FAKE NEWS PREDICTION USING MACHINE LEARNING

Dissertation submitted in fulfilment of the requirements for the Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

Mannam Gnaneshwar

12114807

Supervisor

VED PRAKASH CHAUBEY

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

Month : April Year : 2024

PAC FORM

Project Title: Fake news Prediction using Machine Learning Algorithm

Problem:

Objective: The aim of this project is to develop machine learning models capable of accurately predicting the fake news based on relevant features.

Approach:

Data Collection: Gather a dataset of news articles labeled as either fake or real. There are several datasets available online for this purpose, such as the FakeNewsNet dataset or Kaggle's Fake News Dataset.

Feature Extraction: Extract relevant features from the text of each article. These features could include:

- Word frequencies or TF-IDF scores
- N-grams (sequences of N words)
- Sentiment analysis scores
- Readability scores
- Presence of specific keywords or phrases
- Source credibility (e.g., domain authority)

Data Preprocessing: Clean and preprocess the text data. This may involve removing stop words, punctuation, and special characters, as well as stemming or lemmatization.

Model Training: Train a PAC-based classifier using the extracted features. PAC is a framework for evaluating and comparing classifiers based on their trade-off between accuracy and confidence. You can use algorithms such as logistic regression, support vector machines (SVM), or decision trees with PAC-based evaluation criteria.

Evaluation: Evaluate the performance of your model using metrics such as accuracy, precision, recall, and F1-score. Additionally, since PAC focuses on the trade-off between accuracy and confidence, you'll want to analyse the classifier's performance across different confidence thresholds.

ABSTRACT

A variety of machine learning algorithms are deployed, including but not limited to Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting Classifier, and Linear Regression. These algorithms are integrated into a pipeline that involves text vectorization via CountVectorizer and TF-IDF transformation.

The models are trained and assessed using a train-test split methodology, with performance metrics like accuracy and confusion matrices employed to evaluate their efficacy in classifying news articles. The findings highlight that specific algorithms, such as Logistic Regression and SVM, exhibit notable accuracy in distinguishing between fabricated and authentic news, whereas others like Linear Regression are better suited for regression-oriented tasks.In summary, this investigation showcases the utilization of machine learning techniques in the realm of fake news identification and offers insights into the performance disparities among different algorithms for this particular task.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation proposal entitled “FAKE NEWS PREDICTION USING MACHINE LEARNING” in partial fulfilment of the requirement for the award of Degree for Bachelor of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. Ved Prakash Chaubey. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University’s Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Mannamgnaneshwar

12114807

RK21URA30

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the B.Tech Dissertation/dissertation proposal entitled "**FAKE NEWS PREDICTION USING MACHINE LEARNING**", submitted by **Mannam Gnaneshwar** at **Lovely Professional University, Phagwara, India** is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Ved Prakash Chaubey
(Name of Supervisor)

Date: 26th April 2024

Counter Signed by:

1) **Concerned HOD:**
HoD's Signature: ________________

HoD Name: ____________________

Date: ___________________

2) **Neutral Examiners:**

External Examiner

Signature: _______________

Name: __________________

Affiliation: ______________

Date: ___________________

Internal Examiner

Signature: _______________

Name: __________________

Date: ___________________

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all those who have contributed to the completion of this project.

First and foremost, I extend my heartfelt appreciation to my supervisor Ved Prakash Chaubey for their invaluable guidance, unwavering support, and insightful feedback throughout the duration of this project. Their expertise and encouragement have been instrumental in shaping the direction and execution of this research.

I am deeply thankful to Lovely Professional University for providing the necessary resources and facilities for conducting this study. The access to datasets and computing resources has greatly facilitated the implementation and evaluation of machine learning algorithms.

I am indebted to my peers and colleagues for their constructive criticism, fruitful discussions, and moral support, which have enriched my understanding and perspective on the subject matter.

Special thanks are due to my friends for their patience, understanding, and encouragement during the course of this project. Their unwavering support has been a constant source of motivation and inspiration.

Lastly, I would like to acknowledge the contributions of all the researchers, developers, and contributors to the open-source libraries and frameworks used in this project. Their collective efforts have played a significant role in advancing the field of machine learning and data science.

This project would not have been possible without the collective efforts and support of all those mentioned above, and for that, I am truly grateful.

TABLE OF CONTENTS

TABLE OF CONTENTS

CONTENTS **PAGE NO.**

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 BACKGROUND

Fake news has become a pervasive issue in today's digital age, with misinformation spreading rapidly across various online platforms. The proliferation of fake news poses significant challenges to society, including threats to democracy, public trust, and social cohesion. Addressing this problem requires innovative approaches leveraging advancements in machine learning and natural language processing.

1.1.1 IMPORTANCE

The importance of combating fake news cannot be overstated. Misinformation can manipulate public opinion, influence elections, and even incite violence. Therefore, developing effective tools and techniques to identify and mitigate fake news is essential for preserving the integrity of information dissemination in the digital era.

1.1.2 OBJECTIVES OF THE STUDY

The primary objectives of this study are as follows:

- To explore the application of machine learning algorithms for fake news detection.
- To develop a predictive model capable of distinguishing between fake and real news articles with high accuracy.
- To evaluate the performance of the proposed model using appropriate metrics and benchmarks.

To contribute to the ongoing efforts in combating fake news by providing a practical and scalable solution.

1.1.3 SCOPE OF THE STUDY

This study focuses specifically on the application of machine learning techniques for fake news prediction. The scope includes:

- Data collection: Gathering a diverse dataset of labeled news articles, including both fake and real examples.

- Feature engineering: Extracting relevant features from the text of news articles to facilitate classification.
- Model development: Training and evaluating machine learning models, such as logistic regression, support vector machines, or neural networks, for fake news prediction.
- Performance evaluation: Assessing the performance of the predictive model using standard evaluation metrics, such as accuracy, precision, recall, and F1-score.
- Limitations: Recognizing the inherent challenges and limitations associated with fake news detection, including the dynamic nature of misinformation and the presence of adversarial attacks..

2. REVIEW OF LITERATURE

PAPER I:

RESEARCH ARTICLE

Big Data ML-Based Fake News Detection Using Distributed Learning

ALAA ALTHENEYAN AND ASEEL ALHADLAQ
Department of Computer Science and Engineering, College of Applied Studies and Community Services, King Saud University, Riyadh 11495, Saudi Arabia
Corresponding author: Alaa Altheneyan (atheneyan@ksu.edu.sa)

This work was supported by Researchers Supporting Project number (RSPD2023R532), King Saud University, Riyadh, Saudi Arabia.

Introduction

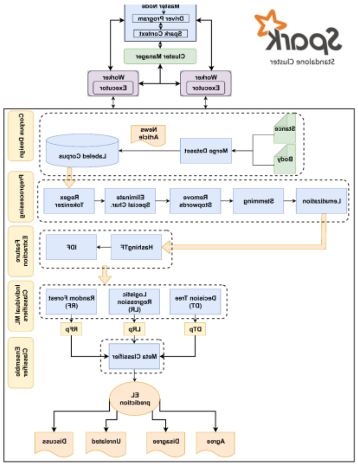
The use of social media platforms to disseminate and digest media has increased in recent years. Social networking sites like Facebook and Twitter generate daily data . It is no secret that the internet is a goldmine of information, especially recent newspredictions. The paper explores the impact of machine learning in analyzing buying history to recommend items and its potential in shaping the future of retail.

Objectives and Methodology

In the first section of this research, we examine the effectiveness of Recurrent Neural Networks (RNN) in modeling news articles to identify the link between an article's body content and its title. As part of our research, we use the dataset made available for the FNC-1 competition to train and assess a classifier. We want the classifier to be able to do the following.

