} \circ \mathbf{M}_{\tilde{\gamma}})_{ij,\tilde{\gamma}}$, $\mathbf{D}_{\mathbf{M}_{\tilde{\gamma}}}$ is a diagonal matrix such that $D_{\mathbf{M}_{\tilde{\gamma}},ii} = \sum_{j \in N} M_{ij,\tilde{\gamma}}$ and $(\circ)$ denotes Hadamard product operation between matrices. For each media category $l \in L$, Z_{jl}^t is generated from a Bernoulli distribution with parameter C_j^t such that $Z_{jl}^t \sim \text{Bernoulli}(C_j^t)$ and C_j^t is the probability of user j liking l . Hence, we assume the prior probability of C_j^t to also be the assumed credibility of a user j and sampled from a Beta distribution with parameters a^t (prior authentic media category counts) and b^t (prior fake media category counts) such that $C_j^t \sim \text{Beta}(a^t, b^t)$. In a similar manner as Eqn. 3, we compute the probability of a link between user i and j given the media matrix ($\mathbf{Z}_j^t$) for each

instant in time t as shown in Eqn. 5

$$\begin{aligned}
p(j=1|\mathbf{Z}_j^t, i) &= p(j=1|i) \cdot p(j=1|\mathbf{Z}_j^t) \\
&= P_{ij}^t \cdot \int p(j=1, C_j^t | \mathbf{Z}_j^t) dC_j^t \\
&= P_{ij}^t \cdot \int p(j=1|C_j^t) \cdot p(C_j^t | \mathbf{Z}_j^t) dC_j^t \\
&= \frac{P_{ij}^t}{\mathbf{B}(a^t, b^t)} \int C_j^{Z_{jl}^t + a^t - 1} \cdot (1 - C_j^t)^{(1 - Z_{jl}^t) + b^t - 1} dC_j^t \\
&= P_{ij}^t \cdot \frac{\mathbf{B}(Z_{jl}^t + a^t, Z_{jl}^t + b^t)}{\mathbf{B}(a^t, b^t)} \\
&= P_{ij}^t \cdot \left(\frac{a^t}{a^t + b^t} \right) \\
&= P_{ij}^t \cdot C_j^t
\end{aligned} \tag{5}$$

where, $\mathbf{B}(\cdot)$ is the Beta function and normalizes the expression in Eqn. 5. Let $p(j=1|\mathbf{Z}_j^t)$ be denoted by $\Omega(Z)$ and $p(j=1|\mathbf{M}_j^t)$ be denoted by $\mathcal{F}(\Omega(Z))$ since $\mathbf{M}_j^t$ is a function of $\mathbf{Z}_j^t$. Using the results from Eqn. 3 and 5, we derive the proportionality between C_j^t and $P_{ij, \tilde{\gamma}}^t$ given $\tilde{\gamma}$ for every timestamp as shown by Eqn. 6.

$$\begin{aligned}
\frac{p(j=1|i) \cdot p(j=1|\mathbf{Z}_j^t)}{C_j^t} &= \frac{p(j=1|i) \cdot p(j=1|\mathbf{M}_{\tilde{\gamma}}^t)}{P_{ij, \tilde{\gamma}}^t} \\
\frac{p(j=1|i) \cdot \Omega(Z)}{C_j^t} &= \frac{p(j=1|i) \cdot \mathcal{F}(\Omega(Z))}{P_{ij, \tilde{\gamma}}^t} \\
P_{ij, \tilde{\gamma}}^t &\propto C_j^t
\end{aligned} \tag{6}$$

Hence, from Eqn. 3, it is evident that we are indeed computing a weighted transition probability with the weight being proportional to the user's credibility. Using these results, we now compute the Shannon's Information Entropy (joint entropy) generated by all users $(i, j, k, \dots, n)$ as the sum of entropy generated by each user at time t .

$$H_{i,j,k,\dots,n,\tilde{\gamma}}^t = - \sum_i^N \sum_j^N (P_{ij}^t \cdot P_{ij, \tilde{\gamma}}^t) \log_2(P_{ij}^t \cdot P_{ij, \tilde{\gamma}}^t) \tag{7}$$

Results and Discussion

Objective

The objective of this study is to use the entropy response of media node-edge homogeneity over time to perform unsupervised classification of fake media. The overview of our fake media classification model is shown in Fig. 1. We have organized this section to discuss hypotheses, hypotheses testing, and how the tests inform the selection of users that separate fake and authentic media in a majority voting classifier to classify fake media.

