Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.754728

DEEP FAKE DETECTION ON SOCIAL MEDIA LEVERAGING DEEP LEARNING AND FAST TEXT EMBEDDINGS FOR IDENTIFYING MACHINE-GENERATED TWEETS

V.V.N.V.Gurusai
KL University, 2201170003ece@gmail.com

M.Ravi Kumar
Associate Professor, KL University

ABSTRACT

Social media opinion manipulators now have more tools at their disposal because to recent developments in natural language processing. access to an additional tool. Furthermore, due to advancements in language modelling, deep neural models now possess greater generative skills, improving their capacity to produce content. Because of in recent years, text-generative models have grown in efficacy, which allows attackers to take use of these incredible skills to fortify social bots and create convincing deep fake posts that sway public opinion. Addressing this issue requires trustworthy and precise methods for detecting deep fake social media postings must be developed. Because of this, research on recognizing computer-generated content on social media networking sites like Twitter is still underway. This work classifies uses Twitter when either human- or bot-generated using word embedding's and a rudimentary deep learning model using the publicly available Twee fake dataset. Using Fast Text word embedding's, using a standard design for a CNN is created to identify deep fake tweets. Many machine learning models were used as reference approaches in this study to show the improved efficacy of the recommended methodology. These baseline techniques incorporated Fast Text, Fast Text sub word embedding's, Term Frequency, and Term Frequency-Inverse Document Frequency. Furthermore, the advantages and effectiveness of the proposed approach are emphasized in comparison with other deep learning models, namely CNN-LSTM Together with LSTM systems, in successfully completing what needs doing. The experiment the results show that the convolutional neural network is suitable for accurately identifying twitter data with a 93% accuracy rate when combined with Fast Text embedding's.

INDEX TERMS; Text categorization, deep fake, machine learning, machine-generated text, and machine learning.

1. INTRODUCTION

Users are able to more easily communicate and share ideas through the use of text, photos, audio, and video on social media sites [1]. Bots are programs that automate the process of publishing, like, and distributing content on social media platforms. These programs utilize methods such as deep fake, video manipulation, search-and-replace, and gap-filling text [2]. Feature representation can be learned by means of a kind of machine learning known as deep learning applied to input data. A hybrid of "deep learning" and "fake," "deep fake" refers to content that is created using artificial intelligence (AI). material that could be deceptive [3]. The production and dissemination of deep fake multimedia on social media have already

caused issues in other domains, including politics, as it has the potential to deceive people into thinking it was made by humans [4]. One possible application of social media is to spread disinformation more easily in an effort to influence people's views and ideas, particularly to sow distrust in democracies [5]. A number of accounts, including cyborg accounts and sock puppets, are utilized for this purpose [6]. However, social bots, which are entirely programmed profiles on social media, mimic human behaviour [7]. The current advancements in natural language generative models, including Grover [9] and the GPT [8], together with the growing usage of bots, have provided the enemy with a way to disseminate misinformation better. An outstanding example of this is the 2017 Net Neutrality case, when the Commission's decision to repeal was greatly influenced by millions of duplicate comments [10]. Addressing the concern that basic text manipulation techniques can lead to the formation of erroneous notions is crucial, as is considering the potential effects of more robust transformer-based models.

2. EXISTING SYSTEM

Deep fake technology, made possible by developments in computer vision, allowed for the effective synthesis of text and manipulation of audio. Computer vision deep fakes frequently employ face manipulation methods such as body re-enactment, identity swapping, mood switching, and whole-face synthesis. A new technique called audio deep fakes can take a text database and combine the voices of several speakers to create five seconds of spoken audio. The language models might be updated thanks to the 2017 enhancements to the transformer and the self-attention mechanism. Modeling language makes employed a variety of statistical and probabilistic methods to ascertain the likelihood of a specific word sequence occurring inside a speech.

