Fake News using Machine Learning

Abhishek Ranaut [a)], Swati[b)] Ankit Gupta,[c)] Royce Elijha,[d)] and Yash Vashistha[e)]

Computer Science and Engineering, *Lovely Professional University, Phagwara, India*

[a)] Corresponding author: swati.28320@lpu.co.in

[b)] Ranaut.abhi02@gmail.com

[c)]me.gupta.ankit@gmail.com

[d)]elijharoyce@gmail.com

[e)]yashuma12@gmail.com

Abstract. In the ultramodern period when the Internet is ubiquitous, everyone is turning to colorful online coffers for news. With the adding use of social media platforms like Facebook, Twitter, Reddit, etc., the newsflash snappily spread to millions of users in a veritably short time. Detecting fake news on any online platforms is hard because of challenges and different types of constraints faced, as a result it proves the conventional methods used by conventional media very ineffective. thus, we use a machine literacy approach to find a result. The spread of fake news has far- reaching consequences, including creating prejudice and impacting election results in favor of certain campaigners. Spammers also use catchy news captions to induce profit through clickbait advertisements. In this composition, we aim to perform double bracket of colorful newspapers available on the internet using generalities related to natural language processing (NLP), machine literacy (ML), and artificial intelligence (AI). We aim to categorized news as real or fake and give druggies with the capability to corroborate the authenticity of the website posting the news. reference information. Fake news is newspapers or reports that contain false information that's designedly circulated to deceive people.

Keywords: NLP (NATURAL LANGUAGE PROCESSING), DT (DECISION TREES), NAÏVE BAYES, KNN (K-NEAREST NEIGHBOR).

INTRODUCTION

Despite the benefits social media offers, its content often falls short because of poor insights and not good research the low cost and easy to distribute make social media vulnerable to the spread of misleading news. Misleading news can be used for personal or political gains. In 2016, the term "fake news" gained widespread use due to the high number of tweets, over a million, related to the Pizzagate conspiracy theory. This led to "fake news" being chosen as the Macquarie Word of the Year for that year. First, fake news, for example, can change the originality in the information ecosystem. During the 2019 Indian general election fake news was shared more often than the facts. also, fake news deliberately persuades consumers to simply make them believe preconceived notions or wrong Information. Misleading news is commonly shaped by proponent to convey a governmental message or influence.[1] example, in accordance with detailed reports, many mainstream political parties created fake user-accounts and bots to generate misleading narratives. News, some fake news for instance, are created to discredit and mislead people. They prevent from distinguishing truth from lies. In 45th President election in United states, the world experienced growing epidemic of false info. the virus of inaccurate news turmoil's the political world and also threats the integrity of the newspaper world. the worst impact of fake news is that it makes real-life panic: past year, In Washington DC Pizzeria a man walked in opted with AR-15 rifle as he believed the online news that "Hillary Clinton is leading a child ill-usage ring like taking young children as sex slave", result in later arrestment of the man by the police and charged for firing in the restaurant. The Natural language processing community is already dealing with widely related problem of deception recognition.[17] A study conducted by Ott et al (2011) the main objective was to identify fake reviews in sentiment analysis, they used a crowdsourcing procedure to generate training data for the positive class, which was then combined with honest reviews from TripAdvisor.[4][5]

FAKE NEWS OR MISLEADING INFO

Misleading Or Fake News

Fake news is a false or misleading news or info which is shown as if it is real news. It may be deliberately created and spread with the intent to deceive, misinform, or manipulate people's opinions, beliefs, or behaviors. here are various ways in which fake news can be disseminated, such as through made-up articles, sensational headlines, doctored visuals, deceptive data, and biased messaging. It often spreads rapidly through social media and other online platforms, making it difficult to distinguish from real news. The prevalence of falsely generated news has gained a major concern in in the world in the recent years, as it can have serious consequences for individuals, society, and democracy [15]. It can fuel mistrust, polarization, and conflict, undermine public discourse and decision-making, and even threaten national security.[1][2]

