NEURALFAKEDETNET - DETECTION AND CLASSIFICATION OF AI GENERATED FAKE NEWS

Poorva Sawant, Parag Rane

Mumbai, India

ABSTRACT

Unreliable and deceiving information is spreading at a great speed these days across the world through social media sources. Fake news is a growing problem in our modern society, and it has become increasingly difficult to distinguish between real and fake news due to the advancement of technology. Fake or misinformation about the latest CORONA pandemic wreaked havoc. Studies conducted during the epidemic COVID that false news might have menaced public health broadly. Detecting and averting the spread of unreliable media content is a delicate problem, especially given the rate at which news can spread online. With the increase in the use of social media platforms; the leading cause for spreading such news can be that fake news can be published and propagated online faster and is also cheaper when compared to traditional news media such as newspapers and television. Online fake news or information which is deliberately designed to deceive readers is mostly commonly manually written; but with the recent progress in natural language generation techniques, models have been built to generate realistic looking 'Fake news'. With the explosion of large language models fake news can be easily created and with proper grammar and sentences. This creates a greater need to handle the fake news identification problem in a different way to not just classify the fake and real news, but also to mark the human-generated and machine-generated (neural) fake news. Considering most of the work that is done in this research area, it is found that only the very complex language models that are used as generators and detectors are able to catch the machine generated fake news. Again, such models have been observed to be performing well on their own generated text, but not quite effective while working with text from other language models. Also, they don't seem to be tested on the human generated fake news. Now if someone uses the language model to generate the news and then change a few elements manually to make it look more real; this kind of fake news might go completely undetected by such models. So, there is a considerable scope to further study and analyze the difference as well as similarities in the human and machine fake news. This study looks at the problem of machine-generated fake news classification as more of a comparative analysis of Human Vs Machine Generated fake news and identify the differences or similarities of the patterns.

KEYWORDS

NLP, Fake News, Generative AI

1. INTRODUCTION

Social media and news outlets publish fake news to increase readership in order to gain advertising revenue, influence opinions and so on. (Wardle. C, 2017). The reach of fake news can best highlighted during the critical times like the current state of pandemic. There seems to be

David C. Wyld et al. (Eds): NLPML, AIAP, SIGL, CRIS, SEC, COSIT, DMA -2023
pp. 41-52, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130804

hardly an area left which is untouched by fake information related to the COVID-19 crisis, be it the origin of the coronavirus, or some unproven prevention and 'cures' we can see loads of false information spreading across the world. The ability to identify fake news is crucial to prevent its spread and maintain the integrity of information. In recent years, many techniques have been developed to detect fake news, including natural language processing, machine learning, and deep learning. These techniques rely on various features such as linguistic and stylistic characteristics, social network analysis, and fact-checking. In this paper, we will provide an overview of these techniques and their performance in detecting fake news.

1.1. Neural Fake News

In order to generate fake news articles, Natural Language Processing (NLP) techniques are being used – this concept is called "Neural Fake News"; which are just as undetectable and harmful.
NLP technique where models learn to identify a missing word or predict the next word in a sentence by understanding the context, is called Language Modelling . Such model is good at understanding different writing styles, grammar, etc. and is capable of generating legitimate looking text that appears authentic to human eye. It's scary to know that such models are being used to confuse people by generating targeted propaganda and spread of false information.

With the ease of availability of advanced pre-trained NLP models like BERT, GROVER and off course the most recent CHATGPT. Anyone can easily download and play around such models. Hence the Risk of such models being exploited to deceive humans and create chaos in society is highly aggravated. There is a threat of such fake or incorrect information affecting the search engines as well and hence there is an urgency to defend such bulk generated and fast spreading disinformation before it gets all over the internet. To humans, the disinformation generated by AI-Models appears trustworthy, and surprisingly even more than human written disinformation. (Zellers et al., 2019) Thus, it turns out to be an important research area to develop a robust verification technique for such generators.