3. PROPOSED SYSTEM

A tagged dataset is retrieved from a publicly available repository and used within the Framework. Both human and automated accounts' tweets are included in the gathered dataset. A number of the batch processing techniques are employed to clean the tweets, improving the wording and making it more comprehensible. With an 80:20 split, the dataset is split between the testing and training sets. By utilizing FastText word embedding, the text is then converted into vectors. Consequently, these three-dimensional models are inputted into the CNN model. During training, the suggested approach is utilized, which integrates a 3-layered CNN with FastText word embedding. Four evaluation metrics—F1-score, Accuracy, Precision, and Recall—are used to determine the efficacy of this approach.When dealing with terms that aren't in the vocabulary, the predetermined vocabulary size of transfer learning models could be problematic. These caveats notwithstanding, the CNN model employed for this investigation remains unaffected.

4. PROPOSED METHODOLOGY

The suggested methods for tweet categorization are detailed in this section. Figure 4 depicts the proposed framework's architecture. Algorithms that use deep learning, such as CNN, may automatically glean useful information from text. Their job is to help the model find hierarchical patterns, local interactions, and long-term connections in the input text so it may extract meaningful representations. It is possible to capture text dependencies by stacking many

CNN layers. In order to classify tweets, this research introduces a CNN model that is based on deep learning. The Framework makes use of a tagged dataset that is retrieved from a publicly accessible repository. Both human and automated accounts' tweets are included in the gathered dataset. To make the text easier to understand and improve its quality, the tweets are cleaned using a number of preprocessing techniques. There are 80:20 splits in the dataset between the training and testing sets. The last step is to convert the text into vectors using Fast Text word embedding. They are subsequently input into the convolutional neural network (CNN) model. For training, we adopt the suggested technique, which integrates Fast Text word embedding with a three-layered CNN. Four evaluation metrics—F1-score, Accuracy, Precision, and Recall—are used to determine the efficacy of this method.

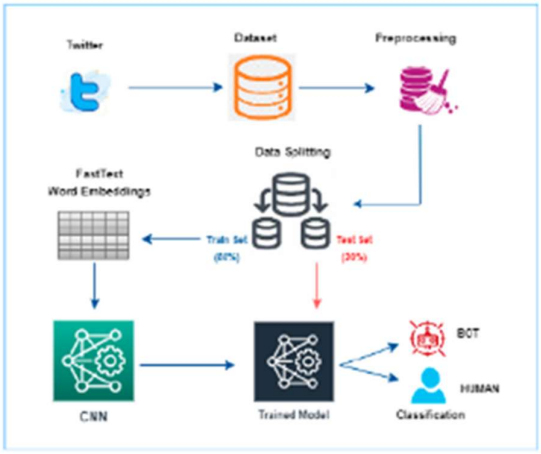

FIGURE 1: Architecture of proposed framework for deep-fake tweet classification

5. PROJECT RELATED WORK

Deepfake approaches first appeared in computer vision [4], and they quickly found success in audio editing and text synthesis [5]. In computer vision, deep fakes frequently include face manipulation methods such as body reenactment, identity swapping, mood switching, and whole-face synthesis. One recent use of audio deep fakes is the generation of spoken audio from a text corpus including the voices of several speakers after only five seconds of listening [7].A refresh of the language models was made feasible by 2017's enhancements to the transformer and the self-attention mechanism. Using various statistical and probabilistic methodologies, language models determine the probability that a given word sequence will appear in a given phrase. Subsequent transformer-based language models, such as GPT and GPT2 [8], enhanced language generation and natural language interpretation. Created in 2019 [11] by Radford et al., the pre-trained language model GPT-2 can autonomously produce coherent, human-like paragraphs of text given a single, brief sentence as input. In the same year that GROVER was invented, authors[9] came up with a new way to analyze and write multi-field documents like journal articles efficiently and effectively. A conditional language paradigm called CTRL was released shortly thereafter [16]. It uses control codes to produce text that has distinct style, content, and behavior for each job. In addition, the text creation process was enhanced by the introduction of OPTIMUS, which contained a variation auto encoder, by researchers [12].

