No.1/2023, JANUARY-MARCH
https://doi.org/10.53477/2284-9378-23-10

BULLETIN
OF "CAROL I" NATIONAL DEFENCE UNIVERSITY

https://buletinul.unap.ro/index.php/en/

A comparison of artificial intelligence models used for fake news detection

Ștefan Emil REPEDE, Ph.D. Student*
Remus BRAD, Ph.D.**
*"Lucian Blaga" University, Sibiu, Romania
e-mail:stefan.repede@ulbsibiu.ro
**"Lucian Blaga" University, Sibiu, Romania
e-mail: remus.brad@ulbsibiu.ro

Abstract

This article aims to compare current state-of-the-art natural language processing models (NLP) fine-tuned for fake news detection based on a set of metrics and asses their effectiveness as a part of a disinformation management structure. The need for a development of this area comes as a response to the overwhelming and unregulated spread of fake news that represents one of the current major difficulties in today`s era. The development of AI technologies has a direct impact over the creation and spreading of misinformation and disinformation as a result of the multiple uses that technology may have. Currently, machine learning techniques are used for the development of large language models (LLM). These developments in science are also used in disinformation campaigns. Related to this matter the concept of disinformation management has arisen as a cybersecurity issue integral in the current cyber threat landscape.

Keywords:

fake news; misinformation; disinformation management; natural language processing; NLP; artificial intelligence; machine learning; cybersecurity.

1. Introduction

Fake news is a serious problem that can mislead and manipulate people into believing false or biased narratives. Fake news is a term that refers to false or misleading information that is presented as factual news. Fake news can have serious consequences for society, such as influencing public opinion, spreading misinformation, and undermining trust in journalism. Therefore, it is important to develop effective methods for detecting and combating fake news. The automatic detection of fake news is a challenging task that requires advanced artificial intelligence (AI) methods to analyze the content and source of news articles. In this research article, we compare different AI methods applied for fake news automatic detection. We use various metrics such as accuracy, precision, recall and F1-score to evaluate the performance of different methods. We review supervised learning techniques such as support vector machines, naive Bayes and decision trees that use labeled data to classify news articles as fake or real. We also discuss deep learning and transformer models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) and BERT that can capture complex features and semantic relations from text data. Furthermore, we explore unsupervised learning techniques such as clustering, topic modeling and anomaly detection that can identify fake news without prior knowledge or labels. Moreover, we examine some rule-based systems that use predefined rules or heuristics to detect fake news based on linguistic or stylistic features. Finally, we present hybrid models that combine different AI methods to achieve better results in fake news detection. The datasets used for testing and validation will focus on the ISOT fake news dataset (University of Victoria 2017) or similar ones. We also discuss the strengths and limitations of each class and provide suggestions for future research directions and usage.

2. A comparison of AI methods applied for Fake News automatic detection

Various AI methods have been used for binary classification tasks concerning the automatic detection of fake news (Kaliyar, Goswami and Narang 2021c). The performance evaluation for fake news datasets is measured using established metrics (Ozbay and Alatas 2020) that enable researchers to compare the performance of different models and identify which methods are most effective at detecting fake news.

3. Metrics used for evaluation

Accuracy, precision, recall, and F1-score are the metrics commonly used to evaluate the performance of classification models (Liu, Ott, et al., RoBERTa: A Robustly Optimized BERT Pretraining Approach 2019), and are described as follows:

3.1 Accuracy – It measures the proportion of correctly classified instances out of the total number of instances. It is calculated as (True Positives + True Negatives) / Total Instances. It is a useful metric when the classes are balanced (Kaliyar, Goswami and Narang 2021c).

3.2 Precision – It measures the proportion of true positives among all instances classified as positive. It is calculated as True Positives / (True Positives + False Positives). Precision measures how accurate the positive predictions are and how often the model correctly identifies true positive instances (Devlin, et al. 2019).

3.3 Recall – It measures the proportion of true positives among all instances that are actually positive. It is calculated as True Positives / (True Positives + False Negatives). Recall measures how well the model can find all the positive instances in the dataset date (Kaliyar, Goswami and Narang 2021a).

3.4 F1-score – It is the harmonic mean of precision and recall, providing a balanced measure between the two metrics. It is calculated as 2 * (Precision * Recall) / (Precision + Recall). The F1-score is a good overall measure of the model's performance, especially when the dataset is imbalanced. In binary classification, true positives (TP) are the number of instances that are correctly classified as positive, false positives (FP) are the number of instances that are incorrectly classified as positive, true negatives (TN) are the number of instances that are correctly classified as negative, and false negatives (FN) are the number of instances that are incorrectly classified as negative (Ozbay and Alatas 2020).
The current state-of-the-art AI methods used in researching this topic may be split into the following categories:

4. Supervised learning techniques used in fake news detection

Supervised learning is an AI method that has been used for fake news classification in different ways. It involves training a machine learning model on a labeled dataset of news articles that are classified as either real or fake. The model learns to identify patterns in the data and generalize to new, unseen examples. The choice of machine learning algorithm depends on the specific characteristics of the dataset, but popular algorithms include methods such as logistic regression, support vector machines, and random forests (Ozbay and Alatas 2020). The following methods are considered for comparison:

4.1 BayesNet – Bayesian network, also known as a Bayes net, is a probabilistic graphical model used for reasoning under uncertainty. It is named after the Reverend Thomas Bayes, an 18th-century British statistician who developed the Bayes theorem. In a Bayesian network (Langley, Wayne and Thompson 1992, 223-228), variables are represented by nodes, and the relationships between them are

represented by directed edges. The nodes can represent either observable or hidden variables, and the edges represent conditional dependencies between them.

