Project Synopsis
On

Fake News Detection

Submitted as a part of course curriculum for

Bachelor of Technology
In
Computer Science

Submitted by

1)Shitiz Rajvanshi
2)Taniya Singh
3)Shubham Goel

Under the Supervision of
Prof. Pardeep Tyagi
(Assistant Professor)

KIET Group of Institutions, Ghaziabad
Department of Computer Science
Dr. A.P.J. Abdul Kalam Technical University

2022-2023

DECLARATION

We hereby declare that this submission is our work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

Signature of Students

Name: 1) Taniya Singh (2000290120174)
2) Shitiz Rajvanshi(2000290120145)
3) Shubham Goel (2000290120158)

Date:

CERTIFICATE

This is to certify that Project Report entitled "**Fake news detection**" which is submitted by Shitiz Rajvanshi, Taniya Singh ,Shubham Goel in partial fulfilment of the requirement for the award of degree B. Tech. in Department of Computer Science of Dr A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Date:

Supervisor Signature
Prof. Pardeep Tyagi
(Assistant Professor)

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the synopsis of the B.Tech Mini Project undertaken during B.Tech. Third Year. We owe a special debt of gratitude to Mr. Pardeep Tyagi, Department of Computer Science, KIET Group of Institutions, Delhi- NCR, Ghaziabad, for his/her constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his/her cognizant efforts that our endeavours have seen the light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Ajay Kumar Shrivastava, Head of the Department of Computer Science, KIET Group of Institutions, Delhi- NCR, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

Last but not the least, we acknowledge our friends for their contribution to the completion of the project.

Signature:

Date : 14-11-2022

Name: 1) Taniya Singh (2000290120174)
2) Shitiz Rajvanshi(2000290120145)
3) Shubham Goel (2000290120158)

ABSTRACT

In this modern era where the internet is everywhere, everyone relies on various online resources for news .Along with the increase in the use of social media platforms like Facebook , Twitter, etc. news spread rapidly among millions of users within a very short span of time. The spread of fake news has far-reaching effects, including the formation of distorted beliefs and the also manipulation of the thinking about a particular topic. As a result, it has become extremely difficult to instantly determine if a piece of information is real or fake by looking at its source, content, and publisher .When it comes to categorizing information, machine learning has been crucial. Therefore, to determine if a specific piece of news is fake or not, we uses a variety of machine learning techniques.

We decided to use various NLP and preprocessing methodologies like tokenization, stop words removal, lemmatization, stemming and machine learning classification algorithms - logistic regression, PAC, ADA, Naive Bayes, SVM, Random forest, XG-boost, decision trees and RNN, to build a model that differentiates between fake news and real news and also analyze the performance of these various classification methodologies to choose the best classifier

TABLE OF CONTENTS

INTRODUCTION

World is changing rapidly. We have a number of advantages of this digital world but it also has its disadvantages as well. There are different issues in this digital world. One of them is fake news. Someone can easily spread a Fake News. Fake news is spread to harm the reputation of a person or an organization. It can be a propaganda against someone that can be a political party or an organization. There are different online platforms where the person can spread the fake news. This includes the Facebook, Twitter etc. Machine learning is the part of artificial intelligence that helps in making the systems that can learn and perform different actions. A variety of machine learning algorithms are available that include the supervised, unsupervised, reinforcement machine learning algorithms. The algorithms first have to be trained with a data set called train data set. After the training, these algorithms can be used to perform different tasks. Online platforms are helpful for users because they can easily access news. But the problem is this gives the opportunity to cyber criminals to spread fake news through these platforms. This news can be proved harmful to a person or society. Readers read the news and start believing it without its verification. Detecting fake news is a big challenge because it is not an easy task. If the fake news is not detected early then people can spread it to others and all the people will start believing it. Individuals, organizations, or political parties can be affected by fake news. People's opinions and decisions are majorly affected by any fake news. Different researchers are working for the detection of fake news. The use of Machine learning is proving helpful in this field. Researchers are using different algorithms to detect the false news. In today's time we can find that the fake news are increasing rapidly. That is why there is a need to detect fake news. The algorithms of machine learning are trained to fulfill this purpose. Machine learning algorithms will detect the fake news automatically once they have trained.
Types of Data in social Media Posts

1. **Text**: Computational linguistics analyzes text, focusing on the genesis of text semantically and methodically. Because many of the posts are written in the form of texts, much work has been carried out into analyzing them.