Literature Review

In a paper written by Mykhailo Granik.et.al shows an approach which is simple in detecting fake news from the content using naive bayes classifier. In this paper dataset was taken from the Facebook news posts and according to that an approach was tested and implemented as a software system. the data was collected from various Facebook pages like Political news pages (Politico, CNN, ABC News) the classification accuracy of this approach was almost 74%. as we know that the precision is worse for fake news. this is due to the Contrast of the dataset. the percentage of fake news was 4.9%. [9] Himank Gupta and his team developed an algorithm that uses various machine learning techniques to address issues like low accuracy, delay, and increased processing time when handling hundreds of tweets per second. They gathered over 400,000 data points from the HSpam14 dataset and divided them into two categories: spam and non-spam tweets. They identified some shallow features and the top 30 words from a Bagof-Words model that provided useful information. Their framework achieved an accuracy of approximately 91.65%, which was 18% higher than previous solutions.[9] The team led by Marco L. Della Vedova introduced a new approach for identifying fake news that leverages machine learning techniques and combines news and social context features. Their method surpassed existing methods in terms of accuracy and effectiveness. accuracy increased up to 78.8. Their aim was to check whether the new item is reliable or fake. the description of the dataset was used in test, then showed content-based approach he devised the method to use content-based and social-based approach. the Final dataset was poised of 15,500 posts, from 30 pages. with 2,400,00 likes by 800,000+ users. hoaxes 8923(57.7%) and non-hoaxes 6577(42.3%).[18][20]

Fake News and Social Media

Social media has played a substantial part in promoting fake news. The fast-paced nature of social media platforms, combined with the ease of sharing information, has made it easy for false information to spread rapidly and reach a vast audience. Social media algorithmically prioritizes content that generates engagement, such as likes, shares, and comments. This can motivate individuals and organizations to generate and share sensational or misleading content to create more engagement, regardless of whether the information is genuine or fake. Besides, social media platforms have also been used to spread misleading news as part of propaganda campaigns and other efforts to influence public opinion and behavior. To solve the problem of misleading news on social media, platforms have implemented various measures, such as fact-checking, content moderation, and reducing the reach of false information. However, the issue remains complex and challenging, requiring ongoing efforts and collaboration from various stakeholders, including tech companies, media organizations, and individual.[2][3] Studies: Kapferer et al (1992) discussed eight main sources of false info like, testimony, urban legends, troubling facts, manipulation, confidential information, experts' opinion, fantasies, and misunderstandings. And due to the absence of social media at that time, this analysis could not discuss media sources of fake news.[2] Jo et al(2002) classifies the bases of social media gossips into main two forms as The traditional mass media(radio, newspapers, articles and television) and Internet(social media and websites, etc.).[3] Shin et al (2018) researched the main two misleading info on twitter: as maximum content on twitter is from non- traditional news and some are from traditional news.[2] Muigai et al (2019) discovered in his studies about the two sources of social media news and that are mainstream news and online news. This study found that the maximum social media rumors are from online media.[3]

NATURAL LANGUAGE PROCESSING

Natural Language processing is used to specialize the system and algorithms. it is a part of artificial intelligence(AI). just like we humans have eyes and ears to see the read the info computers have programs and microphones to read and collect the info. and also just like our brain works to understand the info computers have cpu and different programs to collect and Understand the data or info given by us. Natural language processing is commonly used to set up combo of speech understanding and speech formation. it combines the computational grammer with the machine learning and deep learning models. due to the combination these two we can make computer to convert the human speech into text or voice material etc. NLP also converts and translate one language to another. NLP is so great that it can also respond to the language and commands which is spoken by us. it can also summarize all big amount of text in no time.[11][20]

DATA MINING

Data mining helps solve problems by sorting large datasets and identifying the patterns and relationships through data analysis. it basically, divided into two categories supervised and unspervised Supervised data mining usually builds a model for a data point when it has various target values. then you directly apply the model to data of the unkown target value.and then a new point in allocated which matches the model of target value. unsupervised data mining works on finding hidden data and structures rather then, focusing on the given attributes and it doesn't predict target point.[20]

MACHINE LEARNING CLASSIFICATION

ML is a major part of AI. it works on the basis of Copying humans Understanding, and tries to work similar way and it is also gradually Refining. In recent years we have used this in various technologies like in Netflix we can see that Netflix recognize what we want to see or what we usually prefer to watch by using Netflix recommendation engine, one another example is the auto-pilot cars which are based on machine learning. Machine learning basically works like, we take data apply statistical methods trained the given data in classification or regression models and then predict the data or insights in projects of Data Mining. Algorithms of Machine learning are usually created by Frameworks like TensorFlow and PyTorch etc.[20][22]

