engineering proceedings

MDPI

Proceeding Paper

Deepfake Tweets Detection Using Deep Learning Algorithms †

Hina Kirn [1], Muhammad Anwar [2,*], Ashina Sadiq [1], Hafiz M. Zeeshan [1], Imran Mehmood [1] and Rizwan Aslam Butt [3]

[1] Department of Computer Science and IT, Lahore Leads University, Lahore 54000, Pakistan; hina.kirn@leads.edu.pk (H.K.); ashina.cs@leads.edu.pk (A.S.); hafizzeeshan008@gmail.com (H.M.Z.); imranmehmood47@gmail.com (I.M.)

[2] Department of Information Sciences, Division of Science and Technology, University of Education, Lahore 54000, Pakistan

[3] Department of Electronics Engineering, NED University of Engineering and Technology, Karachi 75270, Pakistan; rizwan.aslam@neduet.edu.pk

* Correspondence: anwar.muhammad@ue.edu.pk

† Presented at the 7th International Electrical Engineering Conference, Karachi, Pakistan, 25–26 March 2022.

Abstract: The simplicity of contact and the significant improvement in records that are easily accessible through the use of web-based broadcasting methods have made it complicated to distinguish between bogus and genuine information. The unchecked distribution of documents by allocation has resulted in the significant growth of misrepresentation. Wherever the dissemination of deceptive material is frequent, the validity of internet broadcasting websites is also being questioned. As a result, it has become an exploratory task to naturally check the data in terms of its source, substance, and supplier to sort it as false or true. Despite some limits, artificial intelligence has assumed many common record groupings. This article examines a variety of deep learning approaches for cutting-edge forms of deception and dissemination. The constraint, techniques, and impromptu inventions that may be achieved by deep learning are also studied.

Keywords: deep learning; fake news; natural language processing; social media

Citation: Kirn, H.; Anwar, M.; Sadiq, A.; Zeeshan, H.M.; Mehmood, I.; Butt, R.A. Deepfake Tweets Detection Using Deep Learning Algorithms. *Eng. Proc.* **2022**, *20*, 2. https://doi.org/10.3390/engproc2022020002

Academic Editor: Saad Ahmed Qazi

Published: 27 July 2022

1. Introduction

During the last several decades, social networking sites-designed to bring people together and encourage them to express their viewpoints over media content such as photographs, videos, sounds, and articles have been most frequently used to influence or modify community views using robots, or computer applications that run a fictional social media account in the same way that an appropriate social user would by "liking", "sharing", and publishing new and existing content. This is commonly used to attack an individual's or object's reputation, as well as to profit from publicity revenues. Furthermore, the word has gradually come to refer to any incorrect information, such as accidental and reflexive methods, and a way for celebrities refer to a report antagonistic to their viewpoints [1].

Deep learning, also known as deep structured learning, is a machine learning architecture that builds structures using artificial neural models. Unsupervised, semi-supervised, and supervised learning are the three types of learning. Deep-learning structural designs, such as deep neural networks, recurrent neural networks, deep reinforcement learning, recurrent neural networks, and convolutional neural networks, are being used in a variety of fields, including computer vision, speech recognition, NLP, machine translation, bioinformatics, drug design, homeopathic image analysis, factual inspection, and board game software, to produce results that are comparable to, and in some cases superior to, human presentation. Deep learning approaches have significant potential because they follow through on their promises. That is not to say there is not excitement about the technology; instead, the buzz is based on genuine breakthroughs being exposed across

a range of artificial intelligence disciplines, from machine learning to natural language processing [2].

2. Literature Review

Kumar et al. [3] conducted a thorough investigation of several facets of fake news. This study looks at several types of fake news, current algorithms for detecting fake news, and future possibilities. Shin et al. [4] studied key ideas from multiple fields to improve the multidisciplinary study of fake news in one of their studies. The authors of this study looked at the subject of fake news from four different angles. Allcott et al. [5] worked on a statistical study to analyse the effect of fake broadcast on public mass media during 2016 United States President General Election as well as the influence on American voters. The report looks at the genuine and non-authentic URLs associated with fake stories in the BuzzFeed dataset. Ahmed et al. [6] worked on the automatic identification of fraudulent material utilizing web review spam in another study. For categorizing bogus news, the authors looked into two alternative feature extraction approaches. Vosoughi et al. [7] have identified prominent traits of rumours. The authors tested their approach scheduled 209 stories comprising 938,806 twitters starting actual incidents such as in 2013 Boston marathon bombings and in 2014 Ferguson riots, as well as the 2014 Ebola outbreak [8].

