Vivek Joshi[1*], Dr. Sanjay Patel[2]

Unveiling Deception: A GAN-Based Unsupervised Learning Approach for Real-Time Generation and Detection of Text-Based Fake News

JES Journal of Electrical Systems

***Abstract**:* Generative artificial intelligence technology advancements have made it easy to generate fake news. Online community platforms like social media have made propagation of such fake news faster and more convenient. We have witnessed the social impact of such fake news in the past few years. In the literature, a Generative Adversarial Network (GAN) is used to detect text-based fake news based on structured data with the supervised learning approaches. However, we have observed that most large-scale online data are unstructured and can not be used with the supervised learning approaches. In this paper, we have used an auto-encoder to select the features from the unstructured data and feed them to GAN.

Keywords: *GAN, NLP, fake news, autoencoders, LatextGAN, Generators, Language Modeling, Machine Learning, LSTM*

1.0. INTRODUCTION

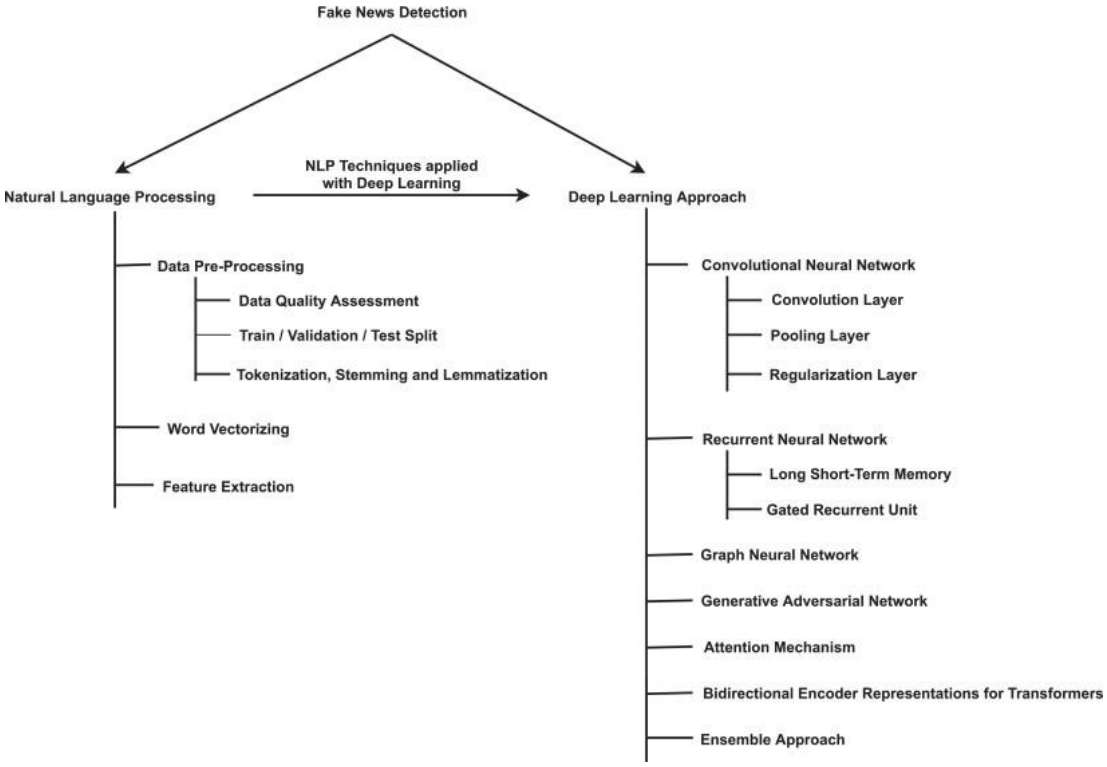
Nowadays, people depend on the Internet and social media platforms to get their news on a daily basis due to low expenses, speedy information access and fast information spreading, because they provide speedy and unfettered dissemination of news.The term ”fake news” describes information that is made up to mislead the public and has a detrimental impact on both the person and the entire society. The fake news intentionally misleads others with inaccurate or biased information for personal gain, which has a negative impact on public opinion and societal stability. The subject of internet false news has grown in prominence, particularly since the 2016 presidential election in the United States[1], [2]. In the opinion of Zhang and Ghorbani[3], misleading political comments and claims can freely manipulate voters. Study shows that rumors are spread more quickly than facts, which have an adverse effect on society politically, socially, and economically. Humans make decisions based on the information available. If the news is fake, then they will make wrong inferences and that will make them less efficient to rely on any data available[4]. The quality of news is distributed on a much lower level than by traditional methods because of the lack of restriction by authorities. There is a lot of noise and fake news in the online information environment.

