Social Network Analysis and Mining (2022) 12:52
https://doi.org/10.1007/s13278-022-00878-9

ORIGINAL ARTICLE

Novel approaches to fake news and fake account detection in OSNs: user social engagement and visual content centric model

Santosh Kumar Uppada[1] · K. Manasa[1] · B. Vidhathri[1] · R. Harini[1] · B. Sivaselvan[1]

Received: 7 September 2021 / Revised: 30 March 2022 / Accepted: 2 April 2022 / Published online: 10 May 2022

Abstract
With an increase in the number of active users on OSNs (Online Social Networks), the propagation of fake news became obvious. OSNs provide a platform for users to interact with others by expressing their opinions, resharing content into different networks, etc. In addition to these, interactions with posts are also collected, termed as social engagement patterns. By taking these social engagement patterns (by analyzing infectious disease spread analogy), SENAD(Social Engagement-based News Authenticity Detection) model is proposed, which detects the authenticity of news articles shared on Twitter based on the authenticity and bias of the users who are engaging with these articles. The proposed SENAD model incorporates the novel idea of authenticity score and factors in user social engagement centric measures such as Following-followers ratio, account age, bias, etc. The proposed model significantly improves fake news and fake account detection, as highlighted by classification accuracy of 93.7%. Images embedded with textual data catch more attention than textual messages and play a vital role in quickly propagating fake news. Images published have distinctive features which need special attention for detecting whether it is real or fake. Images get altered or misused to spread fake news. The framework Credibility Neural Network (CredNN) is proposed to assess the credibility of images on OSNs, by utilizing the spatial properties of CNNs to look for physical alterations in an image as well as analyze if the image reflects a negative sentiment since fake images often exhibit either one or both characteristics. The proposed hybrid idea of combining ELA and Sentiment analysis plays a prominent role in detecting fake images with an accuracy of around 76%.

Keywords Social engagements · Epidemic model · Authenticity score · User bias · Visual information · Deep neural networks · Sentiment analysis

1 Introduction

Online Social Networks(OSNs) is a collection of online communications channels dedicated to community-based input, interaction, content-sharing, and collaboration. Statistics show that there are 4.48 billion social network users worldwide, with an average of 6.6 social media platforms monthly. This value equates to about 56.8% of the current population, with 99% of these users relying exclusively on mobile phones (Brain 2021). This analysis depicts how OSNs profoundly influence social, economic, and political decision-making. However, this brings a fair share of troubles, a prominent one being the rapid spread of fake news.

✉ Santosh Kumar Uppada
coe18d005@iiitdm.ac.in

[1] Indian Institute of Information Technology, Design and Manufacturing, Kancheepuram, Chennai, India

1.1 Fake news

Traditional fake news is considered a form of deliberate disinformation or hoaxes. Fake news is written and published usually to mislead and damage an agency, entity, or person's reputation and gain financially or politically. Creation, publication, and propagation are the basic steps in fake news propagation. Traditional fake news mainly targets consumers by exploiting their vulnerabilities. The success of fake news propagation is often connected with intentionally exaggerated text, written impressively or emotionally, added with compelling images with high user sentiment and clickbait to the links (Zhou and Zafarani 2020). Psychological factors also play a significant role in aiding the spread of fake news (Baptista and Gradim 2020). Users often interact only with certain kinds of news because of how news appears on their feed/homepage. Users also tend to form groups with like-minded people, polarizing their opinions. Two significant

factors make consumers naturally vulnerable to fake news. *Naive realism*, which is user's tendency to believe in the news they perceive from their beliefs or perceptions (based on Theory of Perception or relationalism) and *Confirmation bias* where users prefer to receive information that confirms their existing views (Shu et al. 2017).

Fake News is viewed either as an economic or epidemic model. As per the economic model, the theory of fake news is similar to a two-player strategy game with *publishers* and *consumers* as critical players. Publishers want to maximize their profit by reaching more consumers and their reputation in terms of authenticity, while consumers want to maximize obtaining accurate information and the news that satisfies their prior opinion (Calisir 2021).

It is to be noted that fake news propagation happens when the publisher prefers to maximize profit and consumers prefer to satisfy their prior opinions. COVID-19 also marked a massive spread of fake news all over the world. News spread related to medical advice, misleading figures related to cases and deaths, pseudo-protests against lockdowns, scarcity of basic amenities, and medical equipment. One such news that became popular is related to the riot of 500 lions on roads in Russia, amid lockdown (Hamdan 2020).

Certain psychological and cognitive aspects play a significant role in spreading news. The echo chamber effect observed in social networks shows that certain beliefs and biased information often amplify (Jamieson and Cappella 2008). Confirmation bias makes people trust fake news if it aligns with their pre-existing knowledge, and if users tend to interact with the same news again and again across communities, they believe it blindly, which is termed as Frequency Heurestic (Nickerson 1998). Users who engage with fake news posts can be malicious users who spread the false information intentionally and naive users who participate unintentionally, driven by influence and psychological factors (Zhou et al. 2019).

In the propagation of fake news, it is estimated that posts with images get reshared about 11 times more than those without any visual content (Jin et al. 2016). Thus, visual content is a prime component of fake news, and fake images are often eye-catching and emotional. Thus, it becomes necessary to map such psychological triggers to the characteristics of the image. These psychological patterns are limited to visual appearance, and beyond the standard object-level features (Zhou et al. 2019).

Fake images can be digitally modified to manipulate viewers or misleading images that are authentic, unaltered images used in inappropriate contexts. Images can be used out of context, which includes images of an earlier event getting shared as an event from the current scenario, or even images misrepresented with wrong intent (Qi et al. 2019; Lang 1979). Hence, traditional image sets are not suitable for this task of fake image classification (Jin et al. 2016).

1.2 Fake accounts

People create accounts to share social media data using various social networking platforms. Users tend to create accounts with anonymous or wrong data to propagate Fake news to avoid revealing their identity. Users also tend to create accounts either in the name of some other person (Identity Theft) or intrude into their accounts. Fake accounts creation also has some targeted financial benefits. Fake accounts also got created during incidents such as the Boston Marathon blast and COVID-19 Prime Minister Relief Fund accounts. Bots or automated programs maintain these fake accounts and help in the network's faster and deeper spread of fake news (ABP News Bureau 2020).

Fake accounts always tend to follow and interact with posts of influencer users in the network. Even social platforms like Twitter, Facebook, and Whatsapp delete or freeze these fake accounts through an impersonation policy (Sahoo and Lavanya 2019). During the Boston marathon blast, there are around 32,000 new accounts created just after a few hours of the blast, out of which 20% (6073) accounts have been suspended by Twitter (Gupta et al. 2013). Fake accounts creation creates hoaxes in society and helps in the easy propagation of Fake News. Therefore, fake account detection plays a vital role in detecting fake propagation in social networks (Kondeti et al. 2021).

1.3 Fake news in online social networks—challenges

Content-based and social context-based approaches are the primary methods in fake news detection. Most content-based approaches deal with textual features for fake news detection. For social media content, this approach might not be practical for the following reasons.

- Text-based approaches are language-dependent but social media allows users to post in multiple languages. Usage of traditional language translations might suffer from losing the original meaning.
- The possibility of a user to read every post appearing on their feed is low; it is generally the multimedia content that captures the user's attention. Therefore, fake news publishers tend to use sentiment and catchy images in the social content to grab users' attention.

Usage of visual content in social posts highlights the importance of studying images in content-based approaches to fake news detection. However, content-based approaches alone may not help in the efficient detection of fake news; the intention of publishing fake news

is to mislead the users, and users try to mimic real news to the best possible capacity. Then other possible ways to improvise the detection would be to capture the people's intention to spread the news. As grabbing more attention and reachability are significant goals, users create news with the necessary content and misleading texts to make the news popular and spread faster and more profound into the network. The other approach for fake news detection is the social context-based approach, which depends on the user's psychology or behavior (Zhou and Zafarani 2020).

2 Related work

Fake News propagation targets a person or firm and creates hoaxes in society. The intentional spread of Fake news also has a typical connection with creating accounts with anonymous or fake details. Automated programs or bots are also created to aid in the easy and fast propagation of news more profoundly into the network.

As images propagate faster and have high interaction patterns, users try to propagate fake news using image-based posts. User propagation patterns and user-related attributes such as follower-followers ratio, account being verified, or not also help easy identification of Fake posts. As Fake posts detection has gained significant focus, researchers are working on finding methods to detect fake posts in online social networks. Fake News detection methods target the news propagation patterns, and users involved in the spread. Fake and News can also have its connection with creating fake or anonymous accounts (generally bots) for the faster spread of news without disclosing the identity of the news spreader (Chang 2021). Therefore, fake news detection is a collection of content-based, social context-based, and propagation-based approaches. These approaches help detect the fake accounts typically used in spreading fake news (Zhang and Luximon 2021).