4.2 JRip – Jumping Rule-based Information Processing was developed by W. W. Cohen (Cohen 1995, 115-123). and is a decision tree-based classification algorithm used for machine learning tasks, particularly for classification tasks. It works by creating a set of rules that form a decision tree for the classification of data. The algorithm „jumps" between the rules by selecting the best rule at each node to classify the data. JRIP differs from other decision tree-based algorithms in that it employs a rule-based approach rather than a pure decision tree approach. This means that instead of relying solely on the branching structure of the decision tree, it generates a set of rules that guide the classification process more specifically (Jijo and Abdulazeez 2021, 20-28).

4.3 OneR – Also called One Rule, it is a simple and interpretable classification algorithm proposed by Holt (Holte 1993, 63-90), and it is used for machine learning tasks. It works by identifying the single most significant attribute or feature in a dataset and using it to create a rule for classification. OneR is called „one rule" because it uses only one rule to classify data, making it easy to interpret and explain (Chantar, et al. 2020).

4.4 Decision Stump – A Decision Stump classifier is a simple binary classification algorithm that is often used as a building block for more complex machine learning models (Sammut 2017). Funcționează prin crearea unui arbore de decizie cu un singur nivel, numit „ciot", în. It works by creating a decision tree with only one level, called a „stump," where each node is a decision rule based on a single feature or attribute (Jijo and Abdulazeez 2021, 20-28). The Decision Stump classifier is called a „stump" because it consists of only one level, unlike more complex decision trees with multiple levels (Sammut 2017).

4.5 ZeroR – The ZeroR classifier is a simple, baseline algorithm for classification that always predicts the most frequent class in the training dataset (Devasena, et al. 2011). It is called „ZeroR" because it does not use any input features to make predictions, and instead relies solely on the class distribution of the training data. The ZeroR classifier is often used as a baseline model to compare the performance of other, more complex classifiers.

4.6 SGD – The Stochastic Gradient Descent classifier is an algorithm for training linear classifiers and regression models in machine learning (Chollet 2017, 48-50). It is particularly useful for large datasets, as it updates the model parameters using small batches of data at a time, rather than the entire dataset, which can lead to faster convergence and lower memory requirements. The SGD classifier is commonly used for tasks such as text classification, image classification, and natural language processing.

4.7 CVPS – CV parameter selection (CVPS) refers to the process of selecting the best hyperparameters for a machine learning model using cross-validation (Varma and Simon 2006, 1-8). This technique involves splitting the training data into k folds, training the model with k-1 folds and validating it with the remaining fold. This process is repeated for each fold, and the average performance is used to select the best hyperparameters.

4.8 RFC – The Randomizable Filtered Classifier (RFC) is a machine learning algorithm that combines the concepts of feature selection and classification. It is designed to select a subset of relevant features from the input data before training a classification model (Alam, Ubaid, et al. 2021). By selecting a subset of relevant features, the algorithm can improve the efficiency and accuracy of the classification model (Alam, Ubaid, et al. 2021). Additionally, by training multiple models on different subsets of the data, the algorithm can provide more robust predictions and reduce the risk of overfitting.

4.9 LMT – The Logistic Model Tree (LMT) combines decision trees with logistic regression to build a classification model. It was developed to address the limitations of standard decision trees, which can suffer from overfitting and may not capture complex interactions between input features (Chen, et al. 2017).

4.10 LWL – Locally Weighted Learning (LWL) is a supervised learning algorithm that uses a non-parametric approach to learn the underlying relationship between the input features and output variables (Tuyen, et al. 2021). This allows LWL to capture complex, non-linear relationships in the data while avoiding overfitting.

4.11 CvC – Classification via Clustering is a semi-supervised learning method that uses clustering algorithms to create labels for unlabeled data (Bergsma, et al. 2013). The method works by first clustering the labeled data into different groups based on their features. Then, the unlabeled data points are assigned to the same clusters as the labeled data points. Finally, the most common label within each cluster is assigned to the unlabeled data points within that cluster.

4.12 WIHW – The Weighted Instances Handler Wrapper is a machine learning technique that adjusts the class distribution of a dataset by assigning weights to each instance based on its class label (Khosravi, Khozani and Mao 2021). The WIH Wrapper works by fitting a classifier to the original dataset and then modifying the dataset by assigning weights to each instance based on its class. Instances that are misclassified are given higher weight, while instances that are correctly classified are given lower weight. This process is repeated until the classifier's performance on the modified dataset converges.

4.13 Ridor – This model is a decision tree-based classification algorithm that uses the concept of rule-based induction to improve classification performance

(Lakmali and Haddela 2017). It works by constructing a decision tree in which each node represents a test on an attribute, and each branch represents the outcome of the test. The Ridor model differs from standard decision trees in that it uses a set of rules to determine when to stop partitioning the data into further subgroups. These rules include a minimum number of instances per leaf and a maximum number of rules to be used (Jijo and Abdulazeez 2021, 20-28).

