

FAKE NEWS DETECTION USING NOVEL GAN BASED TECHNIQUE

Mr. Vivek Joshi¹, Dr. Sanjay Patel²

¹Kadi Sarva Vishwavidyalaya, Gujarat, India

¹School of Information Technology, Artificial Intelligence and Cyber Security,
Rashtriya Raksha University, India

²Department of Computer Science and Engineering, Nirma University, India

Abstract

The proliferation of fake news across digital platforms presents a critical threat to information credibility and the integrity of public discourse. This paper proposes a novel fake news detection framework based on Generative Adversarial Networks (GANs). The framework leverages GANs to synthesize news articles while training a discriminator to effectively distinguish between authentic and fabricated content. Through adversarial learning, the discriminator progressively enhances its classification capability by identifying subtle linguistic and contextual cues that separate real news from misinformation. Experimental evaluation demonstrates the robustness of the proposed model in improving detection accuracy, highlighting its potential for deployment in automated verification systems and digital media integrity solutions.

Key Words and Phrases: Generative Adversarial Network (GAN), Natural language Processing, SeqGAN, LatentGAN, Generators, Language Modeling, LSTM

1. Introduction

In today's digital age, people increasingly turn to the internet and social media for their daily news. These platforms are popular due to their low cost, rapid updates, and ease of access, allowing users to stay informed anytime and anywhere. However, alongside these benefits comes a growing concern: the spread of fake news. This refers to intentionally misleading or false information designed to deceive readers, often with harmful consequences for individuals and society at large.

The rise of fake news has become especially noticeable since the 2016 U.S. presidential election. Research has shown that politically charged misinformation can significantly influence public opinion and voter behavior. In fact, false stories tend to spread faster than factual ones, creating widespread confusion and impacting political, social, and economic stability. Since people often base their decisions on the information they consume, exposure to fake news can lead to flawed conclusions and reduce trust in reliable sources.

One of the main challenges with online news is the lack of regulation compared to traditional media. This has led to an overwhelming amount of low-quality content, making it harder to distinguish between truth and falsehood [1]. The problem has been further complicated by

advancements in artificial intelligence, which now enable not only humans but also machines to generate convincing fake news.

To combat this, researchers have developed various tools to verify the authenticity of news articles, including those that analyse images containing headlines and text. However, many of these tools are slow and require significant manual effort. A promising solution lies in the use of Generative Adversarial Networks (GANs), which consist of two components: a generator and a discriminator. The generator produces synthetic data, while the discriminator evaluates it to determine whether it is real or fake. Through this adversarial process, both components improve over time: the generator becomes better at creating realistic fake content, and the discriminator becomes more skilled at detecting it. One of the key strengths of GANs is their ability to generate large amounts of synthetic data. GANs are generally used for fake image or video detection [2]. But we think GAN is particularly useful in fake news detection, where labelled datasets are often limited. By training on a wide variety of examples, GANs can learn to recognize subtle patterns in language, tone, and structure that traditional models might miss. Unlike rule-based systems or standard classifiers, GANs do not rely on predefined features. Instead, they learn directly from the data, identifying linguistic cues, narrative inconsistencies, and emotional triggers that are often used in deceptive content. This makes them especially effective at detecting more sophisticated forms of misinformation that closely resemble legitimate news.

The fake propagation is too fast and too difficult to solve through social media [3] and hence a unique technique has been implanted to get the desired result.

Our approach emphasizes automation and adaptability, relying on unsupervised learning rather than manually labelled data. This makes it a scalable and efficient solution for the dynamic and fast-paced world of online news, offering a powerful alternative to traditional detection methods.

2. Literature Survey

The allure of deep learning stems from its ability to simulate human-like tasks through hierarchical learning, captivating considerable interest. However, achieving high-quality performance in such fields necessitates effective representation learning. In this context, generative adversarial networks (GANs) have emerged as a prominent research focus. The foundational paper authored by Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio lays the groundwork for understanding GANs [9]. As proliferation of data on the web and on personal computers is a direct result of the proliferation of new technologies and gadgets. Most of these pieces of information are collected through a number of different methods (text, image, video) a multimodal model approach for data categorization utilising neuromorphic fuzzy sets has shown good results [10]. Similarly, any multi varied approach with respect to data will more likely give better results.