6. DEEPFAKE TEXT GENERATION METHODS

Deep fake text may be generated using a variety of methods. Some popular generative methods for computer-generated text are summarized below.A Markov chain is a kind of stochastic model that represents a series of states; it iteratively transitions between them with a probability that is completely dependent on the state that is currently being considered. During the text creation process, state tokens are utilized, and the subsequent token or state is selected at random from a set of tokens that follow the one now in use. Token t's selection probability is proportional to the token's follow-up frequency. The RNN keeps track of token data in its accumulated memory and builds the multinomial distribution for selecting the next token using its loop structure. In order for the RNN to produce the next token, the selected token is sent back as input.One possible sampling approach that the RNN+Markov method might use is the following token selection in the Markov Chain. Using the RNN's produced multi-nominal distribution, the next token is really selected at random from among the highest-probability tokens. But we couldn't find any citations that back up our RNN+Markov process idea.

7. MATERIAL AND METHODS

This section covers the machine learning models, deep learning models, feature engineering techniques, and dataset used in the experiments. The pre-scented experimental method is shown in Figure 1.A.

DATASET

This study makes use of the TweepFake [19] dataset, which has 25572 tweets in total. In the dataset, there are 17 human accounts and 23 bot accounts. Each person and body count has its own label. The latter reveals the author of the text.

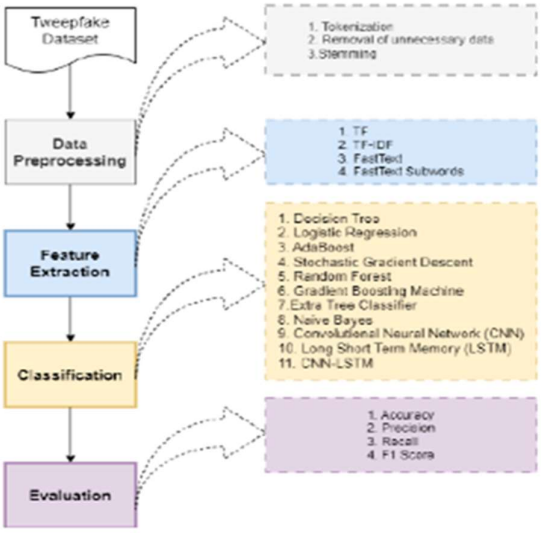

FIGURE 2: Architecture of methodologies adopted for deep-fake tweet classification

There are four possible approaches that may have been used: human (17 accounts, 12786 tweets), GPT-2 (11 accounts, 3861 tweets), RNN (7 accounts, 4181 tweets), or Other (5 accounts, 4876 tweets).Figure 2 displays the count-plot that illustrates the data distribution by account type, and Figure 3 displays the count-plot that illustrates the data distribution by class.

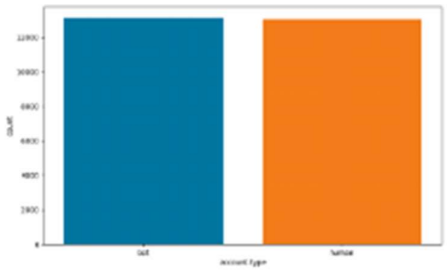

FIGURE 3: Count plot showing account-type data distribution

Data pre-processing

Semi-structured and unstructured useless data are included in datasets. With such unnecessary data, the model's performance might deteriorate and its training time could lengthen. Pre-processing is necessary to maintain computing power and maximize the performance of machine learning models. Preparing the text improves the model's accuracy in forecasting outcomes. Pre-processing includes the following steps: tokenization, case conversion, stopword removal, and number removal.