2.1 Propagation-path-based fake news detection

Existing approaches for fake news detection can be divided into three main categories, based on content, social context-based, which again include stance-based and propagation path-based approaches (Shu et al. 2017). Content-based approaches, which are widely used in fake news detection, rely on linguistic (lexical and syntactical) features that can capture deceptive cues or writing styles (Wynne and Wint 2019; Granik and Mesyura 2017). These focus on linguistic features such as special characters, emotions, symbols, sentiment, positive/negative words, hashtags. However, fake news is intentionally written to mislead readers, which makes it nontrivial to detect based on news content (Castillo et al. 2013). Furthermore, most linguistic features are language-dependent, limiting the generality of these approaches. Tacchini et al. (2017) proved that posts on social media could be classified as hoax/non-hoax based on users who "like" them. Propagation path-based approaches (Liu and Wu 2018; Monti et al. 2019; Kwon et al. 2017) also show a promising research direction based on studying the news proliferation process over time.

Shuo Yang et al. have proposed unsupervised fake news detection where hierarchal user engagement data are used for counterfeit news detection. The model uses second-level user engagement data like retweets, likes, replies, and user opinions for analysis. In general, social media users can be either verified or unverified, and the model relies on the social engagement patterns related to verified users. As verified users have more influence and attention, social engagement patterns related to verified users are only considered. LIAR with 12,800 short news statements and labels verified by PolitiFact.com and BuzzFeed News with 1,627 news articles related to U.S elections from Facebook is used from the analysis. It is observed that the proposed UFD (Unsupervised Fake News Detection) achieved better results on LIAR (75.9% accuracy) and BuzzFeed News (67.9% accuracy) (Yang et al. 2019).

Mahudeswaran et al. proposed a hierarchal propagation pattern for fake news detection. Propagation pattern is taken at micro-and macro-level propagation networks. Micro-level propagation network resembles news and reposts, whereas Macro-level includes more social bots for propagation. FakeNewsNet dataset with data extracted from PolitiFact and GossipCop is used for analysis. Structural features like cascade levels count of bots used for retweets are also considered. In addition, temporal characteristics are also taken into account. It is observed that combining features acquired an accuracy of 84.3% on the PolitiFact and 86.1% on the GossipCop dataset (Shu et al. 2020).

Lu et al. proposed a Graph-aware Co-Attention Networks (GCAN), which works on user characteristics, propagation patterns, the correlation between source tweets and interactions. Dual co-attention mechanisms are employed to capture the correlation between the source of the posts and the user's engagement. Finally, predictions are made based on the interaction patterns for the news. Twitter15 and Twitter 16 are the datasets used for analysis. It is observed that CGAN achieved an accuracy of 87.67% on Twitter15 and 90.84% on the Twitter16 dataset (Lu and Li 2020).

Yang Liu and Yi-Fang proposed a model for detecting fake news on social media through a propagation path using recurrent and convolutional networks. The proposed model helps early fake news detection on social media by classifying propagation paths. Multivariate time series for characterizing users engaged in spreading the news. Recurrent and Convolutional networks are used on the numerical vectors derived from multivariate time series of user propagation.

Model is evaluated on real-time datasets like Weibo, Twitter-15, and Twitter-16. User characteristic data like followers count, IsVerified, friends count. are considered for analysis and observed that the model gave an accuracy of 85% on Twitter and 92% on Weibo dataset (Liu and Wu 2018).

Kia Shu et al. have proposed a model that accounts for social and psychological aspects like confirmation Bias and the Echo chamber effect for fact-checking in online social networks. Tri-relation between publishers, news, and users is analyzed for detecting fake news. BuzzFeed and PolitiFact datasets are used for analysis and observed that TriFN acquired an accuracy of 87.8% on the PolitiFact and 86.4% on BuzzFeed dataset (Shu et al. 2017).

Milan Dordevic et al. proposed a model to identify Fake news from the variables that govern the spread of fake news. Twenty-seven variables that govern the identification of fake news related to users, content, and social network are considered. Properties such as edges, vertices, susceptible, and infected for Users; Timestamp, reference, the source for Content; Crosswire, authentication, and newsgroup are collected for social networks. Variables considered are examined with certain viral events like earthquakes, and Timestamp is used to identify the exact time the event occurred. It is observed that even though identification of these variables helped in combating fake news in some areas, it still failed for some events as prominent variables may vary from one event to another (Dordevic et al. 2020).

2.2 Fake news based on image authenticity

Bayar et al. propose a modified CNN architecture to learn manipulation features of different image editing operations, which aims to address the inadequacy of using traditional forensic techniques—an image can be manipulated in multiple ways, using combinations of tampering methods; thus, a traditional forensic examination would require testing the image for each of the methods (Bayar and Stamm 2016).

Masciari et al. (2020) have proposed a framework for fake news detection by image analysis. The framework works on heterogeneous data. The NLP module uses typical TF-IDF and Google BERT to detect fake text and ELA (Error Level Analysis) and CNN combination to work on counterfeit images. Momentum, RMSProp, and Adam are optimizers used with ReLU as an activation function. For binary classification, Sigmoid and Softmax are used at the last layer. The CNNs are run for ten epochs with an optimal value of 32 as its batch size. FakeNewsNet is the dataset used, which has data extracted from Politifact and GossipCop. The framework also works on the data from the LIAR and PHEME datasets. It is observed that the proposed model acquired an accuracy of 76.5% on the PHEME dataset (similar results observed on other datasets considered) (Masciari et al. 2020).

Jin et al. proposed a Recurrent Neural Network with an attention mechanism (att-RNN) to fuse multimodal features for rumor detection. LSTM is used to obtain a joint representation of text and social context. For visual features, the output of the second to the last layer of the VGG19 architecture pre-trained on ImageNet is used (Jin et al. 2017).

Wang et al. proposed EANN to derive event-invariant features, which consists of a multimodal feature extractor, fake news detector, and an event discriminator (Wang et al. 2018). The work draws the concept of adversarial networks to capture event-invariant features to improve generalisability. As in Jin et al. (2017), the pre-trained VGG19 network for obtaining visual representations.

Dhruv et al. proposed MVAE to learn a shared representation of multimodal information for fake news detection, using a bimodal variational autoencoder with a binary classifier. The optical encoder uses the output of the fully connected layer of the pre-trained VGG19 architecture trained over ImageNet, and these features are used along with the textual features to obtain a shared representation (Khattar et al. 2019).

Qi et al. propose Multi-domain Visual Neural Network(MVNN) to fuse visual information of frequency and pixel domains for fake news detection. They state that re-compressed and tampered images present periodicity in the frequency domain, while the visual impacts and emotional provocations often found in fake news are observed from the pixel domain. For capturing information from the frequency domain, the DCT coefficients of the image are sent as input to a CNN network. For the pixel domain, MVNN contains a CNN-RNN network to extract features of different semantic levels. Both physical and semantic features of images are fused via the fusion sub-network, and an attention mechanism is employed (Qi et al. 2019).

Tharindu et al. have proposed a hierarchal attention memory model for detecting fake and fraudulent faces via neural network memories. Data are considered from FaceForensics, FaceForensics++, FakeFace in the wild datasets. Viola-Jones algorithm is used to detect faces. Neural Turing machine, Dynamic Memory Network, and Tree Networks are used for evaluation. LSTM and ResNet are used as training models. These memories are trained to classify the input images using supervised learning algorithms. Hierarchical Attention Memory Network is combined with memories to store visual evidence, which can be used to identify fake and fraudulent images. For the FaceForensics++ dataset, the proposed model has achieved an accuracy of 84.12% (Fernando et al. 2020).

Hady et al. proposed a model to create and detect Deepfakes using Deep learning. Two autoencoders are used to create fake images, where one autoencoder learns the features of the source image, and the other learns features of the target image. MesoNet CNN is used on

the dataset with 5000 images to detect complex deep fakes. When CNN is trained on the deepfake images, it is observed that CNN detected deepfakes with 80% confidence (Khalil and Maged 2021).

2.3 Related work on fake accounts

Jia et al. (2017) is an algorithm that makes use of homophily to identify fake account users while being fairly resistant to weak homophily. Yushan et al. proposed another model called Liu et al. (2016) which uses machine learning techniques to identify the length of the Sybil path.

There are several algorithms proposed which use machine learning methodologies and approaches like feature-based detection (Boshmaf et al. 2016; Wang et al. 2013; Viswanath et al. 2014), neural networks, SVM (Akyon and Kalfaoglu 2019; Khaled et al. 2018) Sequential Mining Optimization (Galan-Garcia et al. 2016). Faith et al. applied machine learning methods to detect fake accounts on Instagram. Along with it, proposing a genetic-algorithmic approach to handle bias in the dataset (Akyon and Kalfaoglu 2019). Viswanath et al. proposed a method where low dimensionality of user behavior is analyzed and observed that most users can be explained with a limited set of features, and users who are not included here are considered anomalous (Viswanath et al. 2014).

Xiao et al. proposed an algorithm to detect clusters of fake accounts on online social media before they start creating harm or connecting with genuine users (Xiao et al. 2015). It clusters the accounts using a k-means clustering algorithm, determines cluster level features, and then scores the accounts in the cluster and decides cluster labels based on the collective average score of the cluster.

3 Proposed models

The proposed SENAD model identifies the Fake News spread and helps identify Fake accounts created to spread fake news. As part of working with Fake images, CredNN is the framework proposed, where a joint representation of forensic and emotional cues in fake images is learned via an ensemble of deep Convolutional Neural Networks fine-tuned on visual sentiment fake image datasets. Figure 1 depicts the overall workflow of the proposed model.