we introduce an adversarial network-based model. This different model plays a role akin to the police discerning counterfeit currency, while the generative model functions like a group of

forgers producing false currency for clandestine use. This adversarial game compels both teams to enhance their techniques until the forgeries become indistinguishable from their authentic counterparts. Notably, both the generator and discriminator leverage unique neural networks in this competitive process. Generative Adversarial Networks (GANs) are used primarily for analysis of images and motion pictures. The generative approach has been giving good results in identifying and extracting data from images [11]. It performs deep learning transformation which has intended goal to reach to level of perfection in result in which the outcome become indistinguishable from real world result. The GAN is primarily used for image generation and fake image detection [11] but we have used in for text-based approach. We have used sequence of token as an input and used reinforcement techniques for training the generator. The generator generated the sequence of word token then comprised to form words. Once the generator is trained on good amount of data the discriminator will be used to classify from real news to fake news. The main advantage of using deep learning method like GAN for text based detection is it can help in identifying varieties of text based on its writing style, token pattern and GAN helps in creating comprehensive summaries of text based bodies. Another place where generative adversarial model is useful in the context of text is to identify the context from the given text. The context awareness with sentiment analysis provides proper identification of the intent of the writer while writing text. A key advantage lies in the use of a Relational Memory generator, which efficiently extracts long-term dependencies through the application of self-attention layers. Many Gumbel-SoftMax based techniques are constrained by their reliance on standard GAN [13] objectives and require extensive pre-training before adversarial training begins. This dependency can lead to premature collapsing and disrupt the critical balance between the generator and discriminator, ultimately limiting performance.

The objective is to utilize reward signals from the discriminator and relay them to the generator in a manner akin to Reinforcement Learning (RL) techniques [14]. Yu et al. [15] introduced the first RL-based approach, known as Sequence Generative Adversarial Network (SeqGAN). As the pioneering work leveraging RL in this context, SeqGAN established a strong benchmark, despite being primarily compared to traditional Maximum Likelihood Estimation (MLE) training methods. While SeqGAN made meaningful leaps in adversarial-based text generation, the model shows inaccuracy, particularly in generating text with less sufficient diversity and less robustness. In Addition of previous model, the SeqGAN, known as Objective-Reinforced Generative Adversarial Networks (ORGAN) [15], introduces the idea of incorporating a linear combination into the reward function to enhance performance. Arjovsky et al. [16] developed the Wasserstein GAN (WGAN), which utilizes the Earth Mover's distance instead of the Jensen-Shannon divergence to measure the difference between real and generated data. The method was used for dealing with problem related to gradient of text. Which is normal issue faced by deep learning models. Adding Wasserstein displacement into the training process, the method has improved the reward signs between the discriminator and domain-specific objectives. As a corollary the method outperformed traditional baseline methods. Paper authors analysed that proposed method is Advantageous for generating domain-specific data, with reinforcement learning playing a pivotal role in refining the model's process.

The research [17], [18], [19] have shown the capacity of adversarial based models to perform better in data distributions and create fake, picture database with efficiency. But, applying adversarial based models to text generation poses unique challenges due to the discrete nature of text, unlike the continuous data GANs are primarily designed to handle [19]. To outperform the issues mentioned, authors had proposed many optimizing strategies, including Gumbel-Softmax differentiation [13], reinforcement learning and tailored training objectives [15]. The research implores into these advantages, focusing specifically on text generation tasks. RelGAN [20], It introduced a generator leveraging Relational Memory and Gumbel-Softmax relaxation. It paired this with a discriminator utilizing multi-embedded representations to evaluate performance. The creative method had first been demonstrated with the Sequence Generative Adversarial Network (SeqGAN) [10], introduced by Yu et al., which laid the groundwork for future RL-driven advancements. Even after ground-breaking innovation of SeqGAN the method had met with many issues. In order to rectify the issues, Objective-Reinforced Generative Adversarial Networks (ORGAN) [15], an extension of SeqGAN, introduced an enhanced reward mechanism. With using a linear combination in the reward function, ORGAN Showed increased efficiency in creating adversarial and generative data Arjovsky et al. [20] has showed new innovative method the Wasserstein GAN (WGAN), which has taken place of Jensen-Shannon divergence with Earth Mover's distance to better capture discrepancies between real and generated data. The creative approach has solved issued like vanishing gradient. The Wasserstein distance-based training outshines in performance to traditional models like MLE and SeqGAN, creating an effective methodology to follow for text-based GAN mode. The core principle of GANs revolves around Nash equilibrium [21], a concept rooted in game theory.