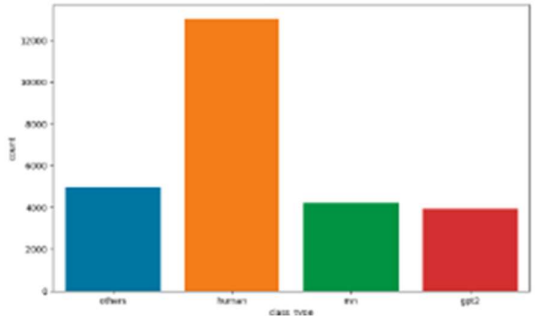

FIGURE 4: Count plot showing class-wise data distribution

8. ALGORITHMS

8.1 Decision tree classifiers

Decision tree classifiers have a wide range of applications. Their unique selling point is the descriptive decision-making knowledge they can extract from the provided data. Using training sets, decision trees may be built. Using an object set (S) from a class C1, C2,..., Ck, the following procedures may be followed to generate this generation:First, there's a leaf in the decision tree for S that is labeled with the class if all the objects in S are of the same class, like Ci.Step 2. Otherwise, let T be some test with possible outcomes O1, O2,…, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2,… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

The K-Nearest Neighbors (KNN) method is an easy-to-understand supervised learning approach that works well for classification and regression. Although simple and easy to use, its main drawback is that it becomes noticeably slower with increasing data consumption. Due to its exceptionally precise prediction capabilities, the KNN algorithm is able to hold its own against the most precise models. Therefore, applications that demand great accuracy without a human-readable model can utilize the KNN technique. The observed distance determines the accuracy of the projections. Comparable pieces of information tend to cluster together. For

KNN to be effective, the underlying assumption must be correct. Similarity, often called closeness, distance, or just similarity, to certain mathematical concepts is captured by KNN.

8.2 Logistic regression Classifiers

To examine the relationship between a group of classification reliant variables and a collection a set of factors that serve as independent variables, logistic regression analysis is employed. When there are just two possible values for the dependent variable, such as yes or no, logistic regression is employed. When there are three or more possible values for the dependent variable (married, single, divorced, or widowed), the phrase logistic regression with multiple outcomes is often reserved for that circumstance. While the dependent variable's data format differs from multiple regression, the method's practical application is comparable. In the realm of categorical response variable analysis, logistic regression and discriminant analysis are rivals. Logistic regression, according to many statisticians, is more versatile and applicable than discriminant analysis when it comes to modeling. The reason behind this is that, unlike discriminant analysis, regression using logistic presume that the independent variables have a consistent distribution. For both categorical and numerical independent variables, this software computes binary and multinomial logistic regression. Not only is the regression equation given, but so are the odds ratios, confidence intervals, probability, and deviance. For diagnostic residuals, we undertake a full study that includes charts and reports. To locate the optimal regression model using the minimum number of independent variables, that is capable of do an independent variable subset selection search. As a tool for finding the best classification threshold, it offers ROC curves and confidence intervals on expected values. You may verify your findings by automatically categorizing rows that aren't utilized in the research.

8.3 Naïve Bayes

An fundamental principle of Within a given class, the presence or absence of one feature does not always indicate the presence or absence of any other feature, according to the naive bayes approach, a supervised learning method.Regardless, it seems effective and forceful. Like other guided learning methods, it gets the job done. There are a number of arguments put out in the literature. An explanation predicated on representation bias is the main focus of this lecture. Many popular classifiers, including the naive bayes, linear support vector machine, logistic regression, and linear discriminant analysis are linear procedures. Inconsistencies arise from the learning bias, which is the process used to estimate to which the classifier is configured. The Naive Bayes algorithm identifies popular in research, although it isn't often used by practitioners who are seeking practical results. From one angle, the researchers discovered that it is simple to implement and use, that determining its parameters is an easy process, that learning occurs quickly even on extremely huge datasets, and that, when compared to other systems, it achieves very excellent accuracy. Users don't benefit from this method since they don't get a model that's straightforward to grasp and utilize. So, we provide the learning process outcomes in an innovative manner. Both the classifier's understanding and its execution are simplified. The first part of this lecture covers the theoretical basis of the naive bayes classifier. After that, we put the method to the test using a Tanagra dataset. We evaluate the model's parameters in comparison to other linear methods' outcomes, such as regression using logistic data, linear discriminant analysis, and linear support vector machines (SVMs). We discover

that the outcomes are rather uniform. When compared to other ways, this helps to clarify why the strategy is so effective. In Part 2, we use Orange 2.0b, Weka 3.6.0, R 2.9.2, Knime 2.1.1, and RapidMiner 4.6.0 on the same dataset. We are more interested in comprehending the outcomes.