2. **Multimedia**: Several types of media are combined in a single post. Audio, video, photos, and graphics may all be included. This is highly appealing, because it captures the attention of the visitors without requiring them to read the content

3. **Hyperlinks**: Hyperlinks allow the post's creator to cross-reference to other sources, gaining viewers' trust by confirming the post's genesis. Cross-reference to other social media networking sites, as well as the embedding of photos, is common practice.

Fake news type

Types of Fake News and patterns that help in detection

1. **Visual Based**: These false news posts make extensive use of graphics as compared to content, which may include manipulated photographs, doctored video, or a combination of the two.

2. **User Generated News**: This sort of falsified news is generated by phoney accounts and is targeted to certain audiences, which might reflect specific age groups, gender, culture, or political affiliations. 3. Knowledge based: These posts provide scientific (so-called) explanations to some unresolved problems, leading people to feel they are genuine. For example, natural therapies for high blood sugar levels in the human body.

4. **Style based**: Pseudo Journalists who impersonate and mimic the style of some accredited journalists write style-based posts.

5. **Stance based**: It is a portrayal of true statements in such a way that its meaning and purpose are altered.

PROBLEM STATEMENT

Fake News have become more prevalent in recent years and with great amount of dynamism in internet and social media, differentiating between facts and opinions, relating to commercial or political has become more difficult.

The internet is used intentionally or accidentally to spread false information.
People and culture have been forever changed by the massive spread of fake news.

We use various NLP and preprocessing methodologies like tokenization, stop words removal, lemmatization, stemming and machine learning classification algorithms - logistic regression, PAC, ADA, Naive Bayes, SVM, Random Forest, XG-boost, decision trees and RNN, to build a model that differentiates between fake news and real news and also analyze the performance of these various classification methodologies to choose the best classifier.

OBJECTIVE

- The main goal is to identify fake news, which is a classic text classification issue. It is needed to build a model that can differentiate between “Real” and “Fake” news
- We decided to gather data, preprocessed the text, and translated our articles into supervised model features.
- Our goal is to develop a model that classifies (using various machine learning algorithms) a given news article as either fake or true.

LITERATURE REVIEW

1) Fake News Detection Using Deep Learning Architecture

- Published by Muhammed Umer, Zainab Imtiyaz And Saleem Ullah at IEEE Access (Volume-8) on 26 August 2020.
- The research mainly focuses on developing the reasoning, this work acquired a dataset from the Fake News Challenges (FNC) website which has four types of stances: agree, disagree, discuss, and unrelated.
- The nonlinear features are fed to PCA and chi-square which provides more contextual features for fake news detection.
- The Experimental results have an accuracy of about 97.8%.
  Keywords Used: NLP, chi-square test, Deep Learning and CNN

2) Propagation-Based Fake News Detection Using Graph Neural Networks with Transformer

- Published by Hayato Matsumoto and Soh Yoshida at 2021 IEEE 20th Global Conference on Consumer Electronics on 12-15 October 2021.
- In this paper, we propose a method for fake news detection using Graph Transformer Network (GTN), which can learn efficient node representations while identifying useful connections between nodes in the original graph.
- Model works on a real world dataset: Twitter news records.
- Keywords Used: Graph Neural Network, GTN.