3.1 SENAD model

Social Engagement-based News Authenticity Detection (SENAD) model is proposed to derive the authenticity score of the user's as-well-as posts that users are interacting. This algorithm addresses the limitations of many fake news detection algorithms by taking into account the authenticity of the users propagating the news and each user's inherent bias. Existing approaches to fake news detection are domain-restricted (training data dependency) and do not explore the social network user engagement patterns and impact the genericity of the Fake news detection models. Most of the Fake news detection algorithms target Fake News or fake Account detections individually, but there is no standard method that calculates the basic authenticity score of the user and posts to detect Fake News and Fake Accounts. The account's score depends on the class of news that the user is sharing, and the account's score also depends on account-related features. The authenticity score of news articles depends on the users interacting and sharing such posts. Authenticity score also relies on the inherent bias of the user for a specific domain or topic. As the authenticity score of the news always gets affected by the user's score, studying metrics that determine the user's score and bias would be helpful.

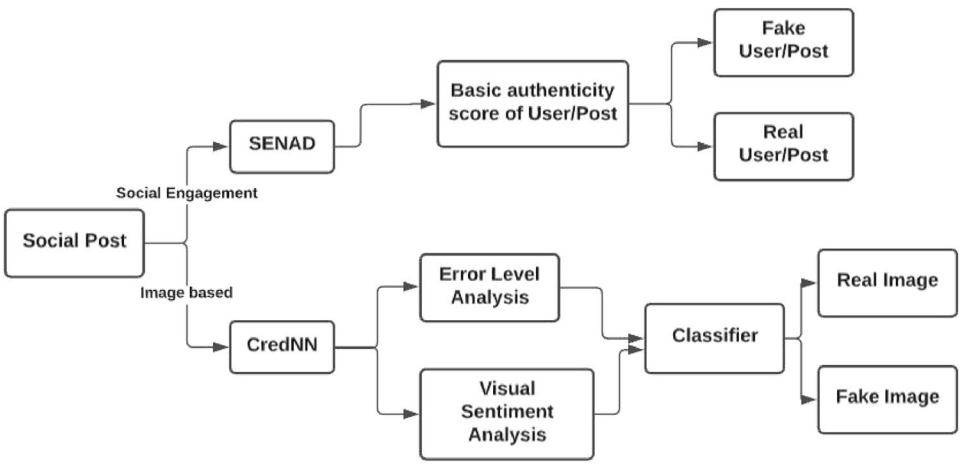

Fig. 1 Proposed model workflow

3.1.1 Dataset and observations

FakeNewsNet (Shu et al. 2018) is the crawler used to retrieve data from GossipCop that fact-checks celebrity posts and PolitiFact that fact-checks posts related to politics. GossipCop also has facts about exciting stories, posts, or news published in magazines, news, or websites. Politifact has news and posts related to political issues (Shu et al. 2018). As there is no labeled data for user accounts, it is assumed that the user features obtained will not change throughout the user activity timeline. Table 1 depicts the average length of text from GossipCop and PolitiFact datasets.

In general, Fake news intention is to gain immediate attention, which might not be the case for other posts. Steady interaction patterns are generally observed for regular posts until they saturate. Interaction patterns of the users with the news posts versus the time are plotted as shown in Fig. 2. It is observed that fake news has maximum interaction during the initial period. On the contrary, there is continuous interaction and saturation of posts over time in the case of real news posts in the Fig. 3.

3.1.2 Determining features for base authenticity score

Social Context-based features help in studying the authenticity of the account. Some of these features are intuitive, whereas some can be obtained by analogy from similar models. Some of the directly intuitive features include

- *Age of account* As discussed earlier malicious accounts include social bots and cyborgs created intentionally for the said purpose, i.e., chances of these accounts being news are exceptionally high. Consequently, old or long-lasting accounts tend to be more trustworthy.
- *Follower, Following ratio* The number of followers and following from a user account themselves carry adequate information about the account. However, the ratio speaks a different feature. This ratio can account for the responsibility of a user. Any one of these values cannot speak by themselves entirely, as the responsibility of a community user shall depend on various features of the community, such as its size, density. Hence, a ratio of these features will give a better insight.
- *Other Features* The credibility and activeness of an account can be studied from a few other account features such as verified(whether or not Twitter verifies the account), protected(whether or not the user account is protected), length of screen name, has description(whether or not the user account has a description), length of the name, favorites count, statuses count.

Table 1 Average length of text

Dataset	Avg. length of text (in words)	Avg. length of text (in characters)
GossipCop_Fake	3336.5	561.4
GossipCop_Real	3512.6	595.8
Politifact_Fake	2286.0	372.0
PolitiFact_Real	13128.3	2264.3

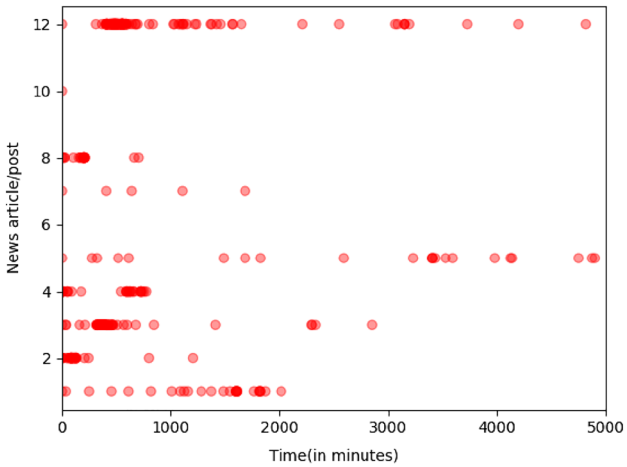

Fig. 2 Interaction of users with fake news post over time

For accounts like Twitter, verified status help in knowing the account's authenticity, and it is observed that users with verified account status will be very cautious in spreading any news. A protected account only allows a limited number of users to view the published messages; hence, the probability of propagating fake news will be less. The bias factor also helps in knowing about the change in users' interaction

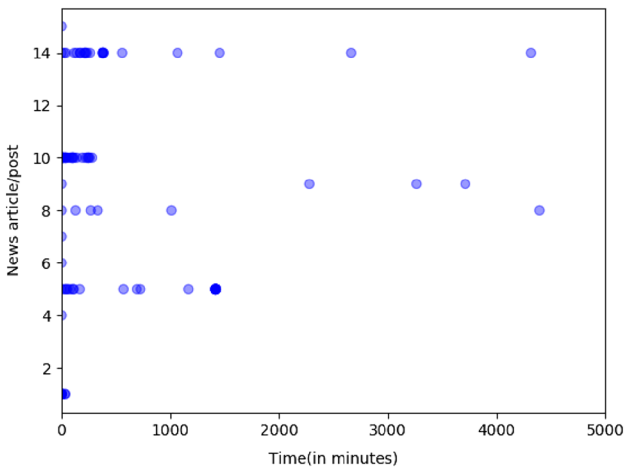

Fig. 3 Interaction of users with real news post over time

patterns. Therefore, these features help in determining the authenticity of the user (Shu et al. 2019).

3.1.3 Bringing the analogy from disease spread model

The spread of fake news in online social media is analyzed using epidemic modeling diseases. Here two basic epidemic models, SIS(Susceptible-Infected-Susceptible) and SIR(Susceptible-Infected-Recovered), are taken into account. The model depicted by the online community, in this case, is more likely to be an SIS model than a SIR. The reason behind this intuition is that, once the user is cured of infection, it does not mean that the user becomes immune to the disease (infection), and the user tends to be susceptible again to the infection (i.e., future hoax posts). The spread of news also depends on the important or prominent person in the group, termed as *Influential Person*. The influence power of the node is referred to the number of nodes a given node can further spread the infection. The influential power of a node in a network is heavily dependent on the term *coreness* of a node. The coreness of the network identifies the interlinked groups of the network. It is typically calculated using k-core decomposition, identifying the maximum number of node connections. Typically, K-core is the maximal group connected to at least k other entries in the group. The k-coreof a graph is the maximal subgraph with minimum degree. Mathematically speaking, the k-core of a graph G is the maximal subgraph, H $\subseteq$ G, such that δ(H) $\geq$ k, where δ (H) is the minimum degree of a vertex in H (Giatsidis et al. 2011).

The interaction of low influential nodes with the post is close to zero in the case of real content posts, whereas a significant number of low influential nodes interact with a fake content post in its early stages. This pattern indicates that low influential power nodes collectively involved in early propagation stages are predictable, i.e., low authenticity score user and the probability of the post intently trying to propagate being fake is high. It is observed that the influential power is directly proportional to the coreness of a node. Users with low influence power try to target or respond to the posts of high influenced users to gain attention and reachability. Identifying such influence nodes helps in identifying intentional spread patterns. The influential power of the node is purely dependent on the term *coreness* of the node (Liu and Wu 2018).

Figures 4 and 5 depict the user engagement patterns of Fake and Real news posts overtime for the coreness of the nodes in the network for the FakeNewsNet dataset. The inferences made from Figs. 4 and 5 are as follows

- The node's size in the graph is proportional to the coreness value of the node. More Low core valued nodes are initially involved in propagation in fake news than the pattern observed in real news.
- It is observed that the pattern followed by the real news is different when compared to fake news. Interaction patterns for fake news are high initially for the fake news, as the intention of fake news itself is to spread it faster and deeper.
- The coreness of the node helped in knowing the spreading pattern of the news. Analyzing the coreness of the network also helps in potentially exploiting automatic fake news detection.