8.4 Random Forest

Random forests, sometimes called construct a vast array of decision trees using a random algorithm for problems like classification and regression during the ensemble learning training phase. When it comes to classification jobs, the majority of Trees select the output from the random forest. In regression tasks, the mean or average prediction for each tree is returned. A decision tree's propensity to overfit its training set can be mitigated by using random decision forests. Despite being less accurate than gradient enhanced trees, random forests often perform better than choice trees. But how effectively they work can depend on how distinctive the data is.

Utilizing Eugene Kleinberg's "stochastic discrimination" method for classification in conjunction with the random subspace methodology, Tin Kam Ho[1] developed the initial random decision forest approach in 1995. Leo Breiman and Adele Cutler, who developed the method further, filed a trademark application for "Random Forests" in 2006. As of the year 2019, the trademark is owned by Minitab, Inc.A collection of decision trees with controlled variance are produced by the extension, which merges random feature selection with Breiman's "bagging" approach. Ho[1] and Amit and Geman[13] separately presented the idea. Businesses often employ random forests as "blackbox" models due to their excellent prediction ability over a wide variety of inputs and inexpensive setting requirements.

8.5 SVM

A decision-making tool that can reliably forecast labels for freshly acquired instances.

are identified using an independent and identically distributed (iid) training dataset. This allows a discriminant machine learning approach to address classification challenges. In contrast to generative machine learning methods that need the construction of conditional probability distributions, discriminant classification functions assign a given data point (x) to one of the many classes involved in the classification activity. For a multidimensional feature space and when just posterior probabilities are required, discriminant algorithms consume fewer computer resources and training data compared to generative processes, which are typically employed for outlier detection in predictions. Mathematically speaking, learning a classifier is like trying to discover the equation for a multidimensional surface that divides the feature space into classes to perfection.

9. IMPLIMATION

Service Provider

A functional username and password provided by the Service Provider is required to access this module. After he's signed in, he may look at datasets and do tests and training. have a look at the anticipated tweet format, get the forecasted datasets, see the trained and tested accuracy in a bar graph, see all the remote users, and see the training and testing results for accuracy.

View and Authorize Users

Here the administrator may see a complete roster of all registered users. In this, the administrator may see user information like name, email address, and address, and permit others to access this data.

Remote User

There are n users for this module. Prior to beginning, the user must complete the registration process. The information a user provides upon registration is saved in the database. After he successfully registers, he will be prompted to log in using his authorized username and password. View Your Profile, Predict Your Tweet Type, and Register and Login are just a few of the options that users may access after successfully login in.

10. RESULTS AND DISCUSSION

Here we go over the results of the experiments that were conducted as part of this study. In order to identify deep fake tweets, this study use deep learning and ML approaches. This study uses eight machine learning models to validate the proposed approach: DT, LR, AC, SGC, RF, GBM, ETC, and NB. These models are described in this way. The hyper parameters that produce the best results on the given dataset are used to apply these models. Optimal hyper parameter selection involves fine-tuning value ranges to provide desired results. In Table 3, you can see the values of the hyper parameters and the tuning range.