3) A Novel Stacking Approach for Accurate Detection of Fake News

- Published by Tao Jiang, Jian Ping Li and Abdus Saboor at IEEE Access (Volume 9) on 1st February 2021.
- Achieved testing success of 99.04% and 96.05% respectively on the ISOT dataset and KDnugget dataset.
- Keywords used: Decision Tree, Logistical Regression, KNN

4) A SMART SYSTEM FOR FAKE NEWS DETECTION USING MACHINE LEARNING

- Published by Uma,Sharma, Sidarth Saran, Shankar M. Patil
- In this paper the proposed model is working well and defining the correctness of results up to 93.6% of accuracy.

- the proposed strategy is a mix of Naive Bayes classifier, Support Vector Machines, and semantic investigation.

5) A smart system for fake news detection using machine learning

- The news websites are publishing the news and provide the source of authentication .It is harmful for the society to believe on the rumors and pretend to be a news. The need is to stop the rumors , and focus on the correct, authenticated news articles.
- In this paper the proposed model is working well and defining the correctness of results upto 93.6% of accuracy. the proposed strategy is a mix of Naive Bayes classifier, Support Vector Machines, and semantic investigation
- Naive Bayes classifier is a supervised machine learning algorithm that uses Bayes theorem .SVM algorithm is used to extract the binary class based on the data given to the model. In the proposed model, the work is to classify the article in two categories either true or false. A Support Vector Machine SVM is a supervised machine learning algorithm that can be used for both regression and classification purpose.

6)Fake News Detection on Social Media using Geometric Deep Learning

- Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, Michael M. Bronstein
- underlying core algorithms are a generalization of classical CNNs to graphs, allowing the fusion of heterogeneous data such as content, user profile and activity ,social graph and news propagation
- Our their model they trained and tested on news stories, verified by professional fact-checking organizations, that were spread on Twitter
- Their experiments indicate that social network structure and propagation are important features allowing highly accurate (92.7% ROC AUC) fake news detection

7) Weak Supervision for Fake News Detection via Reinforcement Learning

- they proposed a novel framework that can leverage user reports as weak supervision for fake news detection.
- The proposed framework works by integrating three components including the annotator, the reinforced selector and the fake news detector
- This was demonstrated in a series of experiments conducted on a WeChat dataset consisting of news articles and user fee.

8)Mykhailo Granik et. al. in their paper shows a simple approach for fake news detection using naïve Bayes classifier.

- This approach was implemented as a software system and tested against a data set of Facebook news posts.
- They were collected from three large Facebook pages each from the right and from the left, as well as three large mainstream political news pages (Politico, CNN, ABC News).
- They achieved classification accuracy of approximately 74%. Classification accuracy for fake news is slightly worse. This may be caused by the skewness of the dataset: only 4.9% of it is fake news.

9) Text classification using association rule with a hybrid concept of Naive bayes classifier and genetic algorithm

- Text classification is the automated assignment of natural language texts to predefined categories based on their content.
- The experimental results show that the proposed system works as a successful text classifier. A system based on the proposed algorithm has been implemented and tested. This paper presents a new algorithm for text classification.
- Instead of using words, word relation i.e association rules is used to derive feature set from pre-classified text documents. Keeping this demand into consideration, new and updated techniques are being developed for the purpose of automated text classification.
- The concept of Naive Bayes Classifier is then used on derived features and finally a concept of Genetic Algorithm has been added for final classification