The creation of fake news is not restricted to humans, but also to intelligent machines, as internet grows. A range of tools are therefore created to determine whether news being published or circulated is true or fake by varying degree of precision.

In the initial work to detect whether the news is fake or real, a traditional supervised approach was implemented with labeled datasets. However, only extracting linguistics features from the text, requires a large labeled dataset, which is not widely available. Another problem is to get manual explanations of the predicted results which is a labor-expensive and time-consuming task. Individual users’ network and chronological behavioral patterns are frequently learned as a result of their use of network and spatial characteristics. The technique’s effectiveness is sometimes reliant on the persistence of users’ network layout and behavioral habits. Also, Other than being difficult to sum up among natural languages, techniques that use information on content would be driven by subject drift.

The below image 1 shows the different taxonomy for fake news detection with a deep learning approach.

[1*]Assistant Professor,School of Information Technology, Artificial Intelligence & Cyber Security, Rashtriya Raksha University, Gandhinagar, India
[2]Assistant Professor, Department of Computer Science, Nirma University, Gujarat, India



To overcome the limitations of the supervised machine learning approach, we have introduced fake news generation and detection utilizing autoencoders and GANs with low amounts of loss. Our formed methodology initially produces highdimensional feature vectors from the unlabeled news dataset. Then, using a generator, fake news is produced using these feature vectors that were put into a generative model. To recognize machine-generated news, we concurrently trained a discriminator. Our method is highly automated and sturdy since it uses unsupervised learning and does not require labeled data to produce or identify fake news.

2.0. LITERATURE SURVEY

Due to the machine's ability to perform tasks that resemble those done by humanity through hierarchical learning, deep learning has caught interest. Nevertheless, in order to get high-quality performance and attain the desired job in such a field, adequate representation learning is required. Generative adversarial networks are an emerging research topic nowadays. The paper written by Ian J. Goodfellow et.al. is the base of understanding of GANs[5]. In the proposed approach, the adversarial network is introduced where a discriminative model builds the ability to distinguish samples from the model distribution from samples from the data distribution. The discriminative model is comparable to the police trying to find forged currency, whereas the generative model might be compared to a group of forgers trying to make false money and utilize it secretly. In the aforementioned usecase, competition forces both teams to refine their techniques until the fakes can no longer be distinguished from the actual thing [5], so both generator and discriminator adopt the currently popular deep neural network[5], [6]. GAN optimization is a minimax game procedure, with the goal of reaching Nash equilibrium[7], whereby the generator is regarded to have approximated the spread of real samples.

The paper on deep convolutions GANs developed the GANs for generating new images from original images which work as feature extractors for supervised tasks. GANs offer an interesting substitute for maximum likelihood methods. With the exception of the output layer, which utilizes the Tanh function, the generator uses ReLU(Rectified Linear Unit) activation. We found that the model was able to learn to saturate and cover the training distribution's color space more quickly when a bounded activation was used. Leaky rectified activation performed well within the discriminator, especially for higher-resolution models [6]. Recently published research [8], [9], and [10] has shown their capacity to learn data distributions and perhaps produce fake image-based data. However, as such networks were created to cope with continuous information rather than discrete data[11], they are unfit for use with text sequences. Some studies have used Gumbel-Softmax differentiation [12], reinforcement learning (RL) [13], or tailored training objectives [14] to attempt to address this issue. The purpose of this study is to examine and talk about the methods listed above that are specifically used for text-generating tasks. To build their architecture, the RelGAN [15] proposed a Relational Memory-based generator, a Gumbel-Softmax relaxation, and a multi-embedded representation discriminator. Their studies evaluated state-of-the-art designs (MLE, SeqGAN, RankGAN, and LeakGAN) to real datasets like COCO Image Captions and EMNLP2017 WMT News as well as synthetic data produced by an Oracle LSTM [16]. Below are the benchmark datasets used in the study of detecting fake news.