Entropy of media node-edge homogeneity

We compute the joint information entropy of the matrix for each value of $\tilde{\gamma}$ and for each t using Eqn. 7. The social media network is directional as it has an in-degree or followee network and an out-degree or follower network. Hence, we hypothesize that the entropy generated by a random variable ($\mathbf{X}$) representing social network users in the in-degree and out-degree network is sampled from the same underlying node-edge distribution, formally stated as follows:

$$H_0: H(\mathbf{X})_{in} = H(\mathbf{X})_{out}.$$

$$H_a: H(\mathbf{X})_{in} \neq H(\mathbf{X})_{out}.$$

where, $H(\mathbf{X})_{in}$ and $H(\mathbf{X})_{out}$ denotes entropy distribution in in-degree and out-degree network respectively. We use H_0 and H_a to denote the null hypothesis and alternate hypothesis following conventional notation. We perform a Kolmogorov–Smirnov (KS) test between the in-degree entropy distribution and out-degree entropy distribution to test our hypothesis. For all hypothesis tests, we use a significance level (α) of 0.05. We find no statistically significant difference between the in-degree entropy distribution and out-degree entropy distribution with a KS statistic = 0.09 and p value = 0.81, thereby failing to reject the null hypothesis. Fig 2 illustrates entropy response and media likes distribution against varying media node-edge homogeneity for the in-degree network across all timestamps such that each timestamp is overlaid against the other (for further details see [Supplementary](#)

Figure 1. A graphical overview of our model. Entropy is computed for target nodes. Homogeneous paths connecting adjacent users are denoted by thick lines.

Figure 2. Normalized entropy response and normalized likes distribution against media node-edge homogeneity in the in-degree network

Information, section 2 for a three dimensional illustration). The comparison draws attention to the correlation between media distribution and entropy distribution in the in-degree network. To check for correlation, we compute the Pearson's correlation coefficient between $\mathbf{M}^t$ and $\mathbf{H}^t$, where $\mathbf{H}^t$ represents network entropy distribution at time t while $\mathbf{M}^t$ represents media node-edge homogeneity of the network at time t . We compute the correlation for all $t \in T$ and for all users in the in-degree and out-degree network. The average correlation across all timestamps is $\rho = 0.82$, suggesting that entropy and media likes distribution are strongly positively correlated. Hence we hypothesize that the entropy in either network is sampled from the same distribution

Figure 3. a)Gaussian KDE of in-degree users who have only liked authentic media, b)Gaussian KDE of in-degree users who have only liked fake media, c)Gaussian KDE of all in-degree users who have liked authentic and fake media, d)GMM approximation of information gain from authentic and fake media

as the fake and authentic media distribution. Formally, we state our hypotheses as follows:

$$H_{01}: H(\mathbf{X}) = T(\mathbf{Z}) \text{ and } H_{02}: H(\mathbf{X}) = F(\mathbf{Z})$$

$$H_{a1}: H(\mathbf{X}) \neq T(\mathbf{Z}) \text{ and } H_{a2}: H(\mathbf{X}) \neq F(\mathbf{Z})$$