Chen et al. [9] suggested a semi-supervised learning model that takes into account machine learning approach and auto-encoders to differentiate allegations as deviations from other trustworthy blog posts [10]. Their suggested approaches were capable of attaining a precision of 92.49% along with F1 total score of 89.16%, according to the testing findings. Yang et al. [11] have developed representative techniques for identifying fake negative information. They discovered numerous fascinating news items at the cutting-edge of research in 2013 regarding the Boston marathon bombings, most of which were instances and had a substantial influence on the stock market.

Shu et al. [12] investigated the relationship between bogus and actual relevant facts on Facebook and Twitter. They employed a hoax-based database in their follow-up analysis, which provides much more reliable forecasts for identifying fake news items by comparing them to recognized media sources through reputable monitoring platforms [13].

The analysis predicts that the techniques mentioned above are not highly accurate while the proposed techniques performed much better than the existing techniques.

3. Methodology

In the study, a technique allows a Twitter user to identify fake news.

3.1. Dataset

A publicly available dataset with more than 10,000 tweets is used based on tweets submitted by Twitter users. Search terms included ‘target’, ‘id’, ‘keyword’, ‘location’, ‘target’, ‘text’. We also chose a random sample of submissive tweets for this experiment. The test dataset can be treated as arbitrary and not subjected to the same treatment as the training dataset. The training data makes up 70% of the whole dataset and 30% was taken as for testing.

3.2. Pre-Processing

In our stage, the critical preprocessing step is feature creation, in which we must represent the sample tweets intractable feature space. Overall, this entails converting each string into a mathematical vector. This should be achievable with SKLearn’s CountVectorizer system, which produces an nK record term grid. K is the number of recognizable words across the n features in our case (fewer stop words and with a maximum feature restriction).

3.3. Small-Scale Evaluation Dataset

We took 116 tweets from the PolitiFact site classified as fake material to create a hand-checked best quality level. These were used to develop affected news tweet models. It is

worth noting that the origins of those tweets are not the same as those used in the planning set. Before building up our classification models, Ness selected the 116 tweets closest to the fake news tweets in the positive class as measured by TF-IDF and cosine proximity and erased those 116 tweets from the planning dataset. Programming a connection point returns a customer's information, such as the number of specialists. We also use the API to make additional estimates that illustrate the customer's Twitter activity, such as the number of tweets retweeted or the number of retweets received. We create 53 client-level features in all. We leverage the Twitter API to trees, support vector machines (SVM), and neural networks as primary classifiers for tweet-level components.

3.4. Assessment

We assess our methodology in various settings. To start with, we perform cross-approval on our uproarious preparation set; second, and more significantly, we train models on the preparation set and also approve them against a constructed highest quality level.

3.5. Text Features

The elements overhead do not reflect the real text of the tweet. For addressing the printed substance of the tweet, we investigated two other options: a pack of words (POW) model utilizing TF-IDF vectors, and a neural Doc2vec model prepared on the corpus. For the last option, we use genism to train models with 100, 200, and 300 aspects, both with DM and DBOW.

3.6. Theme Features

We prepared both a Latent Dirichlet Allocation model (LDA) overall dataset, differing the quantity of subjects somewhere in the range of 10 and 200 in strides of 10, just as a Hierarchical Dirichlet Process (HDP) model, which does not need the choice of various themes.

