Dealing with Fake News on Social Media

Subhalingam D
Department of Mathematics
IIT Delhi
Delhi 110016, India
mt1180770@maths.iitd.ac.in

Krishna Chaitanya Reddy Tamataam
Department of Mathematics
IIT Delhi
Delhi 110016, India
mt6180785@maths.iitd.ac.in

Abstract

In the recent times, there has been a growing concern in the pace at which fake news is propagated through social media. Spread of fake information is highly undesirable, especially in domains like politics, finance, health and research. Hence, dealing with fake news has become a hot topic of research, which is evident from the recent works in the past few years. This paper surveys some of the current state-of-art techniques in the literature to detect fake news and investigates some of the characteristics, performance, limitations of these models.

1 Introduction

With the advancement in technology and improvement in infrastructure, there has been a drastic increase in the amount of information available on the internet and number of users browsing the web in the last decade. Off late, people have switched to platforms like news applications, YouTube, podcasts, social media from traditional sources like television, radio, newspaper to receive updates. Further, with the increase in popularity, improvement in accessibility and convenience, social media has become one of the widely used platforms for sharing news in the recent times. However, such developments have put a big question-mark on the veracity of information, as anyone in the world can post articles on the internet.

1.1 Definition

Giving a perfect definition for *fake news* is itself a challenging task as no universal definition for *fake news* exists in the literature. Some terms associated with *fake news* include *false news* (Vosoughi et al., 2018), *deceptive news* (Allcott and Gentzkow, 2017), *misinformation* (Kucharski, 2016), *disinformation* (Kshetri and Voas, 2017), *rumor* and (Zubiaga et al., 2018). Zhou and Zafarani (2020) proposed a broad definition for fake news as "*fake news is false news*" which was narrowed down to "*fake news is intentionally false news published by a news outlet*". We use a similar notion in this paper.

In literature, there are many studies that were conducted about the structures in which fake news are articulated. Amado et al. (2015) analysed Undeutsch hypothesis, that suggests that content and quality from made-up or fake news differ significantly from statements derived from the memory of real-life experiences. On the other hand, Taylor (1971) worked on Four Factor Theory, which suggests that lies are written in a different style, in terms of manipulative thinking, behavioural arousal, emotions, etc. The Information Manipulation Theory (McCornack, 1992) states that extreme information quantity often leads to deception. Such theories and Hypothesis, based on psychology and the cognitive sciences, play an important role in defining the structure of fake news and devising Machine Learning (ML) algorithms which help in detecting them.

1.2 Examples

In this section, we list down some examples of fake news, majorly in India.

- Fake news was very prevalent during the 2019 Indian general election (Samarth and Snigdha, 2019). Some even called the elections as "*India's first WhatsApp elections*", as WhatsApp was used as a tool of propaganda by many (Billy, 2019).

- 2013 Muzaffarnagar riots, which is known to have claimed over 60 lives and have displaced thousands, was fueled by videos circulated on WhatsApp (Dipankar, 2013).

- There were several *fake* information, that went viral on WhatsApp, stating "*spying technology*" added in the newly introduced Rs. 2000 banknotes, during the demonetisation period.

- There have been multiple instances of pictures from the Syrian and the Iraqi civil wars being passed off as from the Kashmir conflict with the intention of fuelling unrest and backing insurgencies[1].

- There were several fake news that were propagated during COVID-19 targeting a specific religion, at times, in India. A fake video on WhatsApp said to show a Muslim man spitting on bread went viral, calls for a boycott of Muslims grew (Shruti, 2020). There were targeted attacks on Muslims after Tablighi Jamaat event which led to increase in number of COVID-19 positive cases.

- During COVID-19, people became creative and started making their own stories. One of them said "*eating vegetarian food and eliminating meat from your diet could prevent you getting coronavirus*" (Shruti, 2020).

- Claims that "*Barack Obama, the 44th President of the United States, was injured in an explosion*" wiped out $130 billion in stock value (Rapoza, 2017).

1.3 Impact & Challenges

It is evident from examples in Section 1.2 that fake news can affect the individuals as well as society as a whole. It can affect the stock markets, spread hatred among a specific group (like religion), create panic, damage health, disturb the stability of the government and can even be a reason for rejection of a paper in research. Further, there are instances where fake news have turned out to have effects on real-world events. For example, "*Pizzagate*" was actually a fake news from Reddit, which eventually led to a real shooting.