where, $T(\mathbf{Z})$ and $F(\mathbf{Z})$ denotes the authentic and fake media distribution of the random variable ($\mathbf{Z}$) respectively. We find no statistically significant difference between entropy distribution and authentic media distribution (H_{01}) with a KS Statistic = 0.14 and p value = 0.28, thereby failing to reject the null hypothesis H_{01} . We also find no statistically significant difference between entropy distribution and fake media distribution (H_{02}) with a KS statistic=0.12, p value=0.46, thereby failing to reject the null hypothesis H_{02} . Thus, we observe that entropy response of media node-edge homogeneity and media likes distribution are strongly positively correlated. Though entropy response is individually correlated with authentic and fake media distribution, we hypothesize that it characterizes the superposition of both the media distributions. We assume and also empirically show (see Fig. 3a, Fig. 3b and Fig. 3c) that the underlying media likes distribution as well as entropy distribution are sampled from a Gaussian distribution such that media likes distribution is approximated as a linear combination of Gaussians, one describing the spread of fake media likes and the other describing the spread of authentic media likes as shown in Fig. 3d. We compare the entropy and actual media likes distributions using the Jensen-Shannon divergence which measures the similarity between the distributions. Hence, we formulate another pair of hypotheses as follows:

$$H_{01}: I(H; \mathcal{N}(\mu_{\text{authentic}}, \sigma_{\text{authentic}}^2)) = I(H; \mathcal{N}(\mu_{\text{fake}}, \sigma_{\text{fake}}^2)) \text{ and } H_{02}: I(H; \mathcal{N}(\mu_A, \sigma_A^2)) = I(H; \mathcal{N}(\mu_B, \sigma_B^2)).$$

$$H_{a1}: I(H; \mathcal{N}(\mu_{\text{authentic}}, \sigma_{\text{authentic}}^2)) \neq I(H; \mathcal{N}(\mu_{\text{fake}}, \sigma_{\text{fake}}^2)) \text{ and } H_{a2}: I(H; \mathcal{N}(\mu_A, \sigma_A^2)) \neq I(H; \mathcal{N}(\mu_B, \sigma_B^2)).$$

To test our hypotheses, we use the Expectation-Maximization algorithm to estimate the Gaussian parameters and fit the data to the Gaussian components found using Gaussian mixture modeling. We perform a KS test between the information gain from authentic and fake media distribution as well as from the approximated Gaussian components. We reject the null hypothesis (H_{01}) that authentic media likes distribution is statistically significantly different from that of the fake media likes distribution with KS statistic = 1.0, p value=9.99e-16 and the null hypothesis (H_{02}) that approximated authentic media likes distribution is statistically significantly different from that of the approximated fake media distribution with KS statistic = 1.0, p value = 8.61e-83 for the in-degree network.

Now, we describe how our model uses the results from the hypotheses tests to select user opinions as features for the media classifier. Once we compute the joint entropy response of the network to media node-edge homogeneity for varying γ , we approximate joint entropy as a mixture of Gaussians as shown in the flowchart in Fig. 1. We then compute the information gain from actual entropy and either of the Gaussian approximated media distributions and select the distribution with the maximum information gain ($\mathcal{Q}$) as shown in Fig. 1. Then, we define threshold (η) as a percentage of maximum user entropy in the network such that $\mathcal{Q}$ is the set of all users above η . Since the maximum entropy generated by a user is 4.0 bits, >25%

refers to all users who have generated greater than 1.0 bit of entropy. We then compute an aggregated histogram of media likes per category from all users above the threshold. This enables us to aggregate all the media opinions of users in $\mathcal{Q}$ who have been selected as features for the media classifier. Finally, we compute the average media likes per category and predict media categories with media likes above that of the average media likes per category as authentic media, else fake (see Fig. 1).

Table 1. Comparison of evaluation metrics

Method	η	Precision	Recall	Accuracy	F1 Score
Social Network Users	-	-	-	0.50	-
CNN & k-means (k=2) ⁴¹	-	0.67	0.69	0.67	0.68
Ours	(>0)%	0.83	0.50	0.50	0.63
Ours	(>25)%	0.86	0.60	0.58	0.71
Ours	(>50)%	0.88	0.70	0.67	0.78
Ours	(>75)%	1.00	0.60	0.67	0.75