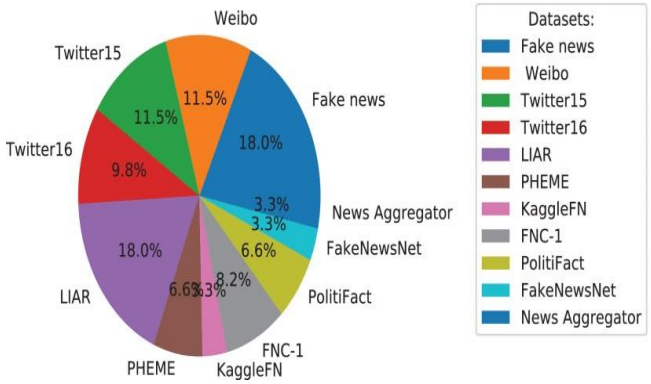


Fig. 2. Dataset Comparison to Detect Fake News

The details of the dataset are given in the below table figure 3
The use of a Relational Memory generator, which more effectively extracts long-term dependencies caused by self-attention layers, is its key benefit. The majority of Gumbel-Softmax-based techniques directly rely on typical GAN objectives and

Dataset	Modality	Size	Labels	Type	URL
Fake news	Text	20,800	Unreliable, reliable	News articles	https://www.kaggle.com/c/fake-news/data.
Weibo [27]	Text & image	40k tweets	Rumor, Non-rumor	Social media data	https://drive.google.com/file/d/14VQ7EWPiFeGzxp3XC2DeEHi-BEisDINn/view
Twitter15 [28]	Propagation trees	1,381 propagation trees, 276,663 users	Unverified, true, false, non-rumor	Social media data	https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0
Twitter16 [28]	Propagation trees	1,181 propagation trees, 173,487 users	Unverified, true, false, non-rumor	Social media data	https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0
LIAR [29]	Text	12.8K	Pants on fire, false, barely true, half-true, mostly true, and true	Political statements	https://paperswithcode.com/dataset/liar
PHEME [30]	Text	5800 tweets	Rumor, Non-rumor	Social media data	https://figshare.com/articles/dataset/PHEME_dataset_of_rumours_and_non-rumours/4010619
FNC-1	Text	75K	Agrees, disagrees, discusses, unrelated	News articles	https://github.com/FakeNewsChallenge/fnc-1
FakeNewsNet [31]	Text	5K	Fake, real	News articles, so-cial media data	https://github.com/KaiDMML/FakeNewsNet
News Aggregator	Text	422,937	Real	News articles	https://www.kaggle.com/uciml/news-aggregator-dataset
Bend the truth [32]	Text	900	Fake, real	News articles	https://github.com/MaazAmjad/Datasets-for-Urdu-news.git
FacebookHoax [33]	Text	15,500	Hoax, non-hoax	scientific news	https://github.com/gabll/some-like-it-hoax/tree/master/dataset
Twitter [34]	Text and Image	992	Rumor, non-rumor	Fact-checked claims	https://github.com/MKLab-ITI/image-verification-corpus/tree/master/mediaeval2015
KaggleFN	Text	13K	Fake	News articles	https://www.kaggle.com/mrisdal/fake-news
FakevsSatire [35]	Text	486	Fake, satire	Political news	https://github.com/jgolbeck/fakenews

Fig. 3. Dataset details used in the study of fake news detection

have a pre-training burden prior to adversarial training, which could lead to premature collapsing and an insufficient generator–discriminator equilibrium. To overcome this problem. The goal is to receive reward signals from the discriminator and send them back to the generator similarly to how a Reinforcement Learning technique would work [17]. The first RL-based work has been introduced by Yu et al.[16], denoted as Sequence Generative Adversarial Network (SeqGAN). SeqGAN was the first RL-based work, and even though it could only be compared with conventional MLE-based training, it set the benchmark high for subsequent RL-based research. Even though SeqGAN addressed. The model lacks consistency for adversarial-based text production, text composed with little diversity and with power. Another SeqGAN extension, denoted as Objective-Reinforced Generative Adversarial Networks (ORGAN)[18] proposed that a linear combination be applied to extend the reward function.
Arjovsky et al.[19] put forward Wasserstein GAN (WGAN) by employing the EarthMover separated as a substitute for the Jensen-Shannon variation for determining the spectrum of difference between real data and the generated data in order to address the vanishing gradient problem. Then there was distance training with Wasserstein. The rewards between the discriminator and the domain-specific objectives, Then there was distance training with Wasserstein. The results of their studies have been superior to conventional baseline approaches, e.g. MLE and SeqGAN, while the authors conclude that domain-based data can be generated by ORGAN and that RL plays an essential role in model learning.

