Do not 'fake it till you make it'! Synopsis of trending fake news detection methodologies using Deep Learning

Rishabh Misra* and Jigyasa Grover*

Twitter, Inc.
{rmisra,jgrover}@twitter.com

Abstract. The modern bloom of social media has propelled a new pattern of information propagation termed *push journalism*, where a certain piece of news is shoved in the faces of as many people as possible with a sliver of hope that it will reach the people who need that information the most. This form of news reporting, especially via social media campaigns has boosted the access and fabrication of bogus reporting, or what is referred to as *fake news*. Fake news, in the form of clickbait, hoax, satire, propaganda, hyperpartisan, deepfakes, or simply unreliable news has the power of influencing its readers to a dangerous extent, predominantly causing political, socio-economic, or psychological harm. In this chapter, we analyze the meaning of fake news in the world of social media, the various forms it can take, what causes its spread, and what are the rudimentary signs of such fake news. We will walk through a comparative study of the state-of-the-art deep learning models to approach the tasks of identifying phony information, verifying the validity of various claims and facts, catching fake content, and so on. The exposition will especially elucidate the adversarial approaches in deep learning to detect counterfeit content that could come in any form like text, images, videos, or audio. In doing so, we establish the importance of generating plausible and understandable explanations for model predictions with a special emphasis on algorithm fairness. With the fact that deep learning methods rely on comparatively larger datasets of top-notch quality, this chapter will also highlight the availability of relevant datasets in this space, as well as share pointers to curate one if needed. Even with sufficient data, however, detection problems in this domain are especially challenging since spammers and fake content generators are working tirelessly to evolve their strategies in parallel to the advancement in detection mechanisms. We will further shed some light on some recent and upcoming trends from the aspect of fake news contributors, and critically evaluate how our current state-of-the-art deep learning techniques fare against those. In closing, we will leave readers with some thoughts on future directions for the development of better and smarter fake news detectors.

* both the authors contributed equally

1 Introduction

The emergence and accelerating growth of social media has altered the human interaction style in tremendous ways. Boosted with the advent of high-speed internet and smart devices, real-time communication has reached a new pedestal involving more and more people every second. Along with its impact on breaking physical barriers for speedy communication, amplifying commerce via promotions and advertisements, enhancing entertainment options, augmenting educational and professional growth, and transforming many more vital applications it has had a crucial effect on the journalism domain. In partnership with social media, a new pattern termed *push journalism* has been garnering interest wherein information tidbits are deliberately propagated to the entirety of the population with an anticipation that it would ultimately converge on the targeted audience.

Oftentimes, in this process, bogus reporting takes over and misinformation or false information is fabricated and disseminated to the masses. Owing to social media, this proliferation is as effortless and rapid as it can get. The inaccurate information that is specifically concocted with an intent to manipulate or deceive people online is called *fake news*. Fake news is a misleading piece of information that is falsely constructed with no relevance to reality, containing no verifiable facts and no reference to credible sources or quotes. It has the potential to influence the readers, and cause reputational damage, thus causing political, socio-economic, and psychological harm. Though fake news is a very subjective and sensitive topic, it should not be confused with pieces of information that do not align with our views.

Fake news can be manifested in various forms, be it a rumor, clickbait, hoax, satire, propaganda, hyperpartisan, or the modern-day deepfakes. Any piece of unreliable or uncertain news published with no verified sources or fact-checking is called a *rumor*. A *clickbait* is an eye-catching, sensationalized piece of content, like an image preview, or a quirky title, created with an aim to garner attention and lure people to a particular web page with little relevance or no meaningful content. *Hoax* is a falsely curated piece of information purposefully made to pass it off as the truth and is different from jokes due to their evil intention. *Satire* or *parody* is exaggerated, ironical content mostly created with an aim to humor people or take a dig at realistic situations, however, some uninformed readers might take it as gospel. Biased or misleading news fundamentally generated to manipulate people and further an agenda, for instance, to promote political ideologies, is termed as *propaganda*. On similar lines, content be it political or socio-economic claiming to be unbiased yet reeking of partial vibes is called *hyperpartisan*. And in contemporary times, we have synthetically generated media, images, or videos, which might or might not bear resemblance to an existing human being called *deepfakes*. This type of fake manipulative audio-visual content is created using deep learning, hence the name, and has the potential to create major havoc. These are just a few examples of the most common forms of fake news and are generally created by unprofessional journalists, people wanting to make money regardless of the content they push out, satirists who want to entertain, or partisans wanting to influence people.

Fig. 1: Snapshots from social media exhibiting misinformation, deepfake, and satire.

With social media becoming ubiquitous, it has become one of the major dispersing platforms for fake news. Since the news on the platform is scarcely targeted and is very subjective in nature, it is ingested by readers that appeals to their emotions. This leads to a manipulated reality for many people, especially the ones who have a hard time distinguishing it from the real news, thereby intensifying societal conflicts. It not only causes mistrust amongst the people, instigates violence but also distracts from the important issues which often are left unresolved due to this. Hence, it is very important to create a robust mechanism to detect and filter these misleading pieces of content from the real news.

Telltale signs of fake content are lack or misquoting of original sources, unknown (or in some cases imitating well-known) publishers or authors, phony websites or apps, poor language in terms of spellings or grammar, incoherent story, exaggerated image previews, etc. However, in this age of information overload, it is cumbersome and almost impossible to manually go over each piece

of content and scrutinize it. This is where the power of deep learning comes into play, to swiftly and accurately identify the counterfeit from factual content with minimal human intervention thereby mitigating the spread of inaccurate information.

In this chapter, we first formally define the problem and explore various contemporary deep learning techniques to handle fake news with due categorization into misinformation, clickbait, satire, and deepfakes. We further discuss the limitations of the deep learning methods along with a few insights into the fairness, interpretability, and accountability of the fake news detection algorithms. In the end, we leave some pointers for the readers regarding emerging trends in this domain.

2 Formal Problem Definition

Now that we are well-versed with the characterization and types of fake news doing rounds on social media, let us explore some ways to detect this as early as possible by plying deep learning techniques. However, before we explore the intricacies of the multiple approaches one can take, we should formally define the problem at hand.

Consider a piece of information, or a news article on social media, denoted by N. It is composed of two major elements, the *publisher details*, say P, and the *content*, say C_N. The *publisher details* comprise the domain where the information or the news article is published, author's name, profile, etc., and other publisher-related attributes. The *content* component is the actual piece of news involving the headline, text, images, videos, tags, and so on. Furthermore, we can signify the *engagements* over N on social media via E_N where E is a set of tuples $\{e_{it}\}$ denoting interaction of a user u_i with the given news article at a given time t and further propagating it via their post p_i. In that case, e_{it} can be represented as a unique combination of (u_i, p_i, t).

Hence, the problem can be formally defined as, given a news article N and all the related attributes like P_N, C_N, and E_N, the task is to predict whether the news article N is a fake news piece or not, i.e. $F : (P, C, E) \rightarrow \{0, 1\}$ such that, $F(n) = 1$, if n is a fake news article and $F(n) = 0$, if it is a genuine news article, where F is the prediction function we aim to learn.