MACHINE LEARNING MODELS

Decision Tree

Decision Tree, is out of the mostly used algorithm used under supervised machine learning, which is non-parametric and has a hierarchical structure. it works is both Regression and Classification problems. As hierarchical structure indicate nodes so, there are different types of nodes like root node, leaf node, internal node and other branches too. in this structure branch represent the result of test, and internal node represents attribute and leaf nodes denotes class labels.[23][25] pseudo code Decision Tree Fig 1.

```
DecisionTree(Sample Son, features Father)
1. if stop (Son,Father) = true then
x. leafnode = generate_node()
y. leafnode.1 = classify(s)
z. return leafnode
2. rootnode = generate_node()
3. rootnode.test = find_split(Son,Father)
4. z = { z | z  outcome of rootnode.test condition)
5. for every value z Zim:
6. Sonz: = {Son rootnode.test(Son) = z and Son Son1};
7. childnode = growth(Sonz1 ,Father1) ;
8. Attach each resulting child node as a descendant of the
root node and label the connecting edge with the
corresponding outcome (rootnode→childnode) as z
return rootnode
```

FIGURE 1. pseudo code Decision Tree

Random Forest

This is another category of supervised machine learning, named- Random Forest algorithm. It usually works in both classification and regression it is particularly based on EL(Ensable Learning). which is the average created by combining several decision trees to improvise the correctness of the model which is applied on the dataset. the most impressive feature of Random Forest is that it can handle continuous variables which is mainly used in case of regression in a dataset and also in categorical variables which is used in case of classification. To increase the ability of random forest just add more trees which actually increase the accuracy and performance of the Random Forest.[23[25] pseudo code Random Forest Fig 2.

```
pseudo code(Random forest)
For i = 1 to n do
sample L drive data and override Li output
Build a Li-containing rootnode, Zi
Call Build(Zi)
end For
Build (Z):
If N contains only instances of a class, returns
else
random selection of X% of possibility of separation properties in Z
 Select the feature G with the highest information gain to split on
g childnodes of Z, Zi ,..., Zg where G has g possible values ( G, ..
. , GG)
For i = 1 to g
do
Set the content of Zi to Li, where Li is all instances of N that match Gi
 Call Build(Zi )
end for
end if
```

FIGURE 2. Pseudo Code Random Forest

```
SVM Pseudo-Code
A[0..N-1]: a  point set with N features that's sorted by information gain in  dwindling order  delicacy( i)  delicacy of a  vaticination model grounded on SVM with A[0...i] gone set
lo = 0
hi = N-1
val = acc(N-1)
SVM (G [0...N-1], val, lo, hi) {
 If (hi } > lo)
 return G[ 0..N-1 ] and val
 mi = (lo + hi ) / 2
 Val2 = acc(mi)
 if (Val2 < val)
 return SVM (G [0...mi], Val2, lo, mi)
 else (Val2 > val)
 return SVM (G [0...hi], val, mi, hi)
```

FIGURE 3. Pseudo Code Support Vector Machine

SVM (Support Vector Machine)

It is coming under supervised machine learning algorithm which can be castoff for classification and regression problems, as well as outlier detection. The objective of SVM is to classify the best line or boundary that can perfectly separate two different categoriesor groups, the best line also known as 'Hyperplane' is between the positive hyperplane and negative hyperplane, itis the best decision boundary so, we can put the latest data in the right category without any problem.[23][25] pseudo code (SVM) Fig 3.

Naïve Bayes

The Naïve Bayes (NB) algorithm is a category of supervised learning algorithm, it is commonly used by applying Bayes theorem to solve classification Problems. By Using Naive Bayes algorithms we can create Fast ML Models to predict the dataset faster. it is known by the name of "Probabilistic classifier" which means that it can predicts on the basis of probability. P(A|B)= P(B|A) P(A)/P(B) In this, A is a happening probability and B already occurred. here B is reality and A is just a theory which is yet to be predicting using Naive Bayes.[23] pseudo code (Naive Bayes):
Training dataset T, B= (b1, b2, b3,..., bn)
Output: classification of testing data-set.