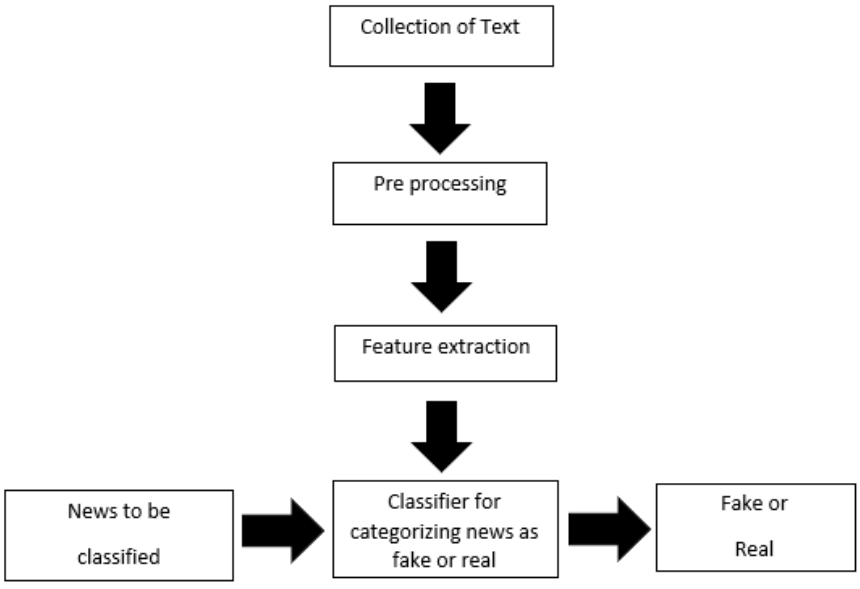
10) Nguyen Vo student of Ho Chi Minh City University of Technology (HCMUT) Cambodia did his research on fake news detection and implemented in 2017**.**

- He used Bi-directional GRU with Attention mechanism in his project fake news detection.
- Yang et al. originally proposed this mechanism. He also used some Deep learning algorithms and tried to implement other deep learning models such that Auto-Encoders, GAN, CNN

PROPOSED METHODOLOGY

The fake news model detection is built using following steps : Text Collection, Text Preprocessing, Feature Extraction and then finally classification using different classifiers.

The proposed system various Machine Learning algorithm for detecting the fake news, The data is divided into two parts (test and train). The train data is trained and is classified into groups with similar datasets. After the data is trained the test data is assigned to the group which has similar characteristics with the group. TF-IDF vector issued to count the number of word and also the number of unique words is counted and at the same time the weights are allotted to each and every word. In this way the unimportant words are not taken into consideration and the accuracy of only important words are matched and detected from the dataset, this saves the time to detect the accuracy of the news



We decided to use various machine learning algorithm for classifying whether a piece of news is false or true.

Algorithm Used:

1)Logistic Regression

2)ADA Aggressive Boost Classifier

3)Passive Classifier

4)XG Boost

5)Random Forest

6)Naïve Bayes

7)SVM

8)Decision Tree

TECHNOLOGY USED

For determining if a given news article is true or false, we have chosen to use machine learning techniques.

Libraries to be used

- Numpy
- Pandas
- Matplotlib
- Scikit-learn
- Tensorflow
- Keras

Algorithms to be Used:

1)Naïve Bayes Classifier: This technique is based on Bayes theorem, which assumes that the presence of a particular feature in a class is independent of the presence of any other feature. It provides way for calculating the posterior probability.
$P(x) = P(c) * P(c)/P(x)$

2) **Random Forest:** Random Forest Algorithm is used to solve both regression and classification problems, making it a diverse model that is widely used by engineers. Pros: Used for regression and classification problems, making it a diverse model. Prevents overfitting of data.

2.1 Bagging (Bootstrap Aggregation) — Decisions trees are very sensitive to the data they are trained on — small changes to the training set can result in significantly different tree structures. Random forest takes advantage of this by allowing each individual tree to randomly sample from the dataset with replacement, resulting in different trees. This process is known as bagging or bootstrapping.

2.2 Feature Randomness — In a normal decision tree, when it is time to split a node, we consider every possible feature and pick the one that produces the most separation between the observations in the left node vs. those in the right node. In contrast, each tree in a random forest can pick only from a random subset of features. This forces even more variation amongst the trees in the model and ultimately results in lower correlation across trees and more diversification

3. **Logistic Regression**: It is used to estimate discrete values (Binary values like true/false) based on given set of independent variables.It predicts the probability of occurrence of an event by fitting data to a logit function. Since, it predicts the probability, its output values lie between 0 and 1 (as expected). Mathematically, the log odds of the outcome are modelled as a linear combination of the predictor variables.
Odds = p/(1-p) = probability of event occurrence / probability of not event occurrence
ln(odds) = ln(p/(1-p))
logit(p)=ln(p/(1-p))= b0+b1X1+b2X2+b3X3....+bkXk

4. Passive Aggressive Classifier: The Passive Aggressive Algorithm is an online algorithm; ideal for classifying massive streams of data (e.g., twitter). It is easy to implement and very fast. It works by taking an example, learning from it and then throwing it away . Such an algorithm remains passive for a correct classification outcome, and turns aggressive in the event of a miscalculation, updating and adjusting. Unlike most other algorithms, it does not converge. Its purpose is to make updates that correct the loss, causing very little change in the norm of the weight vector
we choose to use a website to deploy our model.