4. Results

Despite the enormous number of studies that have been undertaken on fake news detection, there is always an opportunity for experimentation, and new insights into the nature of fake news could lead to more effective and dependable algorithms. Furthermore, this is the first data to indicate the generalization of fake news detection systems to the best of the researchers' knowledge. This research proves that such models perform well on a specific dataset but do not generalize well. Extending the scope of fake news testing methods could bring up new possibilities. The results of both algorithms are shown in Figures 1 and 2:

The use of artificial neural networks to detect fake news appears to be promising. Traditional models may be highly beneficial when augmented with task-unique design techniques. As a result, we produced an overall dataset for the Latent Dirichlet Allocation model (LDA), altering the number of subjects from 10 to 200 in 10-point increments and a Hierarchical Dirichlet Process (HDP) model that does not require the selection of distinct themes.

According to our key findings, the proposed deep learning architecture offers a reasonably high accuracy performance in detecting fake news while the existing technique are not much better in accuracy. Both studies show a performance improvement while LSA perform slightly better than LDA as shown in Table 1.

Table 1. Result Comparison.

Model	Accuracy %	Precision	Recall	F1
LSA	95.72	0.9572	0.9572	0.9572
LDA	95.13	0.9541	0.9541	0.9541

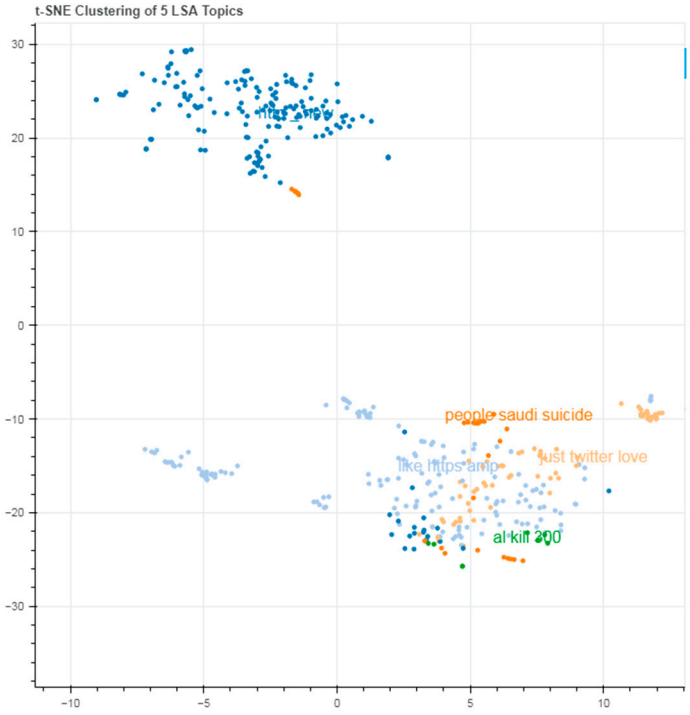

Figure 1. Clustering using LSA.

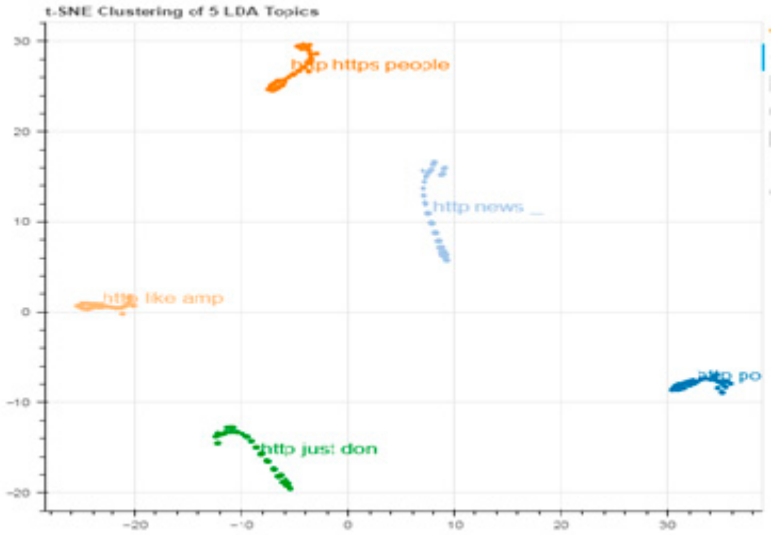

Figure 2. Clustering using LDA.