It was found that about 62 percentage of US people used social media to get their news in 2016 (Gottfried and Shearer, 2016) and Facebook was the widely used platform for propagation of fake news over other news channels (Silverman, 2016). Moreover, large number of people who saw fake news were unable to detect that they were fake (rather believed them) (Silverman and Singer-Vine, 2016). Hence, fake news can also make people to accept biased stories. Thus, detecting fake news is an important issue to be solved.

[1] "*Fact Check: Not CCTV clip of Pulwama blast, old footage from Iraq being pushed on social media*". The Indian Express. 17 February 2019.

2 Literature Review

This paper focuses on the some of the state-of-art models in the literature and some basic models from Machine Learning literature.

2.1 CSI

Ruchansky et al. (2017) proposed Capture, Store and Integrate (CSI) model, which is a hybrid model that classifies fake news and identifies suspicious groups that spread them. In Capture stage, the semantic representation is obtained using an RNN and stored in form of vectors. Store stage is used to compute the vector representation of features of a user and his score (used for classifying the user as suspicious or not) computed by a simple regression neural network. In Integrate stage, the output of Capture and Store are concatenated and the resultant vector is fed to a fully-connected layer, followed by softmax layer.

2.2 dEFEND

Shu et al. (2019a) proposed the **E**xplainable **F**ak**E** **N**ews **D**etection (dEFEND) framework which is made up of a news content encoder, a user comment encoder, a sentence-comment co-attention component, and a fake news prediction component.

Firstly, the news content encoder (which includes both word and sentence encoder) models the linguistic features of news to a latent feature space through a series of hierarchical word-level and sentence-level encoding. A detailed overview of both of these encoders can be obtained from HAN in Section 2.4.

Secondly, the user comment encoder is used for latent feature extraction of the comments through word-level attention networks. It is expected that a large number of people share their opinions or express their emotions by commenting to the post and hence, comments can provide additional semantic information that can be used to improve fake news detection. This is a unique and novel consideration in this model.

Thirdly, the sentence-comment co-attention

component is used to model any interrelation between the news sentences and user comments and learn more feature representations. Through this, we can extract some important features in news content and user comments.

Lastly, the prediction is done by concatenating news content and user comment features and using softmax function.

2.3 GNN

Graph Neural Network (GNN) can be used in detecting fake news (Han et al., 2020) by training with the help of non-euclidean data like propagation of news in social media. In GNN, we maintain an Adjacency Matrix (A) and Feature Matrix (F). Adjacency Matrix is used to represent the graph where the root is the news and other nodes are spreaders of the news (directly or indirectly). The edges represents the spread of news through a tweet. The Feature Matrix is used to assign features like whether (s)he is a verified user, the number of followers, timestamp when the user created his/her account, to each node. These features are more accessible and need not require NLP concepts for analysis.

Han et al. (2020) use Differentiable Pooling (DiffPool) algorithm (Ying et al., 2018), which is specially designed for graph classification problems, which further takes structural details of the graph into account. At each stage, it uses the output H (node embeddings from previous layer) and A, the adjacency matrix, and learns a rough graph.

2.4 HAN

Yang et al. (2016) proposed the Hierarchical Attention Network (HAN) for Document Classification, which can also be extended for fake news classification (Shu et al., 2019a). HAN captures two important properties of the document. Firstly, the 'hierarchical' denotes the way in which document is represented- constructing sentence representations first and then aggregating them into a document representation. This models the idea that words form a sentence and sentences, in turn, form a document. Secondly, the 'attention' refers to the attention mechanisms intended to take into account the importance of a word or sentence in the context. In a natural language, the importance of same word or sentence may be different depending on the context. Indeed, HAN features one attention mechanism at word-level and another at sentence-level, as discussed below.

The HAN is made up of the following components: a word sequence encoder; a word-level attention layer; a sentence encoder and a sentence-level attention layer. The encoders are built using a bidirectional GRU (Bahdanau et al., 2016) which can track the state of sequences and words or sentences are transformed to a vector representation using embedding matrix. The attention layers are used to extract those words (or sentences) that are more important in a sentence (or document) and aggregate the representation of those informative words (or sentences) to form a sentence (or document) vector. Finally, the document vector from sentence-level attention layer is used for classification using a softmax layer and the model is trained using negative log-likelihood loss function.