The GAN does provide a good and efficient way of detecting text based fake news. However, the data set plays a major role in giving precision and efficiency. The dataset we have chosen is diverse ain nature. Most of the research based on fake news have focused solely on politics based fake news we have collected news article form diverse field such as entertainment, fashion, food and art. This diverse form of data gives more precise results.

3. Dataset And Proposed Flow of The System

Our research used the benchmark datasets for studying the detection of fake news. The diverse dataset makes for best input to train GAN based model

Table.1 Dataset details used in the study of fake news detection

Data Set	Modality	Size	Labels	Type
Fake news	text	20,800	unreliable	News Articles
Weibo	Text, Image	40000	Original & fake	posts

Twitter(X)	Propagation tree	27763	Unverified, true false	posts
Fake News Net	Text	20000	Fake & real	News articles
LIAR	Text	12800	Fake & real	Articles data set

We use dataset presented in Table.1 which encompasses a diverse range of modalities, sizes, and labels, each tailored to address different aspects of misinformation detection. The Fake News dataset consists of 20,800 text-based news articles labeled as unreliable, making it suitable for analyzing textual patterns indicative of fake news. The Weibo dataset, comprising 40,000 entries, includes both text and image modalities, capturing original and fake posts from the social media platform, thus enabling multimodal analysis of misinformation. The Twitter (X) dataset focuses on propagation trees, containing 27,763 posts labeled as unverified, true, or false, which is particularly useful for studying the spread and network dynamics of misinformation. The Fake News Net dataset offers 20,000 text-based news articles labeled as fake or real, providing a robust resource for binary classification tasks. Lastly, the LIAR dataset includes 12,800 text-based articles labeled as fake or real, offering a comprehensive collection for training and evaluating models on article-level misinformation detection. Collectively, these datasets provide a rich foundation for research into the detection, classification, and propagation of misinformation across various platforms and modalities. Figure.2 presents a comprehensive architectural diagram, illustrating the system's components and their interconnections. It visually outlines elements such as the dataset of original news headlines, the latent variable space, training and retraining mechanisms, model predictions, and user interaction pathways. This structure highlights the system's flow and constraints. The dataset component, contains approximately 210,000 news headlines sourced from Huff Post [22], covering the period from 2012 to 2022. These headlines are categorized into diverse topics, including Politics, Entertainment, Business, and Sports

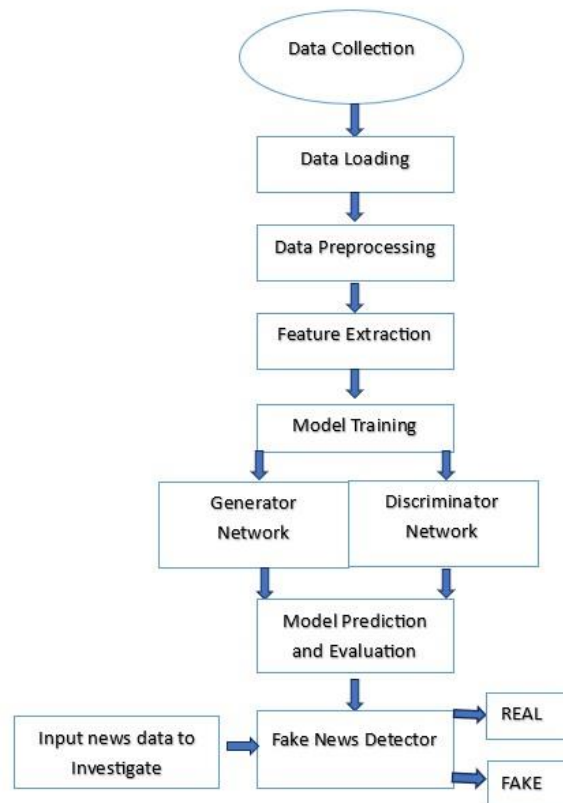


Figure. 2.Flow Diagram of Proposed Architecture

Model training employs two neural networks: one dedicated to generating fake news based on the original headlines and another focused on classifying the headlines as either real or fake. As the models which were trained come to fruition than a comprehensive analysis is done for their predictions which is than shown to end user. The methodology implemented here improves upon the past sequence of steps taken for similar exercise and proves the efficacy to the end user. The figure 2. shows the overall structure to be followed This graphical representation highlights the system's overall structure shows the entire step by step guideline to the proposed system. The flow chart also depicts the dependencies of data at each stage of the flow. This data dependency is crucial in deep learning method like GAN because it becomes the basis on which the generator and discriminator are trained. This clear visual communication helps in understanding and implementation of suggested model. In the proposed approach, the first step involves identifying and collecting relevant datasets of news headlines from diverse sources. Once the data is gathered, it undergoes pre-processing to ensure it is clean and ready for subsequent tasks.