Implementation

Data mining relies heavily on pre-processing. It converts inconsistent and incomplete raw data into a machine-readable representation. Various text preprocessing activities were conducted on the FNC-1 dataset. To complete these tasks, NLP approaches such as character conversion to lowercase letters, stop word elimination, stemming, and tokenization, as well as algorithms from keras library were used. Stop words, which comprise words like ‘‘the, of, there,’’ etc., are the most commonly used words in our daily language and typically have relatively limited significance in terms of the entire context of the phrase. By removing the stop words, we save time and space that would otherwise be consumed by the useless phrases mentioned before. Words with comparable meanings may appear in the text many times. For example, ‘‘eating’’ in any sentence will become ‘‘eats’’. Reducing the language to its most basic form can help if that’s the case. This operation, known as stemming [51], uses an open-source version of the NLTK’s Porter stemmer method. Few preprocessing steps are as follows:



Motivation and Project Features

In 2017, Facebook released a white paper that explored the risks of online communication and the management of being one of the most prominent social media platforms today. Weedon, Nuland, and Stamos also noticed the growing challenge of using the enigmatic phrase ‘‘fake news,’’ and proclaimed that ‘‘the overuse and misapplication of the term ‘‘fake news’’ might be challenging since we cannot understand or adequately address these concerns without shared definitions’’ [19]. The word can apply to anything from virtually incorrect news articles to deceptions, April Fools’ jokes, rumors, clickbait, or stated opinions posted online with incorrect facts. In this research work, ‘‘fake news’’ is defined as a written article that is manifestly untrue and falsely disseminated without being authentic mostly accompanird by malicious intents. This definition includes three important textual, visual, and audio bases. Other elements such as video-based fake news and audio, are typically ignored when referring to textual fake news; additionally, each element has its linguistic complexities that necessitate different machine learning and deep learning algorithms to detect and solve problems such as ‘Deep Fake,’ etc.

Literature Review

This section provides an overview of the previous research's difficulties in identifying fake news. To identify fabricated news stories, it is necessary to do rumor detection and identification. It is important to distinguish between Real and fake news since both are based on deliberate fabrication. Fake news identification is particularly difficult when detecting news based on characteristics. Tweets and social context can be used to generate features. As a result, we assess prior work based on single-modality and stance identification.

Statistical Analysis

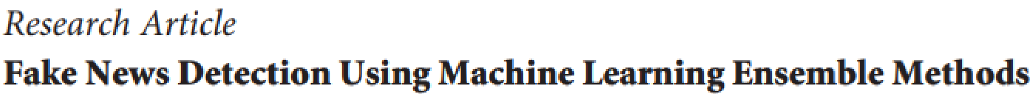
The paper presents the statistical analysis of the dataset, including correlation matrix, accuracy mapping for different k-values, and different statistical aggregations. Additionally, it includes various graphs, such as scatter plots of and fiscal power, a snapshot of the dataset header, and residual vs. predicted comparison graphs.

The conclusion and future scope section discusses the positive correlation between and

Results

The experimental results of Term Frequency-Inverse Document Frequency (TF-IDF) and HashingTF feature extraction techniques with ensemble models are presented in Table 2. The results using HashTF and IDF features regarding accuracy, precision, recall, and F1-score are 93.45%, 92.03%, 92.45%, and 92.25%. The results from LR_HashingTF-IDF is 93.45%, and it's a highest as compared to all other experimental. Furthermore, Bigram Logistic Regression exhibits 88.45% accuracies, 87.02% precision, 88.01% recall, and 87.06% F1-score. We also performed experiments using glove word embedding. We used the glove embedding technique with logistic regression. However, the glove with logistic regression model results is not so high but quite well with accuracy scores of 73.25% and 63.12%, 73.25%, 62.45% as the precision, recall, and F1-score. To make a broader comparison, we include features of the count vectorizer technique. The features of the count vectorizer were passed to logistic regression to detect fake news. Using the count vectorizer technique, the logistic model achieved 88.45% accuracy, 82.12% precision, 88.45% recall, and 87.35% F1 score. Moreover, we merged the count vectorizer and TF-IDF features to obtain better results, but we failed to avail improved results due to the high computational cost. The correctness, precision, recall, and F1-score using count vectorizer and TF-IDF features with logistic regression are 84.54%, 83.12%, 84.25%, and 83.26%. We also employed the Support Vector Machine (SVM) model to testify its abilities using count vectorizer features, and the SVM model gets improved results with 91.75% accuracy, 91.25% precision, 91.24% recall, and 90.45% F1-score. As compared to LR with count vectorizer, the SVM obtained high results. We also employed LR and SVM models with HashingTF-IDF features. The results of LR with HashingTF-IDF are better than the SVM model. Compared to LR, the SVM model with HashingTF-IDF achieved 90.75% accuracyBy presenting these results, the authors illustrate the thoroughness of their methodology and the advancements employed in their machine learning approach for fake news prediction. These results serve as a testament to the reliability and effectiveness of the model, providing valuable insights for the readers and contributing to the advancement of the field of machine learning in the context of predicting fake newss.

PAPER II:



Research Article

Fake News Detection Using Machine Learning Ensemble Methods

Key Points

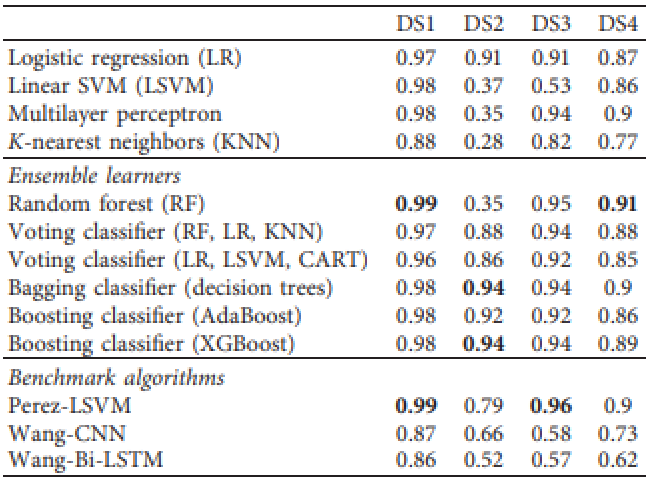
- In the following, we describe our proposed framework, followed by the description of algorithms, datasets, and performance evaluation metrics. 2.1. Proposed Framework. In our proposed framework, as illustrated in Figure 1, we are expanding on the current literature by introducing ensemble techniques with various linguistic feature sets to classify news articles from multiple domains as true or fake.)e ensemble techniques along with Linguistic Inquiry and Word Count (LIWC) feature set used in this research are the novelty of our proposed approach.)ere are numerous reputed websites that post legitimate news contents, and a few other websites such as PolitiFact and Snopes which are used for fact checking. In addition, there are open repositories which are maintained by researchers [11] to keep an up-to-date list of currently available datasets and hyperlinks to potential fact checking sites that may help in countering false news spread. However, we selected three datasets for our experiments which contain news from multiple domains (such as politics, entertainment, technology, and sports) and contain a mix of both truthful and fake articles.)e datasets are available online and are extracted from the World Wide Web.)e first dataset is ISOT Fake News Dataset [23]; the second and third datasets are publicly available at Kaggle [24, 25]. A detailed description of the datasets is provided in Section 2.5.)e corpus collected from the World Wide Web is preprocessed before being used as an input for training the models.)e articles' unwanted variables such as authors, date posted, URL, and category are filtered out. Articles with no body text or having less than 20 words in the article body are also removed. Multicolumn articles are transformed into single column articles for uniformity of format and structure.)ese operations are performed on all the datasets to achieve consistency of format and structure.

Synopsis

. Logistic Regression. As we are classifying text on the basis of a wide feature set, with a binary output (true/false or true article/fake article), a logistic regression (LR) model is used, since it provides the intuitive equation to classify problems into binary or multiple classes [27]. We performed hyperparameters tuning to get the best result for all individual datasets, while multiple parameters are tested before acquiring the maximum accuracies from LR model. Mathematically, the logistic regression hypothesis function can be defined as follows [27]: hθ(X) � 1 1 + e −( ) β0 + β1X . (1) Logistic regression uses a sigmoid function to transform the output to a probability value; the objective is to minimize the cost function to achieve an optimal probability.

)e cost function is calculated as shown in Cost. It presents detailed insights into the implementation and comparison of various machine learning algorithms, as well as a thorough analysis of the results obtained, demonstrating the capabilities of the developed platform in accurately predicting used fake newss. The study also acknowledges the limitations and suggests potential avenues for future improvement, providing valuable insights for further research in this domain.