In this paper, we present analysis of news writing styles of Humans (Fake Vs Real) and AI generated Fake news articles along with detection techniques, including the methods used to detect real news Vs human fake and AI generated fake news and the challenges faced. We also discuss the evaluation metrics used to measure the performance of fake news detection models, along with the limitations of these metrics. Finally, we conclude with a discussion of future research directions for improving fake news detection.

2. METHODOLOGY

Methodology deployed involves key processes such as identification of the suitable news dataset, generating dataset using 'GROVER 'model (for training). We would further do the data pre-processing like tokenization, extracting features like ngrams, the number of characters, number of words, number of complex words, long words, word types, number of paragraphs, number of syllables etc. We would further investigate and derive new features if needed. Additionally, we will use tf-idf vectorization method for extracting features.TF-IDF is an abbreviation for Term Frequency- Inverse Document Frequency and is a very common algorithm to transform text into a meaningful representation of numbers. Once the feature set is ready, we will visualize and perform some exploratory data analysis to gain some insight or patterns from the data. Linear classifiers like SVM are simpler but are still known to generate quite accurate results when trained on the right set of features. Other classifiers that can also be considered are Logistic regression, Naive Bayes as they are also extensively used for the task of text classification. We plan to explore the model performance on various combinations of features to achieve higher accuracy classification. We would thus identify and report the most impactful differentiating features of the Human Vs Machine generated fake news.

Further, we would compare the results of our proposed model to other available models and provide a comparative report based on the parameters like – Accuracy; resource utilization; feature selection; number of features; data sets used and so on. Thus, we propose to take forward and further contribute towards the research to defend against the threat of machine generated fake news.

2.1. Data Selection

To start with the experiments for this research; the first step is identification of the suitable news datasets. A repository for an ongoing data collection project for fake news research at ASU is selected as the basis for fake/real news differentiation. Details as below –

- FakeNewsNet (Fake and Real News dataset)

This is a repository for an ongoing data collection project for fake news research at ASU. Dataset is described and compare FakeNewsNet with other existing datasets in “Fake News Detection on Social Media: A Data Mining Perspective”. JSON version of this dataset is available in GitHub. (Shu et al., 2017) (Shu et al., 2018) (Shu et al., 2019)

- Machine Generated Fake news dataset:

Considering the existing online disinformation; it is worth noticing that adversaries will try to generate targeted content as fake news in order to influence the audience (e.g., some political propaganda or clickbait). Recently introduced large-scale generative models are capable of producing realistic-looking text (Radford et al., 2019), but they fail to produce controllable generations (Hu et al., 2017).3 Keeping this in consideration and to probe the feasibility of realistic-looking neural fake news, authors introduced Grover model, which is capable of producing both realistic and controlled generations.(Zellers et al., 2019).For this research; we have used GROVER model to generate the sample for Neural Fake News. A sample dataset of around 100 + articles is generated which covers various domains like – health, politics, social, climate etc. (Anon, n.d.)

This dataset along with the FakeNewsNet dataset have been considered to perform the experiment to distinguish between human and neural fake news.

2.2. Data Pre-Processing

Most of the times the data available is noisy or is not in the right format to perform operations like predictions / classifications effectively. With text data this problem gets even more prominent and especially when our source is web articles or quotes. Quite a few times the inconsistencies in the results from the NLP applications are due to improper techniques used for text pre-processing and hence it becomes extremely important to choose the correct text pre-processing steps for the problem we need to solve.

Feature Engineering: A core step of a typical NLP problem is featuring engineering which is nothing but converting raw or annotated text into features that give a machine learning model a simple, more focused view of the text. A classification problem in NLP can simply be considered as document or a token level classification task.

- TF-IDF Representation: This is another and quite popular way of feature representation.

TF – Term Frequency – frequency of a term in a given document. IDF - Inverse document frequency. The TF-IDF representation, considers the importance of each word whereas in the bag-of-words model, each word is assumed to be equally important, which may not give proper results.