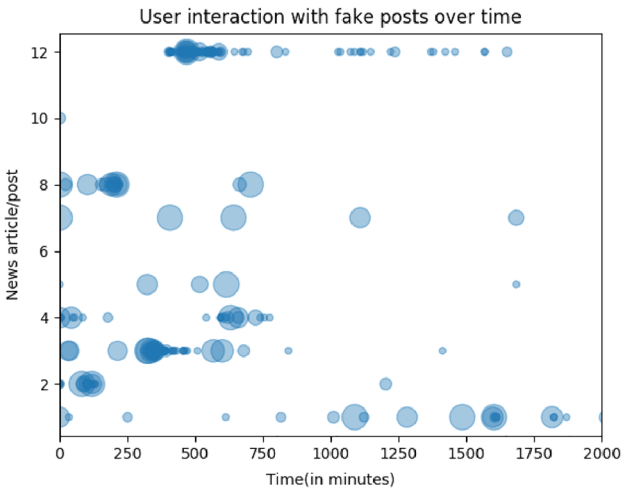

Fig. 4 User engagements with fake news post over time with coreness

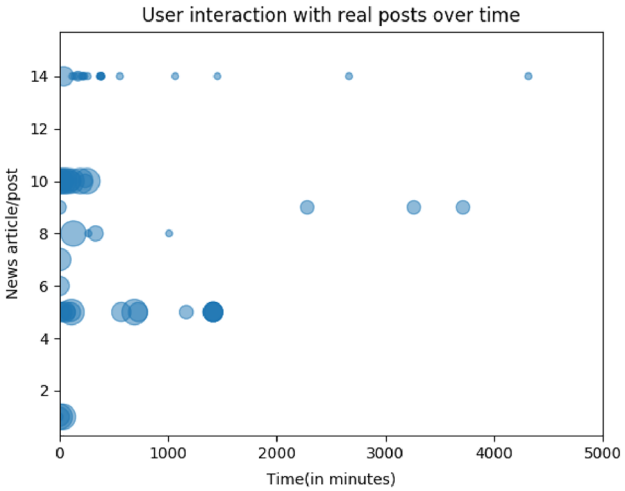

Fig. 5 User engagements with real news post over time with coreness

3.1.4 Base authenticity score

Basic authenticity score of an account is considered as the authenticity of an account considering only the account parameters and not the type of news they are propagating. The parameters of the accounts chosen for authenticity score calculation are (based on the previous inferences and references) age of the account, following/followers ratio, verified, coreness, length of screen name, protected, has_description, length of the name, favourites_count, Friends_count, statuses_count.

As there is no specific ground truth for account authenticity, the K-means algorithm is chosen to cluster the data into real, fake, or ambiguous. After deciding cluster labels, authenticity scores for each user are assigned based on their distance from the cluster's center. Each user has a base score of 50; if the user belongs to a fake cluster, closer the account to the cluster, the authenticity score will get reduced. If the user belongs to a real cluster, the value will increase, and if the user belongs to an ambiguous cluster, the value is either increased or decreased depending on its distance from real and fake clusters.

3.1.5 Determining bias of user

The credibility of the account involved in spreading a news article is not solely enough to determine its authenticity. A single user can be involved in various networks based on his/her various interests with the same account. When a user is spreading a post hoax/ true, it may not be driven by some malicious intent but, in fact, just the user's beliefs.

The user's bias can be observed in various ways, including past user posts and user behavior in the network. Bias is captured as a binary variable, which gives intuition of whether the user is under the person's influence while interacting with any social posts. This typical behavior of the user is captured by applying pattern mining algorithms. Consider the following example where a user u follows users for better understanding. v_1,v_2,v_3...,v_n. Each of these n users can be from different communities that u follows. And it has been observed that: When v_1, v_2, v_3 have tweeted some post,u responds to it. This rule is represented as:

$$v1, v2, v3 \rightarrow u \tag{1}$$

Such patterns are drawn with the help of a simple association rule mining algorithm known as *apriori* (Agrawal et al. 1996). *Apriori* returns rules that have confidence more significant than a mentioned minimum value; This implies that if a user is biased, his/her account shall be present in some rule as a consequent. The antecedent of the rule signifies the persons who have participated in the interactions with the post.

3.1.6 Social engagement-based news authenticity detection (SENAD)

After determining the base authenticity score, the score must be adjusted based on user involvement in news propagation belonging to training data. During this phase,

- Every time a user propagates fake news, the authenticity score of the user is reduced by a factor that is inversely proportional to both the current authenticity score of the user and how early the user is involved in the propagation. This rule follows the basic intuition that reduction is more for already less authentic users. The earlier they are involved in propagating fake news, the more the chances of it being intentional.

$$v \propto \frac{1}{(t * current\_authenticity\_score)} \tag{2}$$

  where t, is the time the user has interacted, starting from the first post, and v is the value by which the current authenticity score of the user is to be reduced.
- Every time a user propagates a true news post, the authenticity score of the account is incremented by a factor that is directly proportional to the current score. This rule is based on the observations made from the earlier descriptive statistics. There is no comparison between the propagation patterns between real and fake news; hence the time of action plays no role here.

$$v \propto current\_authenticity\_score \tag{3}$$

  where v is the value by which current authenticity score of user is to be reduced.

After determining authenticity scores of users existing in a network, score of the news is targeted.

- Every news post initially starts of with a base score 50
- Every time a user propagates the news article the score of the news post is either decremented/ incremented by a value.
- This value is based on following factors :
  - What is the authenticity score of the user who propagated with the post? This score also determines the sign of value. A good user will lead to a positive value, whereas a less authentic account will be negative.

$$sign : (user\_authenticity\_score - 50) \tag{4}$$

$$x \propto \frac{1}{user\_authenticity_s core} \tag{5}$$

where x is the value by which the authenticity score of news will vary, and *sign* determines if it will be an increment/decrement.

– When(how early) did the user get involved with the propagation

$$x \propto \frac{1}{t} \tag{6}$$

where t indicates how long after the first post has the current user interacted with the post.

– Was the user biased when propagating the news. This is obtained by checking if there exists a rule where the user exists in a rule, and all the conditions of the rule are met i.e., all the antecedents present in the rule have already propagated the news post before the user.

– If the user was biased during propagation then the value is incremented accordingly based on whether it is positive/ negative.

$$(sign = +)\ and\ (user\ is\ biased) \implies = x * 1.2 \tag{7}$$

$$(sign = -)\ and\ (user\ is\ biased) \implies = x * 0.8 \tag{8}$$

Upon sampling on the values, by adjusting the values to attain the least error rate, the proportionality constant is set to 1.2 for the authentic user who is biased and a value of 0.8 for the less authentic biased user. Once the news label is determined for the new post, the authenticity scores of all the users involved in propagating the particular news should be updated. This way, the cycle can continue in the network, identifying low-authentic accounts and posts.

It should be noted that this value keeps on changing as a new interaction between user and news post occurs. However, it is to be kept in mind that a new interaction's effect on the news score keeps reducing with time. Figure 6 depicts the Workflow of the SENAD model, and a similar Workflow applies in identifying Fake users.

3.1.7 Results

Initially, the SENAD model assumed a neutral value (zero) whenever a new user or post was created. Depending on further interactions, if the score is negative, it is treated as fake, and if the value is positive, it is treated as real. The entire system is normalized to a scale of 1 to 100 to make the system confined to a positive scale. Here value 50 is set to be neutral. On analyzing the datasets and varying the proportionality constants, it is observed that the scores for the posts lie mostly in the range of 40 to 60. Therefore, the ranges are fixed to be 40 and 60. Here values between 40 to 50 are treated as "likely to be fake," between 50 to 60 as "likely to be real," value below 40 is fake and greater than 50 as real.

All the accounts with a low score may not be bots, and some may be of users spreading biased views that blindly spread any news that seems to fit their interest without verification. In the case of news, however, based on labels, the score is divided into ranges that help categorize it. The following are the labels assigned:

– *Mostly fake :* News score < 40
– *Likely fake :* News score in the range 40–50
– *Likely genuine :* News score in the range 50–60 (50 inclusive)
– *Mostly genuine :* News score > 60

However, to study the procedure's efficiency, binary classification is implemented, where a score < 50 implies fake and a score > 50 implies true post. The dataset obtained contains only fake and authentic labels for news articles.

There might arise cases where True news gets predicted as hoaxes or fake news (while aiming at higher accuracy). Therefore, while dealing with fake news detection, one should equally achieve good precision while maintaining good accuracy values.

The proposed algorithm was run using 5-fold cross-validation, where the dataset is divided into 80% for training and 20% for testing. Table 2 specifies the training and testing samples for the SENAD model.

Table 3 shows the performance of the proposed model and that of a few baseline models in the task of fake news detection on Twitter, PolitiFact, and GossipCop. State-of-art models like PPC_RNN+CNN, TriFN, and HPFN also work on the propagation patterns and the news content and interactions. These models use similar properties like follower's count, friend's count, length of user name, et cetera to detect fake posts. In addition, the TriFN and HPFN models have data scrapped from GossipCop and PolitiFact, as used in the proposed model.