4.14 MLP – The Multi-Layer Perceptron (MLP) algorithm is a type of feedforward neural network that is commonly used in supervised learning tasks such as classification and regression and was proposed by Rosenblatt in 1950 (Ozbay and Alatas 2020). It consists of multiple layers of nodes or neurons, with each neuron in a layer connected to all neurons in the previous layer. The input layer receives the input data, and the output layer produces the final output or prediction. The hidden layers between the input and output layers perform non-linear transformations of the input data to extract meaningful features (Botalb, et al. 2018, 1-18).

4.15 OLM – Ordinal Learning Model (OLM) is a type of supervised learning algorithm used for ordinal regression problems proposed by Ben-David et al. (Ben-David, Sterling and Pao 1989, 45-49). In an ordinal regression problem, the target variable has a natural ordering, such as a rating from 1 to 5, rather than being nominal or binary.

4.16 SimpleCart – Simple CART (Classification and Regression Trees) was first proposed by Leo Breiman in 1984 (Breiman, Friedman, et al. 2017) and it is a decision tree algorithm that recursively partitions the data into subsets based on the values of the input features to minimize the impurity of the target variable (Loh 2011, 14-23).

4.17 ASC – The Attribute Selected Classifier (ASC) is a supervised learning algorithm that combines feature selection with a classification algorithm to improve classification accuracy and reduce the computational complexity of the model (Gnanambal, et al. 2018, 3640-3644). ASC works by first selecting a subset of the most relevant features from the input data using a feature selection method such as information gain, gain ratio, or chi-squared test.

4.18 J48 – This algorithm is regularly the favored model for classification applications. J48 is a decision tree algorithm and an implementation of the C4.5 algorithm (Bhargava, et al. 2013). It works by recursively partitioning the data into subsets based on the values of the input features to minimize the entropy or information gain of the target variable.

4.19 SMO – Sequential Minimal Optimization (SMO) is an algorithm mainly used to strengthen the training of support vector machines (SVMs) for binary classification tasks (Ozbay and Alatas 2020) and was initially introduced in 1998 by Platt (Platt 1998).

4.20 Bagging – Is short for Bootstrap Aggregating and is an ensemble learning technique for improving the stability and accuracy of machine learning models. It works by training multiple instances of the same algorithm on different subsets of the training data, and then combining their predictions through a voting or averaging mechanism (Breiman 1996, 123-140).

4.21 Decision Tree – It is a type of supervised learning algorithm used for both classification and regression tasks. It is a non-parametric model that recursively splits the data into subsets based on the values of input features to predict the value of the target variable (Jijo and Abdulazeez 2021, 20-28).

4.22 IBk – The "IBK" algorithm (Instance-Based K-Nearest Neighbor) is a machine learning algorithm used for classification and regression tasks that belongs to the family of lazy learning algorithms, where the model is trained by storing the entire training dataset and making predictions based on the similarity between new input data and the stored training instances (Moayedi, et al. 2019).

4.23 KLR – Kernel Logistic Regression (KLR) is a supervised learning algorithm used for classification tasks. It is an extension of the traditional logistic regression algorithm that uses a kernel function to transform the input data into a higher dimensional space, allowing for the modeling of nonlinear relationships between features (Zhu and Hastie 2005, 185-205).

4.24 Performance comparison: Without getting into more detailed specifics on how the models were fine-tuned for fake news detection, the performance of the described

TABLE 1 Claimed performance of the supervised AI algorithms described in chapter 3.2, trained and evaluated using the ISOT Fake News data set according to FA Ozbay and B. Alatas (Ozbay and Alatas 2020). The highest scores are underlined.

Model	Accuracy	Precision	Recall	F-measure
BayesNet	0,586	0,587	0,586	0,586
JRip	0,607	0,611	0,588	0,599
OneR	0,559	0,567	0,560	0,547
Decision Stump	0,564	0,574	0,564	0,549
ZeroR	0,501	0,501	**1.000**	0,667
SGD	0,589	0,590	0,583	0,586
CVPS	0,501	0,501	**1.000**	0,667
RFC	0,526	0,525	0,534	0,530
LMT	0,607	0,604	0,616	0,610
LWL	0,570	0,573	0,570	0,566
CvC	0,553	0,556	0,526	0,541
WIHW	0,501	0,501	**1.000**	0,667
Ridor	0,557	0,563	0,558	0,549
MLP	0,565	0,565	0,571	0,568
OLM	0,516	0,540	0,516	0,430
SimpleCart	0,604	0,607	0,586	0,597
ASC	0,588	0,598	0,534	0,564
J48	0,558	0,558	0,563	0,560
SMO	0,534	0,536	0,489	0,512
Bagging	0,598	0,603	0,576	0,589
Decision Tree	**0,968**	**0,963**	0,973	**0,968**
IBk	0,551	0,551	0,551	0,550
KLR	0,606	0,605	0,614	0,609

supervised AI algorithms has been compared using the ISOT Fake News Dataset by a research team from the Department of Software Engineering of Firat University from Elazig, Turkey (Ozbay and Alatas 2020), with the results shown in Table 1.