The TF-IDF weight of a given term in a document can be calculated by below formula:

$$TF_{a,t} = \frac{\text{Frequency of Term'a' in document't'}}{\text{Total Number of Terms in document't'}}$$

$$IDF_a = \log \frac{\text{Total Number of documents}}{\text{Total documents that have term'a'}}$$

Now, the tf-idf score for any term in a document is calculated as the product of above two terms:

$$TF-IDF=TF_{a,t}*IDF_a$$

The terms that occur frequently in a given document are assigned higher weights, also these words are exclusively used in this particular document or in other words occur very rarely in any other document in the given set. Same way, the words which are commonly used across all the documents are assigned lower weightage.

To summarise the research methodology; data (collected and generated from the GROVER model) is processed with the help of pre-processing technique. For this research the recommended pre-processing techniques can be Lowercasing followed by Stemming (porter's stemmer); Stopwords removal; Normalization and Noise removal.

Processed data is then represented in its tf-idf forms (feature engineering). For this study Linear classification models are a good starting point and can provide a good baseline for further experimentation. We propose to train the Naïve Bayes; SVM and Logistic Regression classifiers on the identified features.

Further the results and analysis sections would explain in detail about the important classifier features and also provide a comparative analysis of the performance of the 3 models suggested.
This research will provide a baseline analysis about the differentiators and similarities for machine and human generated fake news by providing a simple classification model to classify fake news.

3. ANALYSIS

Analysis is a very important aspect to decide on the approach of problem solving. It is as important as the solution itself. At times the analysis can provide some important insights which may highlight some new research areas or further scope of research for the given problem. Writing patterns can vary from person to person and so can vary from human to machine. Even we can expect difference in the generated text from different NLG models; similar to how humans from different linguistic background will have some unique writing styles. Hence, understanding and exploring data to get some insights is very important for the problem of human vs machine text classification, before implementing the solutions. We first start with exploring and sampling the data that we would be using for conducting experiments. For this research we have selected the dataset which consists of human generated real and fake news and using GROVER fake news generation model we have generated the machine generated fake news dataset.

In order to understand our data better we explored the writing styles of all three news categories that we have considered for our experiments – human written ‘Real’ and ‘Fake’ news and ‘AI written Fake news’.

- Total Length of the article – While we compared the overall length of all the news articles; it is interesting to see that most of the articles be it human written or machine written fake news articles are below 5000 characters. As we look at the histograms below, we can easily recognize that majority of the human written articles especially the legitimate ones are shorter while almost all the machine written are approximately same length. Real news dataset has shorter articles except a few outliers which are quite long beyond 1000 to 2000 characters. It is quite obvious to note that the average length of articles written by this NLG model (GROVER) has a specific writing pattern and has an article length range set as greater than 2000 and less than 5000. On the contrary, length of Human written articles varies and even though most are in range from 1000 to 3000, we also see a few outliers.

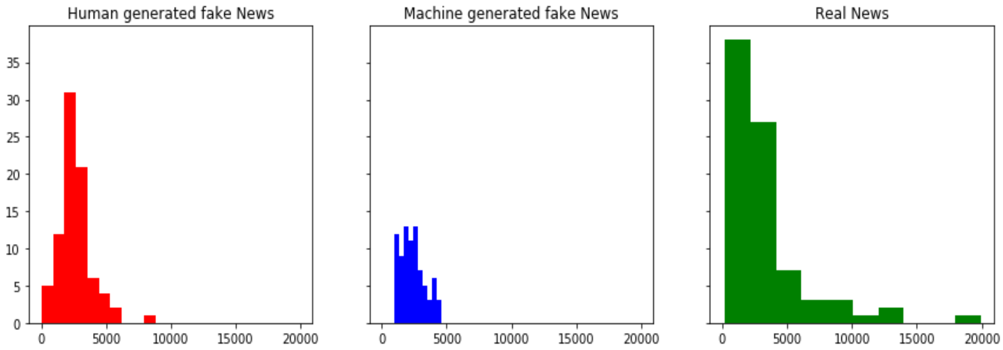

Figure 1. Length of the article pre category.

- Number of sentences per article – To further examine the writing patterns, count of sentences were compared. These counts look almost similar and on an average all the articles have around 10 to 20 sentences. Maximum number of sentences are around 40 with a few outliers for real news dataset where some articles stretch to about 120 sentences.