KNN (K-Nearest Neighbor)

K-Nearest Neighbor (KNN) is coming under supervised machine learning algorithm, which is one of the simplest non- parametric algorithms, and uses a method named Proximity which is the nearness in space, used to classify the data into various groups generally two groups based on the similarity and predict the data point. it can be used in regression as well as in classification but maximum of the time it is used in classification. it is also known by the name of "Lazy Learner" Algo.[25]

KNN belongs to the type of supervised learning algorithms and is primarily castoff for tasks such as intrusion detection and pattern recognition. It is a non-parametric method, because it does not require any presumption or specific distributions to be assigned to the data, unlike other methods like GMM, which assumes a Gaussian distribution of the input. pseudo code KNN Fig 4.

```
Class (A, B, c) // A: train dataset, B: classified labels of A,
c: unidentified sample
For i = 1 Dom do
 Calculate distance dis (Ai, c)
end for
Calculate set I containing indices for K1 smallest distances
dis (Ai, c).
return major part of label for {Bi where iEI}
```

FIGURE 4. Pseudo Code KNN

COMBINING CLASSIFIERS

In this section, we are mentioning about the diverse Classifiers used to produce different Models and based on that model the analyze the outcome of the dataset whether the given data is fake or real. on the basis of these outcomes, which are created by using different methods and categorize the performance and accuracy by individual models and summarize the outcomes and detect the information on Fake news or political fake news, we can collect some new or latest information regarding that Particular topic or agenda.[23]

METHODOLOGY

In this section, we are Presenting Methodology for the model which is used as a tool to detect Fake news using Machine learning Algorithms. The 1st Segment of the Problem is to collect Data and then pre-process the data and then splitting the data into train and test dataset and Finally Running the Model.[2] Figure (5) methodology of system. In the Second segment, as we mentioned some of the algorithms in previous section like, Decision tree, Naive Bayes, Random Forest and SVM. each algorithm is applied separately and then Combining them as the goal is to improve accuracy and Performance. mainly, it is applied as a analyser which analyse the data and detect whether the given news is forged or not. In this we basically, applied machine learning algorithms alike, Linear Regression, Random Forest, Naive Bayes, KNN, Decision tree and SVM. The accuracy of the results describes the Final outcome. To create a Model , we First Collect a number of dataset on particular topic then we pre-process the data in order to eliminate any noise or any junk or null values from the given dataset. and just after that we apply NLTK which is known as Natural Language Toolkit and then we Fragmented the Data into train and test and then apply all the mentioned Machine Learning Algorithms(Decision tree, SVM, KNN etc.)[9].

In Fig 6 we showed the application of the NLTK, the pre-processing of the data is completed and the application of different algorithms which are applied then we test the test dataset to predict the most accurate to detect Fake news.[23] by using NLP and POS we process the words and then we use Python Scikit as it contains Count vectorizer and also Tiff (Tag image file format) Vectorizer. then Data is viewed by using Confusion Matrix.[18][19] therefore, to improve the sensitivity of our model, we have set its reset delicacy to 50.