3.2 CredNN

Reviewing the problem of fake news online and its widespread negative effects, it is found that the visual contents of such fake news significantly contribute to the spread of the news. This pattern has been attributed to the emotional impact and negative sentiments the image evokes in the viewer. By offering a perception of reality, the fake news image influences the viewer to share the news, leading to the fake news being spread across the online social network (Yang et al. 2018). Traditional forensic methods and most fake news work focus on fusing multiple modalities miss to

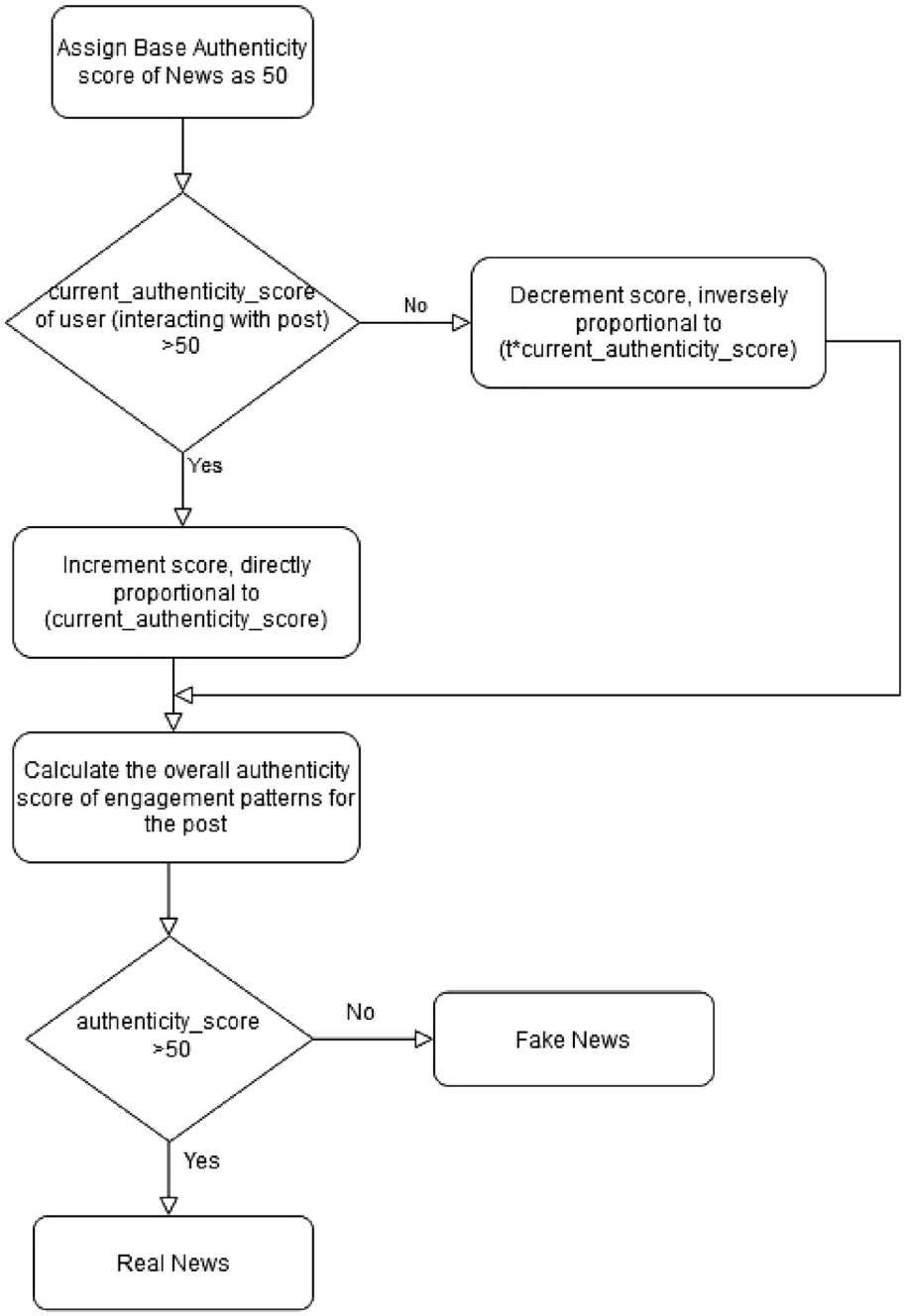

Fig. 6 SENAD model workflow diagram

Table 2 Statistics of dataset for training and testing

Samples	Training	Testing
GossipCop_Fake	4270	1070
GossipCop_Real	6100	1530
PolitiFact_Fake	380	96
PolitiFact_Real	630	162

Table 3 Performance of model compared to baseline models

Model	Accuracy	Precision	Recall	F1 score
PPC_RNN+CNN	86.3	83.0	81.0	82.0
TriFN	87.8	86.7	89.3	88.0
HPNF	86.1	85.4	86.9	86.2
SENAD	**93.7**	**92.6**	**95.0**	**93.7**

Bold values indicate the proposed model measures

model and capture visual characteristics unique to fake news images. Given the significant impact of visual information in fake news posts, the objective is to develop a framework focusing on the visual modality and model features unique to fake news images to identify such images in online social networks.

3.2.1 Challenges

In this pursuit, the initial challenges identified in developing such a framework are as follows:

- Obtaining a large dataset that contains the ground truth labels of posts (fake or real) is complex since manual fact-checking is exceptionally time-consuming.
- Fake images are of two kinds—tampered and misleading, and capturing features for both variants of fake images from genuine ones is challenging because fake images are deliberately made very 'believable,' i.e., they pretend to be genuine.
- For tampered fake images, the original genuine version is unavailable; hence there will not be any reference image to compare with the potentially fake image and identify alterations and manipulations.
- Similarly, it is to be analyzed if an image is misleading without any information on the original scenario or context in which it was used.
- Even authentic images may contain physical alterations (such as changing brightness or contrast, resizing and re-scaling the image, filters to sharpen or blur the image), which are not malicious and misleading. Hence, not all physically altered images can be 'fake.'
- Images uploaded online are often of varying sizes and quality, originating from different camera models. Thus, the framework developed needs to be robust to these variations.
- New events and trends emerge daily on OSNs, and images used in these platforms have their visual features closely tied with the event. Thus, the framework needs to be generalizable to new emerging events never seen earlier.

3.2.2 Scope for transfer learning

The Convolutional Neural Networks (CNN) are greatly favored for extracting characteristics of images because of their spatial properties. They are well suited to learning complicated semantics in fake images. They typically have several convolutional and fully connected layers, requiring millions of parameters to be learned during the supervised training process. This mandates a large labeled set of training information (Ranganathan 2021). A major challenge in fake news detection is the unavailability of a large labeled dataset of fake and real posts online. Each of the posts is to be fact-checked to assign ground truth labels, which requires a lot more information and time.

Hence, transfer learning is a viable alternative. Transfer learning is a machine learning technique by which a model which performs a particular task is tuned to perform another related task. Networks like VGG16, MobileNet, GoogLeNet perform exceedingly well on the ImageNet dataset, with over 1000 classes. The number of parameters in these models is in the order of millions. Therefore, these pre-trained weights can learn additional features for fake image efficient classification with a limited labeled dataset.

3.2.3 Benchmark datasets used for fake image detection

Two widely used fake news multimedia datasets are available—one collected from Twitter (Boididou et al. 2015) and another from a popular Chinese social network called Weibo, built-in (Jin et al. 2017. However, owing to a large number of duplicates in the Twitter dataset, the number of distinct images are about 500, thus making the dataset too small to use for neural network training.

Therefore, like in Qi et al. (2019), only the dataset collected from Weibo (Jin et al. 2017) is used to evaluate the proposed model. In this dataset, the fake news posts are crawled from May 2012 to January 2016. These posts were verified by the official rumor debunking system of Weibo. Weibo dataset has images which are flagged as true, verified by Xinhua news agency (Ye et al. 2021).

As done in previous work (Jin et al. 2017; Khattar et al. 2019; Qi et al. 2019; Wang et al. 2018), hashing algorithms were explored to identify nearly close images. The perceptual hashing (pHash) algorithm was used to remove the near-duplicated images in this dataset. This algorithm matches images with variations in measures like scaling, aspect ratios, minor coloring differences, which are matched as similar images. An example of two similar images identified by this technique is illustrated in Fig. 7. Figure 7a has an original sample image of the currency notes, and Fig. 7b has a set of images of currency notes that have been changed concerning the color of the notes and their denomination. pHash algorithm, when employed on these images, has been matched as similar images.

After removing duplicates and nearly similar images, the dataset is split into training, validation, and test sets in the ratio of 7:1:2, as in Wang et al. (2018), Qi et al. (2019). The dataset contains 9708 images in total, in .jpg format. The statistics of the dataset are stated in Table 4.

Most real images tend to be of good quality. From the analysis of the dataset, it is observed that re-compression artifacts are more prominent in fake news images than in real images. This pattern is because fake images tend to be

Fig. 7 Fake images identified as similar through pHash

(a) Original Image of Currency notes **(b)** Fabricated images with respect to color and denomination of notes

Table 4 Statistics of Weibo dataset

Samples	Training	Validation	Testing
Fake Images	3397	485	972
Real Images	3397	485	972
Total	6794	970	1944

uploaded and downloaded several times on social network platforms.