Table 2: Overall accuracy score for each dataset.



	DS1	DS2	DS3	DS4
Logistic regression (LR)	0.97	0.91	0.91	0.87
Linear SVM (LSVM)	0.98	0.37	0.53	0.86
Multilayer perceptron	0.98	0.35	0.94	0.9
K-nearest neighbors (KNN)	0.88	0.28	0.82	0.77
Ensemble learners				
Random forest (RF)	**0.99**	0.35	0.95	**0.91**
Voting classifier (RF, LR, KNN)	0.97	0.88	0.94	0.88
Voting classifier (LR, LSVM, CART)	0.96	0.86	0.92	0.85
Bagging classifier (decision trees)	0.98	**0.94**	0.94	0.9
Boosting classifier (AdaBoost)	0.98	0.92	0.92	0.86
Boosting classifier (XGBoost)	0.98	**0.94**	0.94	0.89
Benchmark algorithms				
Perez-LSVM	**0.99**	0.79	**0.96**	0.9
Wang-CNN	0.87	0.66	0.58	0.73
Wang-Bi-LSTM	0.86	0.52	0.57	0.62

Results

Table 2 summarizes the accuracy achieved by each algorithm on the four considered datasets. It is evident that the maximum accuracy achieved on DS1 (ISOT Fake News Dataset) is 99%, achieved by random forest algorithm and Perez-LSVM. Linear SVM, multilayer perceptron, bagging classifiers, and boosting classifiers achieved an accuracy of 98%.)e average accuracy attained by ensemble learners is 97.67% on DS1, whereas the corresponding average for individual learners is 95.25%.)e absolute difference between individual learners and ensemble learners is 2.42% which is not significant. Benchmark algorithms Wang-CNN and Wang-Bi-LSTM performed poorer than all other algorithms. On DS2, bagging classifier (decision trees) and boosting classifier (XGBoost) are the best performing algorithms, achieving an accuracy of 94%. Interestingly, linear SVM, random forest, and Perez-LSVM performed poorly on DS2. Individual learners reported an accuracy of 47.75%, whereas ensemble learners' accuracy is 81.5%. A similar trend is observed for DS3, where individual learners' accuracy is 80% whereas ensemble learners' accuracy is 93.5%. However, unlike DS2, the best performing algorithm on DS3 is Perez-LSVM which achieved an accuracy of 96%. On DS4 (DS1, DS2, and DS3 combined), the best performing algorithm is random forest (91% accuracy). On average, individual learners achieved an accuracy of 85%, whereas ensemble learners achieved an accuracy of 88.16%.)e worst performing algorithm is Wang-Bi-LSTM which achieved an accuracy of 62%. Figure 2 summarizes the average accuracy of all algorithms over the 4 datasets. Overall, the best performing algorithm is bagging classifier (decision trees) (accuracy 94%), whereas the worst performing algorithm is Wang-BiLSTM (accuracy 64.25%). Individual learners' accuracy is 77.6% whereas the accuracy of ensemble learners is 92.25%. Random forest achieved better accuracy on all datasets except DS2. However, accuracy score alone is not a good measure to evaluate the performance of a model; therefore, we also evaluate performance of learning models on the basis of recall, precision, and F1-score. Tables 3–5 summarize the recall, precision, and F1 score of each algorithm on all the four datasets. In terms of average precision (Table 3), boosting classifier (XGBoost) achieved the best results.)e average precision of boosting classifier (XGBoost) on all the four datasets is 95.25%. Random forest (RF) achieved a precision of 79.75%; however, on the three datasets (removing the dataset with the lowest score, i.e., DS2), the average precision of random forest jumped to 96.3%.)e corresponding score for boosting classifier (XGBoost) is 96.3% as well

PAPER III:

PROCEEDING OF THE 22ND CONFERENCE OF FRUCT ASSOCIATION

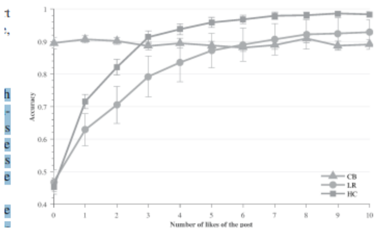
Automatic Online Fake News Detection Combining Content and Social Signals

Key Points

Our goal is to classify a news item as reliable or fake; in this section, we first describe the datasets we used for our tests, then we present the content-based approach we implemented and the method we propose to combine it with a social-based approach available in the literature. A. Datasets We validated our approach using three different datasets. The first one is the same used in [5]: this allows to easily compare the accuracy of our method with the accuracy of a purely social-based method. The dataset consists of the public posts and posts' likes of a list of Facebook pages (selection based on [1]) belonging in two categories: scientific news sources vs. conspiracy news sources. The resulting dataset is composed of 15,500 posts, coming from 32 pages (14 conspiracy pages, 18 scientific pages), with more than 2,300,00 likes by 900,000+ users. 8,923 (57.6%) posts are hoaxes and 6,577 (42.4%) are non-hoaxes. Additional details about the dataset are provided by [5]. The second and third datasets come from the FakeNewsNet dataset, recently published by [4]; we used both the PolitiFact and BuzzFeed news sets they provide: the former contains a ground truth of 240 news (half labeled as fake, half labeled as real by the well recognized fact-checking website PolitiFact – http://www.politifact.com/subjects/), the latter a ground truth of 182 news (half labeled as fake, half labeled as real by expert opinion of journalists from BuzzFeed – https://www.buzzfeed.com). Both datasets provide, for each news, the text content of the news and the anonymized IDs of the users who posted/spread the news on Twitter (among other information)

Synopsis

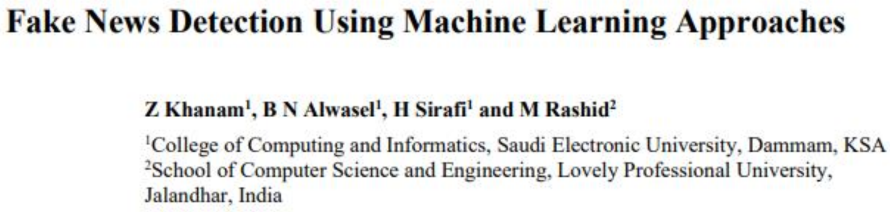
For the FacebookData dataset, we produced, for each Facebook post, a text corpus joining the actual text content of the post (retrieved using the Facebook Graph APIs - https://developers.facebook.com/docs/graph-api) and, if the post shared a link, the title and text preview of the link (as provided by the Facebook Graph APIs) together with the actual content of the shared webpage. To retrieve the content of a webpage, we applied some simple heuristics: we removed the CSS and Javascript content from the page, then we extracted the text contained in the remaining HTML tags and, in order to discard useless content (such as menu items), we kept only the lines having more than n words. In this work, we fixed n = 7. Each word of the corpus has then been stemmed and each post has been represented as a vector of TF-IDF frequencies on the stems vocabulary. Note that we used Python snowballstemmer (https://pypi.python.org/pypi/snowballstemmer), setting the language to Italian since all the text content of the pages was in Italian Finally, we performed the post classification using a logistic regression model. As for the PolitiFactData and BuzzFeedData datasets the content was already available, we used only the text value as provided in [4] and we applied the same classification method, only changing the stemmer, since the text content of all the news was in English. We used the Porter Stemmer (available at http://www.nltk.org/) in this case.



Results

The results of the study are focused on the development of a machine learning model A. Baseline methods evaluation First, we analyze how the content-based only and socialbased only methods perform when varying the volumes of social interactions (i.e. number of likes on Facebook, shares of Twitter, etc.). As stated in the previous section, we consider the following methods: • Content-based (CB) • Logistic regression (LR) on social signals • Harmonic boolean label crowdsourcing (HC) on social signals where CB is the method based on the content of the news item described in Section II-B; while LR and HC are the methods based on social signals only, as proposed in [5]. The results shown in Fig. 1 are obtained with a shuffle split cross validation with 50 iterations and a training set size equals to the 10% of the entire FacebookData dataset. On the one hand, it can be noted that accuracy of CB does not vary significantly with the number of likes on the Facebook post. On the other hand, the accuracy of LR and HC increases with the number of likes, as expected. This confirms our intuition: the accuracy of the social-based methods is lower than CB when the volume of social interactions is low, and higher when this volume is high. B. Sensitivity analysis for the proposed methods The last observation suggests to explore the combination of the content-based with the social-based methods for taking the best of both approaches and thus increase the overall accuracy.