8. Conclusion

In this age of disinformation and bogus content, deep fake text detection has become an important and difficult topic. A deep fake text detection algorithm was proposed as a solution to this challenge, and this study aimed to evaluate its efficacy. An examination of a collection of tweets from both people and bots is conducted using feature engineering approaches in conjunction with several machine learning and deep learning strategies. Fast Text and Fast Text subwords are word embedding methods, while TF-IDF and Tf are well-known feature extraction methods. With an accuracy score of 0.93, the suggested technique showed promise in properly identifying deepfake text, thanks to its combination of CNN and Fast Text algorithms. In addition, the suggested method's outcomes are contrasted with those of alternative cutting-edge transfer learning models that have been the subject of prior research. In light of the advantages of a CNN model in terms of computing performance, usability, and handling non-dictionary items, this inquiry makes use of a CNN model structure. The suggested method is a good choice for text identification jobs because of these benefits.

REFERENCES

[1] J. P. Verma and S. Agrawal, ‘‘Big data analytics: Challenges and applications for text, audio, video, and social media data,’’ Int. J. Soft Comput., Artif. Intell. Appl., vol. 5, no. 1, pp. 41–51, Feb. 2016.

[2] H. Siddiqui, E. Healy, and A. Olmsted, ‘‘Bot or not,’’ in Proc. 12th Int. Conf. Internet Technol. Secured Trans. (ICITST), Dec. 2017, pp. 462–463.

[3] M. Westerlund, ‘‘The emergence of deepfake technology: A review,’’ Technol. Innov. Manage. Rev., vol. 9, no. 11, pp. 39–52, Jan. 2019.

[4] J. Ternovski, J. Kalla, and P. M. Aronow, ‘‘Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments,’’ Ph.D. dissertation, Dept. Political Sci., Yale Univ., 2021.
[5] S. Vosoughi, D. Roy, and S. Aral, ‘‘The spread of true and false news online,’’ Science, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.
[6] S. Bradshaw, H. Bailey, and P. N. Howard, ‘‘Industrialized disinformation: 2020 global inventory of organized social media manipulation, Comput. Propaganda Project Oxford Internet Inst., Univ. Oxford, Oxford, U.K., Tech. Rep., 2021.
[7] C. Grimme, M. Preuss, L. Adam, and H. Trautmann, ‘‘Social bots: Humanlike by means of human control?’’ Big Data, vol. 5, no. 4, pp. 279–293, Dec. 2017.
[8] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, ‘‘GPT understands, 2021, arXiv:2103.10385.
[9] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, ‘‘Defending against neural fake news,’’ in Proc. 33rd Int. Conf. Neural Inf. Process. Syst. (NIPS), Dec. 2019, pp. 9054–9065, Art. no. 812.
[10] L. Beckman, ‘‘The inconsistent application of internet regulations and suggestions for the future,’’ Nova Law Rev., vol. 46, no. 2, p. 277, 2021, Art. no. 2.
[11] J.-S. Lee and J. Hsiang, ‘‘Patent claim generation by fine-tuning OpenAI GPT-2,’’ World Pat. Inf., vol. 62, Sep. 2020, Art. no. 101983.
[12] R. Dale, ‘‘GPT-3: What’s it good for?’’ Natural Lang. Eng., vol. 27, no. 1, pp. 113–118, 2021.
[13] W. D. Heaven, ‘‘A GPT-3 bot posted comments on Reddit for a week and no one noticed,’’ MIT Technol. Rev., Cambridge, MA, USA, Tech. Rep., Nov. 2020, p. 2020, vol. 24. [Online]. Available: www.technologyreview.com
[14] S. Gehrmann, H. Strobelt, and A. M. Rush, ‘‘GLTR: Statistical detection and visualization of generated text,’’ 2019, arXiv:1906.04043. [15] D. I. Adelani, H. Mai, F. Fang, H. H. Nguyen, J. Yamagishi, and I. Echizen, ‘‘Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection,’’ in Proc. 34th Int. Conf. Adv. Inf. Netw. Appl. (AINA). Cham, Switzerland: Springer, 2020, pp. 1341–1354.
[16] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, ‘‘Grover—A state-of-the-art defense against neural fake news,’’ in Proc. Adv. Neural Inf. Process. Syst., vol. 32. Curran Associates, 2019. [Online]. Available: http://papers.nips.cc/paper/9106-defending-againstneural- fake-news.pdf