That being, fake news detection is a straightforward case of binary classification problem in the machine learning world where the outcome of the prediction function is either 0, or 1. Or to say, the given piece of news can either be genuine or fake. Shu et al. [1] define fake news detection as a *distortion bias* based on previous research media bias theory which is usually modeled as a binary classification problem since fake news is nothing but a distortion bias introduced by the publisher in a world of genuinity.

3 Deep Learning Techniques For Fake News Detection

Being an emerging area of research, fake news detection is garnering a lot of traction and there is a continuous development of new tools and technologies to combat this social evil of bogus reporting. Past research and pragmatic studies have shown that deep learning techniques have been quite successful in creating a predictive model to identify fake news with high rates of accuracy. Let us walk through a few different approaches, with an attempt to categorize them according to the types of fake news they help tackle.

3.1 Misinformation

Social media is empowering novel forms of communication for global reach and along the way unleashing innovative journalism strategies. This in turn has accelerated the dispersion of false and inaccurate content, also referred to as *misinformation*. These fallacious pieces of news might be shared inadvertently due to lack of awareness or simply if the fabricated content is too convincing. However, if the same phony news is created and shared with malicious intent, it gets termed as *disinformation*. Misinformation online is a pressing public issue in political, socio-economical, and many other domains and is known to cause real-world consequences. Personalized ranking algorithms are further aggravating the issue by promoting this sort of misleading, sensational, and conspiratorial content. With tremendous amounts of data being churned every second, it is important to move on from human fact-checking mechanisms to automated artificially intelligent mechanisms to detect and adjourn the spread of misinformation online.

Advancing from knowledge-based detection mechanisms, style-based detection which analyzes the content of the news article is making headway. Zhou et al. [2] describe knowledge-based detection methods like the ones that flag fake news by cross-checking the knowledge dissipated in the given news article with facts. Whereas, style-based detection methods are the ones that focus on how the content is actually written. Here, we emphasize style-based detection methods since they assess the intention to spread misinformation.

Fusion of Neural Networks A benchmark study by Khan et al. [3] points to the superior performance of deep learning models over traditional machine learning models in detecting fake news, albeit requiring large-sized datasets. In that direction, Wang et al. [4] contributed the largest public dataset at the time, *LIAR*, which has politicians' statements along with the label of whether they were genuine or misinformation. In their work, they also propose a hybrid neural network framework to integrate both text and speaker metadata. The architecture has a convolutional layer to capture speaker metadata with standard max pooling on the latent space followed by a bidirectional LSTM layer. These embeddings are then concatenated with the max pooled textual representation from another CNN and then fed to a fully connected layer with a softmax activation function to generate the final prediction. In another instance, Singhania

et al. [5] propose a *Three-level Hierarchical Attention Network (3HAN)*, a level each for words, sentences, and the headline of a news article, thereby effectively representing the input news article, by processing the article in a hierarchical bottom-up manner. The experiment yields 96.77% accuracy when evaluated on a real-world dataset. The visualization of the attention layer helps with qualitative analysis and provides an insight that fake news articles use an inverted pyramid writing style (i.e. distributing information in decreasing importance).

Expanding Data Signals In an extensive linguistic analysis of fake news titled *'Truth of Varying Shades'*, Rashkin et al. [6] found that misinformation uses more first-person and second-person pronouns, more subjectives, superlatives, and modal adverbs, and less assertive words. Their work uses a simple LSTM model with pre-trained embeddings, however, the experiment demonstrates that crafted linguistic features provide a lot of value.

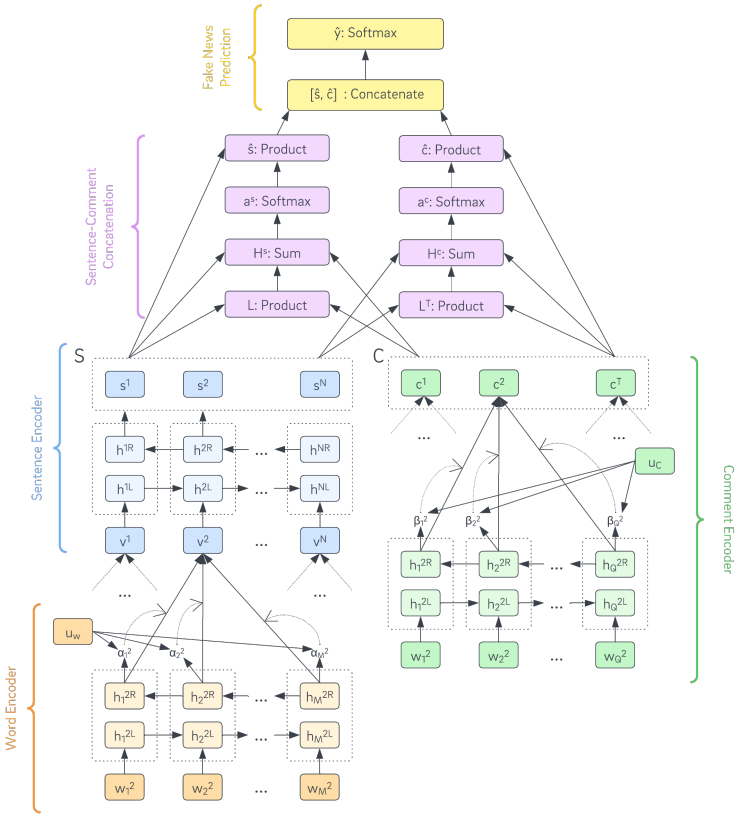

Fig. 2: dEFEND framework: news content encoder, user comment encoder, sentence-comment co-attention component, and fake news prediction component.

Based on the learnings from the past works, Shu et al. [7] in their study *'dEFEND'* reason that the explainability of a fake news detector is critical, and

how a user comments on news articles on social media can be utilized to facilitate that. The authors propose having a news content encoder (at the sentence level) as well as a user comments encoder, which are then fed to a co-attention sub-network to exploit both news content and user comments. The experiment setup indicates that this technique not only outperforms state-of-the-art fake news detection methods by at least 5.33% in F1 score but also (concurrently) identifies top-k user comments that explain why a news piece is fake.

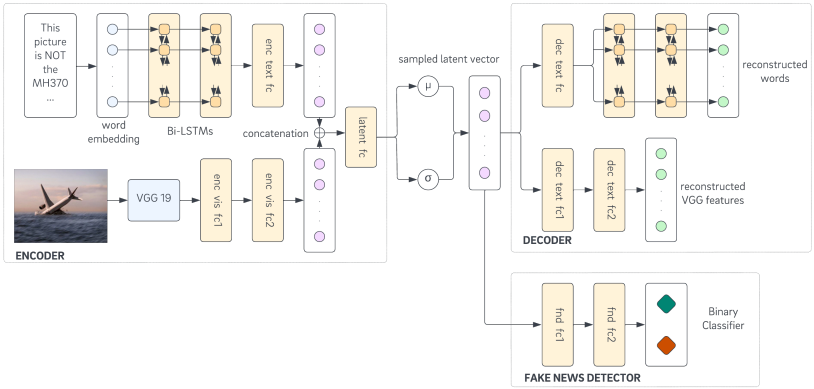

Fig. 3: Components of MVAE framework: Encoder, Decoder and Fake News Detector.