Since it has been shown that there exists a positive correlation between users and credibility, we test our model on the users from only the in-degree network. Apart from selecting users based on our model, we validate our work by benchmarking against i) all users from the network as a classifier, and ii) Durall et al. method⁴¹ which is a state of the art unsupervised media classifier. From Table 1, we observe that using the combined human decision is similar to a random coin flip. Since we assume that media likes refers to user authentication³⁵, other evaluation metrics such as F1 score, precision and recall, cannot be calculated from just media likes. In replicating Durall’s method, we compute feature embeddings for all media used in this experiment, and consider this the test set. We use the convolutional neural network (CNN) architecture proposed by Durall et al. to generate media embeddings for the test set, which we use as input to the k-means clustering method with $k = 2$. We find the F1 score of the unsupervised classifier to be outperformed by our model, as shown in Table 1. We find that selectively filtering users on the basis of a network independent measure such as entropy is sufficient to classify fake media with an accuracy higher than random chance and unsupervised state of the art model. In a network where users are initially unknown to one another, we find that it is possible to approximate the underlying objective credibility using our approach. Further, by computing the entropy response of the in-degree users using media node-edge homogeneity, we compute the entropy within the network. We approximate entropy as a superposition of fake and authentic media likes of users. Since the distribution of fake media is different from authentic media, we are able to use the Gaussian approximated cluster with the highest information gain to utilize user opinions as features to classify fake media.

Figure 4. Sliding window versus evaluation accuracy

Time invariance of the proposed model

To study if our model is time-invariant to information classification, we perform an empirical experiment by introducing a sliding window parameter h . In Fig. 4, we illustrate the average performance of the classifier across all timestamps for every sliding window parameter in the range [2,50] and a stride parameter $s = 3$ which corresponds to the movement of the time window across time. We find that varying the time window varies the F1 score of the classifier with increasing time windows leading to increased F1, precision, recall and accuracy scores. When the sliding window parameter is increased by a value of 1 from 3-50, we approach the highest media classification accuracy and F1 score as shown in Table. 1. Since a time window of 2 corresponds to a period of 40 minutes in our study, a period of 17 hours is required to achieve an F1 score of 0.68 which is comparable to the performance of the state of the art unsupervised media classifier. Since social media networks have been operational for hundreds of thousands of hours, accessing prior time should not be an issue. In comparison, the state of the art unsupervised approach proposed by Durall et al. requires neural network fine-tuning as new media is uploaded to the network or in the worst case, re-training, given the development of new media doctoring tools. Hence, while our model is not time invariant, we show that it is possible for user-user and user-media interactions of a network over a period of time to be able to quantify the entropy needed to filter fake media.

Further, if we consider the relationship between time and entropy, we find that user activity in different timestamps often differs as a result of varying interaction with media (some network interactions happen at various timestamps with huge period of inactivity), causing entropy distribution of individuals to fluctuate with high uncertainty. However, when we consider 50 timestamps as shown in Fig. 4, the uncertainty due to user entropy fluctuations across different timestamps averages out, yielding consistent metric scores. This further highlights the dependence between time and evolution of entropy.

Conclusions

In this study, we design and conduct a controlled social media experiment with unknown human subjects in a network similar to Twitter. To characterize fake media using a network independent metric, we propose using media polarity and node-edge homogeneity, which are well-known detriments affecting the spread of fake media as a metric to compute the similarity between users in the network. In social media networks, interactions such as liking and sharing media, and forming/removing connections with other users, lead to the fluctuation of entropy over time. The entropy response of the in-degree users is captured using media node-edge homogeneity to compute the entropy within the network. In this paper, we compute entropy as a function of media polarized node-edge homogeneity to characterize the spread of fake and authentic media media. Using our findings, we describe a majority voting classifier to classify online media using selective user opinions. However, our model is limited by the size of the moving window with the moving window size equal to that of the whole study improving the F1 score of the classifier. This shows that new users who join the network at different times are valuable for media classification. As future work, selective user opinions can be combined with unsupervised media classifiers to further improve fake media classification in social media networks.

Apart from our findings, we highlight some of our assumptions that can possibly be grounded in theory using existing work or addressed as future work. We assume that no user knows each other prior to joining the network (random user network), media likes equal validation of authenticity, our two day study linearly maps to a longer term study. Finally, future work may consider exploring the generalizability to other media types such as text and audio.