3.2.4 Error level analysis

Both VGG16 and VGG19 architectures pretrained on ImageNet were fine-tuned with the dataset for initial experiments. While fine-tuning VGG19 yields better precision scores (~70%) when compared to that of VGG16 (~67%), the recall scores for fake images for both designs are almost equal(~66%). In the pursuit of improving the model's performance and increasing the recall of fake images, a change of CNN architecture though useful, did not bring about significant improvements. One way to deal with this issue is to transform the input images to the model in a way that aids in picking up distinguishing features between fake and real images. This analysis can help to improve the recall scores of fake images. To achieve this, Error Level Analysis was explored.

Error Level Analysis (ELA) is a forensic technique that analyses compression artifacts in images compressed with lossy techniques such as JPEG. It helps identify regions in the image which have different compression levels. Certain sections of the image may have significantly different error levels because those sections may have been subjected to the same lossy compression a different number of times or a different type of lossy compression. Therefore, a difference in error level in different sections of an image indicates that it is likely that the image is edited. ELA works by intentionally resaving the image at a known error level (e.g., at 90%) and computing the difference between the images. The error level potential is lowered with each resave, resulting in a 'darker' ELA result. If there is no change that means that the cell has reached the local minima for error at that quality level. If the picture is modified, the regions with no other error (stable) become unstable because of those alterations (Krawetz and Solutions 2007).

Figure 8 shows an original, unedited image captured by a smartphone camera. The figure is resaved at 95% quality, and its Error Level Analysis shows that most colors have been compressed well. (Darker ELA values mean lower error levels). Figure 9 shows an altered version of original image in Fig. 8—some books have been copied. The corresponding Error Level Analysis shows higher error levels at tampered regions (Jeronymo et al. 2017).

3.2.5 Transfer learning with ELA

Error Level Analysis(ELA) helps identify digitally altered images since the error levels throughout such images are not uniform. ELA outputs (the image which is the difference between the image and its resaved version) are computed for each image in the entire dataset (resaved at 90% quality) and used for performing transfer learning on VGG16 architecture. Only the convolutional part of VGG16 is instantiated, and the model is run on the training and validation data once. The 'bottleneck' features (those before the fully connected layers) are recorded. Now, the recorded outputs are used to train the fully connected classifier. This classifier component (with trained weights) is then loaded on top of the convolutional part of VGG16(with pre-trained weights from ImageNet), and the model is fine-tuned with the ELA

Fig. 8 Original image and its ELA output

Fig. 9 Edited image and its ELA output

Table 5 Transfer learning with ELA

Samples	Precision (%)	Recall (%)	F1-score (%)
Fake	67	79	73
Real	75	62	68

outputs of the fake image dataset using the Stochastic Gradient Descent algorithm for optimizing the loss. The classification scores are presented in Table 5.

It is observed that most fake images are correctly identified as fake. It is also observed that transfer learning with ELA outputs of images performs significantly better than that done previously on VGG19 without such transformations (with about +6% higher overall precision and +13% more excellent recall on fake images). These results further validate the initial hypothesis that using ELA outputs of images for transfer learning leads to the learning of better distinguishing features between fake and authentic images.

3.2.6 Enhancing model's generalisability

One of the significant challenges of a fake image detection system is performing efficiently on upcoming events and trends. For learning event-invariant features during the transfer learning process, the dataset needs to be split into train, validation, and test sets to ensure no event overlap between them. Likewise, splitting a dataset is challenging because the images are labeled only to signify fake or real and not inherently grouped by their event. In order to learn

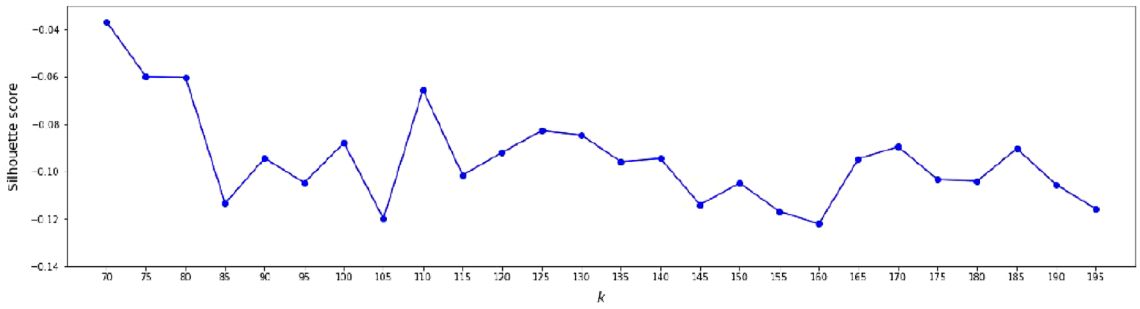

Fig. 10 Silhouette score-clusters(k)

Event-invariant features using transfer learning, the dataset needs to be split into the train, validation, and test sets. Proper care is taken to ensure that no event overlap between them, i.e., images belonging to the same event (e.g., earthquake in a city) do not occur in two or more sets.

There is also no information about the number of events represented by the images in the dataset, and manual grouping is ineffective for a dataset of 9.7k images. The MVNN model (Qi et al. 2019) from the literature splits the data into 200 clusters, using K-means algorithm. For grouping the images into events, MiniBatch K-means is run, which is less computationally expensive when compared to traditional K-means, with the quality of the results only slightly lower than from K-Means. Small images are removed to maintain the quality of the dataset.

MiniBatch k-means is tried with values less than 200. Silhouette score is calculated for different values of k, and it is found that the score is high for 70 clusters (score of −0.0367) compared to 200. A decreasing trend of the score is observed below 70 clusters. Therefore, the optimal number of clusters is set to be 70. The dataset is then split into training, validation, and test sets in the ratio 7:1:2. This processed version of the dataset is referred to as Dataset 2.0 and that used previously as Dataset 1.0 (Fig. 10).

Even though there is no prior information about which images correspond to a particular event or the total number of events, the techniques employed help arrive at optimal groups of images and split the data into training and evaluation sets without any event overlapping. This aspect promises better generalisability of the model in real-world settings.

3.2.7 Augmenting transfer learning on ELA outputs with visual sentiment analysis

Previously, Error Level Analysis (ELA) was performed on all images in the dataset before using them for transfer learning. This analysis improved the identification of fake images since edits made to an image are reflected in a higher error level rate in the ELA analysis. However, it must be noted that ELA analysis is not tailored to bring out cues to identify misleading images (which are also a type of fake images) since these types of images do not contain tampering or alterations.

Table 6 Statistics of Twitter dataset

Samples	Training	Validation
Negative Images	241	60
positive Images	466	115
Total	707	175

Table 7 VGG16 for sentiment analysis

	Precision (%)	Recall (%)	F1-score (%)
Negative	83	72	77
positive	86	92	89

Since fake images, both tampered with and misleading, tend to evoke strong negative sentiments, analyzing the polarity of the image could be helpful.

In order to learn features to analyze the sentiment of an image, transfer learning on VGG16 (a similar procedure followed for ELA images) is done with a dataset containing positive and negative images collected from Twitter by authors of past work in visual sentiment analysis (You et al. 2015). The statistics of the dataset are described in Table 6. This data are used in the training and validation of the CredNN model.

3.2.8 Transfer learning for visual sentiment analysis

As done previously for transfer learning from ELA outputs, the VGG16 architecture is fine-tuned on the Twitter dataset mentioned above, and the evaluation results as shown in Table 7. It is observed that the model has high precision in identifying positive and negative images. As most fake images reflect negative sentiments, these results motivate the use of visual sentiment analysis in identifying fake images

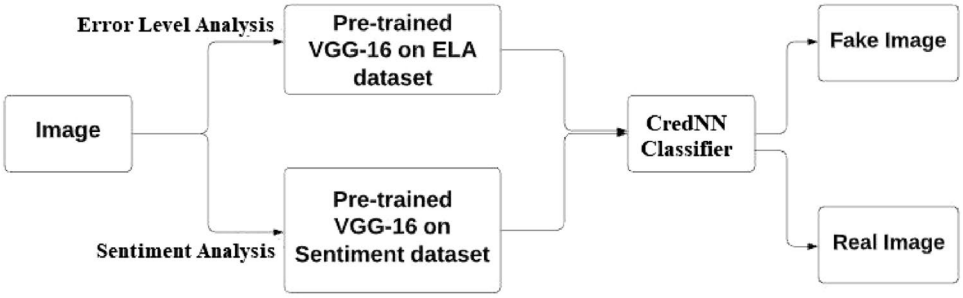

Fig. 11 Proposed CredNN classifier

(Zaeem et al. 2020). This analysis can also help improve the confidence that an image is flagged as fake. The same pattern is observed even for tampered images with alterations and high visual impact (e.g., depicting bloodshed, extreme pain, et cetera).