In the process of converting raw text into token which is called tokenization we perform classification of text in to n-gram sub words. steps we mentioned is very important for data preparation as the generator and discriminator will rely Soley in to this token for training. The process also involves appending keyword before and after tokens to identify individual tokens. The differentiation of words in to tokens become the pillar on which crucial task are to be

performed. Another benefit of tokenization is that it helps in reducing noise from unlabelled data. The second methodology that is word2vec the we have used requires to perform tokenization in order to get efficient outcome. word2vec has been significantly useful specially while using Convolution Neural Network (CNN) [23]. Another reason to word2vec is it supports Bag of Word (B.o.W) [24] Classification of text-based data.

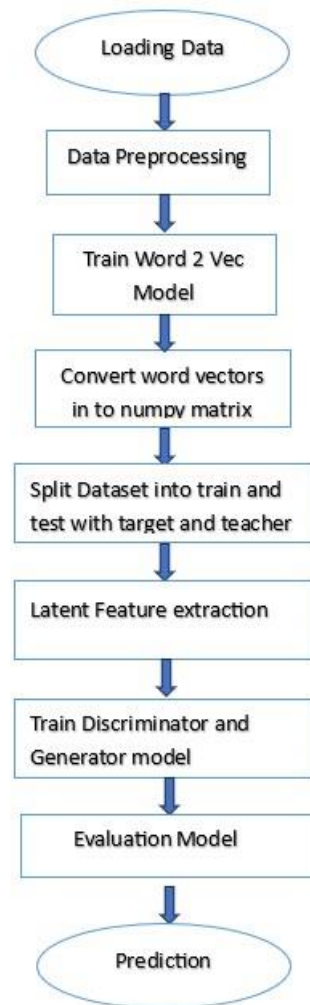


Figure. 3.Detailed Flow chart

A. Feature Extraction

As this is one of the most critical stage for the proposed model we make sure to minimize the mistakes as much as possible. The use of Autoencoders [25] is one of the most significant steps in the entire process. Autoencoders are powerful tools for feature extraction in text processing. The ability of read compressed and mindful content for the text makes autoencoders one of the best of feature extraction methods. It has capability to understand the and recognize features from latent vectors of the texts .As the each cycle of data processing happens the autoencoders ability of extracting the meaning feature extraction becomes more prominent and precise. This also helps in noise removal and reduction awhile extracting feature from given dataset. It works very well with unlabelled data. Another feature of it helps identifying

anomalies in the dataset used. Other than autoencoder we use latent vectors in feature extraction. With this we use Encoder. The main purpose of encoder is to transform higher dimensional data into latent vectors. The function of encoders is to identify underlying patterns from higher dimensional data. The encoder's also function as identification of essential information from the input while removing noise from the details. As an output to encoder the latent space which includes the latent vector resides in this space. The latent vector is a limited size numerical representation of input data usually in compressed form. The latent vector recognizes the internal patterns and condense it in meaningful yet useful form. In Adversarial model such as GAN the latent vectors act as an input to the generator. The generator maps random vector to sample data. Input as random noise vector V and mapping takes place in generator from V to sample data.

B. Model Training

The training of the output of the autoencoders[25] that is latent vectors which are collected from latent space. As shown in figure 3. We follow series of steps to get desired output. For getting useful output we have deployed word2vec as mentioned in the previous paragraph and in addition we use skip gram model which is another version of word2vec. By using word2vec we get dense vectors in a sequential feature space. The method predicts the next words for the input provided as latent vectors. This model is very efficient when used word2vec[26] in predicting the next word and identifying semantic relationship between words. These embeddings are essential for producing text that is contextually and semantically coherent. For Training the proposed model we would require two neural networks for output. As the placement of first neural network will create the fake news headlines based on the unlabelled data set provided as input. The second neural network will be trained to identify fake news and hence at each cycle of training will improve the efficiency of identification of fake news. The set of two neural network training and repeating cycle requires high end hardware as the size of labelled data is very significant. This approach would require bidirectional layer and processing. A bidirectional Stacked LSTM(Long Short Term memory)[27] is suited for the proposed method. This is particularly useful as we are using sequential data for model training. The LSTM layers allow the model to capture complex data patterns from the given dataset. Each layer adds a new level of abstraction, enabling the generator to produce sequences with richer and more precise content. The method also includes bidirectional processing in which as we are traversing input sequences in both forward and reverse directions, the model learns from preceding and succeeding contexts simultaneously. This bidirectional ability increases the model's contextual understanding, improving coherence and fluency in generated text. For comprehending the meaning and context which is very important for predicting the next word and creating a meaningful sentence and paragraph. Integrating this bidirectional method which is stacked with LSTMs enables the encoder to process input sequences holistically. It ensures a more comprehensive ability to get text's context. While using these methods, the system achieves precise and meaningful text generation, characterized by semantic context, contextual accuracy, and enhanced comprehension of the text. This approach also decreases the need for many passes which are essential for neural network based models. Through the data,