3.0. PROPOSED APPROACH

The basic idea of GAN comes from game theory's Nash equilibrium. It assumes two players: generator and discriminator. The generator's goal is to learn what is the spread of real data, whereas the discriminator's goal is to accurately discern

whether the information being supplied comes from genuine information or from the generator. To win this match, the two players must consistently optimize themselves to enhance their generation and discrimination abilities, respectively. The optimization process's goal is to find a Nash equilibrium among the two different participants[20].

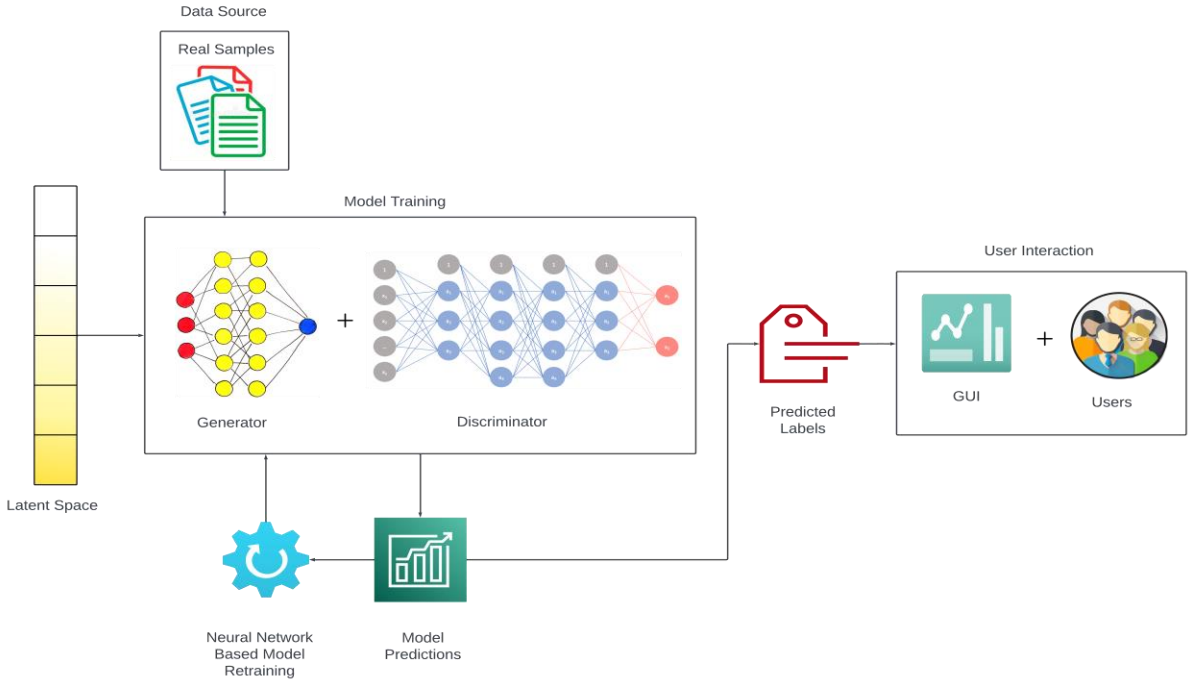


Fig. 4. Proposed Architecture

The proposed architecture diagram 4 depicts a graphical representation of the parts utilized to construct the project by abstracting the overall system outline, linkages, and restrictions. It includes various components like dataset with original news headlines, latent variable space, model training and retraining, model predictions, and user interaction. To train the model, the data[21] consists of around 210K news headlines collected from HuffPost[22] between 2012 and 2022. It is divided into different categorical news headlines like Politics, Entertainment, Business, Sports, etc. Model training has two different neural network models, one generates the fake news from the original headlines, and another neural network is made to identify whether the data is fake or real. Once the models are trained, the model predictions are done to check how a model is performing. The predictions are sent to the end user.