3.2.9 CredNN architecture

A new architecture, named CredNN—Credibility Neural Network consisting of two CNN sub-networks, is introduced to identify fake images. CredNN is an ensemble construction of CNN architectures—with ELA technique to help identify images high digital alterations, and visual sentiment analysis to learn features that distinguish an image with negative sentiment from that which induces positive emotions—thereby identify misleading and also tampered fake images with high confidence. The design of CredNN is as depicted in Fig. 11. The fine-tuned VGG16 architectures trained on ELA images and that trained on visual sentiment analysis datasets are used for this construction. The ELA sub-network consists of the layers (all excluding final classification layers) from the VGG16 model fine-tuned on ELA images, and similarly, the visual sentiment analysis sub-network consists of layers from the model fine-tuned on the sentiment analysis dataset.

3.2.10 Results

CredNN is loaded with the best weights obtained from 2 models which were trained independently—fine-tuning VGG16 on ELA images on Dataset 2.0 and fine-tuning VGG16 for sentiment analysis. CredNN is fine-tuned again on Dataset 2.0. As input for CredNN classifier, one branch consists of the layers from the VGG16 model fine-tuned on ELA images, and the other fine-tuned on the sentiment analysis dataset. All layers from these models except the last classification layers are included in this ensemble construction, called CredNN.

CredNN is run in five phases, with Adam optimizer, a learning rate of 0.001, and run for different epochs. The proposed CredNN uses *concatenate* to combine features from both branches and dense layers to decrease the validation loss. The best results are obtained when the model is run for 20 epochs and with a learning rate of 0.0001. The evaluation scores CredNN achieves on the test set of Dataset 2.0 is as depicted in Table 8.

Table 8 CredNN on dataset 2.0

	Accuracy	Precision	Recall	F1-Score
CredNN	**76.28%**	**78.02%**	**78.38%**	**78.21%**

Bold values indicate the proposed model measures

Table 9 Evaluation of CredNN model

	Precision (%)	Recall (%)	F1-score (%)
Fake	74	74	74
Real	78	78	78

CredNN achieves an accuracy of 76.3% and about 78% on precision, recall, and F1-scores. Further evaluation scores are in Table 9.

The evaluation scores obtained from CredNN are comparable to the state-of-the-art, with the evaluation scores reported on similar fake images datasets collected from Weibo (Qi et al. 2019) as shown in Table 10.

CredNN has a higher margin of ∼**7**% on F1-scores of pre-trained VGG, fine-tuned VGG, and autoencoder-based baselines, while CNN-RNN-based MVNN achieves 83.2% on the similar dataset. CredNN's lower margin with MVNN could be because of the additional steps taken in CredNN to ensure no event overlap and inclusion of visual sentiment of the images. The construction leverages the novelty of both types of analysis to make better predictions on the 'fakeness'/genuineness of the images.

4 Inferences

The following are the inferences drawn after analyzing the results obtained by applying SENAD on the FakeNewsNet and Twitter dataset.

Table 10 Results of other methods on similar Weibo dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Pre-trained VGG	72.1	66.9	73.8	70.2
Fine-tuned VGG	75.4	74	68.9	71.4
ConvAE (Masci et al. 2011)	73.4	68.5	74.4	71.3
MVNN (Qi et al. 2019)	84.6	80.9	85.7	83.2

- One prominent advantage of this method is that it is dependent on the users propagating it. The more the users involved, the better the accuracy achieved, and minimal to zero users leads to an ambiguous result.

  However, this is acceptable when dealing with the negative impacts of fake news in the community. This impact is observed when the post has spread through a large part of the community. The impact is not considered when the number of users that interacted with the article is low, as its impact on the community is minimalism.
- It is observed that scoring the news articles with base authenticity score alone was less efficient than training the account authenticity with news articles first and then training news articles score. It is inferred that the account parameters alone are insufficient to determine the news authenticity.
- While analyzing the news scores, it is observed that most of the real news articles were in the likely genuine category. This pattern is because the spread of real news takes longer to start spreading, and since our delta score is inversely proportional to time, the existing score becomes lower.
- While analyzing the wrongly classified true articles, it is observed that the articles are spread only during early stages and had very minimal interaction later, essentially putting them in an ambiguous category.

The following inferences were made for CredNN with a fake image dataset (Weibo dataset):

- In CredNN, additional steps were taken to process the dataset to find an optimal number of clusters to ensure no event overlap; that is, images from the same event do not occur in both training and test sets. Though this may have led to lower evaluation scores on validation and test sets, the generalisability of the model to new events is improved.
- Additionally, the dataset used for sentiment analysis is from Twitter, different from the fake images dataset collected from Weibo and much smaller. More appropriate sentiment cues can be learned if the sentiment analysis dataset is from the same platform (Weibo) and larger in size.
- As future work, visual sentiment analysis dataset can be created by obtaining labels on the fake images, and CredNN can be fine-tuned accordingly. This analysis is much less time-consuming than the fact-checking required for larger fake image datasets. Amazon Mechanical Turk is one crowd sourcing platform that has images tagged with its sentiment.

5 Error analysis

The error rate is calculated from the confusion matrix upon analyzing the misclassifications. As the proposed models SENAD and CredNN deal with identifying Fake Posts/Users/Images, identifying the samples to be Fake will be treated as "Positive cases" and identifying samples as Real to be "Negative cases". From the confusion matrix, samples that are either falsely classified as "Positive" (not Fake) or falsely classified as "Negative" (Fake) are used in calculating the error rate.

In the SENAD model, users and posts are identified as fake depending on the interaction patterns. There is a scope of misclassification if the article has fewer engagement patterns. It was observed that scoring the news articles with base authenticity score alone was less efficient than training the account authenticity with news articles first and then training news articles score. We can infer that the account parameters alone are insufficient to determine the news authenticity. While analyzing the wrongly classified true articles, it is observed that these articles were spread only during the early stages and had minimal interaction later, essentially putting them in an ambiguous category. The error rate of the proposed model is obtained to be 0.063.

In the case of the CredNN model, initially, VGG-16 is finetuned to work with the image dataset. Here it is observed that VGG-16 recorded an error rate of 0.33. At the later state, VGG-16 is employed on ELA images, and it is observed that the error rate is 0.294. The same model, working on the visual sentiment data, recorded an error rate of 0.148. Finally, the proposed CredNN (ensemble of both ELA and Visual Sentiment data) recorded an error rate of 0.237. As ELA works with tampered images, it is tedious to identify the intentionally tampered images that are compressed, downloaded, and uploaded multiple

times. Finding an image dataset with visual sentiment data tagged to it is also challenging.

6 Conclusion and future work

This paper proposes a method of identifying *fake news* and *fake images* on OSNs. Fake account detection is also used as a metric for identifying the user engagement patterns by scoring the users and articles. This paper proposes ways to work with fake images, which are now used as the main tool for propagating fake news. User engagements are studied to understand how user engagements can help determine the authenticity of news on social media. We bring in new features such as the coreness of a node, responsibility of an account, and user bias, which are based on user behavior. Base authenticity score, the user's bias, is used as a metric for identifying fake accounts and fake news. The proposed SENAD method has achieved an accuracy of 93.5% and with higher precision and recall rate. For fake news image detection, features unique to fake news images are modeled by a combination of forensic techniques and image polarity analysis used to train Convolutional Neural Networks to capture object level and even semantic level details from the images. An enhanced transfer learning approach is proposed, which leverages the capabilities of Error Level Analysis (ELA) in identifying tampered fake images. This analysis significantly improves the recall of fake images, with a score of 79% when compared to recall values of nearly 65% obtained from the previous designs. The paper proposes CredNN (Credibility Neural Network), an ensemble construction of CNN architectures trained to capture features based on Error Level and Visual Sentiment Analysis. The proposed CredNN model achieves an accuracy of 76.3% and about 78% on precision, recall, and F1-scores on the dataset processed specifically for better generalisability. As a part of future study, the plan is to broaden the scope of bias from binary variable to variable bounded by range based on the confidence of the rules involved. Even user bias concerning the topics and publishers can be determined to improve the efficiency of fake news and fake account detection. The plan is to fine-tune the CredNN model and the number of convolutional blocks for fake image detection to improve computational efficiency in identifying fake/real images. Also, the plan is to explore attention mechanisms to ensure that different visual features are appropriately highlighted. The research is now focused on improving the performance measures with NLP level analysis of user posts such as POS Tagging, Name Entity Recognition, and powerful deep learning models for fake image detection.