significantly improving the training accuracy. The efficiency of capturing a broader context in both directions enhances the model's understanding of the input data. This results in improved training efficiency and more effective performance in text generation. Another benefit of using in the proposed model is to get proper sentiment analysis of both generated and detected text. The proposed model aims to identify fake news to the high degree of accuracy with analysing semantic and syntactic meaning of the text .The generator and identifiers part of the GAN requires training of both

4. Result

As the training of model is reaching to completion point , the accuracy and precision of the model has to be tested . The testing is achieved using evaluation prediction , training, testing, and validation datasets. The BLEU(Bilingual Evaluation Understudy)[28] score is employed as a metric to assess the model's accuracy by comparing the actual text with the generated text. This method is used to commonly evaluate the machine capacity to generate quality text. It measure how closure the machine generated text is to the real text .The methodology deployed by BLEU For calculating score is by comparing n-grams in the candidate text compare to the reference text. To give the precise score BLEU imposes a brevity penalty is the text is too short in order to avoid inflated score. To start the BLEU calculation process for the research we started by taking 1500 random samples from the generated dataset which resulted in BLEU score of 0.31. Than we increased the input from 1500 to 5000 which in turn increased the BLEU score to 0.32 which didn't make much difference on the quality of the score. After that we increased the sample size to 25000 which shown increase of BLEU score by 0.43. The increase in sample size improves the quality of BLEU score because larger the sample size larger comprehensiveness of the scoring mechanism Another factor which improves with larger data sample is data variability. This in text samples increase the accuracy of evaluation hence better score. As we increase the number to 50000 to 75000 the accuracy peaks at 0.56 and for 75000 it remains at 0.58 because as the sample data size increases we average out over the data set. As data set grows contribution of smaller sentences decreases in overall score. A higher BLEU-4[28] score signifies a greater resemblance between the generated text and the actual text, reflecting the model's capability to produce accurate and coherent outputs. However, the previously developed discriminator model goes beyond this metric by reliably distinguishing between real and generated text with remarkable precision and minimal error. This robustness enhances the proposed approach, making it a highly effective tool for detecting false news.

Table 2. Evaluation of BLEU Score

Standard BLEU Score	Quality	Sample dataset	BLEU Score of Sample dataset
More than 0.7	Excellent	500	0.31
0.7-0.5	Good	5000	0.32
0.5-0.3	Average	25000	0.43

Less than 0.3	Poor	50000	0.56
---------------	------	-------	------

5. Conclusion And Future Work

The BLEU method used for analysis is an efficient way of identifying the accuracy and precision of data generation to the reference model. The model has shown that as we increase the sample data size the accuracy of model increases, and it shows how effective it has performed. By analyzing the result, we conclude that GAN based model can be effective and precise method to identify fake news detection using specially with autoencoders as shown in our proposed methodology. The drawback of this model is that it hits plateau when only small set of data or small text size is used. In future we will add language analytics-based scoring metrics to analyze the accuracy of model.