As highlighted in Table 1, the Decision Tree algorithm returns the best metrics (96.8% accuracy score) when confronted with fake news binary classification tasks. The result is also confirmed by other authors that achieved an accuracy score of over 95% (Lyu and Lo 2020). The basic idea behind a decision tree is to create a tree-like model that represents a set of decisions and their possible consequences. The reason why the decision tree model performed better than the other supervised models for a fake news may be that it can handle high-dimensional and sparse data effectively. Decision trees can handle categorical data in a precise way, which is common in NLP tasks, where words or phrases are often used as features multiple (Jijo and Abdulazeez 2021, 20-28).

5. Deep learning and transformer models used in fake news detection

Deep Learning is another popular AI method for fake news classification. Deep learning and transformer models can contribute to fake news detection by allowing machines to learn complex patterns and features from large amounts of text data (Young, et al. 2018, 55-75). These models can handle the complexity and variability of language, and can identify the subtle linguistic cues and patterns that distinguish fake news from real news. Such models involve training a neural network model on a large dataset of news articles with labels indicating whether each article is real or fake (Chollet 2017):

5.1 XLNet – Called eXtreme Learning NETwork model, it a state-of-the-art pre-trained language model that was introduced in 2019 and it is claimed to have achieved high metrics in binary tasks involving fake news balanced datasets (Gautam, Venktesh and Masud 2021). The model is based on the Transformer architecture, which is a type of neural network that is well-suited for natural language processing tasks. What sets XLNet apart from other pre-trained models is its use of an autoregressive method that allows for bidirectional context modeling, which helps the model to better understand the context and relationships between words in a sentence. This approach allows XLNet to achieve state-of-the-art performance on a wide range of natural language tasks, including language modeling, question answering, and sentiment analysis (Gundapu and Mamidi 2021). Take for example a text like „The President made an announcement that the new policy would benefit all Americans, but experts have criticized the plan as harmful to the economy." This contains several linguistic cues that are associated with fake news, including the use of positive language („benefit all Americans") followed by negative language („criticized the plan as harmful"). An autoregressive model like XLNet captures

these subtle patterns by considering the bidirectional context of each word in the sentence, allowing it to identify the relationships between words and phrases that are indicative of fake news.

5.2 BERT and ALBERT – Other models have been based on Google`s BERT (Bidirectional Encoder Representations from Transformers) model (Devlin, et al. 2019), which is a pre-trained deep learning model that has achieved state-of-the-art performance on a wide range of natural language processing tasks, including question answering, sentiment analysis, and language translation. BERT is a transformer-based model that is trained on a large corpus of text data, allowing it to learn rich representations of language that can be fine-tuned for specific tasks. BERT or variants like ALBERT (A Lite BERT model) (Gundapu and Mamidi 2021) have been claimed to be highly effective in tasks such as natural language understanding and text classification, tasks similar to fake news binary classification.

5.3 RoBERTa – The Robustly Optimized BERT Approach (RoBERTa) model was introduced in 2019 (Liu, et al. 2019) and is based on the BERT (Bidirectional Encoder Representations from Transformers) architecture but was trained on a much larger corpus of data than BERT, with an extended training duration and improved training techniques. This allows RoBERTa to better capture complex relationships and patterns in natural language text, resulting in improved performance on a wide range of NLP tasks, including fake news classification. RoBERTa is fine tuned for entity classification and has been claimed to have superior metrics when applied on fake and real news datasets (Liu, et al. 2019).

5.4 FakeBERT – One of the earlier models based on BERT and fine-tuned for fake news detection tasks was called FakeBERT (Kaliyar, Goswami and Narang 2021c) and it uses a data augmentation technique called back-translation. This involves translating real news articles into another language, and then translating them back into the original language using a machine translation system. This process helps to generate additional training data and increase its diversity, which can improve the model's accuracy and ability to detect subtle variations in the text. Back-translation can be useful for fake news detection by generating synthetic data for training models. This synthetic data can be used to augment real datasets of labeled news articles, helping to improve the performance of NLP models trained for fake news detection.

5.5 DeepFakE and EchoFakeD – Other authors used a DNN model, or Deep Neural Network model and fine-tuned it for fake news classification tasks. Models like DeepFakE DeepFakE (Kaliyar, Goswami and Narang 2021a, 1015-1037) or EchoFakeD (Kaliyar, Goswami and Narang 2021b, 8597–8613), which were trained on BuzzFeed and PolitiFact fake news datasets, have been claimed achieve accuracy scores between 88% and 98%. A DNN is a type of artificial neural network that is composed of multiple layers of interconnected nodes, or neurons. These layers

allow the network to extract and learn increasingly complex features from the input data, enabling it to make more accurate predictions or classifications. DNN models consist of an input layer, one or more hidden layers, and an output layer. The hidden layers contain the majority of the neurons and are responsible for processing and transforming the input data. Each neuron in a DNN model receives input from multiple neurons in the previous layer, and uses an activation function to transform the input before passing it on to the next layer (Kaliyar, Goswami and Narang 2021c, 11765–11788). By training on a domain-specific dataset (like fake news), DNNs can learn to identify patterns and features that are specific to that domain or language, improving their accuracy and effectiveness for detecting the different classes.