Additionally, Khattar et al. [8] further augment the data signals by using visual information along with textual information to detect misinformation. Their proposed framework, termed *Multimodal Variational Autoencoder (MVAE)*, uses a bi-modal variational auto-encoder to learn the representations from textual and visual data, which is then fed to a fully-connected feed-forward neural network for the task of fake news detection. This setup is evaluated on datasets from *Weibo* and *Twitter* and the results show that across the two datasets, on average this model outperforms state-of-the-art methods by $\sim 6\%$ in accuracy and $\sim 5\%$ in F1 scores.

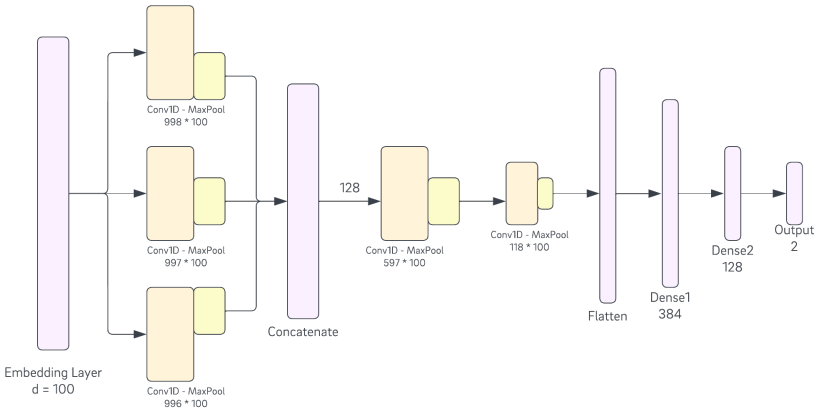

Fig. 4: FakeBERT Architecture.

Advancements with Language Models Presently, advanced pre-trained language models like BERT, ELECTRA, ELMo are receiving sizable attention for several natural language tasks including text classification, and rightly so, the misinformation detection domain has seen some work in this direction as well. Jwa et al. [9] propose using a BERT-based model to detect fake news that analyzes the relationship between the headline and the body of the news article. Since BERT is pre-trained on generic data, the authors incorporate news domain data for further fine-tuning and see a considerable improvement in the performance. The deep-contextualizing nature of BERT improves the F1 score by 0.14 over the previous state-of-the-art models. The *FakeBERT* model proposed by Kaliyar et. al. [10] further suggests improvement by feeding BERT representations to three parallel blocks of 1d-CNN having different kernel-sized convolutional layers with different filters for better feature extraction. The authors illustrate an accuracy of 98.90% which is a 4% improvement over the baseline approaches and is a promising direction for the fake news detectors' development.

3.2 Clickbait

In social media, *clickbait* is a text or a link with exaggerated or eye-catching headlines that lure a reader to *'click'* on it. Clickbait is a nuisance in the online user experience since it exploits a reader's curiosity and lures them to poor quality or inaccurate information. It is one of the most common types of fake news and is often used for advertisements, where mass content generators make money on phony content using a click-based model to optimize it. Clickbait takes advantage of some vulnerable users and abuses the purpose of *user-generated content*. This form of deception is considered a fraudulent activity on social media and is frowned upon in the news reporting circles since it leads to obstruction of real news propagation. Hence it is important to combat this category of fake news by detecting and flagging clickbaits early on in the network.

Post 2016 US presidential elections, this domain is getting a lot of attention, as experts debated that clickbait headlines on social media and other fake news might have influenced the decision making. Some of the earlier works [12] characterize clickbait using certain linguistic cues like *Unresolved Pronouns*, *Forward Referencing*, *Backward Referencing*, and *Action Words*. However, since headlines contain limited information only, it becomes challenging to encode these features in the model. Forging ahead, we dive into a few deep learning techniques that take advantage of the knowledge from past research and have proven to effectively detect clickbait.

Utilizing Textual Data In the past, CNNs have demonstrated effectiveness in various sentence classification tasks hence it is instinctive to try them out for the clickbait detection task as well. Contrary to the traditional machine learning approaches, CNN obviates the need to curate meaningful features, which might or might not be as helpful. One of the early works in this domain uses CNN

with pre-trained *word2vec* embeddings to extract meaningful information from the headlines of the news articles which is further used for prediction [13]. This model, termed *TextCNN*, performs remarkably well compared to its antecedent techniques, however, there is a huge room for improvement with respect to clickbait's domain.

In their attempt to optimize the TextCNN model for clickbait detection, Zheng et al. [14] recognize that different types of clickbait articles tend to use different ways to draw users' attention. For their approach, coined as *ClickbaitCNN*, the authors collect headlines from four famous Chinese news websites that fall into four article types: news, blogs, *BBSs*, and *WeChats*. They propose using a new word-embedding layer that takes both overall and the article type-related word meanings into consideration. They also propose a new loss function to regulate the influence of article type-related meaning of the word. It is found that employing these techniques improves the performance over TextCNN by 2-3% in terms of precision and recall.

Encoding Sequential Information Though CNN-based techniques have proven helpful for language-related tasks, such as this, they are limited by their nature since they can not leverage sequential information. The clickbait detection problem could benefit a lot from RNN especially if posed as a multi-class classification problem since they can encode sequential/contextual information well.

On these lines, Zhou et al. [15] propose a self-attentive RNN based model to infer the levels of importance of the text tokens in predicting clickbait. The data is sourced from *Twitter* and manually annotated as either *'not clickbaiting'*, *'slightly clickbaiting'*, *'considerably clickbaiting'*, or *'heavily clickbaiting'*. *GloVe* embeddings are used to effectively learn from the representations and dropout regularization is applied to the outputs of the word embedding layer, on the outputs of the *bidirectional Gated Recurrent Unit (biGRU)* encoding layer, as well as on the outputs of the self-attentive layer. Experimental runs indicate an improvement over the ClickbaitCNN by 4% in terms of F1 score with very low computational cost.

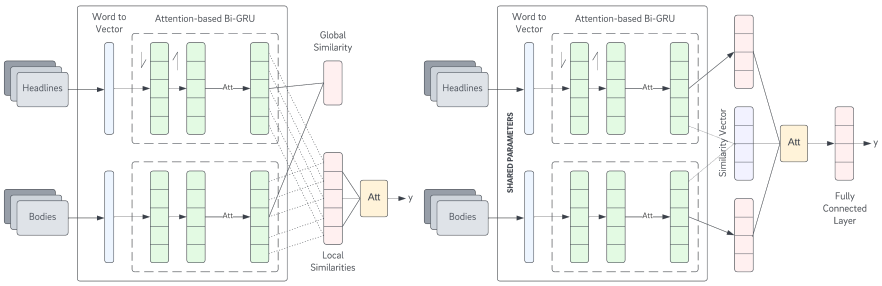

Fig. 5: L: Learning of the global and local similarities. R: Combined method for final prediction.

Dong et. al. [16] push the state-of-the-art further by exploiting the relationship between the misleading titles and the content, which is found to give

important clues for solving this problem. The authors propose a deep attentive similarity model to capture both global and local similarities of the pair of inputs (i.e. title and content). The representations of the textual input are obtained by using an attention-based biGRU model, similar to Zhou et al. [15]. The global similarity is learned via cosine similarity between the title and the content by minimizing it for mismatched pairs. The local similarity on the other hand is computed on pieces of text selected using block size and strides, again by minimizing the value for mismatched pairs. These similarities are concatenated with the latent representations and fed into a fully connected layer to produce the output. This approach leads to an improvement over Zhou et al.'s self-attentive RNN model by $\sim 4\%$ in terms of F1 score.