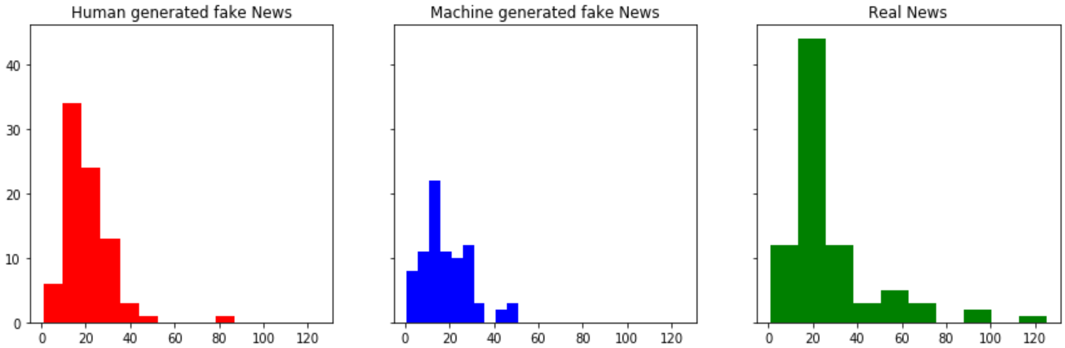

Figure 2. Number of sentences per article.

- Word Count – While doing some basic checks on the dataset; we checked for the distinct word count in the news articles of each category. As you can see in the below histograms; for human written fake news, majority of the articles use around 200 to 300 distinct words and maximum goes to about 600 words for a very low cunt of articles. In case of Human written real news, quite a high number of articles have about 200 distinct words. While most have less than 700 distinct words; one or 2 articles seem to be quite longer with 1000+ distinct words. These are the same longer articles as we can see from the sentence count and length of the articles. Now about the machine written fake news dataset, distinct word counts seem to go from 100 to 380 approximately. We don't see a spike as we do in the human written articles which indicates that we have a variety of articles with different word counts but ranging in a standard range. (no outliers). There seem to be some thresholds set by the model of having minimum 100 distinct words and maximum not more than 400 words.

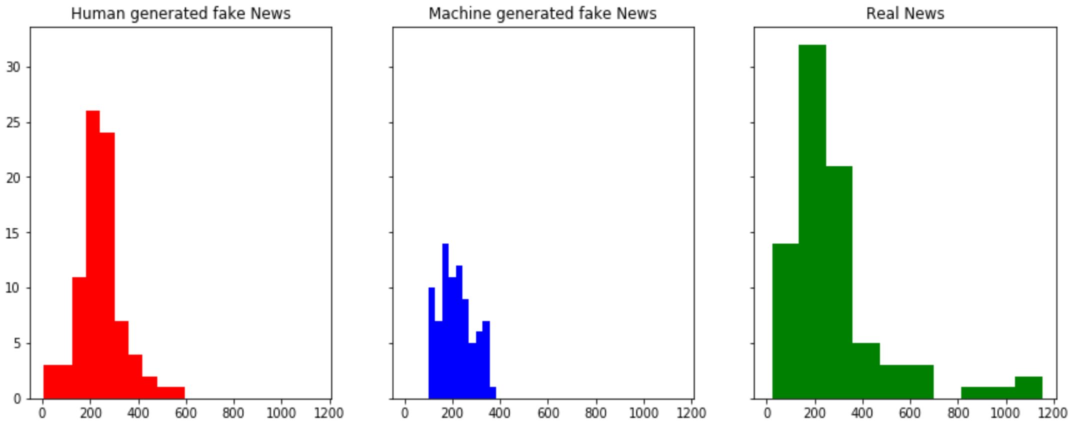

Figure 3. Distinct word count per article.