5.6 LSTM-RRN and BiLSTM-RNN – Some researchers (Bahadad, Saxena and Kamal 2019) experimented with architectures based on Long Short-Term Memory Recurrent Neural Networks (LSTM-RRN) or Bidirectional Long Short-Term Memory Recurrent Neural Network (BiLSTM-RNN) with some degree of success after fine-tuning it despite the fact that a LSTM-RNN model is a type of deep neural network architecture that is designed to process sequential data, such as time-series data or natural language text. The LSTM layer allows the network to retain information from previous inputs over a long period of time, making it well-suited for tasks like fake news detection that require understanding of the context or history of the data. Previously, LSTM-RNNs have been claimed to be effective for tasks such as language modeling, sentiment analysis, and machine translation, but have only recently been used even for fake news classification (Bahadad, Saxena and Kamal 2019, 74-82).

5.7 CNN – The models used include CNN (Luan and Lin 2019) or Convolutional Neural Networks which represent a particular type of deep neural network architecture that is commonly used for image and video recognition tasks. The key innovation of the CNN is the use of convolutional layers, which apply a set of filters to the input image, allowing the network to learn important features and patterns at different spatial scales. In addition to convolutional layers, a typical CNN also includes pooling layers, which reduce the spatial size of the feature maps, and fully connected layers, which perform the final classification. CNNs are able to identify patterns and features in text data by convolving a set of filters over the input text sequence. This process allows the network to capture local dependencies between adjacent words and phrases in the text, which is important for detecting subtle linguistic cues that distinguish fake news articles from real news articles (Luan and Lin 2019).

The claimed performance of the above presented models has been compared using the same metrics as the supervised learning techniques, with the results shown in Table 2.

TABLE 2 Claimed performance of deep learning and transformer models for automatic fake news detection tasks on ISOT, BuzzFeed and PolitiFact fake news datasets. The highest claimed scores are underlined.

Model	Accuracy	Precision	Recall	F-measure
ROBERTa	**0,9996**	**0,9997**	**0,9994**	**0,9996**
LSTM-RRN	0,9697	0,97	0,97	0,97
BiLSTM - RRN	0,9875	0,97	0,97	0,97
ALBERT	0,9780	0,9781	0,9781	0,9780
FakeBERT	0,9874	0,99	0,99	0,99
DeepFakE	0,8864	0,8210	0,8460	0,8404
EchoFakeD	0,9230	0,9047	0,8636	0,8837
BERT	0,9813	0,9813	0,9813	0,9813
XLNet	0,9785	0,9787	0,9789	0,9785
CNN	0,9698	0,9698	0,9698	0,9698

5.8 Discussion: According to the selected metrics, The RoBERTa model has the best results across the table, with the mention that, for such binary classification tasks, machine learning models seem to achieve high metrics all around, The RoBERTa model achieves high accuracy on fake news detection by leveraging its strong language representation capabilities and the ability to effectively capture semantic relationships between words and phrases. For example, consider the following headline: "Scientists discover new treatment for cancer that works in 100% of cases". A human reader may immediately be skeptical of this headline, as it seems too good to be true. However, a machine learning model that is trained on bag-of-words or simple word embedding representations may not be able to capture the nuances of the language and may incorrectly classify this article as real. In contrast, the RoBERTa model is able to analyze the full context of the headline and identify the subtle cues that suggest that the article is fake, such as the use of hyperbolic language and the lack of scientific evidence to support the claim.

6. Unsupervised learning techniques used in fake news detection

Unsupervised learning is an AI method that can be used for fake news classification when labeled data is not available. Unsupervised learning techniques for fake news detection do not require labeled data to train the model, but instead rely on identifying patterns and relationships in the data to classify new instances as either real or fake news false (Gangireddy, Deepak, et al. 2020, 75-83). One common unsupervised approach is clustering, where similar news articles are grouped together based on their content and language patterns. This can help identify clusters of news articles that are similar in style and content, which can help distinguish between real and fake news (Celebi and Aydin 2018, 164-170). Another unsupervised approach is topic modeling, which identifies topics and themes within a corpus of text (Li, et al. 2021). Topic modeling can help identify topics that are common in fake news articles, such as conspiracy theories, clickbait headlines, and sensationalism. Anomaly detection is another unsupervised technique, where the model learns to identify instances that deviate significantly from the norm (Celebi and Aydin 2018, 23-28). This can be useful in detecting fake news articles that contain unusual

language patterns or syntax. Such methods have achieved when trained and tested on the PolitiFact dataset (PolitiFact 2017) accuracy scores between 0.81 and 0.82 (Gangireddy, Deepak, et al. 2020).

6.1 Discussion: Unsupervised techniques may prove crucial in identifying ongoing disinformation campaigns on social media or similar online platforms and can be combined with supervised techniques to improve performance (Celebi and Aydin 2018).

7. Rule-based systems for fake news detection

Rule-based systems are another AI method that can be used for fake news classification. Rule-based systems involve defining a set of rules or heuristics that can be used to identify fake news articles based on specific characteristics, such as the use of emotionally charged language or the presence of logical fallacies. Rule-based systems can be less accurate than supervised or deep learning methods, but they can be effective when the task is relatively simple and labeled data is not available (Yuliani, et al. 2019).