Linguistic Analysis of Headlines Furthermore, Naeem et al. [11] exhibit the success of modeling the intrinsic characteristics of clickbait for knowledge discovery and using it for decision making. The said knowledge discovery is done by performing a linguistic analysis of the news headlines using the *Part of Speech Analysis Module (POSAM)*. The idea is to understand the underlying structure and syntax and accordingly adjust the structure of the LSTM module for the conclusive classification. POSAM is a variation of the original n-gram classifier based on *Part of Speech (POS)* tagger, which reveals a stark difference in the occurrences of *WH-Determiners* and *Personal Pronouns* amongst others in clickbait text. Moreover, the observation that POS tokens that create an information gap exist in the latter half of the sentence motivates the decision to have a loopback of five words that effectively double the weights of the second half. This framework is evaluated on a newly collected dataset from *Reddit* and produces an F1 score of 0.973.

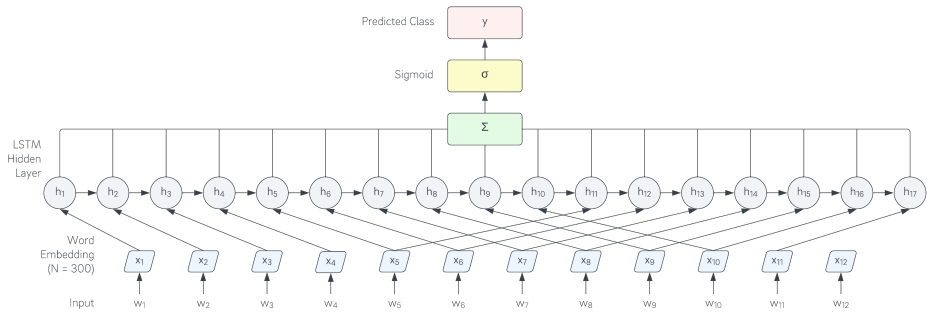

Fig. 6: LSTM using POSAM uses a single layer with a time stamp of 10 units and a loopback of half the words in an average headline to focus on the words that occur its last half.

3.3 Satire

Satire is a very interesting genre that involves ridicule, humor, irony, or exaggeration and is primarily considered a form of entertainment. However, oftentimes

readers might not be able to recognize it if the cues are too subtle to be identified or they lack relevant context. In cases where people perceive it as the truth, satire becomes a type of fake news thereby spreading inaccurate information and causing mistrust in society. As reported in the *Guardian*, it was found that regardless of the ridiculous content, people not only believe satirical content but also share it in their networks frequently.

The task of detecting satire in an article, which mimics genuine news, is understandably a binary classification problem. Past research in the domain of fake news might not directly apply to satire detection since it is a distinct domain with its unique traits. For instance, the works by Xiao et al. [17], Dong et al. [18], and Ge et al. [19] focus on tackling misinformation and fake news by tending to discover the truth through the knowledge base and truthfulness estimation, however since there is no ground truth for satire generally, these techniques might not work for this category of fake news. Owing to the similarity to the problem of deception detection, we can utilize its solutions involving analysis of psycholinguistic features [20], writing analysis [21], and cultural differences [22] to tackle the satire detection problem. Though it should be noted that these techniques consider features at the document-level however it is observed that satirical cues are found only in certain paragraphs thus indicating that document-level features might even be superfluous for this use case. In this section particularly, we will discuss a few deep learning techniques that leverage past research in devising clever frameworks to effectively address the problem at hand

Exploiting Lexical and User Signals One of the early works by Amir et. al. [23] exploits the fact that the way satirical content is delivered depends heavily on the writer. Therefore, their CNN-based deep learning framework learns and uses the user embeddings in addition to the lexical signals without the need for manual feature engineering to tackle the problem effectively. The textual embeddings which take advantage of the lexical signals are produced by employing multiple CNNs with different filter sizes (to generate multiple feature maps) on pre-trained word embeddings. Whereas, the user embeddings encode latent aspects of users and capture homophily. Both these embeddings are concatenated to provide holistic contextual information and fed to a fully connected network, which ultimately produces the output. This approach was evaluated on a dataset derived from *Twitter*, collected using hashtag-based supervision, and demonstrated an edge over the state-of-the-art approaches (with a 2% accuracy increase) which leveraged an extensive set of carefully crafted features.

Improving Data and Model Quality Datasets collected from *Twitter* tend to have noisy labels, and language along with the fact that the tweets might not be self-contained, that is there is a trend of conversations in replies or the content being broken into threads. To address these shortcomings, Misra et al. [24] collect a high-quality dataset by leveraging two popular web sources: *HuffPost* (for the real news) and *The Onion* (for the satirical news). Furthermore, the authors

improve the model quality by recognizing that RNNs work better in encoding sequential information, hence they propose a hybrid framework where textual embeddings are learned via both CNN and self-attentive bidirectional LSTM. Then embeddings from both methods are concatenated and ultimately fed to fully connected layers for final prediction. Apart from increasing classification accuracy by 5%, the work also visualizes the attention layer on various headlines for improved model interpretability. Qualitative results show that the network emphasizes the co-occurrence of incongruent word phrases within each sentence such as *'oppressing other people'* & *'insane k-pop sh*t during opening ceremony'*, which are important cues to detect satire for us, humans, too.

Utilizing Paragraph-level Linguistic Features Work by Yang et. al. [25] showcases that satirical cues are often reflected in certain parts of the document, hence a hierarchical neural network with an attention mechanism to extract paragraph-level linguistic features is successful in tackling the problem of detection satire. They propose using four levels of features: *character-level features* that are extracted using CNN as they recognize morphological information and name entities well, *word-level features* are generated by applying bi-GRU on top of character-level representations, *paragraph-level features* are generated by applying bi-GRU on word representations concatenated with engineered linguistic features, and *document-level features* are generated using attention on paragraph-level representations concatenated with engineered linguistic features. The linguistic features include psycholinguistic, writing stylistic, readability, and structural features from the given text. The qualitative evaluation suggests readability features support the final classification while psycholinguistic, writing stylistic, and structural features are beneficial at the paragraph level. In addition, this exposition reveals that the writing of satirical news tends to be emotional and imaginative.

Source-agnostic Adversarial Training The approaches discussed so far are built upon corpora labeled automatically based on the source of the article. McHardy et. al. [26] hypothesize that this encourages the models to learn characteristics from the publication sources, rather than characteristics of satire to some degree, e.g. it is satirical if from *The Onion* and it is real if from *HuffPost*, thus leading to poor generalization performance on unseen publication sources. The authors propose having an adversarial component to control for the confounding variable of the publication source. The framework consists of three parts: feature extractor, satire detector, and publication identifier. The *feature extractor* takes pre-trained word embeddings and feeds them into bidirectional LSTM with attention. *Satire detector* and *publication identifier* have a softmax layer to output the prediction of respective tasks. Since the goal is to control for the confounding variable of publication sources, the training is done considering the publication identifier as an adversary, i.e. classifier's parameters are updated to optimize the publication identification while the parameters of the shared feature extractor are updated to fool the publication identifier. This technique was

evaluated on a dataset collected from 4 genuine news and 11 satirical German web sources, and the qualitative analysis shows that the adversarial component enables the model to pay attention to linguistic characteristics of satire.