2.5 n-gram/Topic models

Oriola (2020) proposed some base models using n-gram and Topic models and a hybrid of these. In n-gram model, sequential words are combined into a single entity and the occurrences of such entities or counted. Topic model is used to find out the topic that each document talks about. Latent Dirichlet Allocation (LDA) (Yan et al., 2019) is one such topic model in which documents in a corpus are represented as belonging to one of the randomly generated topics. Hybrid models are a combination of these features. In this paper, we focus on $n-gram$ and n-gram & Topic models.

2.6 Naive Bayes

This is a basic model for fake news detection. The main assumption that it takes that all words in an article as independent of each other and similar kind of words occur in fake news. then it simply estimates the probability of article (bag of words) being fake or not using the Bayes Theorem. $P(F/W) = (P(W/F)*P(F))/(P(W/F)*P(F) + P(W/T)*P(T))$ and $P(W/F)$ is given by nfw/nw where nfw is # fake news articles containing word w and nw is # news articles containing the word w.

2.7 SVM

A Support Vector Machine (SVM), that is used commonly in classification problems, can also be used for fake news detection. The optimal goal is that the SVM will find a hyper-plane that divides the dataset into two groups. The hyper-plane can be intuitively thought of a boundary based on

which the class (fake or not) an article falls into is predicted.

2.8 TCNN-URG

Qian et al. (2018) proposed Two-Level Convolutional Neural Network with User Response Generator (TCNN-URG), which can be used to detect fake news before it spreads. It has two parts in it. The TCNN learns features in two level condenses word-level information into sentence level information and applying CNN over this sentence level representation to analyse the semantic meaning of an article. The User Response Generator (URG) is a probabilistic model which is constructed from Neural Networks, trained by data of previous user responses. The model gives a hard label whether the text is fake or not while the URG generates soft semantic labels to guide the CNN training.

TCNN uses the average of all one hot representation of a words as a sentence. The article is represented as concatenations of all the sentence representations. A filter is applied to a group of n sentences to extract the semantic information. Following that, a max pooling layer is applied to give the first part of next layer. The second part comes from URG, which generally contains encoder and decoder. This layer is then linked to a fully-connected layer and a soft-max function is used for the classification.

3 Analysis

We study the models listed in Section 2 and give our opinions for some of the questions we ask on their characteristics, performance and limitations, in this section.

3.1 Performance

How do these models perform in the task of fake news detection?

We utilise a popular fake news benchmark called *FakeNewsNet*, which is a public multidimensional data repository containing two datasets with news(both political,social) and dynamic information. (Shu et al., 2017; Shu et al., 2019b). The datasets were compiled based on the news articles collected from *PolitiFact*[2] (political news) and *GossipCop*[3] (entertainment news).

We use the following metrics to compare performance of various models.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1-Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

From the values listed in Table 1 (Han et al., 2020; Oriola, 2020; Shu et al., 2019a; Shu et al., 2019b), it is evident that dEFEND outperforms all the model in PolitiFact dataset, however the Graph-based model performs well in GossipCop dataset in terms of accuracy as well as F1-score.

In Section 3.2, we give our opinions on which among Precision and Recall is important. In Section 3.3, we compare these models based on their efficiency to train, the size of dataset and maintenance.

3.2 Precision-Recall Trade off

Which is more desirable-precision or recall-or do we really need a balance between the two?

It is often desirable to have a balance between the two (i.e. have high F1 score) but we will move forward considering that we have to choose one among the two. This is highly subjective and we only present our views in this section.

To have a model with high precision (and low recall) would mean all news it predicts fake is actually fake, but it also predicts some fake news as not fake. On the other hand, a model with high recall (and low precision) would mean the model also predicts a lot of news which are not actually fake as fake.

We might want to have higher recall as it is fine to flag some real news as fake until all the fake news gets flagged as fake. Moreover, it is not desirable to have a model that misses to flag a fake news as fake because this can still have an adverse effect (from point of view of the public) when compared to the case of flagging some extra real news as fake. For instance, we can prevent the stock market from crashing if we are able to stop propagation of fake news.