These AI methods are useful for fake news classification because they are effective at identifying patterns and structures in the data. For example, a Rule-based Model used to detect Arabic Fake News propagation during Covid-19 achieved a 79,7% accuracy (Alotaibi and Alhammad 2022), which was trained on a dataset of 5015111 tweets and is quite a great success keeping in mind the limitations arising from the difficulty of processing Arabic language.

7.1 Discussion: Rule-based systems seem to be less effective than other techniques in identifying more nuanced or complex forms of fake news that do not fit neatly into pre-defined categories and should be used in combination with models that use labeled data in order to achieve great reliance in a disinformation management model.

8. Hybrid models used for fake news detection

Hybrid models are a combination of two or more AI methods. For example, a rule-based system can be combined with a supervised or unsupervised learning method to improve the performance of the classification task. Hybrid models can be more accurate and effective than single-method models because they can leverage the strengths of multiple methods.

8.1 Hybrid CNN-RNN – A hybrid CNN-RNN model was proposed by Nasir et all. (Nasir, Khan and Varlamis 2021), with limited success (0.5 accuracy). Such an architecture should combine the strengths of both CNNs and RNNs to capture both the local and global context of the input data, while also modeling the temporal

dependencies and context of the input sequence. The CNN component extracts high-level features from the input data, while the RNN component models the temporal dependencies of the sequence. The final hidden state of the RNN component is then used for classification (Nasir, Khan and Varlamis 2021).

8.2 CSI – Another hybrid model called CaptureScoreIntegrate (CSI) (Ruchansky, Seo and Liu 2017) that used datasets gathered from Twitter and Weibo achieved promising success. The model is composed of 3 parts: Capture (includes capturing the news article content and features using an RRN), Score (which computes the score for the source of the article) and Integrate (which classifies the results).

8.3 SVM-RNN-BI-GT – Another study proposed a hybrid model where SVM and RNN with bidirectional GRUs are incorporated in leveraging news content and user comments in fake news (Albahar 2021) on a PolitiFact dataset (PolitiFact 2017).

8.4 HAN – Another proposed hybrid model for fake news detection is HAN (Hierarchical Attention Network) (Albahar 2021). and has a hierarchical structure that mirrors the hierarchical structure of the news presented in the dataset and has two levels of attention mechanisms applied at the word and sentence-level, enabling it to attend differentially to more and less important content when constructing the false/true text representation (Yang, et al. 2016, 1480-1489).

8.5 HSA–BLSTM – Abbreviated from the Hierarchical Social Attention–Bidirectional Long Short-Term Memory was tested on datasets gathered from Twitter and Weibo (Albahar 2021) after being initially used to detect rumors by leveraging hierarchical representations at different levels and the social contexts (Guo, et al. 2018, 943-951).

8.6 TCNN–URG – the Transfer Convolutional Neural Network– User Response Generator (TCNN–URG) is usually used to improve the quality of responses generated by social media chatbots or virtual assistants (Qian, et al. 2018, 3834-3840). It is composed of a CNN and a conditional variational autoencoder and was also tested for fake news automatic text detection on a PolitiFact fake news dataset (Albahar 2021).

8.7 The results for hybrid models claimed by the above-mentioned research teams are presented in Table 3, using the same metric system as in Tables 1 and 2, in order to provide a comparison between similar types of approaches on a binary fake news detection task.

8.8 Discussion: As shown in Table 3, according to the accuracy, precision and F-measure metrics, the best model performing model seems to be the CSI model on a Weibo (a Chinese microblogging site similar to Twitter) dataset (Ruchansky, Seo and Liu 2017). The CSI model is separated into 3 parts, which allows CSI to output

Model	Accuracy	Precision	Recall	F-measure
CNN-RNN hibrid	0,5	0,5	0,5	0,5
CSI-Twitter	0,892	0,9	0,8	0,894
CSI-Weibo	**0,953**	**0,953**	0,953	**0,954**
SVM-RNN-BI-GT	0,912	0,910	**0,961**	0,932
Han	0,837	0,824	0,941	0,810
TCNN–URG	0,712	0,711	0,860	0,860
HSA-BLSTM	0,846	0,894	0,868	0,881

TABLE 3 Claimed performance of hybrid models for automatic fake news detection tasks on ISOT, PolitiFact, Twitter and Weibo fake news datasets. The highest claimed scores are underlined.

a prediction for users and articles independently while combining the information for classification. The experiments were conducted by the research team on two real-world obtained datasets (Weibo and Twitter), which demonstrated the accuracy of the CSI model in classifying fake news articles.

Because fake news detection is a complex task that requires the analysis of various features such as linguistic, temporal, and user-related information, hybrid models may overcome the limitations of a single approach and achieve better performance even if at the current date they are behind other models in their metrics. In order to apply such research in a real, daily operating system, one should take into consideration that deep learning models such as BERT or CNN can capture complex patterns in the text, but they may not consider the temporal or user-related features that are important in fake news detection (Kaur, Boparai and Singh 2019, 2388-2392). Similar, rule-based systems can leverage domain knowledge to detect suspicious patterns in the news, but they may not generalize well to unseen examples and thus making them less accurate than supervised learning approaches that have been previously trained on a large variety of examples. Since fake news detection is a constantly evolving problem and new techniques or models may be required to detect emerging types of fake news, future hybrid models can be flexible and adaptable, allowing the integration of new models or features as they become available, leading to more accurate and robust fake news detection (Thaher, et al. 2021).