- Parts of speech tags – POS tags or parts of speech tagging is a fundamental and quite effective syntactic analysis method. A word can be tagged as a noun, verb, adjective, adverb, preposition etc. depending upon its role in the sentence. POS tagging is assigning such tags like noun, pronoun, verb, adjective etc. correctly in a sentence. There are 36 POS tags in NLTK; out of which the 3 most commonly used in our dataset are
- Common Nouns "NN" – these are mostly names of some general objects like 'Box', 'Paper', 'Cat','Animal','Bird' etc. All our articles seem to widely use common nouns. Human written texts (both Fake and Real) seem to use around same frequency for common nouns while in case of machine generated fake news this frequency is quite high.
- Proper Nouns "NNP" – these are specific names of the objects /places/persons like – 'Mumbai','John','Japan','Sumit' etc. Here we can observe that the real news dataset uses maximum number of proper nouns; which justifies the fact that these are true news carrying legitimate information. While Fake news articles use a low number of proper nouns as it can be difficult to create fake identities or locations. Still if we compare the human and machine written fake news; humans seem to do a better job coming up with false names of places/persons.
- Preposition "IN" – these are connectors which link nouns, pronouns and phrases to other words in a sentence, e.g. – 'above','across','in' etc. Prepositions appear at the third rank of top 3 commonly used POS tags. While nothing much is seen to vary, in terms of

count, we can still say that humans make more use of connectors while writing fake news.

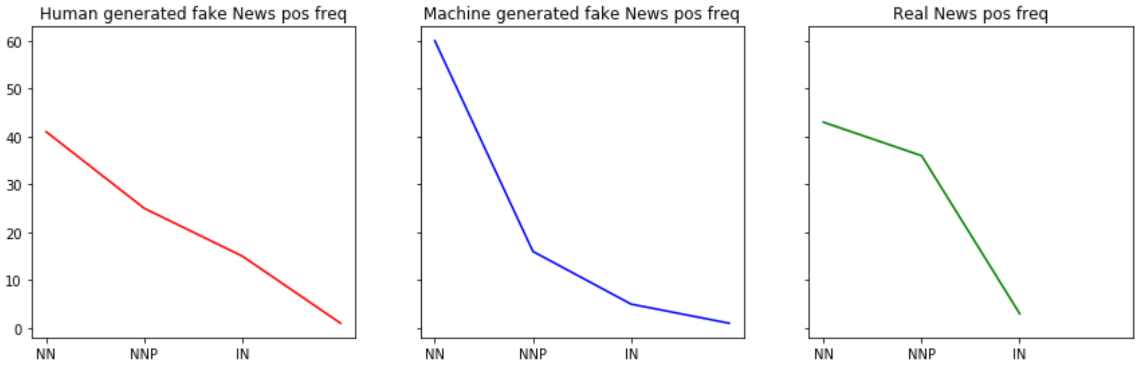

Figure 4. Most commonly used POS tags.

As a quick summary of the Analysis done in this research; 3 categories of datasets are considered – Human written Real news dataset; Human written Fake News dataset and AI /machine written Fake news dataset. For our research and EDA, only the test articles for all 3 categories are considered and all other features like source or domain, images etc have been ignored (which can be considered for the future scope of this research). Data cleansing and EDA performed on these datasets; give some valuable insights about the writing styles of each category of news.

Comparing the lengths shows that the majority of the human and machine generated articles are almost of same length, while human written real news articles have a few very long ones. Similarly, the count of sentences also shows similar trends as of the lengths, which is quite expected. As we further analyse, usage of distinct words per article comes out to be quite low in machine generated fake news – this indicates a limited word corpus used by the generation model. Syntactic analysis performed shows almost similar trend for POS tags for all 3 categories of articles, although the use of common nouns is high for all 3 categories, machine generated text shows a very high count here. The next most popularly used tag is proper noun. In this case the usage of proper noun is significantly higher than the other two fake news categories.

4. Results and Discussion

Once we complete some basic data cleaning and data pre-processing we proceed to feature extraction. While conducting exploratory data analysis; we observed that there is no drastic difference in the writing style of the three categories of news. For our experiments, TF-IDF vectorization method is chosen along with unigram and bigram features. We feed these TF-IDF vectorized data for n-grams to some linear classification models like – Naïve Bayes; Logistic Regression and Linear Support Vector Machine. Sections below will discuss in detail about the results obtained from our different iterations of our experiments.