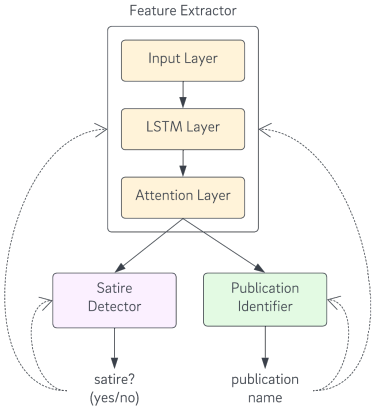

Fig. 7: Source-agnostic Adversarial Training Model.

3.4 Deepfake

With the advancement in image processing and machine learning technology, it has become so much easier to generate images, audio, and video of situations that did not happen in reality, which is quite daunting. These synthetic audio-visual content, called *deepfakes* are leading to the propagation of misinformation on social media, for instance - celebrity pornography, tweaked videos of political leaders to induce conflict, hoax calls, and so on. It should be noted that the idea of faking content is not novel, it has existed for a while now via *'Photoshop'* and regular video editing, however, deepfakes are such a powerful convincing set of media that it is almost impossible to tell the fact from the fiction. The creation of these phony images and videos is done via deep learning technology using the sophisticated *Generative Adversarial Networks (GANs)*. *CycleGAN* by Zhu et al. [27] is a popular technique that uses GANs to generate a new image with the same characteristics as that of the input image. The key characteristic of CycleGAN is that it applies a cycle loss function that enables it to learn the latent features and performs an image-to-image translation without the need for a paired example, thus falling under the unsupervised learning paradigm. The success of this technique can be owed to a humongous amount of data available for training on the web these days in the form of images and videos, which are now a popular medium of content, thanks to social media networks like *TikTok* and *Instagram*.

Evidently, deepfakes have a lot of potential to create havoc in our society by seeding and boosting social conflict, fraud, and revenge in society. It is thus of utmost importance to construct ways to accurately detect and flag this set of counterfeit images and videos as early on in the propagation journey as possible.

Ironically, deep learning has proven to be one of the more precise methodologies in performing this act of distinction between genuine and fake content online.

Deepfake Image Detection To distinguish phony images from real ones, it is no surprise that one of the successful techniques is the use of GANs.

Plying Deep CNN In one of the simpler approaches based on face recognition techniques, Nhu-Tai et al. [28] use a deep CNN to tackle the deepfake image detection problem. First, they normalize the faces to a frontal view followed by the deep feature extraction to derive and normalize facial features. Then in the face matching process to distinguish between the real and fake images, they create a suitable model to calculate the distance between the facial feature vectors for face identification and verification. This is topped with a fine-tuning mechanism by adjusting the weights of the classifier layer to calibrate the extracted features as per the forensics data provided by the *National Research Foundation of Korea (NRF)*. The proposed method performs decently well on the given forensic data with an 80% accuracy.

Boosting Performance with Ensemble Going the extra mile, Tariq et al. [29] design an ensemble model in *'GAN is a Friend or Foe?'* with three different shallow CNNs complete with L2 kernel regularizer of 0.0001, batch normalization, max pooling, and dropout which they refer to as the *ShallowNet*. Different variations of the ShallowNet are created with different layer settings mentioned below. V2 and V3 variations are similar in depth, however, shallower than V1, which leads to lower training durations. However, the introduction of the max pooling layer in V3 yields better performance on lower resolution images, on which V1 was noticeably poor. The ShallowNet achieves ∼99.99% accuracy and outperforms the well-known GAN-generated image detection neural networks like the VGG16, XceptionNet, and NASNet in terms of accuracy and AU-ROC. The results also indicate that as the resolution of the deepfakes goes down, it becomes harder to detect them as the model performance dips a bit.

ShallowNetV1	ShallowNetV2	ShallowNetV3
C-N-R-D-C-N-R-D-C-N-R-M-D	C-R-D-C-R -D-C-R-D	C-R-D-C-R -D-C-R-M-D
C-N-R-D-C-N-R-D-C-N-R-M-D		
C-N-R-D-C-N-R-D-C-N-R-M-D	C-R-D-C-R -D-C-R-D	C-R-D-C-R -D-C-R-M-D
C-N-R-D-C-N-R-D-C-N-R-M-D		
C-N-R-D-C-N-R-D-C-N-R-M-D	C-R-D-C-R -D	C-R-D-C-R -D
C-N-R-D-C-N-R-D		
F-De-R-N -D-De-S	F-De-R-N -D-De-S	F-De-R-N -D-De-S

Fig. 8: Variations of the ShallowNet. Each row represents a block in the architecture. (Note: C=Conv2D, N=Batch Normalization, R=ReLU, D=Dropout, M=MaxPooling, F=Flatten,De=Dense & S=Sigmoid.)

Enhancing Model Generalization Ability using Image Preprocessing One of the major drawbacks of the above-stated techniques is their gen-

eralizability or to put it simply, the use of the same dataset for training and evaluating the model which is not well-suited in the real world scenario. To battle against the detection of an unseen variety of deepfakes and amplify the generalization ability of the deepfake detection models, Xuan et al. [30] propose a novel method of image preprocessing, namely *Gaussian Blur* and *Gaussian Noise*, in the training phase. Contrary to the generic methods, the idea is to destroy the unstable low-level high-frequency noise cues by adding Gaussian Blur and Gaussian Noise to the training images only to improve pixel-level statistical similarity between real images and fake images. This propels the model to learn more intrinsic and meaningful features, instead of simply learning the style of the fake image generating model. Instead of using a complex model, they use a fairly simple *Deep Convolutional Generative Adversarial Network (DCGAN)* as the discriminator. The experimental results demonstrate the effectiveness of the proposed technique in enhancing the generalization ability although not by a lot owing to the inherent difficulty of this problem.

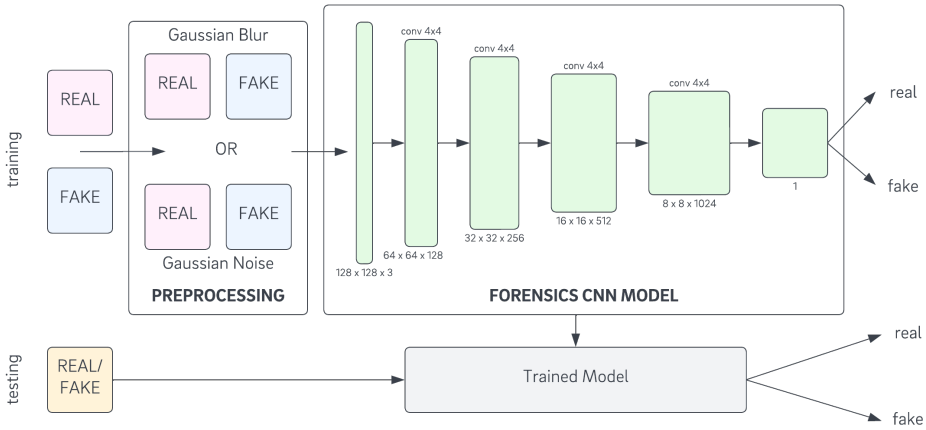

Fig. 9: Architecture of Deep Convolutional Generative Adversarial Network.