9. Conclusion

As shown in the article, different AI machine learning methods return a wide range of metrics when dealing with a binary classification task involving real and fake news. This shows that there is still enough room for improvement in this area. After the comparison of the metrics shown in tables 1, 2 and 3 and taking into consideration the accuracy scores of the unsupervised and rule-based models, the fine-tuned ML model "RoBERTa" claims to achieve the best metrics on a ISOT fake news dataset (99,96% accuracy score, 99,97% precision, 99,94% recall and 99,96% F-score). The model has been developed by Facebook AI. It is based on the BERT model but it

is trained on a larger corpus of text data and includes additional pre-training techniques to improve its accuracy. RoBERTa's performance on fake news makes it a strong candidate for inclusion within a broader disinformation management integrated system. Additionally, RoBERTa can be fine-tuned for specific domains, such as politics or health, which is important in disinformation management, where the subject matter can be highly specialized.

We must point out that RoBERTa is one of the models that has a large and active community of users and developers, which means that it is well-supported and frequently updated with new features and improvements. This makes it easier to integrate into a larger system and to stay up-to-date with the latest research in the field of NLP. Because, RoBERTa's pre-trained weights and associated models are freely available, it is accessible to a wider range of users, regardless of their resources or technical expertise granting it a combination of high performance, flexibility, and accessibility for being included in a complementary software system.

References

Alam, M.T., S. Ubaid, S.S. Sohail, M. Nadeem, S. Hussain, and J. Siddiqui. 2021. "Comparative Analysis of Machine Learning based filtering techniques using MovieLens." *Procedia Computer Science 194* 2010-2017.

Albahar, Marwan. 2021. "A hybrid model for fake news detection: Leveraging news content and user comments in fake news." *IET Information Security.* doi:https://doi.org/10.1049/ise2.12021.

Alotaibi, Fatimah L, and Muna M. Alhammad. 2022. "Using a Rule-based Model to Detect Arabic Fake News Propagation during Covid-19." *International Journal of Advanced Computer Science and Applications.* doi:10.14569/IJACSA.2022.0130114.

Bahadad, Pritika, Preeti Saxena, and Raj Kamal. 2019. "Fake News Detection using Bi-directional LSTM-Recurrent Neural nETWORK." *Procedia Computer Science 165*: 74–82. doi:https://doi.org/10.1016/j.procs.2020.01.072.

Ben-David, A., L. Sterling, and Y.H. Pao. 1989. "Learning and classification of monotonic ordinal concepts." *Computational Intelligence* 5 (1): 45-49. doi:https://doi.org/10.1111/j.1467-8640.1989.tb00314.x.

Bergsma, S., M. Dredze, B. Van Durme, T. Wilson, and D. Yarowsky. 2013. "Broadly improving user classification via communication-based name and location clustering on twitter." *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies.*

Bhargava, N., G. Sharma, R. Bhargava, and M. Mathuria. 2013. "Decision tree analysis on j48 algorithm for data mining." *Proceedings of international journal of advanced research in computer science and software engineering.*

Botalb, A., M. Moinuddin, U.M. Al-Saggaf, and S.S. Ali. 2018. "Contrasting convolutional neural network (CNN) with multi-layer perceptron (MLP) for big data analysis." *International conference on intelligent and advanced system (ICIAS).* 1-18.

Breiman, L. 1996. "Bagging predictors." *Machine Learning* 24 (2): 123-140. doi:10.1023/A:1018054314350.

Breiman, L., J.H. Friedman, R. A. Olshen, and C. Stone. 2017. *Classification and regression trees.* New York: Routledge. doi:https://doi.org/10.1201/9781315139470.

Celebi, M. Emre, and Kemal Aydin. 2018. *Unsupervised Learning Algorithms.* doi:https://doi.org/10.1007/978-3-319-24211-8.

Chantar, H., M. Mafarja, H. Alsawalqah, A.A. Heidari, I. Aljarah, and H. Faris. 2020. "Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification." *Neural Comput. Appl 32* 12201–12220. doi:https://doi.org/10.1007/s00521-019-04368-6.

Chen, W., X. Xie, J. Wang, B. Pradhan, H. Hong, D.T. Bui, and J. Ma. 2017. "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility." *Catena.* http://dx.doi.org/ 10.1016/j.catena.2016.11.032.

Chollet, François. 2017. *Deep Learning with Python.* New York: Manning.

Cohen, William W. 1995. "Fast effective rule induction." *Machine learning proceedings, 12th anual conference.* Morgan Kaufmann. 115-123.

Devasena, C.L., T. Sumathi, V.V. Gomathi, and M.Hemalatha. 2011. "Effectiveness evaluation of rule based classifiers for the classification of iris data set." *Bonfring International Journal of Man Machine Interface 1.*

Devlin, J., M.W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." doi:https://doi.org/10.48550/arXiv.1810.04805.

Gangireddy, Siva Charan Reddy, P. Deepak, Cheng Long, and Tanmoy Chakraborty. 2020. "Unsupervised Fake News Detection: A Graph-based Approach." *Proceedings of the 31st ACM Conference on Hypertext and Social Media (HT '20)* 75-83. doi:https://doi.org/10.1145/3372923.3404783.