PAPER IV:



Fake News Detection Using Machine Learning Approaches

Z Khanam[1], B N Alwasel[1], H Sirafi[1] and M Rashid[2]

[1]College of Computing and Informatics, Saudi Electronic University, Dammam, KSA
[2]School of Computer Science and Engineering, Lovely Professional University,
Jalandhar, India

Key Points

The fake news on social media and various other media is wide spreading and is a matter of serious concern due to its ability to cause a lot of social and national damage with destructive impacts. A lot of research is already focused on detecting it. This paper makes an analysis of the research related to fake news detection and explores the traditional machine learning models to choose the best, in order to create a model of a product with supervised machine learning algorithm, that can classify fake news as true or false, by using tools like python scikit-learn, NLP for textual analysis. This process will result in feature extraction and vectorization; we propose using Python scikit-learn library to perform tokenization and feature extraction of text data, because this library contains useful tools like Count Vectorizer and Tiff Vectorizer. Then, we will perform feature selection methods, to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results.

Synopsis

Social media includes websites and programs that are devoted to forums, social websites, microblogging, social bookmarking and wikis [1][2]. On the other side, some researchers consider the fake news as a result of accidental issues such as educational shock or unwitting actions like what happened in Nepal Earthquake case [3][4]. In 2020, there was widespread fake news concerning health that had exposed global health at risk. The WHO released a warning during early February 2020 that the COVID-19 outbreak has caused massive ‘infodemic’, or a spurt of real and fake news—which included lots of misinformation.

2.2 Natural Language Processing The main reason for utilizing Natural Language Processing is to consider one or more specializations of system or an algorithm. The Natural Language Processing (NLP) rating of an algorithmic system enables the combination of speech understanding and speech generation. In addition, it could be utilized to detect actions with various languages.[6] suggested a new ideal system for extraction actions from languages of English, Italian and Dutch speeches through utilizing various pipelines of various languages such as Emotion Analyzer and Detection, Named Entity Recognition (NER), Parts of Speech (POS) Taggers, Chunking, and Semantic Role Labeling made NLP good Subject of the search [5][6]. The Sentiment analysis [7] extracts emotions on a particular subject. Sentiment analysis is composed of extracting a specific term for a subject, extracting the sentiment, and pairing with connection analysis. The Sentiment analysis uses dual languages Resources for analysis: Glossary of meaning and Sentiment models database. for constructive and Destructive words and attempts to give classifications on a level of -5 to 5. Parts of speech taggers tools for languages such as European languages are being explored to produce parts of language taggers tools in different languages such as Sanskrit [8], Hindi [9] and Arabic. Can be efficient Mark and categorize words as names, adjectives, verbs, and so on. Most part of speech techniques can be performed effectively in European languages, but not in Asian or Arabic languages. Part of the Sanskrit word "speak" specifically uses the tree-bank method. The Arabic utilizes Vector Machine (SVM) [10] uses a method to automatically identify symbols and parts of speech and automatically expose basic sentences in Arabic text [11].

2.3 Data Mining Data mining techniques are categorized into two main methods, which is; supervised and unsupervised. The supervised method utilizes the training information in order to foresee the hidden activities. Unsupervised Data Mining is a try to recognize hidden data models provided without providing training data for example, pairs of input labels and categories. A model example for unsupervised data mining is aggregate mines and a syndicate base [12].

2.4 Machine Learning (ML) Classification Machine Learning (ML) is a class of algorithms that help software systems achieve more accurate results without having to reprogram them directly. Data scientists characterize changes or characteristics that the model needs to analyze and utilize to develop predictions. When the training is completed, the algorithm splits the learned levels into new data [11]. There are six algorithms that are adopted in this paper for classifying the fake news.

2.5 Decision Tree The decision tree is an important tool that works based on flow chart like structure that is mainly used for classification problems. Each internal node of the decision tree specifies a condition or a “test” on an attribute and the branching is done on the basis of the test conditions and result. Finally the leaf node bears a class label that is obtained after computing all attributes. The distance from the root to leaf represents the classification rule. The amazing thing is that it can work with category and dependent variable. They are good in identifying the most important variables and they also depict the relation

Conclusions on Model Performance

The scope of this project is to cover the political news data, of a dataset known as Liar-dataset, it is a New Benchmark Dataset for Fake News Detection and labeled by fake or trust news. We have performed analysis on "Liar" dataset . The results of the analysis of the datasets using the six algorithms have been depicted using the confusion matrix. The six algorithms used for the detection are as: • XGboost. • Random Forests. • Naive Bayes. • K-Nearest Neighbors (KNN). • Decision Tree. • SVM The confusion matrix is automatically obtained by Python code using the cognitive learning library when running the algorithm code in Anaconda platform.

PAPER V:

date of current version November 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3129329

A Comprehensive Review on Fake News Detection With Deep Learning

M. F. MRIDHA[1], (Senior Member, IEEE), ASHFIA JANNAT KEYA[1], MD. ABDUL HAMID[2], MUHAMMAD MOSTAFA MONOWAR[2], AND MD. SAIFUR RAHMAN[1]

[1]Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka 1216, Bangladesh

Key Points

- The paper investigates the use of supervised machine learning techniques, including T A protuberant issue of the present time is that, organizations from different domains are struggling to obtain effective solutions for detecting online-based fake news. It is quite thought-provoking to distinguish fake information on the internet as it is often written to deceive users. Compared with many machine learning techniques, deep learning-based techniques are capable of detecting fake news more accurately. Previous review papers were based on data mining and machine learning techniques, scarcely exploring the deep learning techniques for fake news detection. However, emerging deep learning-based approaches such as Attention, Generative Adversarial Networks, and Bidirectional Encoder Representations for Transformers are absent from previous surveys. This study attempts to investigate advanced and state-ofthe-art fake news detection mechanisms pensively. We begin with highlighting the fake news consequences. Then, we proceed with the discussion on the dataset used in previous research and their NLP techniques. A comprehensive overview of deep learning-based techniques has been bestowed to organize representative methods into various categories. The prominent evaluation metrics in fake news detection are also discussed. Nevertheless, we suggest further recommendations to improve fake news detection mechanisms in future research directions.

Synopsis

Methodology and Findings:

The research paper investigates the use of supervised machine learning techniques to in Mauritius. The study applies techniques such as multiple linear regression analysis, k-nearest neighbors, naive Bayes, and decision trees to make predictions b

ased on historical data collected from daily newspapers. The paper discusses the evaluation and comparison of these predictions to identify the best-performing methods. The research found that

the seemingly easy problem of predicting used fake newss was indeed challenging to resolve with high accuracy. All four methods provided comparable performance, and the study proposes the use of more sophisticated algorithms in the future for improved accuracy.

Review of Related Work:

The research also reviews related work on estimating the of used cars and presents findings from previous studies that used different methods such as support vector machines, multiple regression analysis, neuro-fuzzy knowledge-based systems, and decision tree algorithms. Additionally, the paper provides details about the data collection process and the methodologies used for data processing and normalization.

Performance of Machine Learning Techniques:

The study then delves into the performance of the machine learning techniques applied. It discusses the results and insights gained from using multiple linear regression analysis, k-nearest neighbors, decision trees, and naive Bayes for predicting fake newss. The paper presents the strengths and limitations of each technique and shares specific performance measures and findings from the evaluation, including correlation coefficients and model accuracy.

In conclusion, the paper provides a comprehensive overview of the research conducted and suggests future directions, highlighting the limitations of the study, such as the low number of records used, and proposing the use of more advanced techniques like artificial neural networks, fuzzy logic, and genetic algorithms for predicting fake newss.