Real-Fake Pairwise Learning So far, deepfake detection has been considered a binary classification problem, with an aim to compartmentalize a given image either as fake or genuine. Along with, previous studies have been focused on fake images which were partially reconstructed from a real image, i.e. those models can not be used to detect a full-fledged synthetic image that bears no resemblance to any being living or dead. Understandably, the latter is a hard problem since gathering a training set composed of only GAN-generated images is cumbersome. In other words, the supervised learning technique of learning from past GAN-generated images might not lead to generalizable models, since the detector will not be able to recognize an image that it might not have seen in the past during the training process. To overcome this problem, Hsu et al. [31] proposed a pairwise learning approach over a modified CNN, which they refer to as the *Common Fake Feature Network (CFFN)*.

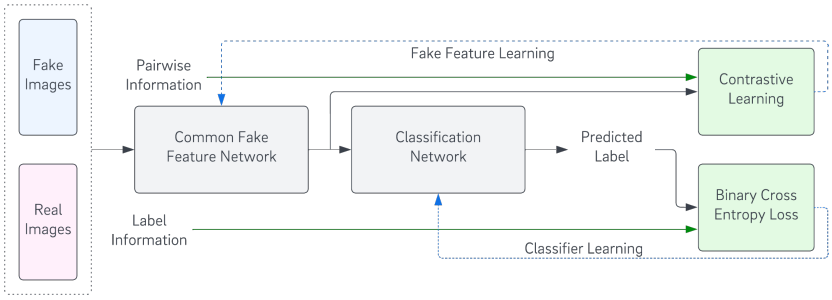

Fig. 10: CFFN based two-step learning approach.

In their approach, the authors pair the fake and real images and use the pairwise information to construct the contrastive loss which helps the proposed CFFN to learn discriminative *common fake features (CFFs)*. CFFs can then be used further to discriminate between a fake and a genuine image. The CFFN is a Siamese network integrated with the DenseNet. Contrary to regular CNNs which are usually single-streamed, CFFN is a dual-streamed network having the ability to ingest pairwise input for CFF learning. Since CNNs use only high-level feature representation to detect fake from the real, the cross-layer feature concatenation in CFFN further helps in capturing fine-grained feature representation, which is where the CFFs of fake face images exist. This technique involves a two-step learning policy since it uses contrastive loss to learn the CFFs and then the classifier is optimized by minimizing the cross-entropy loss. Experimental results indicate that CFFNs had a higher generalization ability and effectiveness than the other methods.

Deepfake Video Detection At present, there has been a lot of major build-out in artificially generated videos with such sophisticated lip movements and eye syncs that it is hard to tell whether it is fake or not. And due to the loss of frame information after video compression, it is not viable to apply deepfake image detection mechanisms on each frame to unravel the fakeness.

Leveraging the Physiological Signal of Eye-Blink Li et al. [32] make a very valid point by utilizing the physiological signal of eye blinking to detect fake videos in their research *'In Ictu Oculi'*. Their novel approach employs *Long-term Recurrent Convolutional Neural Networks (LRCN)* which is a combination of CNN and RNN to capture the phenomenological and temporal regularities in the process of eye blinking. The idea is that on average resting blinking rate is 17 blinks/min which can go up to 26 blinks/min or as low as 4.5 blinks/min depending on the intensity of the conversation, however, it was found that synthetically generated face videos lacked eye blinking function, since training datasets seldom contain faces with closed eyes. The proposed LRCN model comprises three chief components. The feature extraction part converts the input eye region into discriminative features, followed by sequence learning which is implemented with RNN with LSTM cells to increase the memory capacity of the RNN model and

avoid gradient vanishing while backpropagating. Ultimately, the state prediction component takes the output of the LSTM to generate the probability of the eye being open or closed using which we can plot an eye blinking time series. This method was evaluated using the *Closed Eyes (CEW)* dataset and the *DeepFake* generated videos. The experiments indicate that even though regular CNN was able to predict the state of the eye exceptionally well for an image, it lacks temporal knowledge which is where LRCN comes into play by taking advantage of the long-term dynamics to effectively predict eye state, which is more smooth and accurate.

Outside the Limitations of Visible Signals Building upon the use of biological signals, Ciftci et al. [33] approach this problem in depth by researching the effectiveness of signals like heart rate, blood flow which might not be visible to the human eye. Their research *'How Do the Hearts of Deep Fakes Beat?'* proposes extraction of *photoplethysmography (PPG)* signals from the image to detect the change in skin reflectance over time when the blood flows through the veins which would ultimately help us flag the fake video. The idea is to find a face in each frame of the video using a face detector and extract regions on the face that have as many stable PPG signals as possible. The power spectral density of raw value in the PPG cells in the different time windows are then computed which helps the classifier, which is a shallow CNN, flag it as fake or not. This setup is experimented with on public datasets like *CelebDF*, *Face2Face*, *FaceSwap*, etc. which contain fake videos, and is able to achieve about 97.29% accuracy which is considerable.

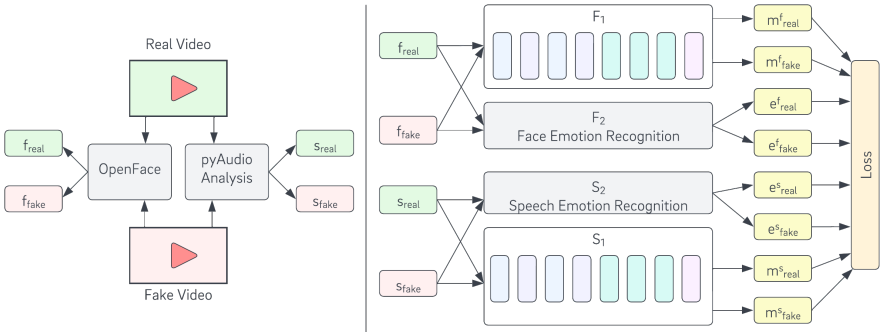

Fig. 11: Multimodal DeepFake Detection: Training Routine.

Employing Emotive Cues In addition to the biological signals like the blinking of the eye and blood flow, there are other signals like the sync between the audio and video which can help detect if a video is fake or not. In their study *'Emotions Don't Lie'*, Mittal et al. [34] suggest the idea to fish for facial cues, speech cues, background context, hand gestures, and body posture, and orientation from a video and further analyze these modalities to identify a fake video from the real one. The experiment uses a Siamese network-based architecture that takes the input of a real video and its counterpart deepfake video and obtains the modality along with the perceived emotion embedding vectors for

the face and speech of the subject. The embeddings are then used to compute the triplet loss function to maximize the similarity between modalities for the real video and minimize it for the fake video. The said approach is tested on two deepfake identification benchmark datasets, *DeepfakeTIMIT* and *DFDC*, and yielded an accuracy of 96.6% and 84.4% respectively.