4.1. Computing Models for classification

For this research we conduct various experiments with different combination of datasets. We computed results from linear classifiers like – Naïve Bayes; Logistic Regression and Support Vector Machine. Linear classifiers are simpler to implement and are also quite efficient in performance as compared to others. Also, in case of a simple text classification problem; linear models have produced suitable results. Thus, considering these parameters; most extensively used linear classifiers are chosen for our problem.

Using these classifiers we performed 2 iterations with TF-IDF feature set on unigrams and bigrams respectively on 4 combination of datasets –

- Human written Fake news and Human written Real news
- Human written Real news and Machine written Fake news.
- Human written Fake news and Machine written Fake news.
- Human written Fake news, Human written Real news and Machine written Fake news.

As a summary of our results, Model evaluations and performance comparison is done for 4 types of data combination. Two iterations with unigram and bigram tf-idf vectorized features each have been considered for all 3 linear classification models, namely – Naïve Bayes; Logistic Regression and Linear Support Vector Machine.

Performance measures like accuracy; precision; recall; f1 score have been examined and accordingly the best performing model for each group is identified.
Below table can be referred as a quick reference of the best performing models per group and their corresponding accuracies –

Table 1. Accuracy Summary

Human Written Fake news & Human written Real news	
SVM	74

Human written fake news & Machine Written Fake News	
Logistic Regression	98

Human written real news & Machine Written Fake News	
SVM	100

Human written real news & Human written fake news & Machine Written Fake News	
SVM	78

Overall, the results clearly show that linear classifiers like SVM and Logistic regression do a reasonable job in classifying the human and machine-written news. Primary aim of the research is to pick out the machine or AI-generated fake news, and the classifiers have effectively picked out the machine fake news. As discussed in this chapter simple linear classifiers if trained on appropriate features show a good performance for the problem at hand. Where humans seem to often fail in capturing machine-generated fake news; the classifiers considered in this research have shown very good sensitivity towards the machine-generated fake news.

Data annotation is the biggest task before model training it requires SME support. Data annotation is a subjective task, and different annotators may have different interpretations or labelling conventions. This can lead to inconsistency in the labelling and affect the overall quality of the training data. Data annotation can be a time-consuming and expensive process, especially when the data is complex or requires specialized domain knowledge. Data annotation can also introduce bias into the training data, especially when the annotators have preconceived notions or prejudices that influence their labelling decisions. As datasets become larger and more complex, it can become increasingly difficult to annotate data accurately and efficiently, leading to decreased data quality and model performance. Ensuring the accuracy and consistency of the annotated data can be challenging, and quality control measures are required to minimize errors and inconsistencies. But to overcome these challenges we are utilizing active learning, which is iterative POS tagging process. Active learning can significantly reduce the amount of labelled data required for model training, which can lead to significant cost and time savings. By selecting the most informative samples for labelling, active learning can help to improve the accuracy of machine learning models, especially when data is limited.

By selecting the most informative samples for labelling, active learning can reduce the overall annotation cost, as annotators only need to label the most relevant samples. Active learning can also involve domain experts in the labelling process, which can lead to more accurate labelling and better model performance.

In active learning for POS tagging, the model starts with a small set of labelled data, and then iteratively selects the most uncertain or informative examples for annotation by a human expert. The expert annotates these examples, and the model is retrained on the updated labelled data. This process continues until the model achieves a desired level of accuracy or the cost of annotation becomes too high.

One common approach for selecting informative examples in active learning for POS tagging is to use uncertainty sampling. Uncertainty sampling selects the examples for which the model is least certain about the correct label. For example, if the model is uncertain whether a word should be tagged as a noun or a verb, it may select that example for annotation.