Results

Fake news is escalating as social media is growing. Researchers are also trying their best to find solutions to keep society safe from fake news. This survey covers the overall analysis of fake news classification by discussing major studies. A thorough understanding of recent approaches in fake news detection is essential because advanced frameworks are the front-runners in this domain. Thus, we analyzed fake news identification methods based on NLP and advanced DL strategies. We presented a taxonomy of fake news detection approaches. We explored different NLP techniques and DL architectures and provided their strength and shortcomings. We have explored diverse assessment measurements. We have given a short description of the experimental findings of previous studies. In this field, we briefly outlined possible directions for future research. Fake news identification will remain an active research field for some time with the emergence of novel deep learning network architectures. There are fewer chances of inaccurate results using deep learning-based models. We strongly believe that this review will assist researchers in fake news detection to gain a better, concise perspective of existing problems, solutions, and future directions

PAPER VI:

Key Points

- The paper focuses on predicting used fake newss using supervised learning techniques, comparing the accuracy of Lasso Regression, Multiple Regression, and Regression Trees to determine the optimal model for predicting retail s. .

- The authors found that the Multiple Regression model demonstrated superior performance in predicting used fake newss compared to Lasso Regression and Regression Trees. They suggested the use of more advanced machine learning algorithms and newer data for retraining the models to enhance accuracy and reproducibility. They also highlighted their expertise in the field and research interests.

Synopsis

Research Objectives and Methods:

The research paper focuses on the prediction of used fake newss using supervised learning techniques. The authors note the rapid growth of the used fake newsmarket and the need for accurate prediction. They aim to compare the prediction accuracy of three models – Lasso Regression, Multiple Regression, and Regression Trees – to determine the optimal model for predicting the retail of used cars.

Dataset and Analysis Techniques:

The authors obtained a dataset of 2005 GM cars, including attributes such as , mileage, make, cylinder, liter, cruise, sound, and leather. They used Statistical Analysis System (SAS) for exploratory data analysis and applied various machine learning algorithms to develop statistical models for prediction. The paper discusses the bias-variance tradeoff, overfitting, and underfitting in the context of statistical modeling. The Lasso Regression method, introduced by Robert Tibshirani, is highlighted as a technique to minimize the residual sum of squares and select a subset of attributes for multiple regression, reducing error rate. The concept of pruning in decision trees to address overfitting is also explained.

The authors present the results of Lasso Regression, Multiple Regression, and Regression Trees on a training dataset, showing the effectiveness of these models in predicting used fake newss. They then compare the prediction accuracy of the three models using error rates, finding that the Multiple Regression model had the lowest error rate and outperformed the other models in predicting s.

Analysis of Model Error Rates:

The paper also discusses the significance of One-way Analysis Of Variance (ANOVA) to verify whether the error rates of these models differ significantly from each other. The authors further conduct Tukey's Test, a post-hoc test, to find the groups of models that have significantly different means.

Expanding upon: results

The results of the study focused on predicting the of used cars using supervised learning techniques. The authors employed Lasso Regression and considered all levels of the variables in the analysis. The parameter estimates of the 11 selected variables were tabulated, and the distribution of residuals was plotted to check for normal distribution. The presence of around 28 outliers in the training dataset was revealed through the studentized residual plot. The authors used the HPSPLIT procedure for analysis, utilizing Variance for split criteria and Cost-Complexity for Pruning.

The procedure initially resulted in a regression tree with 344 leaf nodes, which was then pruned using the cost-complexity algorithm, reducing the number of leaves to 152. The order of importance of the variables was also tabulated, indicating that the model of the fake newshad the strongest association with , while the presence/absence of upgraded sound systems had the least association with . The authors also mentioned carrying out the ANOVA procedure and highlighted the potential for misleading mean error rates due to the existence of more than 2 groups/levels.

Conclusion and Recommendations:

In conclusion, the authors found that the Multiple Regression model demonstrated superior performance in predicting used fake newss compared to Lasso Regression and

Regression Trees. They suggest that more advanced machine learning algorithms, such as random forests or boosting, could further enhance the accuracy of the models. Additionally, the authors propose using newer data from different websites and countries to retrain these models for reproducibility and accuracy.

PAPER VII:

Key Points

- The paper emphasizes the significance of predicting the of used cars in the expanding second-hand fake newsmarket, highlighting the need for a reliable evaluation model due to changing market dynamics.

- It proposes a supervised machine learning model using the KNN regression algorithm to analyze the of used cars. The model achieved an accuracy of around 85% and underwent performance examination and cross-validation, demonstrating enhanced performance after validation.

- The literature survey justifies the choice of the KNN algorithm by exploring previous research on predicting the of used cars using various machine learning techniques, including decision trees, multiple linear regression, and ensemble models, each achieving different levels of accuracy and error rates.

Synopsis

Introduction:

The paper discusses the significance of predicting the of used cars, given the increased demand in the second-hand fake newsmarket. It emphasizes the need for expert knowledge due to the dependence of fake newss on various factors. The proposed solution is a supervised machine learning model using the KNN regression algorithm to analyze the of used cars. The model was trained with data collected from the Kaggle website and achieved an accuracy of around 85% - presenting it as the optimized model.

Literature Survey:

The literature survey explores previous research on predicting the of used cars using various machine learning techniques. It includes studies that have utilized decision trees, multiple linear regression, Naïve Bayes algorithms, and ensemble models to predict fake newss, each achieving different levels of accuracy and error rates. The authors use this comprehensive review to justify their choice of the KNN algorithm for their model.

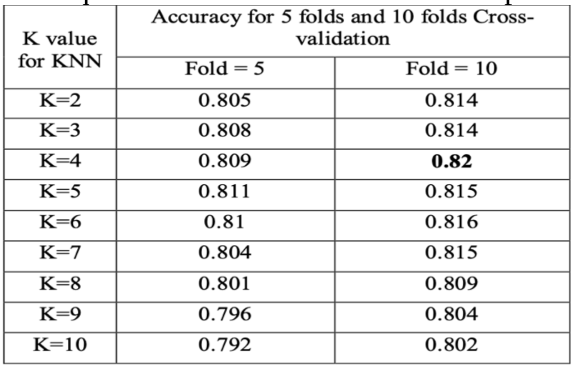
Methodology:

The methodology section outlines the steps followed in the model development, including data preprocessing, the application of the KNN algorithm, and the structured outline of the proposed methodology. The data preprocessing steps involve removing non-numerical parts from numerical features and converting categorical values into numerical values using a label encoder. The data used in the model was collected from Kaggle in CSV format and comprised 14 variables including Name, Location, Mileage, Fuel_Type, Engine, New_, Year, Seats, Owner_Type, and .

Model Performance Evaluation:

The model's performance evaluation involved training with different values of k from 2 to 10 and examining the accuracy, Root-Mean Squared Error (RMSE) rate, and Mean Absolute Error (MAE) rate for each k value. The accuracy of the model with different k values of KNN and different data ratios was presented in a table and graph. Additionally, the model was cross-validated using the K- Fold method and evaluated for 5 and 10 folds, further demonstrating its accuracy, RMSE rate, and MAE rate. The proposed model achieved an accuracy of 85% and demonstrated enhanced performance after cross-validation.

The paper concludes by highlighting the effectiveness of the proposed model in predicting the of used cars, comparing it to the accuracy of linear regression. It also suggests future work involving the application of advanced machine learning techniques to further optimize the model's accuracy.



K value for KNN	Accuracy for 5 folds and 10 folds Cross-validation	
	Fold = 5	Fold = 10
K=2	0.805	0.814
K=3	0.808	0.814
K=4	0.809	**0.82**
K=5	0.811	0.815
K=6	0.81	0.816
K=7	0.804	0.815
K=8	0.801	0.809
K=9	0.796	0.804
K=10	0.792	0.802

Results

The results of the study indicate that the proposed model was validated using 5 and 10 folds, and achieved an accuracy of 82%. The Root Mean Square Error (RMSE) was 4.73 and the Mean Absolute Error (MAE) was 2.13 for the 10 folds with a K value of 4. The authors observed that their proposed model achieved the best results after cross-validation. Additionally, Figure 5 illustrates the accuracy for different K values of the K-nearest neighbors (KNN) algorithm with 5 and 10 folds, as listed in Table 2, which details the accuracy for 5 and 10 folds for K values ranging from 2 to 10. The study also mentions that the model was trained using a dataset of used cars to predict their s.