References

- [1] J. R. Babu, V. Venkataiah, T. L. Narayana, C. R. Rao, C. D. Raj, R. C. H. S. N. P. Sairam, and A. B., "Analyzing the Effectiveness of ML and DL in Combatting Personal Condition Misinformation on Social Media," *IAENG International Journal of Computer Science*, vol. 52, no. 1, pp. 258–267, Jan. 2025.
- [2] G. Padmashree and A. K. Karunakar, "Ensemble of Machine Learning Classifiers for Detecting Deepfake Videos using Deep Feature," *IAENG International Journal of Computer Science*, vol. 50, no. 4, Dec. 2023. [Revised online: Dec. 15, 2023].
- [3] Y. Ding and A. Muhammadhaji, "Dynamic Analysis and Control of a Rumor Propagation Model with Saturation Incidence in Online Social Networks," *IAENG International Journal of Applied Mathematics*, vol. 55, no. 7, pp. 1955–1971, Jul. 2025.
- [4] Stanford study examines fake news and the 2016 presidential election," *Stanford News*, Jan. 18, 2017. [Online]. Available: <https://news.stanford.edu/stories/2017/01/stanford-studyexamines-fake-news-2016-presidential-election>. [Accessed: September 21, 2024]
- [5] 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.
- [6] T. Brown and B. Shetty, "A Turn of the Page for Newsweek," *The Daily Beast*, Oct. 18, 2012. [Online]. Available: <https://www.thedailybeast.com/articles/2012/10/18/aturn-of-the-page-for-newsweek.html>. [Accessed: September 21, 2024].
- [7] A. Loth, M. Kappes, and M.-O. Pahl, "Blessing or curse? A survey on the Impact of Generative AI on Fake News," 2024 IEEE International Conference on Artificial Intelligence and Applications (ICAIA), 2024, pp. 1-8.
- [8] A. Ahmed, I. Traore, and S. Saad, "Detecting Fake News Using Machine Learning: A Supervised Approach," 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, 2018, pp. 460-464.

- [9] Chaudhary, L., Singh, B. Gumbel-SoftMax based graph convolution network approach for community detection. *Int. j. inf. tecnol.* 15, 3063–3070 (2023). <https://doi.org/10.1007/s41870-023-01347-y>.
- [10] Digi, M. TwitterGAN: robust spam detection in twitter using novel generative adversarial networks. *Int. j. inf. tecnol.* 15, 3103–3111 (2023). <https://doi.org/10.1007/s41870-023-01352-1>.
- [11] Marriwala, N.K., Shukla, V.K., William, P. et al. Detection of viral messages in twitter using context-based sentiment analysis framework. *Int. j. inf. tecnol.* 16, 5069–5075 (2024). <https://doi.org/10.1007/s41870-024-02084-6>
- [12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," arXiv preprint arXiv:1406.2661, 2014.
- [13] Wajid, M.A., Zafar, A. & Wajid, M.S. A deep learning approach for image and text classification using neutrosophy. *Int. j. inf. tecnol.* 16, 853–859 (2024). <https://doi.org/10.1007/s41870-023-01529-8>
- [14] Cherian, A.K., Vaidhehi, M., Arshey, M. et al. Generative adversarial networks with stochastic gradient descent with momentum algorithm for video-based facial expression. *Int. j. inf. tecnol.* 16, 3703–3722 (2024). <https://doi.org/10.1007/s41870-024-02043-1>
- [15] 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," *IEEE Access*, vol. 9, pp. 156151–156170, 2021.
- [16] E. Jang, S. Gu, and B. Poole, "Categorical Reparameterization with GumbelSoftmax," arXiv preprint arXiv:1611.01144, 2016.
- [17] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2017). Generative adversarial nets. *Communications of the ACM*, 60(11), 67-72. <https://doi.org/10.1145/3137597.3137600>
- [18] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, San Francisco, CA, USA, Feb. 2017, pp. 2852-2858
- [19] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," arXiv preprint arXiv:1701.07875, 2017. <https://arxiv.org/abs/1701.07875>.
- [20] 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.
- [21] K. Wang and X. Wan, "Automatic generation of sentimental texts via mixture adversarial networks," *Artificial Intelligence*, vol. 275, pp. 540–558, 2019.
- [22] 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.
- [23] W. Nie, N. Narodytska, and A. Patel, "RelGAN: Relational generative adversarial networks for text generation," May 2023.

- [24] J. Nash, "Equilibrium Points in N-Person Games," *Proceedings of the National Academy of Sciences*, vol. 36, no. 1, pp. 48-49, Jan. 1950. doi: 10.1073/pnas.36.1.48.
- [25] Misra, R. (2017). News Category Dataset. Kaggle. Retrieved from <https://www.kaggle.com/datasets/rmisra/news-category-dataset>.
- [26] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8, 53 (2021).
- [27] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513-523, 1988. doi: 10.1016/0306-4573(88)90021-0
- [28] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, "Autoencoders and their applications in machine learning: a survey," *Artificial Intelligence Review*, vol. 57, no. 28, Feb. 2024
- [29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013.
- [30] A. Graves, J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM networks," in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, vol. 4, pp. 2047-2052, 2005.
- [31] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA, July 2002, pp. 311-318.