4.0. PROPOSED FLOW OF THE SYSTEM

The following flow diagram 5 illustrates the planned sequence of operations and interactions within the proposed system. It serves as a visual representation of the system's overarching structure and the interplay of its key components. This proposed flow is designed to optimize efficiency, enhance user experiences, and achieve specific objectives that the system aims to fulfill. In the subsequent sections, we will delve deeper into each component of the flow, elucidating their roles and relationships to provide a comprehensive understanding of the system's functionality and its ability to meet the desired outcomes.

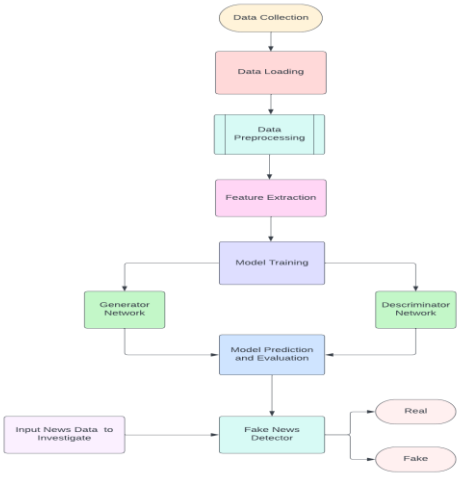


Fig. 5. Proposed Flow chart

A. Data Preparation and Preprocessing

As described above, the data preparation is done by finding and collecting the relevant dataset of news headlines from different data sources.

For the data pre-processing step, the standard text preprocessing techniques are performed to remove punctuations and artifacts, replacing numbers with special tokens and special characters with some words. This will help to clean and prepare insightful textual data for further tasks to perform.

Remove punctuation: The periods, commas, exclamation marks, etc. are removed from the data as they are of no use in providing meaningful insight.

***Handling artifacts*:** The unwanted characters, symbols, or special characters existing in the dataset, are handled by removing it or replacing it with meaningful words.

***Replacing numbers with special tokens*:** The numbers appearing in the texts are replaced by special tokens like < *NUM* > to treat them as a single entity rather than different numbers.

Tokenization is applied to work on textual data, which splits text into smaller units like words or subwords. It is important to work a model with discrete units of texts. Once tokens are created from paragraphs, the words having frequency < 1 are removed from the corpus to optimize the corpus size and reduce overhead.

< *Start* > and < *End* > sentence tokens are added to each sentence to indicate the beginning and ending of the sequence. This is done to work with seq-to-seq models.

B. Feature Extraction

To extract the features from a given text corpus, autoencoders are used as they are designed to learn a compressed representation of data. They are helpful in learning hidden representations of data called latent vectors or embeddings which capture the most important and relevant information from the given data. The whole space is called latent space. Here is how it works in feature extraction:

Encoder: takes input data and maps it to lower-dimensional representation in the vector space called latent vector. It learns to capture the necessary information from the data.

Latent space: The latent vector resides in the space, which is called latent space. It is a lower dimensional space where data is projected. The size of the latent space can be adjusted as the value of the hyperparameter.

Decoder: It takes the latent vector as an input and reconstructs the original input vector. The goal of a decoder is to create as near to the original input data.

C. Training Model

Different models are trained to achieve the goal. First, the skip-gram model which is the word2vec model is used to represent words as dense vectors in continuous feature space. It is trained for vector embeddings to predict the next word from the given input as it works on predicting the next word with maximum likelihood approximation. From the resulting word embeddings, the semantics between the words can be captured, which is important in the generation of text semantically.

Model training of GANs employs two distinct neural network models: one that generates false news from actual headlines and another that determines if the data is true or not.

D. Model Predictions and Evaluations

Once the model is trained, it is important to check the correctness of the model. So the model predictions are performed by providing a train, test, and validation set to evaluate the model.

E. User Interaction

After successful training and evaluation of the system, it is time to give input to the fake news detector on whether it is properly able to identify the news from the user end. It will give the result that the news is real or fake.

5.0. EXPERIMENTAL RESULT

In the figure 6, the time taken to train each epoch, training, and testing loss for each epoch is shown which helps to get an idea about whether a model is overfitting or underfitting the data and how well it generalizes on data. Along with it, the generated fake news is shown. The proposed system will work like this. At the end, the autoencoder model with its parameters is shown which helps to make inferences or generate text. Overall, the above image helps to monitor the model's inference and display the model's final parameters.