5. Conclusions

Artificial intelligence advancements have opened up numerous possibilities for detecting fake news on Twitter. Based on natural language processing (NLP), the suggested method for detecting fraudulent tweets has demonstrated highly effective results, with an accuracy of over 95%. The findings of the proposed methodology were as expected. The accuracy of the LSA and LDA models is nearly identical, which is extremely impressive. Furthermore, this study can help us determine the purity of social media platforms by detecting incorrect information uploaded by anyone. Based on data availability, the accuracy can be altered—the greater the dataset, the greater the accuracy. We will look at more advanced neural network architectures in the future, in addition to CNN and RNN.

Author Contributions: Conceptualization, H.K. and M.A.; methodology, A.S.; validation, H.M.Z.; writing—original draft preparation, I.M.; writing—review and editing, R.A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data can be obtained from the corresponding author on request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kumar, S.; Shah, N. False Information on Web and Social Media: A Survey. *arXiv* **2018**, arXiv:1804.08559.
2. Naseem, S.; Alhudhaif, A.; Anwar, M.; Qureshi, K.N.; Jeon, G. Artificial General Intelligence based Rational Behavior Detection Using Cognitive Correlates for Tracking Online Harms. *Pers. Ubiquitous Comput. vol.* **2022**, *2022*, 1–19. [CrossRef]
3. Sadiq, A.; Anwar, M.; Butt, R.A.; Masud, F.; Shahzad, M.K.; Naseem, S.; Younas, M. A review of phishing attacks and countermeasures for internet of things-based smart business applications in industry 4.0. *Hum. Behav. Emerg. Technol.* **2021**, *3*, 854–864. [CrossRef]
4. Shin, J.; Jian, L.; Driscoll, K.; Bar, F. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Comput. Hum. Behav.* **2018**, *83*, 278–287. [CrossRef]
5. Anwar, M.; Abdullah, A.H.; Altameem, A.; Qureshi, K.N.; Masud, F.; Faheem, M.; Cao, Y.; Kharel, R. Green communication for wireless body area networks: Energy aware link efficient routing approach. *Sensors* **2018**, *18*, 3237. [CrossRef] [PubMed]
6. Allcott, H.; Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **2017**, *31*, 211–236. [CrossRef]
7. Anwar, M.; Abdullah, A.H.; Butt, R.A.; Ashraf, M.W.; Qureshi, K.N.; Ullah, F. Securing data communication in wireless body area networks using digital signatures. *Tech. J.* **2018**, *23*, 50–55.
8. Ganame, K.; Allaire, M.; Zagdene, G.; Boudar, O. Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. *First Int. Conf. Intell. Secur. Dependable Syst. Distrib. Cloud Environ.* **2017**, *10618*, 169–181. [CrossRef]
9. Anwar, M.; Masud, F.; Butt, R.A.; Idrus, S.M.; Ahmad, M.N.; Bajuri, M.Y. Traffic Priority-Aware Medical Data Dissemination Scheme for IoT Based WBASN Healthcare Applications. *Comput. Mater. Contin.* **2022**, *71*, 4443–4456. [CrossRef]
10. Vosoughi, S.; Mohsenvand, M.; Roy, D. Rumor gauge: Predicting the veracity of rumors on twitter. *ACM Trans. Knowl. Discov. Data* **2017**, *11*, 1–36. [CrossRef]
11. Anwar, M.; Abdullah, A.H.; Saedudin, R.R.; Masud, F.; Ullah, F. CAMP: Congestion avoidance and mitigation protocol for wireless body area networks. *Int. J. Integr. Eng.* **2018**, *10*, 59–65. [CrossRef]
12. Chen, W.; Zhang, Y.; Yeo, C.K.; Lau, C.T.; Lee, B.S. Unsupervised rumor detection based on users' behaviors using neural networks. *Pattern Recognit. Lett.* **2018**, *105*, 226–233. [CrossRef]
13. Anwar, M.; Abdullah, A.H.; Qureshi, K.N.; Majid, A.H. Wireless body area networks for healthcare applications: An overview. *Telkomnika* **2017**, *15*, 1088–1095.