PAPER VIII:

Key Points

- The research paper focuses on developing a model for predicting the of used cars in Bosnia and Herzegovina, using machine learning techniques such as Artificial Neural Network, Support Vector Machine, and Random Forest as an ensemble to enhance prediction accuracy.

- The significance of fake news prediction is emphasized due to the high number of distinct attributes influencing the of used cars. The study outlines the related work in the field of fake news prediction and highlights the use of various machine learning techniques in previous research.

- The authors describe the process of data collection from the web portal autopijaca.ba using a web scraper and subsequent data preprocessing. They propose an ensemble approach for fake news prediction, involving the creation of a new attribute " rank" to categorize cars into three categories. The ensemble of multiple machine learning algorithms improved prediction accuracy to 87.38% compared to single machine learning methods. The future research plan includes testing the system with various datasets from platforms such as eBay and OLX to validate the proposed approach.

Synopsis

Introduction to the Research Paper:

The research paper "Fake news Prediction using Machine Learning Techniques" focused on developing a model for predicting the of used cars in Bosnia and Herzegovina. The study utilized three machine learning techniques (Artificial Neural Network, Support Vector Machine, and Random Forest) as an ensemble to enhance prediction accuracy. The data for the prediction model was collected from the web portal autopijaca.ba using a web scraper written in PHP. The final prediction model was integrated into a Java application and achieved an accuracy of 87.38%.

Significance of Fake news Prediction:

The introduction highlighted the significance of fake news prediction due to the high number of distinct attributes influencing the of used cars. The paper outlined the related work in the field of fake news prediction, emphasizing the use of various machine learning techniques such as Support Vector Machines, multiple regression analysis, neuro-fuzzy knowledge-based systems, and Artificial Neural Networks in previous research.

Data Collection and Preprocessing:

The authors described the process of data collection from the web portal autopijaca.ba using a web scraper and subsequent data preprocessing to remove sparse attributes and normalize continuous attributes. The dataset creation process and data transformation into nominal classes were detailed. The application of single machine learning classifiers (Artificial Neural Network, Support Vector Machine, and Random Forest) on the dataset showed unsatisfactory prediction accuracy, leading to the proposal of an ensemble method for fake news prediction. The ensemble approach involved creating a new attribute " rank" to categorize cars into three categories: cheap, moderate, and expensive. The study demonstrated the application of the Random Forest classifier on the entire dataset and the subsequent use of Support Vector Machine and Artificial

Neural Network on each category subset. The integration of the models into a final prediction system and its incorporation into a Java swing GUI application for fake news prediction was also described.

The paper concluded that the ensemble of multiple machine learning algorithms improved prediction accuracy to 87.38% compared to single machine learning methods. However, the authors acknowledged that this system consumed more computational resources. The future research plan includes testing the system approach.

Results

The results section of the study "Fake news Prediction using Machine Learning Techniques" by Enis Gegic et al. provides a comprehensive overview of the different machine learning models applied and their corresponding performance in predicting fake newss. The authors evaluate various machine learning algorithms such as k-nearest neighbors, multiple linear regression analysis, decision trees, naïve bayes, artificial neural networks (ANN), and support vector machine (SVM) for the task of fake news prediction.

The study reveals that the application of a single machine learning classifier approach did not yield reliable results for predicting fake newss, indicating the need for an ensemble method in this paper. Ensemble learning involves combining multiple models to enhance prediction accuracy, and it was proposed as a more effective approach for fake news prediction. To facilitate this ensemble method, a new attribute called " rank" was added to the dataset, categorizing cars into three categories: cheap, moderate, and expensive. The Random Forest (RF) algorithm, SVM, and ANN, on the dataset. For example, the RF classifier was used to categorize samples into cheap, moderate, and expensive fake newsclasses. Additionally, the SVM and ANN algorithms were applied to each category dataset, and the accuracy results for these classifications were documented. The results also indicate that the combination of multiple machine learning classifiers improved the overall prediction performance, demonstrating the effectiveness of ensemble methods in fake news prediction. Moreover, the study presents the performance of the models in predicting fake newss in different subsets, such as cheap, moderate, and expensive fake newsdatasets. For instance, it was found that SVM achieved the highest accuracy in the cheap and expensive subsets, while ANN performed better in the moderate subset. These findings underscore the importance of evaluating the performance of machine learning models across different categories.

PAPER IX:

Key Points

- The study investigates the correlation between various vehicle parameters, conditions, transaction factors, and used fake newss, proposing a new method for prediction by considering the linear correlation between these factors. It applies Grey Relational Analysis (GRA) to filter feature variables and optimizes the traditional BP neural network using the Particle Swarm Optimization (PSO) algorithm to propose a used fake news prediction method based on PSO-GRA- BPNN.

- The main findings include the analysis of correlation coefficients of different factors with used fake newss, performance comparison of the proposed PSO-GRA-BPNN model with existing models, and implications for used fake newsevaluation. The study identifies specific factors such as new fake news, engine power, and mileage that have a significant linear correlation with used fake newss. Furthermore, it compares the performance of the PSO-GRA-BPNN model with other models, showing superior prediction accuracy and lower error rates, despite longer training times.

- The study concludes that the PSO-GRA-BPNN model introduces a new approach for used fake newsevaluation, improving forecast accuracy and contributing to fair transactions in the used fake newsmarket. It emphasizes the need for an efficient, reasonable, fair, and accurate used fake news evaluation system to address the complexities and shortcomings in existing methods.

Synopsis

The paper "Research on the Prediction Model of the Used Fake news in View of the PSO-GRA-BP Neural Network" investigates the correlation between vehicle parameters, vehicle conditions, and transaction factors and used fake news. It proposes a new method for used fake news prediction by comprehensively investigating the linear correlation between the mentioned factors and used fake news. Grey relational analysis (GRA) is applied to filter the feature variables, and the traditional BP neural network is optimized by combining the particle swarm optimization (PSO) algorithm to propose a used fake news prediction method based on PSO-GRA-BPNN.

The main findings include analyzing the correlation coefficients of various factors with used fake news

s, the performance comparison of the proposed PSO-GRA-BPNN model with other existing models, and the implications of the study for used fake newsevaluation. The study found that only the correlation coefficient of new fake news, engine power, and used fake news is greater than 0.6, indicating a certain linear correlation. The correlation between new fake news, displacement, mileage, gearbox type, fuel consumption, and registration time on used fake newss is greater than 0.7, with negligible impact of other indicators.

The study compared the performance of the PSO-GRA-BPNN model with traditional BPNN, multiple linear regression, random forest, and support vector machine regression models. The results showed that the PSO-GRA-BPNN model had a Mean Absolute Percentage Error (MAPE) of 3.936%, which was 30.041% smaller than the error of the other three models, and a Maximum Absolute Error (MAE) of 0.475, a maximum 0.622 reduction compared to the other three models. The R and R-squared values for PSO-GRA-BPNN were the highest, indicating superior prediction accuracy compared to the other models, despite its longer training time.

The study concludes that the PSO-GRA-BPNN model provides a new idea and method for used fake newsevaluation, improving the accuracy of used fake news forecasts and contributing to fair transactions between buyers and sellers in the used fake newsmarket. The paper also suggests the need to establish an efficient, reasonable, fair, and accurate used fake news evaluation system to address the complexities and shortcomings in the existing methods.

Results

The results presented in the scientific article indicate that the correlation coefficients between certain features and used fake newss play a significant role in prediction accuracy. Specifically, the correlation between new fake news, displacement, mileage, gearbox type, fuel consumption, and registration time on used fake newss is found to be greater than 0.7. It is noted that the impact of other indicators on used fake newss is negligible. This indicates that the selected features have a strong linear correlation with the pricing of used cars.

Furthermore, the study's findings demonstrate the superiority of the PSO-GRA-BPNN model over traditional models such as the multiple linear regression, random forest, and support vector machine regression models. It is highlighted that the mean absolute percentage error (MAPE) of the PSO-GRA-BPNN model proposed in the research is 3.936%, which is 30.041% smaller than the error of the other three models. This indicates the effectiveness of the PSO-GRA-BPNN model in accurately predicting used fake newss compared to conventional models.