From the figure 7, the loss between training and testing can be analyzed. It can be seen that, as the epochs count increases, the difference between training and testing loss increases, which is an indication of overfitting. With an increase in the number

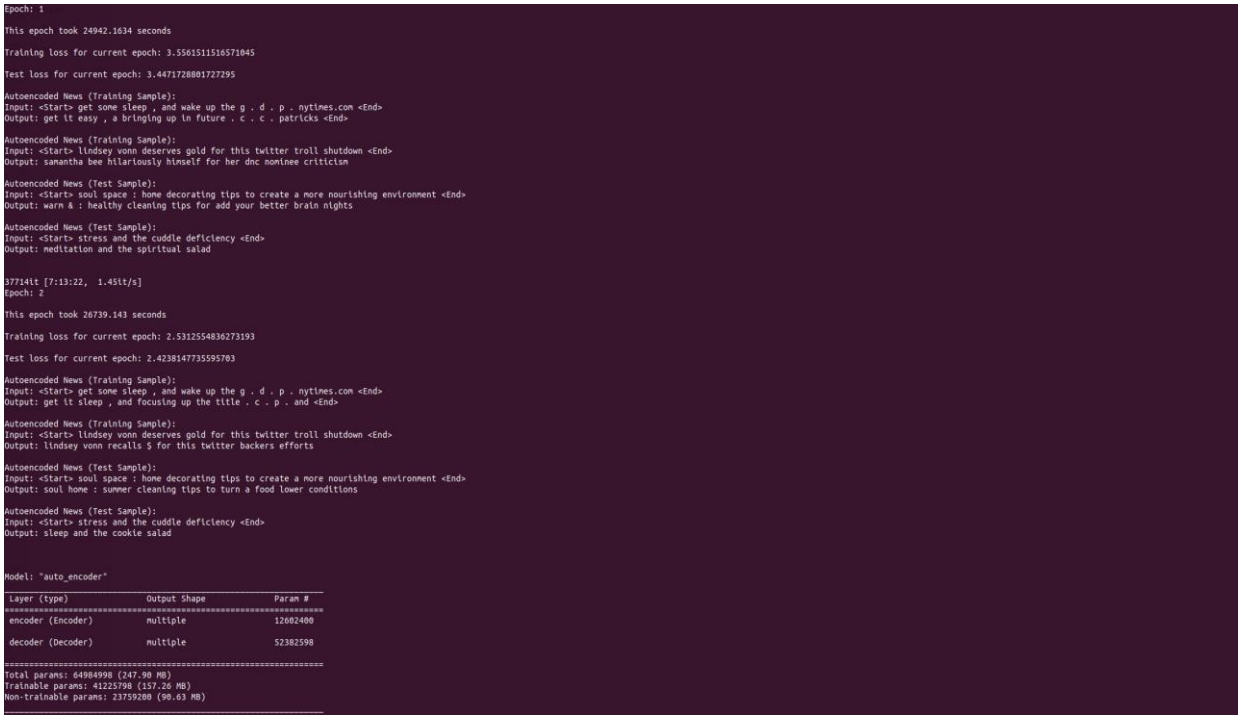


Fig. 6. Snapshot of training the model

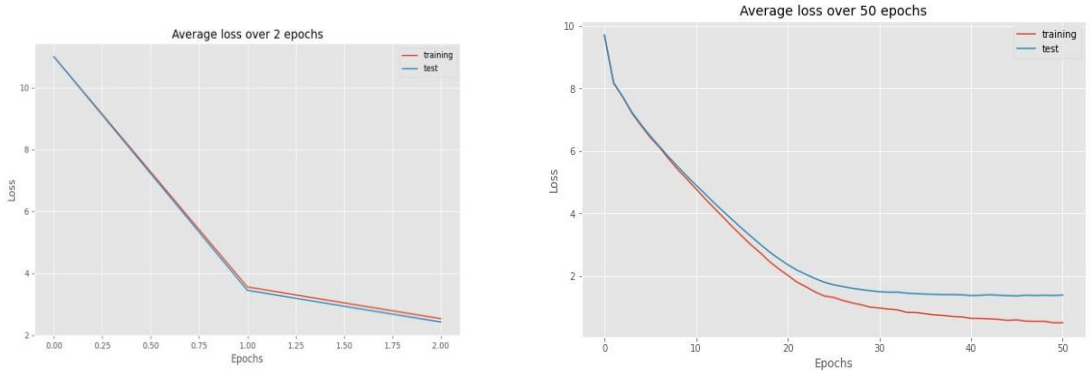


Fig. 7. Model training and testing loss

of epochs, the model tries to memorize data instead of finding underlying patterns because the model has more opportunities to fine-tune the parameters. The monitoring of this loss will further help to identify when the model's performance is degrading and intervention is needed.