In fact, a similar methodology is used by many social media giants to prevent propagation of fake news. Sometimes, users (or pages) with real information can get blocked or a real post can get removed on grounds of suspicion. In such cases, there are provisions to get the post (or account)

[2] https://www.politifact.com/
[3] https://www.gossipcop.com/

Model	PolitiFact				GossipCop			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
CSI	0.827	0.847	0.897	0.871	0.772	0.732	0.638	0.682
dEFEND	0.904	0.902	0.956	0.928	0.808	0.729	0.782	0.755
GNN	0.803	0.806	0.801	0.801	0.841	0.820	0.831	0.825
HAN	0.837	0.824	0.896	0.860	0.742	0.655	0.689	0.672
Naive Bayes	0.617	0.674	0.630	0.651	0.624	0.631	0.669	0.649
n-gram	0.80	0.79	0.78	0.78	0.82	0.75	0.79	0.77
n-gram + Topic	0.77	0.76	0.76	0.76	0.82	0.75	0.78	0.76
SVM	0.580	0.611	0.717	0.659	0.497	0.511	0.713	0.595
TCNN-URG	0.712	0.711	0.941	0.810	0.736	0.715	0.521	0.603

Table 1: Performance of models on FakeNewsNet

back by contacting the concerned team directly. Hence, we would say recall metric can be slightly preferred over precision in detection of fake news.

3.3 Efficiency

What are the features that a dataset should have? How does the size affect the training efficiency?

The complexity and depth of a model decides the number of features or size of dataset required. Simpler models like SVM can have high accuracy and perform extremely well on smaller datasets. They can also handle high dimensional spaces effectively and turn out to be memory efficient. However, they have difficulty with large datasets because of noise.

On the other hand, models like CSI, GNN, HAN and dEFEND have multiple components and often require larger dataset for training and fine-tuning to improve performance. However, such complexity can be welcomed because of the significant improvement in performance.

3.4 Capturing semantic and syntactic cues

Do the models capture the relationship between two words in a sentence? Is it desirable such a feature?

In a natural language, there is some relationship between words in a sentence. For example, "*Barack*" will most probably be followed by "*Obama*", rather than "*Chaitanya*". Hence, it is desirable for the models to consider such properties, rather than just treating the words as bag of words.

The n-gram model takes this into account. Models like HAN and dEFEND use attention mechanisms to treat some words or sentences as important, depending on the context. dEFEND also makes use of comments to get some idea of the influence of the words.

However, simpler models like Naive Bayes and SVM simply treats document as bag of words.

3.5 Leverage Additional Resources

Do the models take into account additional data available like source of the post or reactions by the public on the post, for classification?

Some models like GNN, CSI and TCNN-URG consider the source of the information. The whole idea of GNN is to capture non-textual features about the news (mostly about how it propagates). It observes the activity of users like what are the news source they tweeted and retweeted. It also takes into account users data like number of followers, verified status of the account, etc. In TCNN-URG, URG (User Response Generator) accesses the user responses (like likes and comments) to generate User responses to a new article. CSI has a component of user scores for suspicious user accounts.

On the other hand, dEFEND makes use of comments posted by various people using a co-attention mechanism as mentioned in Section 3.7.

It can be concluded from Table 1 that taking into account such information has improved the performance when compared to basic models. Intuitively, some sources have a very low probability of posting fake news (discussed more in Section 3.6) or reactions from comments can give a clue on the nature of the news.

3.6 Reliable today... Reliable Tomorrow?

Should a source that is considered reliable remain reliable all the time (in the future)?

When a source is considered reliable, we expect it to remain reliable in the future also. However, in real life, there can be instances where media or news channels start becoming biased to a story or want to put up false yet catchy posts to get attention and improve their rating. For example, several news channels posted on Twitter that "*President Kovind gains 3 million new followers in an hour*"[4]. However, the fact is that official Twitter account of the President is considered as digital asset that belongs to the government and there is o digital transition when the occupant of a position changes in Twitter, to preserve the digital history of the previous occupant. In short, President Kovind had actually inherited the followers of President Mukherjee.

3.7 Labelling based on user reaction

Can the model classify news based on the users' reaction to the post? Is it a good idea to do this?

dEFEND is the only model that takes into account the comments of the people in classification of fake news.

Comments might contain textual information such as sensational reactions, skeptical opinions, etc., which can be related to content of original news article. Thus, comments *may* contain some useful semantic information and can help to improve fake news detection critically.

One might think that there might be cases where a news can deceive people easily and people start believing in such an article (Silverman and Singer-Vine, 2016), or a group of people (majority) might react *positively* to target another group of people (minority). However, Shu et al. (2019a) showed that fake news classification performance improved when both comments and news content are taken into account simultaneously. It is also evident from the values in Table 1.