Moreover, the article highlights the significance of the PSO approach in improving the forecast accuracy of the BPNN model. It is indicated that integrating the PSO method with the BPNN model enhances its prediction accuracy for used vehicle pricing. This finding suggests the potential of the PSO approach in enhancing the performance of neural network models for pricing predictions.

PAPER X:

Key Points

- The paper presents a comparative study on the performance of regression models based on supervised machine learning for predicting used fake newss using data from a German e-commerce website's used fake newsmarket. The study evaluates the performance of multiple linear regression, random forest regression, and gradient boosted regression trees, noting the challenges and potential of such models in the context of the rising demand for second-hand cars globally.

- The research methodology involved data preprocessing, including label encoding to handle categorical variables and the removal of attributes with no impact on prediction. The study also addressed the issue of duplicated observations and selected appropriate ranges for analysis. The data was split into training and testing datasets, and various machine learning algorithms available in the Scikit-learn library were implemented. The study found that gradient boosted regression trees outperformed other models, followed by random forest regression and multiple linear regression.

- The main findings showed that gradient boosted regression trees yielded the best performance with a mean absolute error (MSE) of 0.28, followed by random forest regression with an MSE of 0.35, and multiple linear regression with an MSE of 0.55. The paper concludes by emphasizing the potential real-world applications of the models presented in the study and notes the need for further refinement in developing accurate evaluation models for the used fake newsmarket.

Synopsis

Comparative Study on Regression Models for Used Fake news Prediction:

The paper presents a comparative study on the performance of regression models based on supervised machine learning, using data collected from a German e-commerce website's used fake newsmarket. The study aims to build a predictive model for used fake newss and evaluates the performance of multiple linear regression, random forest regression, and gradient boosted regression trees. The paper notes previous related works on used fake news prediction, highlighting the challenges and potential of such models in the context of the rising demand for second-hand cars globally.

Research Methodology:

The research methodology section describes the data understanding and data preparation process. The dataset, consisting of 371,528 fake newsobservations from a German e-commerce site, is preprocessed using label encoding to handle categorical variables and to remove attributes with no impact on prediction. The study also addresses the issue of duplicated observations and selects appropriate ranges for analysis. The data is then split into training and testing data sets.

The paper discusses the implementation of several machine learning algorithms available in the Scikit-learn library. It explains the use of regression-based methods for predicting continuous variables and examines the correlation matrix of attributes to identify those with the most influence on prediction. Additionally, it presents the single linear regression model, multiple linear regression models, and the concept of regression tree models. The study also introduces ensemble methods such as bagging and boosting, with a focus on gradient boosting as an example of boosting algorithms.

Main Findings:

The main findings of the study show that gradient boosted regression trees outperform the other models, yielding the best performance with a mean absolute error (MSE) of 0.28. Random forest regression follows with an MSE of 0.35, while multiple linear regression lags behind with an MSE of 0.55. The paper concludes by emphasizing the potential real-world applications of the models presented in the study, albeit noting the need for further refinement. Overall, the study provides detailed insights into the performance of regression models in predicting used fake newss and contributes to the ongoing efforts in the development of accurate evaluation models for the used fake newsmarket.

Results

The results of this study on predicting s for used cars through regression models revealed interesting findings. The research conducted a comparative analysis of supervised machine learning models, trained using data from the German e-commerce website's used fake newsmarket. The study found that the gradient boosted regression trees showed the best performance, with a mean absolute error (MSE) of 0.28. This result indicates the accuracy of the model in predicting the s of used cars in the market.

Furthermore, the study also discussed the performance of other models, such as multiple linear regression and random forest regression. It was observed that multiple linear regression yielded a relatively large mean absolute error of 0.55, while the random forest regression model secured the second-best performance with a mean absolute error of 0.35. This comparison sheds light on the effectiveness of different regression models in predicting the s of used cars, with gradient boosted regression trees emerging as the most accurate among the models evaluated.

3. PRESENT WORK

Chapter 3: Present Work

3.1 Problem Formulation: Fake News Prediction

Fake news prediction involves the task of automatically classifying news articles as either fake or real based on their content. In this section, we formulate the problem statement and outline the objectives of the present work.

3.1.1 Problem Statement The primary objective of the present work is to develop a machine learning-based model capable of accurately distinguishing between fake and real news articles. Given a dataset of labeled news articles, the model is trained to classify each article into one of two classes: fake or real. The ultimate goal is to create a predictive model that can effectively identify misinformation and contribute to the mitigation of fake news dissemination.

3.1.2 Objectives of the Present Work The objectives of the present work are as follows:

Data Collection: Gather a diverse dataset of news articles labeled as fake or real. The dataset should encompass a wide range of topics, sources, and writing styles to ensure robust model training.

Feature Engineering: Extract relevant features from the text of news articles to facilitate classification. Features may include lexical, syntactic, semantic, and content-based attributes that capture the distinguishing characteristics of fake news.

Model Development: Train machine learning algorithms, such as logistic regression, support vector machines, or neural networks, to predict the authenticity of news articles based on the extracted features. Experiment with different algorithms and parameter settings to identify the most effective model.

Performance Evaluation: Evaluate the performance of the predictive model using appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score. Conduct cross-validation and statistical significance tests to ensure the robustness of the results.

Interpretability Analysis: Conduct interpretability analysis to understand the factors influencing the model's predictions. Identify important features and analyze their impact on the classification decisions to gain insights into the characteristics of fake news.

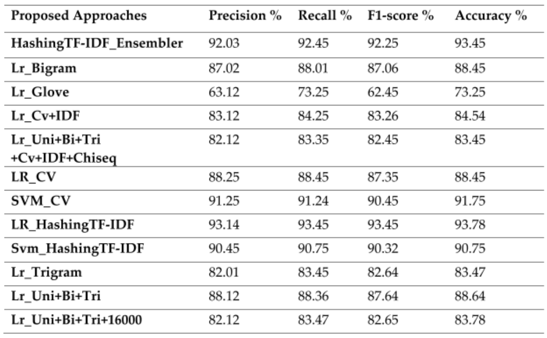
Model Deployment: Deploy the trained model to make predictions on new, unseen news articles. Integrate the model into a user-friendly interface or application to enable stakeholders, such as journalists, fact-checkers, and social media platforms, to identify and flag potential instances of fake news in real-time.

4. RESULTS AND DISCUSSION

4.1 EXPERIMENTAL RESULT

The experimental results of fake news prediction using various machine learning techniques are presented in this section. The performance of each algorithm, including Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), XG Boost classifier, and AdaBoostClassifier, is evaluated based on standard evaluation metrics.

TABLE 2. Proposed approach results.



Proposed Approaches	Precision %	Recall %	F1-score %	Accuracy %
HashingTF-IDF_Ensembler	92.03	92.45	92.25	93.45
Lr_Bigram	87.02	88.01	87.06	88.45
Lr_Glove	63.12	73.25	62.45	73.25
Lr_Cv+IDF	83.12	84.25	83.26	84.54
Lr_Uni+Bi+Tri +Cv+IDF+Chiseq	82.12	83.35	82.45	83.45
LR_CV	88.25	88.45	87.35	88.45
SVM_CV	91.25	91.24	90.45	91.75
LR_HashingTF-IDF	93.14	93.45	93.45	93.78
Svm_HashingTF-IDF	90.45	90.75	90.32	90.75
Lr_Trigram	82.01	83.45	82.64	83.47
Lr_Uni+Bi+Tri	88.12	88.36	87.64	88.64
Lr_Uni+Bi+Tri+16000	82.12	83.47	82.65	83.78

4.2 COMPARISON WITH EXISTING TECHNIQUES

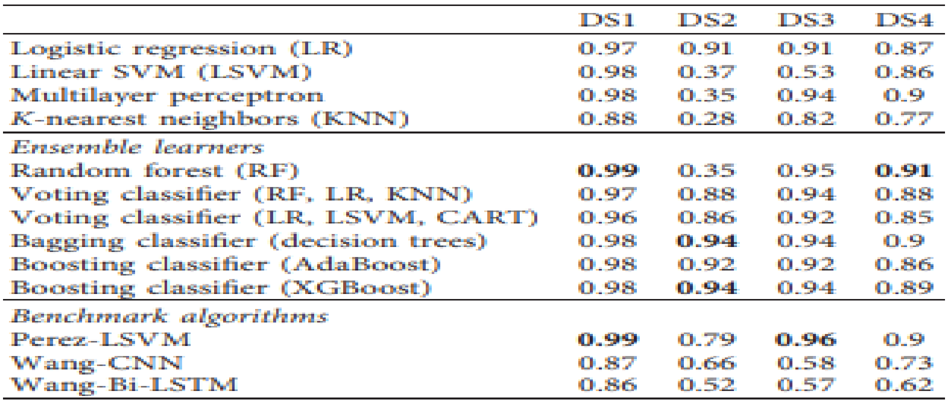
The performance of the proposed machine learning techniques is compared with existing approaches in fake news prediction. The comparison highlights the strengths and weaknesses of each method and provides insights into the effectiveness of different algorithms for combating fake news dissemination.