5. Conclusions and Recommendations

5.1. Conclusion and Discussion

This research was broadly divided in 3 parts; Identifying and computing data for experiments; Exploring and getting insights from the writing style of human and machine generated, real and fake articles and finally to build and evaluate classification models and understand their results.
For this research, we used a small sample of human written real and fake news articles form ‘Buzz feed Dataset’. In order to get the sample of AI generated news articles; we considered GROVER (AI model for fake news generation) as source and generated a sample of fake news articles. These articles had a variety of domains like political, crime, media, healthcare etc. Exploratory data analysis showed quite useful insights such as –

- Length of majority of the human and machine generated articles is almost same, while human written real news articles have a few very long ones.
- Usage of distinct words per article is quite low in machine generated fake news – this indicates a limited word corpus used by the generation model.
- Syntactic analysis shows almost similar trend for POS tags for all 3 categories of articles, although the use of common nouns is high for all 3 categories, machine generated text shows a very high count here.
- Lesser use of proper nouns in case both human and machine generated articles shows the lack of facts or evidence in the news and that these articles keep away from providing specifics of the news or message they convey.

While we have the above quite useful insights from the writing styles of all the three categories of articles; these features are still not enough to completely define a particular news category. Hence for the classification purpose tf-idf vectorized features on n-grams tokens were considered.

Thus, it is good to conclude that simple linear algorithms can work quite well to pick out the machine or AI generated fake news articles and separate them out form other mix of human written fake/real articles.

5.2. Contribution Towards Knowledge

There has been quite some work in the area of fake news identification but the problem of identifying machine generated neural fake news is still unsolved. Various models have been tested for human written fake and real news identification. As fake news is designed to deceive human targets; and use of AI for generating such news makes the problem quite complex to solve. Humans mostly classify the machine generated fake news as legitimate and hence there arises a need to address this problem using ML- techniques itself. There has been considerable work on Natural language generation and machine generated text verification (Hashimoto et al., Apr 2019) for the quality and diversity of the text; however, these verifiers have not shown desired results when it comes to classifying Human Vs Machine generated text.

This research contributes further to the findings and approach for solving the problem of AI generated fake news identification. It should be noted that even though the writing style of Human and machine generated news are very similar; it is worth noticing the articles generated by an AI model has some peculiarities in terms for some standard range of the distinct word count, length of the article etc. Also, as we further explored the writing styles on Parts of Speech tags; high usage of Common nouns in machine written and very low use of proper nouns in case of fake news should be noticed and can be explored further for other NLG models which are used for writing fake news.

Most of the complex language generation models like BERT, GROVER have shown high accuracy in identifying the fake news articles generated by the model itself but fail to achieve the same accuracy for articles generated from other models. The results of this research show that linear models like SVM can perform very well in picking out the machine generated fake news articles and separate them out from Human written real news. Although for our experiment we have considered only the GROVER model generated news articles, we can extend this research and train the model with datasets of different models and a variety of domains. Thus, we can state that linear models, if trained on appropriate features and datasets can be used for the purpose of automatic fake news detection and recognizing the source as – human or machine.

5.3. Future Scope and Recommendations

The results of this research have shown a good performance of our classifier, however there is scope for improvement. Our models have considered TF-IDF of bigrams to distinguish between the human written fake and machine written fake from human written real news. Even though TF-IDF features seem to perform better, there is a possibility of overfitting to specific topics/terms for the sample dataset. It should also be taken into consideration that for machine written dataset we have considered articles from a single AI model, whereas there are quite few models available which can effectively generate fake news which is good enough to deceive human eye. Hence its recommended that we should collect data from various such models on diverse topics and train our model on such dataset. Also, with a vectorized approach it becomes quite difficult to understand which individual features are most important, thus limiting our analysis and prevent broader generalizability.

For further research, writing styles of such different AI fake news generators should be studied and explored to see if we get comparable results to what we found as a part of this research. Other vectorization techniques, or altogether different feature set should also be explored for building classifiers. This research was more focussed on the Language and text classification methods, however various other features like images or videos in the articles, date / time, source of the article like a specific publication, website or domains might as well provide valuable information for the classifiers.

Fake news or misinformation can be extremely dangerous; hence automated techniques to not just identify but also to trace the source and stop such news from spreading becomes extremely critical. Thus, as a future scope; tracing the origin of a machine generated fake news to the language model that it has originated from (e.g. BERT or GPT or GROVER etc.) will prove to be an important step towards the goal of tracing and blocking such misinformation from spreading.