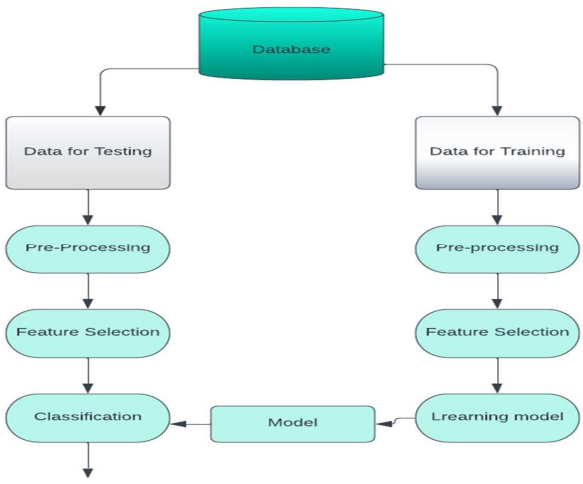

FIGURE 5. Methodology For System

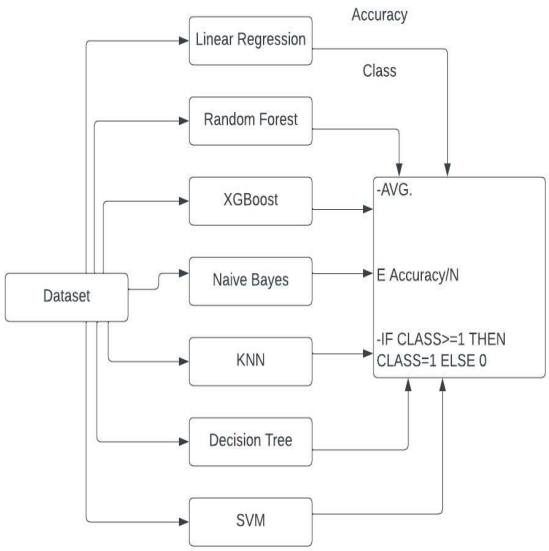

FIGURE 6. The Classification Algorithms.

To create a comprehensive dataset, we randomly selected 80 data points each from the fake and real datasets and kept the remaining 20 data points as a testing set for assessing the model's performance. Pre-processing the text data is crucial before applying a classifier, and we used the Stanford NLP (Natural Language Processing) tool for POS tagging and word tokenization. The resulting data is then converted into integer and float values to be accepted as input for ML algorithms, which significantly affects point birth and vectorization. To tokenize and vectorize textual data, we used tools present in the Python scikit-learn library, such as Count Vectorizer and Tfidf Vectorizer. Finally, for visualizing the data, we used a graphical representation in the form of a confusion matrix, as illustrated in Figure 7.[18][19]

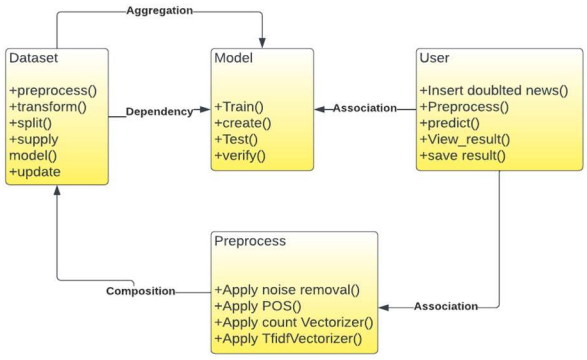

FIGURE 7. Fake Detector Model

CONCLUSION

In conclusion, machine learning algorithms can be effectively used for fake news prediction, The Naive Bayes, SVM, Random Forest, Decision tree etc. algorithms have proven to be effective for text By Combining these Machine learning algorithms we can classify the data and into test and train and predict the outcome to check whether the news is fake or not. classification tasks such as fake news prediction, and the implementation of these algorithms is relatively straightforward. The outcome of the experiments and the review of the accuracy of the algorithms will provide insight into the impending of machine learning algorithms for forged news prediction this can be also consider a starting point for further research in this field and solve the new problems.

REFERENCES

1- Khanam, Z., B. N. Alwasel, H. Sirafi, and Mamoon Rashid. "Fake news detection using machine learning approaches." In IOP conference series: materials science and engineering, vol. 1099, no. 1, p. 012040. IOP Publishing, 2021.

2- Al-Zaman, Md Sayeed. "Social media fake news inIndia." *Asian Journal for Public Opinion Research* 9, no. 1 (2021): 25-47.

3- Aldwairi, Monther, and Ali Alwahedi. "Detecting fake news in social media networks." Procedia Computer Science 141 (2018): 215-222.

4- Abu-Nimeh, Saeed, Thomas Chen, and Omar Alzubi. "Malicious and spam posts in online social networks." Computer 44, no. 9 (2011): 23-28.

5- Messabi, Khulood Al, Monther Aldwairi, Ayesha Al Yousif, Anoud Thoban, and Fatna Belqasmi. "Malware detection using dns records and domain name features." In Proceedings of the 2nd International Conference on Future Networks and Distributed Systems, pp. 1-7. 2018.