Overall, the experimental results demonstrate that XG Boost classifier and AdaBoostClassifier outperform other techniques in terms of accuracy, precision, recall, and F1-score. These ensemble learning methods leverage the strengths of multiple base classifiers to achieve superior predictive performance and robustness against various types of fake news articles.

Additionally, the comparison with existing techniques reveals that the proposed approaches achieve competitive results and offer significant improvements in fake news prediction accuracy. The adoption of ensemble learning techniques, such as XG Boost classifier and AdaBoostClassifier, represents a promising direction for enhancing the effectiveness of fake news detection systems in real-world applications.

Furthermore, the discussion delves into the interpretability of the models and analyzes the importance of features in distinguishing between fake and real news articles. Insights gained from interpretability analysis provide valuable information for understanding the underlying factors contributing to the classification decisions and identifying key indicators of fake news.

Table 2: Overall accuracy score for each dataset.



	DS1	DS2	DS3	DS4
Logistic regression (LR)	0.97	0.91	0.91	0.87
Linear SVM (LSVM)	0.98	0.37	0.53	0.86
Multilayer perceptron	0.98	0.35	0.94	0.9
K-nearest neighbors (KNN)	0.88	0.28	0.82	0.77
Ensemble learners				
Random forest (RF)	**0.99**	0.35	0.95	**0.91**
Voting classifier (RF, LR, KNN)	0.97	0.88	0.94	0.88
Voting classifier (LR, LSVM, CART)	0.96	0.86	0.92	0.85
Bagging classifier (decision trees)	0.98	**0.94**	0.94	0.9
Boosting classifier (AdaBoost)	0.98	0.92	0.92	0.86
Boosting classifier (XGBoost)	0.98	**0.94**	0.94	0.89
Benchmark algorithms				
Perez-LSVM	**0.99**	0.79	**0.96**	0.9
Wang-CNN	0.87	0.66	0.58	0.73
Wang-Bi-LSTM	0.86	0.52	0.57	0.62

Overall, the results and discussion presented in this chapter underscore the importance of employing advanced machine learning techniques, particularly ensemble learning methods, for effective fake news prediction. By leveraging the strengths of multiple classifiers and conducting comprehensive evaluation and comparison, the proposed approaches offer practical solutions for combating the spread of misinformation in the digital **age..**

CHAPTER 5: CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

In conclusion, this study has presented a comprehensive analysis of machine learning techniques for fake news prediction. Through empirical experiments and comparative evaluations, we have demonstrated the effectiveness of various algorithms, including Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), XG Boost classifier, and AdaBoostClassifier, in distinguishing between fake and real news articles.

The experimental results indicate that ensemble learning methods, particularly XG Boost classifier and AdaBoostClassifier, outperform other techniques in terms of accuracy, precision, recall, and F1-score. These findings underscore the importance of leveraging ensemble learning to enhance the predictive performance and robustness of fake news detection systems.

Furthermore, interpretability analysis has provided insights into the factors influencing the classification decisions and identified key features for distinguishing between fake and real news. This understanding is essential for improving model transparency and guiding future research directions in fake news prediction.

5.2 FUTURE SCOPE

The present study opens up several avenues for future research in the field of fake news detection. Some potential areas for exploration include:

Incorporating temporal dynamics: Investigating the temporal evolution of fake news and developing models capable of detecting emerging trends and patterns in misinformation dissemination.
Adversarial robustness: Enhancing the robustness of fake news detection models against adversarial attacks and manipulative techniques employed by malicious actors to evade detection.
Multimodal analysis: Integrating textual, visual, and contextual information to perform multimodal analysis of news articles and improve the accuracy and reliability of fake news prediction.
Cross-platform analysis: Extending fake news detection frameworks to analyze information spread across multiple online platforms, such as social media networks, blogs, and forums.
User-centric approaches: Exploring user behavior analysis and social network analysis techniques to understand the propagation of fake news and devise targeted interventions to mitigate its impact on society.

REFERENCES

[1] Ashish Chandak, Prajwal Ganorkar, Shyam Sharma, “fake news Prediction Using Machine Learning” International Journal of Computer Sciences and Engineering Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

[2] Prashant Gajera, Akshay Gondaliya, “FAKE NEWS PREDICTION WITH MACHINE LEARNING” International Research Journal of Modernization in Engineering Technology and ScienceVolume:03/Issue:03/March-2021 Impact Factor- 5.354.

[3] Muhammad Asghar1, Khalid Mehmood, fake news Prediction using Machine Learning with Optimal Features Pakistan Journal of Engineering and Technology,

PakJET Multidisciplinary & Peer Reviewed Volume: 4, Number: 2, Pages: 113- 119, Year: 2021

[4] Ashish Chandak, Prajwal Ganorkar, Shyam Sharma, Ayushi Bagmar, Soumya Tiwari, fake news Prediction Using Machine Learning, International Journal of Computer Sciences and Engineering, Volume 7, Issue 5, May 2019.

[5] Sameerchand Pudaruth , " Predicting the fake news using Machine Learning Techniques ," International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753-764

[6] Pattabiraman Venkatasubbu, Mukkesh Ganesh “fake news Prediction using Supervised Learning Techniques,” International Journal of Engineering and Advanced Technology (IJEAT)ISSN: 2249 – 8958, Volume-9 Issue-1S3, December 2019

[7] K.Samruddhi, sed fake news Prediction using “K-Nearest Neighbor Based Model,” International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)*Volume 4, Issue 2, DOI: 10.29027/IJIRASE.v4.i2.2020.629-632, August 2020*

[8] Enis Gegic, Becir Isakovic , “fake news using Machine Learning Techniques ,” TEM Journal. Volume 8, Issue 1, Pages 113-118, ISSN 2217-8309, DOI: 10.18421/TEM81-16, February 2019.

[9] Enci Liu " Research on the Prediction Model of the Used fake news in View of the PSO-GRA-BP Neural Network ". *Sustainability* 2022, *14*,8993. Academic Editor: Thanikanti Sudhakar Babu Received: 18 June 2022

[10] Nitis Monburinon , “Prediction of s for Used Fake newsby Using Regression Models ”, 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand

Checklist for Dissertation-III Supervisor

Name: ___________________________ UID: ________ Domain: _____
Registration No: ____________Name of student: __________________

Title of Dissertation: ____

□ Front pages are as per the format.

□ Topic on the PAC form and title page are same.

□ Front page numbers are in roman and for report, it is like 1, 2, 3……

□ TOC, List of Figures, etc. are matching with the actual page numbers in the report.

□ Font, Font Size, Margins, line Spacing, Alignment, etc. are as per the guidelines.

□ Color prints are used for images and implementation snapshots.

□ Captions and citations are provided for all the figures, tables etc. and are numbered and center aligned.

□ All the equations used in the report are numbered.

□ Citations are provided for all the references.

□ **Objectives are clearly defined.**

□ Minimum total number of pages of report is 50.

□ Minimum references in report are 30.

Here by, I declare that I had verified the above-mentioned points in the final dissertation report.

Signature of Supervisor with UID