REFERENCES

[1] Anjali, M. & Jivani, G. (n.d.). A Comparative Study of Stemming Algorithms. Available from: www.ijcta.com.

[2] Anon (n.d.). GPT-2: 1.5B Release. Available from: https://openai.com/blog/gpt-2-1-5b-release/.

[3] Anon (n.d.). Grover - A State-of-the-Art Defense against Neural Fake News. Available from: https://grover.allenai.org/.

[4] Anon (n.d.). Linear classifier - Wikipedia. Available from: https://en.wikipedia.org/wiki/Linear_classifier.

[5] Anon (n.d.). Understanding Support Vector Machines(SVM) algorithm (along with code). Available from: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[6] Beresneva, D. (2016). Computer-generated text detection using machine learning: A systematic review. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 9612. p.pp. 421–426.

[7] Christian, J., Cruz, B., Tan, J.A. & Cheng, C. (2018). Localization of Fake News Detection via Multitask Transfer Learning.

[8] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Mlm). Available from: http://arxiv.org/abs/1810.04805.

[9] Gilda, S. (2018). Evaluating machine learning algorithms for fake news detection. IEEE Student Conference on Research and Development: Inspiring Technology for Humanity, SCOReD 2017 - Proceedings. 2018-Janua. p.pp. 110–115.

[10] Goldani, M.H., Momtazi, S. & Safabakhsh, R. (n.d.). Detecting Fake News with Capsule Neural Networks.

[11] Hashimoto, T., Zhang, H. & Liang, P. (2019). Unifying Human and Statistical Evaluation for Natural Language Generation. 2. p.pp. 1689–1701.

[12] Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. (2017). Bag of tricks for efficient text classification. 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference. 2. p.pp. 427–431.

[13] Ni, B. (2020). Score Matching.

[14] Pérez-Rosas, V., Kleinberg, B., Lefevre, A. & Mihalcea, R. (2017). Automatic Detection of Fake News. Available from: http://arxiv.org/abs/1708.07104.

[15] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2018). Language Models are Unsupervised Multitask Learners.

[16] Rashkin, H., Choi, E., Jang, J.Y., Volkova, S. & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings. p.pp. 2931–2937.

[17] Schuster, T., Schuster, R., Shah, D.J. & Barzilay, R. (2019). Are We Safe Yet? The Limitations of Distributional Features for Fake News Detection. p.pp. 1–9. Available from: http://arxiv.org/abs/1908.09805.

[18] Shu, K., Mahudeswaran, D., Wang, S., Lee, D. & Liu, H. (2018). FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. (September). Available from: http://arxiv.org/abs/1809.01286.

[19] Shu, K., Sliva, A., Wang, S., Tang, J. & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. (i). Available from: http://arxiv.org/abs/1708.01967.

[20] Shu, K., Wang, S. & Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining. (December). p.pp. 312–320.

[21] Thorne, J., Vlachos, A., Christodoulopoulos, C. & Mittal, A. (2018). FEVER: a Large-scale Dataset for Fact Extraction and VERification. p.pp. 809–819.

[22] Traylor, T., Straub, J., Gurmeet & Snell, N. (2019). Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator. Proceedings - 13th IEEE International Conference on Semantic Computing, ICSC 2019. p.pp. 445–449.

[23] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F. & Choi, Y. (2019). Defending Against Neural Fake News. p.pp. 1–21. Available from: http://arxiv.org/abs/1905.12616.

[24] Zhou, X. & Zafarani, R. (2018). Fake News: A Survey of Research, Detection Methods, and Opportunities. Available from: http://arxiv.org/abs/1812.00315.

[25] arXiv:2011.00767v2 **[cs.CL]**

[26] https://doi.org/10.1613/jair.295

[27] https://aclanthology.org/W07-1516.pdf

AUTHORS

Ms. Poorva Sawant, 16 years of experience in Data and AI. Developed and operationalised AI models for various clients to provide solutions for business problems with her expertise in Analytics and ML/AI

Mr. Parag Rane, 16 years of experience in ML/AI. Developed, Deployed, and operationalised innovative AI solutions for various clients with his expertise in Machine Learning and Artificial Intelligence.