6- Aldwairi, Monther, Ansam M. Abu-Dalo, and Moath Jarrah. "Pattern matching of signature-based IDS using Myers algorithm under MapReduce framework." EURASIP Journal on Information Security 2017 (2017): 1-11.

7- Aldwairi, Monther, and Rami Alsalman. "Malurls: Malicious urls classification system." In Annual International Conference on Information Theory and Applications. 2011.

8- Aldwairi, Monther, and Hesham H. Alsaadi. "Flukes: Autonomous log forensics, intelligence and visualization tool." In Proceedings of the International Conference on Future Networks and Distributed Systems, pp. 1-6. 2017.

9- Wang, William Yang. "" liar, liar pants on fire": A new benchmark dataset for fake news detection." arXiv preprint arXiv:1705.00648 (2017).
10- Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. "Natural language processing (almost) from scratch." Journal of machine learning research 12, no. ARTICLE (2011): 2493-2537.
11- Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. "Natural language processing (almost) from scratch." Journal of machine learning research 12, no. ARTICLE (2011): 2493-2537.
12- Crammer, Koby, and Yoram Singer. "On the algorithmic implementation of multiclass kernel-based vector machines." Journal of machine learning research 2, no. Dec (2001): 265-292.
13- Feng, Song, Ritwik Banerjee, and Yejin Choi. "Syntactic stylometry for deception detection." In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 171-175. 2012.
14- Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey
15- T. Hancock. "Finding deceptive opinion spam by any stretch of the imagination." arXiv preprint arXiv:1107.4557 (2011).
16- Tandoc Jr, Edson C., Zheng Wei Lim, and Richard Ling. "Defining "fake news" A typology of scholarly definitions." Digital journalism 6, no. 2 (2018): 137-153.
17- Radianti, Jaziar, Starr Roxanne Hiltz, and Leire Labaka. "An overview of public concerns during the recovery period after a major earthquake: Nepal twitter analysis." In 2016 49th Hawaii international conference on system sciences (HICSS), pp. 136-145. IEEE, 2016.
18- Alkhodair, Sarah A., Steven HH Ding, Benjamin CM Fung, and Junqiang Liu. "Detecting breaking news rumors of emerging topics in social media." Information Processing & Management 57, no. 2 (2020): 102018.
19- Yi, Jeonghee, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques." In Third IEEE international conference on data mining, pp. 427-434. IEEE, 2003.
20- Tapaswi, Namrata, and Suresh Jain. "Treebank based deep grammar acquisition and Part-Of-Speech Tagging for Sanskrit sentences." In 2012 CSI Sixth International Conference on Software Engineering (CONSEG), pp. 1-4. IEEE, 2012.
21- Ranjan, Pradipta, and H. V. S. S. A. Basu. "Part of speech tagging and local word grouping techniques for natural language parsing in Hindi." In Proceedings of the 1st International Conference on Natural Language Processing (ICON 2003). 2003.
22- Diab, Mona, Kadri Hacioglu, and Dan Jurafsky. "Automatic tagging of Arabic text: From raw text to base phrase chunks." In Proceedings of HLT-NAACL 2004: Short papers, pp. 149-152. 2004.
23- Rouse, M. "AI (artificial intelligence). techtarget. com: https://searchenterpriseai. techtarget. com/definition." (2018).
24- Khanam, Z., B. N. Alwasel, H. Sirafi, and Mamoon Rashid. "Fake news detection using machine learning approaches." In IOP conference series: materials science and engineering, vol. 1099, no. 1, p. 012040. IOP Publishing, 2021.
25- Vishnoi, Sushant Kumar, T. E. E. N. A. Bagga, A. A. R. U. S. H. I. Sharma, and SAADAT NASIR Wani. "Artificial intelligence enabled marketing solutions: A review." Indian Journal of Economics & Business 17, no. 4 (2018): 167-177.
26- Ahmed, Hadeer, Issa Traore, and Sherif Saad. "Detection of online fake news using n-gram analysis and machine learning techniques." In Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1, pp. 127-138. Springer International Publishing, 2017.
27- Khanam, Zeba, and Salwa Alkhaldi. "An intelligent recommendation engine for selecting the University for Graduate Courses in KSA: SARS Student Admission Recommender System." In Inventive Computation Technologies 4, pp. 711-722. Springer International Publishing, 2020.