Data availability

The dataset created and analysed in this work is uploaded to and available in a public repository.

References

1. Li, M., Wang, X., Gao, K. & Zhang, S. A survey on information diffusion in online social networks: models and methods. *Information* **8**, 118 (2017).
2. Marra, F., Gagnaniello, D., Cozzolino, D. & Verdoliva, L. Detection of GAN-Generated Fake Images over Social Networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (IEEE, 2018).
3. Lazer, D. M. J. *et al.* The science of fake news. *Science* **359**, 1094–1096 (2018).
4. Bergström, A. & Jervelycke Belfrage, M. News in Social Media: Incidental consumption and the role of opinion leaders. *Digit. Journalism* **6**, 583–598 (2018).
5. Chesney, R. & Citron, D. Deep Fakes: A Looming Crisis for National Security, Democracy and Privacy. *The Lawfare Blog* (2018).

6. Li, L. *et al.* Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. *IEEE Transactions on computational social systems* **7**, 556–562 (2020).
7. Vaidyanathan, G. News Feature: Finding a vaccine for misinformation. *Proc. Natl. Acad. Sci.* (2020).
8. Shu, K., Sliva, A., Wang, S., Tang, J. & Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor. Newsl.* **19**, 22–36 (2017).
9. Fouss, F., Francoise, K., Yen, L., Pirotte, A. & Saelens, M. An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification. *Neural networks* **31**, 53–72 (2012).
10. Parks, L. Dirty Data: Content Moderation, Regulatory Outsourcing, and The Cleaners. *Film Q.* **73**, 11–18 (2019).
11. Etlinger, S. What's So Difficult about Social Media Platform Governance? *Model. for Platf. Gov.* **20** (2019).
12. De Domenico, M., Lima, A., Mougél, P. & Musolesi, M. The anatomy of a scientific rumor. *Sci. reports* **3**, 2980 (2013).
13. Filia, A. *et al.* Ongoing outbreak with well over 4,000 measles cases in Italy from January to end August 2017- what is making elimination so difficult? *Eurosurveillance* **22**, 30614 (2017).
14. Datta, S. S. *et al.* Progress and challenges in measles and rubella elimination in the WHO European Region. *Vaccine* **36**, 5408–5415 (2018).
15. Tambuscio, M., Ruffo, G., Flammini, A. & Menczer, F. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proceedings of the 24th international conference on World Wide Web*, 977–982 (2015).
16. Gupta, M., Zhao, P. & Han, J. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, 153–164 (SIAM, 2012).
17. Johnson, T. J. & Kaye, B. K. Reasons to believe: Influence of credibility on motivations for using social networks. *Comput. human behavior* **50**, 544–555 (2015).
18. Friggeri, A., Adamic, L. A., Eckles, D. & Cheng, J. Rumor Cascades. In *ICWSM* (2014).
19. Del Vicario, M. *et al.* The spreading of misinformation online. *Proc. Natl. Acad. Sci.* **113**, 554–559 (2016).
20. Bovet, A. & Makse, H. A. Influence of fake news in Twitter during the 2016 US presidential election. *Nat. communications* **10**, 1–14 (2019).
21. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
22. Stefanone, M. A., Vollmer, M. & Covert, J. M. In News We Trust? Examining Credibility and Sharing Behaviors of Fake News. In *Proceedings of the 10th International Conference on Social Media and Society*, 136–147 (2019).
23. Klayman, J. & Ha, Y.-W. Confirmation, disconfirmation, and information in hypothesis testing. *Psychol. review* **94**, 211 (1987).
24. Lou, C. & Yuan, S. Influencer marketing: how message value and credibility affect consumer trust of branded content on social media. *J. Interact. Advert.* **19**, 58–73 (2019).
25. Bandura, A. Social cognitive theory: An agentic perspective. *Annu. review psychology* **52**, 1–26 (2001).
26. Golbeck, J. & Hendler, J. Filmtrust: Movie recommendations using trust in web-based social networks. In *Proceedings of the IEEE Consumer communications and networking conference*, vol. 96, 282–286 (Citeseer, 2006).
27. Briscoe, E. J., Appling, D. S., Mappus IV, R. L. & Hayes, H. Determining credibility from social network structure. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 1418–1424 (2013).
28. Scheufele, D. A. & Krause, N. M. Science audiences, misinformation, and fake news. *Proc. Natl. Acad. Sci.* **116**, 7662–7669 (2019).
29. Van den Berg, P. & Wenseleers, T. Uncertainty about social interactions leads to the evolution of social heuristics. *Nat. communications* **9**, 1–7 (2018).
30. Zhao, K., Karsai, M. & Bianconi, G. Entropy of dynamical social networks. *PloS one* **6**, e28116 (2011).
31. Hentschel, C., Wiradarma, T. P. & Sack, H. Fine tuning CNNs with scarce training data—Adapting ImageNet to art epoch classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, 3693–3697 (IEEE, 2016).
32. Rossler, A. *et al.* Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, 1–11 (2019).