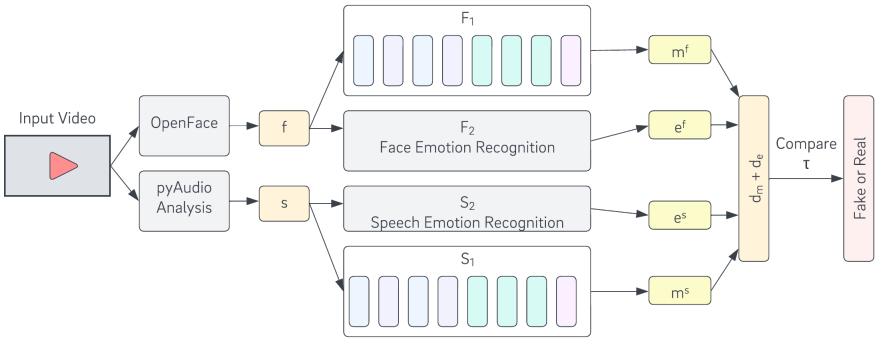

Fig. 12: Multimodal DeepFake Detection: Testing Routine.

Going Beyond a Single Frame So far, the trend has been to do a frame-by-frame analysis to dig for clues that could indicate if a given video is artificially generated or not. Lima et al. [35] propose incorporating temporal aspects into various action recognition methods to apply for deepfake detection, which demonstrates an edge over the contemporary frame-based mechanisms. The intuition is that these synthetic videos lack temporal coherence due to various tempering. *CelebDF v2* dataset is used for this setup and the data is preprocessed by cropping beyond the face (considering the tampering is done only on the face), thus eliminating noise from the data. Experimental results exhibit the advantageous nature of incorporating temporal information in various architectures over the frame-based baselines. R3D network outperforms other methods and it consists of a sequence of residual networks which introduce shortcut connections bypassing signals between layers.

Undoubtedly, there is a massive surge in the fabrication of phony images and video, and with such perfection. The tools for the creation of deepfakes are becoming more and more accessible, and social media is providing a prominent pedestal for their propagation. The above-mentioned research in deep learning for deepfake detection are just a few of the many work streams going on in parallel to fight this social evil.

4 Limitations of Deep Learning Approaches

Though deep learning models have proven to be successful in flagging fake news by utilizing textual and temporal attributes to an extent, fake news detection remains a tough challenge to tackle. With sophisticated feature extraction methods and state-of-the-art architecture, it is hard to distinguish genuine information

from counterfeit without cross-referencing, fact-checking, or additional information. This is because the fake content is being generated with such poise to deceive even the well-informed readers.

Most of the techniques we have discussed in this chapter adopt *Natural Language Processing (NLP)* to analyze the writing styles and lexical patterns of news articles to flag as fake news or not. These methodologies are quite shallow in their analysis since they check whether the news articles adhere to the standard styles generally used by professional journalists or not. Zhou et al. [36] in their study suggest using crowdsourced fact-based knowledge checks on top of these NLP-based fake news detection techniques to achieve a more robust mechanism. With their experiments on *Fakebox*, a fake news detector, they highlight the vulnerabilities of the NLP-based fake news detection methods since without deeper semantic knowledge they can lead to inaccurate results. They are also highly likely to be fooled by malicious users with adversarial attacks via fact distortion, content exaggeration, linking of two independent events, exchange of subject-object, and so on. These kinds of attacks are much more subtle since they don't change the overall writing style of news articles and thus have the potential to evade similarity detection.

5 Fairness and Interpretability

There is a growing focus on deep learning-based approaches to research for more novel, accurate, and robust fake news detection techniques. In this attempt, one territory that is lagging behind the rapid development is the concentration on rational and ethical considerations. Since the societal repercussions of unfair fake news detection models are far more severe than comparative pieces of work, say spam filters or speech recognition, the ethics and fairness aspects of fake news detection also warrant some attention.

Deepak et al. [37] analyze the ethical considerations of the fake news detection methodologies across the key dimensions: mismatch of values, data-driven nature, and domain attributes. The context of the content matters a lot while approaching this problem, hence it is but obvious that there is a tautness between the accuracy of the news and the fairness aspect of the model. The rules of conduct learned by a model might be different from that of the society since politics, legality, and emotions is a spectrum and oftentimes there is no black and white view. Since the deep learning techniques to detect fake news, be it supervised or unsupervised, are nothing but a statistical model built from the *past data*, they end up encoding assumptions based on the bygone events which might or might not be relevant in the present age. There can also be instances where the decision timelines and reversals of laws or any quotes are not factored in based on the architecture of the model. Hence this implicit assumption of the static nature of the context of the fake news is another dimension that can hinder a wholly fair and ethical insight. Another aspect that makes this problem unique is that fake news is a *universal* attribute, there is hardly any instance where it might be genuine for a certain set of people and phony for others. Therefore,

any kind of inconsistencies in the results are not acceptable especially if viewed from an ethical point of view.

Despite significant advancements in fake news detection methodologies, there remain corners that are yet to be fully discerned. Interpretability of fake news detection techniques is one such domain since fake news detection is a wide spectrum problem depending on the type of fake news we are dealing with. Interpretable models can help with debugging and model validation along with assisting users in identifying bias in algorithms by putting forward an explanation as to why and how a decision was made. Moheseni et al. [38] propose the *algorithmic interpretability* - ability to visualize model parameters to inspect model behavior, *human interpretability* - transparency for end-users with understandable explanations of hows and whys, and *supporting evidence* - verified claims related to the news, as the three dimensions which can help improve fake news detection research.

6 Emerging Trends

Investigation in fake news detection is gaining attention and remains a growing area of research and development, especially amongst ML practitioners and scientists.

Given the sensitive and high-stakes nature of the fake news domain, accountability of fake news detection algorithms is another emerging area of exploration. This has become essential due to possible adversarial attacks via fact distortion, content exaggeration, linking of two independent events, or exchange of subject-object, which can deviate even the more rugged techniques from their normal mode of operation due to lower quality data. Bogaert et al. [39] study the robustness of the fake news detection against fabricated adversarial examples in *'Can Fake News Detection be Accountable?'*. To tackle the issues, one suggestion is to have an adversarial component built into the detection frameworks, automate fake news generation by morphing genuine news, and allow detection models to better discern the syntactic and semantic patterns. For accountability of such methods in the long term, Bogaert et al. suggest developing ways to evaluate such countermeasures.

Recently, the popular social network, *Twitter*, introduced a *'Read before you retweet!'* prompt in an attempt to promote informed discussion amongst its users. Since the headlines can be often misleading, this prompt encourages users to read the whole article before sharing it, in case it turns out to be yet another fake news article. These types of prompts are a smart way to encourage media literacy and control the viral spread of fake news on social media. In another effort to combat fake news, *Twitter* introduced a community-driven effort called *Birdwatch* that allows people to identify information in tweets they believe is misleading and write notes that provide informative context. Such data could be really valuable for algorithmic accountability and help in improving the fake news detection methods.

Ahead of the elections, a law in Singapore called *Protection from Online Falsehoods and Manipulation Act (POFMA)*, now pushes social media companies to adhere to their repressive law that aims to combat fake news by putting warnings next to posts the authorities deem to be false, and in extreme cases get them taken down. Though these laws were considered a tool for potential censorship, the authorities claim that since social media companies put profit above principle by attracting eyeballs via fake news propagation, this is a necessary action to curb this nuisance. In compliance with POFMA, Twitter now has an additional *'Legally Required Notice'* flag to indicate misinformation.