4 Some Countermeasures taken

- Facebook is working with more than sixty fact checking organizations that have been rating and reviewing content in many global languages.Some of the fact verifying websites FB partnered with in India are BOOM and Webqoof

- From 2018, the number of forwards to a person in WhatsApp is limited to five which is done as a countermeasure of the incident of more than thirty killings caused due to rumors spread on WhatsApp

- Facebook applies labels to that have which are determined as fake by its fact verification department and tries not to display such content During the pandemic, Facebook displayed warnings on posts related to COVID-19

- The blue verified badge on Twitter, Instagram, Facebook helps people know that an account of public interest is authentic or verified.

5 Conclusion

In this paper, we discussed some of the state-of-art fake news detection models that are in practice, and we compared their performance and studied their characteristics. In terms of performance, we observed that dEFEND outperformed all the models in terms of accuracy and F1-score for PolitiFact dataset (which was based on political news) while GNN performed better in case of GossipCop dataset (which was based on entertainment news). This could mean that fake news detection on political news, which is considered to be followed by specific group of people, improved with the utilisation of user comments data, while fake news detection in entertainment news, which are followed by a larger group of people from diverse background, require information on the source and users who engage on them. This would mean a fake news detection model has to be domain specific. However, our works were limited to surveying the models and more work is required with this regard before coming to such a conclusion.

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, May.

Bárbara Amado, Ramon Arce, and Francisca Fariña. 2015. Undeutsch hypothesis and criteria based content analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context*, 7:3–12, 01.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

[4]More examples in https://www.altnews.in/top-fake-news-stories-circulated-indian-media-2017/

Yi Han, Shanika Karunasekera, and Christopher Leckie. 2020. Graph neural networks with continual learning for fake news detection from social media.

N. Kshetri and J. Voas. 2017. The economics of "fake news". *IT Professional*, 19(6):8–12.

Adam Kucharski. 2016. Post-truth: Study epidemiology of fake news. *Nature*, 540:525–525, 12.

Steven A. McCornack. 1992. Information manipulation theory. *Communication Monographs*, 59(1):1–16.

O. Oriola. 2020. Exploring n-gram, word embedding and topic models for content-based fake news detection in fakenewsnet evaluation. *International Journal of Computer Applications*, 176:24–29.

Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. 2018. Neural user response generator: Fake news detection with collective user intelligence. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3834–3840. International Joint Conferences on Artificial Intelligence Organization, 7.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Nov.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019a. Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019b. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media.

James C. Taylor. 1971. An empirical examination of a four-factor theory of leadership using smallest space analysis. *Organizational Behavior and Human Performance*, 6(3):249 – 266.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Ruidong Yan, Yi Li, Weili Wu, Deying Li, and Yongcai Wang. 2019. Rumor blocking through online link deletion on social networks. *ACM Trans. Knowl. Discov. Data*, 13(2), March.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June. Association for Computational Linguistics.

Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. 2018. Hierarchical graph representation learning with differentiable pooling. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 4800–4810. Curran Associates, Inc.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news. *ACM Computing Surveys*, 53(5):1–40, Oct.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2), February.

References

Gottfried Jeffrey, and Elisa Shearer. 2016. News Use across Social Media Platforms 2016. *Pew Research Center, May 26*.

Silverman Craig. 2016. *This Analysis Shows how Fake Election News Stories Outperformed Real News on Facebook*. BuzzFeed News, Nov 16.

Silverman Craig, and Jeremy Singer-Vine. 2016. *Most Americans Who See Fake News Believe It, New Survey Says*. BuzzFeed News, Dec 6.

Samarth Bansal, and Snigdha Poonam 2019. Misinformation Is Endangering India's Election. *The Atlantic*. ISSN 1072-7825.

Billy Perrigo 2019. How Whatsapp Is Fueling Fake News Ahead of India's Elections. *Time*. 25 Jan.

Dipankar Ghose, and Apurva 2013. Muzaffarnagar rioters used WhatsApp to fan flames, find police – Indian Express. *The Indian Express*. 12 Sept.

Kenneth Rapoza 2017. Can 'fake news' impact the stock market?. *www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/*.

Shruti Menon. 2013. Coronavirus: The human cost of fake news in India. *BBC Reality Check*. 30 Jun.