(Python)Flask

Frontend: HTML, CSS, JavaScript , jQuery

Backend: Django

ANALYSIS

We can evaluate machine learning algorithm using various metrics like

1. Accuracy = $\frac{|T\ P|+|T\ N|}{|T\ P|+|TN|+|F\ P|+|F\ N|}$

2. Precision= $\frac{|T\ P|}{|T\ P|+|F\ P|}$

3. Recall= $\frac{|T\ P|}{|T\ P|+|F\ N|}$

4. F1-Score= $2 * \frac{Precisionn * Recall}{Precision + Recall}$

we will evaluate and analyze the result based on these metrics for different datasets.

EXPECTED OUTCOME AND FUTURE SCOPE

Fake news is a problem that is only becoming worse and more complicated.
We, therefore, decided to create our Fake News Detection system, which accepts user input and categorizes it as true, to stop the situation. we expect the accuracy may vary from 93 to 98.9 % as we have decided to use various algorithms for our model designing. and we have chosen to deploy our model over website.
This web-based application will assess a news article's credibility so that consumers can only link to the authentic news articles. Further in the future, it could be expanded by using various more efficent machine learning as well as deep learning algorithms to get much more effective accuracy.

REFERENCES

[1] N. Smitha and R. Bharath, "Performance Comparison of Machine Learning Classifiers for Fake News Detection," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 696-700.

[2] A. Benamira, B. Devillers, E. Lesot, A. K. Ray, M. Saadi and F. D. Malliaros, "Semi-Supervised Learning and Graph Neural Networks for Fake News Detection, 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2019, pp. 568-569.

[3] S. 1. Manzoor, J. Singla and Nikita, "Fake News Detection Using Machine Learning approaches: A Systematic Review, 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 230-234.

[4] A. Kesarwani, S. S. Chauhan and A. R. Nair, "Fake News Detection on Social Media using K-Nearest Neighbor Classifier," 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE). 2020, pp. 1-4.

[5] K. Poddar, G. B. Amali D. and K. S. Umadevi, "Comparison of Various Machine Learning Models for Accurate Detection of Fake News," 2019 Innovations in Power and Advanced Computing Technologies (I-PACT), 2019, pp. 1-5.

[6] S. B. Parikh and P. K. Atrey, "Media-Rich Fake News Detection: A Survey," 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, 2018, pp. 436-441

[7] A. Martínez-Garcia,S.Morris,M.Tscholl,F.Tracy, and P. Carmichael, "Case-basedlearning,pedagogicalinnovation,andsemanticweb technologies," IEEE Trans.Learn. Technol., vol. 5, no. 2, pp. 104–116,2012

[8] T.Granskogen and J.A. Gulla,"Fake news detection: Network data from social media used to predict fakes,"CEUR Workshop Proc., vol. 2041, no. 1, pp. 59–66, 2017.

[9] S.B.Parikh ,V.Patil, and P.K.Atrey, "On the Origin ,Proliferation and Toneof FakeNews,"Proc.-2ndInt.Conf.Multimed.Inf. Process. Retrieval, MIPR 2019, pp. 135–140,2019

[10] Y. Seo, D. Seo, and C. S. Jeong, "FaNDeR: Fake News Detection Model Using MediaReliability," IEEE Reg. 10 Annu. Int. Conf.Proceedings/TENCON, vol. 2018–October, no. October, pp. 1834–1838, 2019.