33. Alhindi, T., Petridis, S. & Muresan, S. Where is your Evidence: Improving Fact-checking by Justification Modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 85–90 (2018).
34. Yang, D. *et al.* True and fake information spreading over the Facebook. *Phys. A: Stat. Mech. its Appl.* **505**, 984–994 (2018).
35. Yang, S. *et al.* Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 5644–5651 (2019).
36. Rand, D. G., Arbesman, S. & Christakis, N. A. Dynamic social networks promote cooperation in experiments with humans. *Proc. Natl. Acad. Sci.* **108**, 19193–19198 (2011).
37. Pennycook, G. & Rand, D. G. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci.* **116**, 2521–2526 (2019).
38. Hansen, D., Shneiderman, B., Smith, M. A. & Himelboim, I. *Analyzing Social Media Networks with NodeXL: Insights from a Connected World* (Elsevier Science, 2019).
39. Resnick, P., Kuwabara, K., Zeckhauser, R. & Friedman, E. Reputation systems. *Commun. ACM* **43**, 45–48 (2000).
40. Josang, A. & Ismail, R. The beta reputation system. In *Proceedings of the 15th bled electronic commerce conference*, vol. 5, 2502–2511 (2002).
41. Durall, R., Keuper, M. & Keuper, J. Watch your Up-Convolution: CNN Based Generative Deep Neural Networks are Failing to Reproduce Spectral Distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7890–7899 (2020).

Figure Legend

1	A graphical overview of our model. Entropy is computed for target nodes. Homogeneous paths connecting adjacent users are denoted by thick lines.	5
2	Normalized entropy response and normalized likes distribution against media node-edge homogeneity in the in-degree network	5
3	a)Gaussian KDE of in-degree users who have only liked authentic media, b)Gaussian KDE of in-degree users who have only liked fake media, c)Gaussian KDE of all in-degree users who have liked authentic and fake media, d)GMM approximation of information gain from authentic and fake media	6
4	Sliding window versus evaluation accuracy	7

Table Legend

1	Comparison of evaluation metrics	7

Author contributions statement

S.K.A.P. and C.T. designed research; S.K.A.P. and C.T. performed research; S.K.A.P. analyzed data; and S.K.A.P. and C.T. wrote the paper. S.K.A.P. and C.T. reviewed the manuscript.

Additional information

Competing interests The authors declare no competing interests.

Supplementary information is available for this paper at:

Correspondence and requests for materials should be addressed to C.T.

Figures

Figure 1

A graphical overview of our model. Entropy is computed for target nodes. Homogeneous paths connecting adjacent users are denoted by thick lines.

Figure 2

Normalized entropy response and normalized likes distribution against media node-edge homogeneity in the in-degree network

Figure 3

a) Gaussian KDE of in-degree users who have only liked authentic media, b) Gaussian KDE of in-degree users who have only liked fake media, c) Gaussian KDE of all in-degree users who have liked authentic and fake media, d) GMM approximation of information gain from authentic and fake media

Figure 4

Sliding window versus evaluation accuracy

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupportingInformationClassificationofunlabelledonlinemedia.pdf](#)