An interesting take on this topic is to analyze the propagation of fake news on social media, and one of the major elements in the network is the users. Lopez et al. [40] provide a compelling argument of how certain users can be flagged as malicious based on their history and thus indicating the intention of spreading fake news. Since *Twitter* is one of the popular social media networks where people actively share updates, the authors worked on profiling the users as fake news spreaders or not, for now focusing on English and Spanish tweeters only. This is an area having scope to research further, for instance, based on multiple languages, other demographics, temporal or psychology-related features, and so on.

References

1. Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu: Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter 19, no. 1: 22-36. (2017)
2. Zhou, X., and Zafarani, R.: A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (2020)
3. Khan, Junaed Younus, Khondaker, Md. Tawkat Islam, Afroz, Sadia, Uddin, Gias, Iqbal, Anindya: A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications (2021)
4. Wang, William Yang: "liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648 (2017)
5. Singhania, Sneha and Fernandez, Nigel and Rao, Shrisha: 3HAN: A Deep Neural Network for Fake News Detection. 10.1007/978-3-319-70096-0_59 (2017)
6. Rashkin, Hannah, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi: Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of the 2017 conference on empirical methods in natural language processing, pp. 2931-2937. (2017)
7. Shu, Kai, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu: defend: Explainable fake news detection. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 395-405. (2019)
8. Khattar, Dhruv, Goud, Jaipal Singh, Gupta, Manish, and Varma, Vasudeva: MVAE: Multimodal Variational Autoencoder for Fake News Detection. In The World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, 2915–2921. (2019)
9. Jwa, H., Oh, D., Park, K., Kang, J.M., Lim, H.: exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT). Applied Sciences 9(19):4062. (2019)

10. Kaliyar, R.K., Goswami, A. and Narang, P. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimed Tools Appl 80, 11765–11788 (2021)
11. Naeem, B., Khan, A., Beg, M. O., and Mujtaba, H.: A deep learning framework for clickbait detection on social area network using natural language cues. Journal of Computational Social Science, 1-13. (2020)
12. Yimin Chen, Niall J. Conroy, and Victoria L. Rubin: Misleading Online Content: Recognizing Clickbait as "False News". In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, 15–19. (2015)
13. Kim, Yoon: Convolutional neural networks for sentence classification. EMNLP. (2014)
14. Zheng, Hai-Tao, Jin-Yuan Chen, Xin Yao, Arun K. Sangaiah, Yong Jiang, and Cong-Zhi Zhao: Clickbait Convolutional Neural Network. Symmetry 10, no. 5: 138. (2018)
15. Zhou, Yiwei: Clickbait detection in tweets using self-attentive network. arXiv preprint arXiv:1710.05364 (2017)
16. Dong, Manqing, Lina Yao, Xianzhi Wang, Boualem Benatallah, and Chaoran Huang: Similarity-aware deep attentive model for clickbait detection. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). (2019)
17. Houping Xiao, Jing Gao, Qi Li, Fenglong Ma, Lu Su, Yunlong Feng, and Aidong Zhang: Towards Confidence in the Truth: A Bootstrapping based Truth Discovery Approach. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1935–1944. (2016)
18. Dong, Xin Luna, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang: Knowledge-based trust: Estimating the trustworthiness of web sources. arXiv preprint arXiv:1502.03519. (2015)
19. Liang Ge, Jing Gao, Xiaoyi Li, and Aidong Zhang: Multi-source deep learning for information trustworthiness estimation. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 766–774. (2013)
20. Li, Jiwei, Myle Ott, Claire Cardie, and Eduard Hovy: Towards a general rule for identifying deceptive opinion spam. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1566-1576. (2014)
21. Feng, Song, Banerjee, Ritwik, Yejin, Choi: Syntactic Stylometry for Deception Detection. 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference. 2. 171-175. (2012)
22. Taylor, Paul, Larner, Samuel, Conchie, Stacey, Zee, Sophie: Cross-Cultural Deception Detection. Detecting Deception: Current Challenges and Cognitive Approaches. 175-201. 10.1002/9781118510001.ch8. (2015)
23. Amir, Silvio, Byron C. Wallace, Hao Lyu, and Paula Carvalho Mário J. Silva: Modelling context with user embeddings for sarcasm detection in social media. arXiv preprint arXiv:1607.00976. (2016)
24. Misra, Rishabh and Arora, Prahal: Sarcasm detection using hybrid neural network. arXiv preprint arXiv:1908.07414. (2019)
25. Yang, Fan, Arjun Mukherjee, Eduard Dragut: Satirical news detection and analysis using attention mechanism and linguistic features. arXiv preprint arXiv:1709.01189. (2017)
26. McHardy, Robert, Heike Adel, and Roman Klinger: Adversarial training for satire detection: Controlling for confounding variables. arXiv preprint arXiv:1902.11145. (2019)

27. Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros: Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pp. 2223-2232. (2017)
28. Do Nhu, Tai, Na, In, Kim, S.H.: Forensics Face Detection From GANs Using Convolutional Neural Network. (2018)
29. Tariq, Shahroz, Lee, Sangyup, Kim, Hoyoung, Shin, Youjin, and Woo, Simon: Detecting Both Machine and Human Created Fake Face Images In the Wild. 81-87. 10.1145/3267357.3267367. (2018)
30. Xuan, Xinsheng, Bo Peng, Wei Wang, and Jing Dong: On the generalization of GAN image forensics. In Chinese conference on biometric recognition, pp. 134-141. (2019)
31. Hsu, Chih-Chung, Yi-Xiu Zhuang, and Chia-Yen Lee: "Deep Fake Image Detection Based on Pairwise Learning" Applied Sciences 10, no. 1: 370. https://doi.org/10.3390/app10010370 (2020)
32. Li, Yuezun, Ming-Ching Chang, and Siwei Lyu: In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1-7. IEEE, (2018)
33. Ciftci, Umur Aybars, Ilke Demir, and Lijun Yin: How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals. In 2020 IEEE International Joint Conference on Biometrics (IJCB), pp. 1-10. (2020)
34. Mittal, Trisha, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha: Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues. In Proceedings of the 28th ACM international conference on multimedia, pp. 2823-2832. (2020)
35. de Lima, Oscar, Sean Franklin, Shreshtha Basu, Blake Karwoski, and Annet George: Deepfake detection using spatiotemporal convolutional networks. arXiv preprint arXiv:2006.14749. (2020)
36. Zhou, Zhixuan, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu: Fake news detection via NLP is vulnerable to adversarial attacks. arXiv preprint arXiv:1901.09657. (2019)
37. P, Deepak, Chakraborty, Tanmoy, Long, Cheng, and G, Santhosh: Ethical Considerations in Data-Driven Fake News Detection. 10.1007/978-3-030-62696-9_10. (2021)
38. Mohseni, Sina, Eric Ragan, and Xia Hu: Open issues in combating fake news: Interpretability as an opportunity. arXiv preprint arXiv:1904.03016. (2019)
39. Bogaert, Jérémie, Quentin Carbonnelle, Antonin Descampe, and François-Xavier Standaert: Can Fake News Detection be Accountable? The Adversarial Examples Challenge. In 41st WIC Symposium on Information Theory in the Benelux. (2021)
40. López, Álvaro, and Pasqual Martí: Profiling Fake News Spreaders on Twitter. In CLEF (Working Notes). (2020)