References

ABP News Bureau (2020) PM-CARES fund fraud operating on facebook exposed, retrieved from https://news.abplive.com/news/india/fake-facebook-account-under-pm-modis-name-asking-for-relief-funds-1195373

Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI (1996) Fast discovery of association rules. Adv Knowl Discov Data Min 12(1):307–328

Akyon FC, Kalfaoglu E (2019) Instagram fake and automated account detection. In: 2019 Innovations in intelligent systems and applications conference (ASYU), pp 1–7

Baptista JP, Gradim A (2020) Understanding fake news consumption: a review. Soc Sci 9(10):185

Bayar B, Stamm MC (2016) A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM workshop on information hiding and multimedia security, pp 5–10

Boididou C, Andreadou K, Papadopoulos S, Dang-Nguyen DT, Boato G, Riegler M, Kompatsiaris Y et al (2015) Verifying multimedia use at mediaeval. MediaEval 3(3):7

Boshmaf Y, Logothetis D, Siganos G, Lería J, Lorenzo J, Ripeanu M, Beznosov K, Halawa H (2016) Integro: Leveraging victim prediction for robust fake account detection in large scale OSNs. Comput Secur 61:142–168

Brain Dean (2021) Statistics: How many people use Social Media in 2021? Retrieved from Backlinko, https://backlinko.com/social-media-users

Calisir V (2021) Disinformation, post-truth, and naive realism in COVID-19: melting the truth. In: Handbook of research on representing health and medicine in modern media. IGI Global, pp 200–215

Castillo C, Mendoza M, Poblete B (2013) Predicting information credibility in time-sensitive social media. Internet Res Electron Netw Appl Policy 23:560–588

Chang AB (2021) Using machine learning models to detect fake news, bots, and rumors on social media. Thesis and Dissertations, Arizona State University, Library

Dordevic M, Pourghomi P, Safieddine F (2020) Identifying fake news from the variables that governs the spread of fake news. In: 2020 15th international workshop on semantic and social media adaptation and personalization (SMA). IEEE, pp 1–6

Fernando T, Fookes C, Denman S, Sridharan S (2020) Detection of fake and fraudulent faces via neural memory networks. IEEE Trans Inf Forensics Secur 16:1973–1988

Galan-Garcia P, de la Puerta JG, Gomez CL, Santos I, Garcia Bringas P (2016) Supervised machine learning for the detection of troll profiles in Twitter social network: application to a real case of cyber bullying. Log J IGPL 24(1):42–53

Giatsidis C, Thilikos DM, Vazirgiannis M (2011) Evaluating cooperation in communities with the k-core structure. In: 2011 International conference on advances in social networks analysis and mining. IEEE, pp 87–93

Granik M, Mesyura V (2017) Fake news detection using Naive Bayes classifier. In: 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON), pp 900–903

Gupta A, Lamba H, Kumaraguru P (2013) \$1.00 per rt# bostonmarathon# prayforboston: Analyzing fake content on twitter. In: 2013 APWG eCrime researchers summit. IEEE, pp 1–12

Hamdan YB (2020) Faultless decision making for false information in online: a systematic approach. J Soft Comput Paradigm 2(04):226–235

Jamieson K, Cappella J (2008) Rush Limbaugh and the Conservative Media Establishment, Echo Chamber

Jeronymo DC, Borges YCC, dos Santos Coelho L (2017) Image forgery detection by semi-automatic wavelet soft-thresholding with error level analysis. Expert Syst Appl 85:348–356

Jia J, Wang B, Gong NZ (2017) Random walk based fake account detection in online social networks. In: 2017 47th annual IEEE/IFIP international conference on dependable systems and networks (DSN), pp 273–284

Jin Z, Cao J, Zhang Y, Zhou J, Tian Q (2016) Novel visual and statistical image features for microblogs news verification. IEEE Trans Multimed 19(3):598–608

Jin Z, Cao J, Guo H, Zhang Y, Luo J (2017) Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM international conference on Multimedia, pp 795–816

Jin Z, Cao J, Luo J, Zhang Y (2016) Image credibility analysis with effective domain transferred deep networks, ArXiv arXiv:abs/1611.05328

Khaled S, El-Tazi N, Mokhtar HMO (2018) Detecting fake accounts on social media. In: 2018 IEEE international conference on Big Data (Big Data), pp 3672–3681

Khalil HA, Maged SA (2021) Deepfakes creation and detection using deep learning. In: 2021 International mobile, intelligent, and ubiquitous computing conference (MIUCC). IEEE, pp 1–4

Khattar D, Goud JS, Gupta M, Varma V (2019) Mvae: multimodal variational autoencoder for fake news detection. In: The world wide web conference, pp 2915–2921

Kondeti P, Yerramreddy LP, Pradhan A, Swain G (2021) Fake account detection using machine learning. In: Evolutionary computing and mobile sustainable networks. Springer, Singapore, pp 791–802

Krawetz N, Solutions HF (2007) A picture's worth, Hacker Factor. Solutions 6(2):2

Kwon S, Cha M, Jung K (2017) Rumor detection over varying time windows. PLoS ONE 12(1):1–19

Lang PJ (1979) A bio-informational theory of emotional imagery. Psychophysiology 16(6):495–512

Liu Y, Ji S, Mittal P (2016) Smartwalk: enhancing social network security via adaptive random walks. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, CCS'16. Association for Computing Machinery, New York, NY, USA, pp 492–503

Liu Y, Wu YB (2018) Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. AAAI Press, pp 354–361

Lu Y-J, Li C-T (2020) GCAN: graph-aware co-attention networks for explainable fake news detection on social media. arXiv preprint arXiv:2004.11648

Masciari E, Moscato V, Picariello A, Sperlí G (2020) Detecting fake news by image analysis. In: Proceedings of the 24th symposium on international database engineering & Applications, pp 1–5

Masci J, Meier U, Cireşan D, Schmidhuber J (2011) Stacked convolutional auto-encoders for hierarchical feature extraction. In: International conference on artificial neural networks. Springer, pp 52–59

Monti F, Frasca F, Eynard D, Mannion D, Bronstein MM (2019) Fake news detection on social media using geometric deep learning. CoRR arXiv:abs/1902.06673

Nickerson RS (1998) Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol 2(2):175–220

Qi P, Cao J, Yang T, Guo J, Li J (2019) Exploiting multi-domain visual information for fake news detection. In: 2019 IEEE international conference on data mining (ICDM), pp 518–527

Ranganathan G (2021) A study to find facts behind preprocessing on deep learning algorithms. J Innov Image Process 3(01):66–74

Sahoo DPK, Lavanya K (2019) Identification of malicious accounts in Facebook. Int J Eng Adv Technol 9(1):2917–2921. https://doi.org/10.35940/ijeat.A1228.109119

Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor Newsl 19(1):22–36

Shu K, Mahudeswaran D, Wang S, Lee D, Liu H (2018) Fakenewsnet: a data repository with news content, social context and dynamic information for studying fake news on social media, arXiv preprint arXiv:1809.01286

Shu K, Mahudeswaran D, Wang S, Liu H (2020) Hierarchical propagation networks for fake news detection: investigation and exploitation. In: Proceedings of the international AAAI conference on web and social media, vol 14, pp 626–637

Shu K, Wang S, Liu H (2017) Exploiting tri-relationship for fake news detection. Association for the Advancement of Artificial Intelligence, arXiv preprint arXiv:1712.07709

Shu K, Wang S, Liu H (2019) Beyond news contents: the role of social context for fake news detection. In: Proceedings of the twelfth ACM international conference on web search and data mining, pp 312–320

Tacchini E, Ballarin G, Vedova MLD, Moret S, de Alfaro L (2017) Some like it hoax: automated fake news detection in social networks. CoRR arXiv:abs/1704.07506

Viswanath B, Bashir MA, Crovella M, Guha S, Gummadi KP, Krishnamurthy B, Mislove A (2014) Towards detecting anomalous user behavior in online social networks. In: Proceedings of the 23rd USENIX conference on security symposium, SEC'14. USENIX Association, USA, pp 223–238

Wang G, Konolige T, Wilson C, Wang X, Zheng H, Zhao BY (2013) You are how you click: clickstream analysis for sybil detection. In: Proceedings of the 22nd USENIX conference on security, SEC'13. USENIX Association, USA, pp 241–256

Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, Su L, Gao J (2018) Eann: Event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp 849–857

Wynne HE, Wint ZZ (2019) Content-based fake news detection using n-gram models. In: Proceedings of the 21st international conference on information integration and web-based applications & services, iiWAS2019. Association for Computing Machinery, New York, NY, USA, pp 669–673

Xiao C, Freeman DM, Hwa T (2015) Detecting clusters of fake accounts in online social networks. In: Proceedings of the 8th ACM workshop on artificial intelligence and security, AISec'15. Association for Computing Machinery, New York, NY, USA, pp 91–101

Yang S, Shu K, Wang S, Gu R, Wu F, Liu H (2019) Unsupervised fake news detection on social media: a generative approach. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, no. 01, pp 5644–5651

Yang Y, Zheng L, Zhang J, Cui Q, Li Z, Yu PS (2018) TI-CNN: convolutional neural networks for fake news detection. arXiv preprint arXiv:1806.00749

Ye W, Liu Z, Pan L (2021) Who are the celebrities? Identifying vital users on Sina Weibo microblogging network. Knowl Based Syst 231:107438

You Q, Luo J, Jin H, Yang J (2015) Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: Twenty-ninth AAAI conference on artificial intelligence

Zaeem RN, Li C, Barber KS (2020) On sentiment of online fake news. In: 2020 IEEE/ACM international conference on

advances in social networks analysis and mining (ASONAM). IEEE, pp 760–767
Zhang J, Luximon Y (2021) Interaction design for security based on social context. Int J Hum Compu Stud 154:102675
Zhou X, Zafarani R (2020) A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv 53(5):1–40
Zhou X, Zafarani R, Shu K, Liu H (2019) Fake news: fundamental theories, detection strategies and challenges, In: Proceedings of the twelfth ACM international conference on web search and data mining, WSDM'19. Association for Computing Machinery, New York, NY, USA, pp 836–837