6.0. CONCLUSION AND FUTURE WORK

This study introduces a unique unsupervised learning strategy based on GANs for producing and identifying fake news items. The results show that employing adversarial generative models to handle the issues of fake news detection without depending on labeled datasets is feasible and successful. The suggested approach not only improves the reliability of fake news detection but also sheds light on the methods used by bad actors to create fake news.

In the next phase, the model will be trained on more number of epochs to reduce the loss and overfitting of data. The model will be tuned with hyperparameter tuning and applying regularization techniques (i.e. L1, L2), early stopping techniques, and increasing size of the training dataset. The versatility of this strategy makes it an important weapon in the continuing war against fake news as disinformation continues to change.

REFERENCES

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–36, May 2017. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211

[2] T. Rasool, W. H. Butt, A. Shaukat, and M. U. Akram, "Multi-label fake news detection using multi-layered supervised learning," in *Proceedings of the*

[3] *2019 11th International Conference on Computer and Automation Engineering*, ser. ICCAE 2019. New York, NY, USA: Association for Computing Machinery, 2019, p. 73–77. [Online]. Available: https://doi.org/10.1145/3313991.3314008

[4] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing Management*, vol. 57, no. 2, p. 102025, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457318306794
[5] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, and M. S. Rahman, "A comprehensive review on fake news detection with deep learning,"
[6] *IEEE Access*, vol. 9, pp. 156151–156170, 2021.
[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.
[8] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016.
[9] J. Patel, M. Pandya, and V. Shah, "Review on generative adversarial networks," vol. 4, p. 1230, 07 2018.
[10] J. Chen, Y. Wu, C. Jia, H. Zheng, and G. Huang, "Customizable text generation via conditional text generative adversarial network," *Neurocomputing*, vol. 416, pp. 125–135, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219308744
[11] K. Wang and X. Wan, "Automatic generation of sentimental texts via mixture adversarial networks," *Artificial Intelligence*, vol. 275, pp. 540–558, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370218306088
[12] Y. Li, Q. Pan, S. Wang, T. Yang, and E. Cambria, "A generative model for category text generation," *Information Sciences*, vol. 450, pp. 301–315, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025518302366
[13] G. Rizzo and T. H. M. Van, "Adversarial text generation with context adapted global knowledge and a self-attentive discriminator," *Information*
[14] *Processing Management*, vol. 57, no. 6, p. 102217, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457319303541
[15] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," 2017.
[16] R. Sutton and A. Barto, *Reinforcement Learning, second edition: An Introduction*, ser. Adaptive Computation and Machine Learning series. MIT Press, 2018. [Online]. Available: https://books.google.co.in/books?id=uWV0DwAAQBAJ
[17] T. Che, Y. Li, R. Zhang, R. D. Hjelm, W. Li, Y. Song, and Y. Bengio, "Maximum-likelihood augmented discrete generative adversarial networks," 2017. [15] W. Nie, N. Narodytska, and A. Patel, "Relgan: Relational generative adversarial networks for text generation," May 2023. [Online]. Available: https://openreview.net/forum?id=rJedV3R5tm
[18] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/10804
[19] G. H. de Rosa and J. P. Papa, "A survey on text generation using generative adversarial networks," *Pattern Recognition*, vol. 119, p. 108098, 2021. [Online]. Available: https:// www. Sciencedirect . com/ science /article /pii/S0031320321002855
[20] G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias, and A. Aspuru-Guzik, "Objective-reinforced generative adversarial networks (organ) for sequence generation models," 2018.
[21] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017.
[22] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, "Generative adversarial networks: introduction and outlook," *IEEE/CAA Journal of*
[23] *Automatica Sinica*, vol. 4, no. 4, pp. 588–598, 2017.
[24] R. Misra, "News category dataset," *arXiv preprint arXiv:2209.11429*, 2022.
[25] [Online]. Available: https://www.huffpost.com/