Gautam, Akansha, V. Venktesh, and Sarah Masud. 2021. "Fake news detection system using xlnet model with topic distributions: Constraint@ aaai2021 shared task." *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event.*

Gnanambal, S., M. Thangaraj, V.T. Meenatchi, and V. Gayathri. 2018. "Classification algorithms with attribute selection: an evaluation study using WEKA." *International Journal of Advanced Networking and Applications* 3640-3644.

Gundapu, Sunil, and Radhika Mamidi. 2021. "Transformer based Automatic COVID-19 Fake News Detection System." *International Institute of Information Technology.*

Guo, H., J. Cao, Y. Zhang, J. Guo, and J. Li. 2018. "Rumor Detection with Hierarchical Social Attention Network." *Proceedings of the 27th ACM international conference on information and knowledge management.* doi:https://doi.org/10.1145/3269206.3271709.

Holte, Robert C. 1993. "Very simple classification rules perform well on most commonly used data sets." *Machine learning 11.* 63-90.

Jijo, B.T., and A.M. Abdulazeez. 2021. "Classification based on decision tree algorithm for machine learning." *Journal of Applied Science and Technology Trends (JASTT)* 20-28.

Kaliyar, Rohit Kumar, Anurag Goswami, and Pratik Narang. 2021a. "DeepFakE: improving fake news detection using tensor decomposition-based deep neural network." *Journal of Supercomputing* 77 (2): 1015-1037. doi:10.1007/s11227-020-03294-y.

—. 2021b. "EchoFakeD: improving fake news detection in social media." *Neural Computing and Applications* 33: 8597–8613. doi:https://doi.org/10.1007/s00521-020-05611-1(0123456789().,-volV)(0123456789(). ,- volV).

—. 2021c. "FakeBERT: Fake news detection in social media with a BERT- based deep learning approach." *Multimedia Tools and Applications* (80): 11765–11788. doi:10.1007/s11042-020-10183-2.

Kaur, Prabhjot, Rajdavinder Singh Boparai, and Dilbag Singh. 2019. "Hybrid Text Classification Method for Fake News Detection." *International Journal of Engineering and Advanced Technology (IJEAT)* 8 (5): 2388-2392.

Khosravi, Khabat, Zohreh Sheikh Khozani, and Luca Mao. 2021. "A comparison between advanced hybrid machine learning algorithms and empirical equations applied to abutment scour depth prediction." *Journal of Hydrology.* doi:https://doi.org/10.1016/j.jhydrol.2021.126100.

Lakmali, K.B.N., and P.S. Haddela. 2017. "Effectiveness of rule-based classifiers in Sinhala text categorization." *National Information Technology Conference (NITC).* Colombo, Sri Lanka. doi:10.1109/NITC.2017.8285655.

Langley, Pat, Iba Wayne, and Kevin Thompson. 1992. "An analysis of Bayesian classifiers." *Proceedings of the Tenth National Conference of Artificial Intelligence.* California. 223-228.

Li, Dun, Haimei Guo, Zhenfei Wang, and Zhiyun Zheng. 2021. "Unsupervised Fake News Detection Based on Autoencoder." *Access.* doi:https://doi.org/10.1109/ACCESS.2021.3058809.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *ArXiv.* doi:https://doi.org/10.48550/arXiv.1907.11692.

Loh, Wei-Yin. 2011. "Classification and regression trees." *WIREs Data Mining Knowl Discov* 14-23. doi:10.1002/widm.8.

Luan, Yuandong, and Shaofu Lin. 2019. "Research on Text Classification Based on CNN and LSTM." *International Conference on Artificial Intelligence and Computer Applications (ICAICA).* doi:https://doi.org/10.1109/ICAICA.2019.8873454.

Lyu, Shikun, and Dan Chia-Tien Lo. 2020. "Fake News Detection by Decision Tree." *SoutheastCon.* doi:https://doi.org/10.1109/SoutheastCon44009.2020.9249688.

Moayedi, H., D. Tien Bui, B. Kalantar, and L. Kok Foong. 2019. "Machine-Learning-Based Classification Approaches toward Recognizing Slope Stability Failure." *Applied Sciences* 9 (21). doi:https://doi.org/10.3390/app9214638.

Nasir, J.A., O.S. Khan, and I. Varlamis. 2021. "Fake news detection: A hybrid CNN-RNN based deep learning approach." *International Journal of Information Management Data Insights.* doi:10.1016/j.jjimei.2020.100007.

Ozbay, Feyza Altunbey, and Bilal Alatas. 2020. "Fake news detection within online social media using supervised artificial intelligence algorithms." doi:https://doi.org/10.1016/j.physa.2019.123174.

Platt, John. 1998. *Sequential minimal optimization: A fast algorithm for training support vector machines.* Technical Report MSR-TR-98-14, Microsoft Research.

PolitiFact. 2017. https://www.politifact.com/.

Qian, F., C. Gong, K. Sharma, and Y. Liu. 2018. "Neural User Response Generator: Fake News Detection with Collective User Intelligence." *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence.*

Ruchansky, Natali, Sungyong Seo, and Yan Liu. 2017. "CSI: A Hybrid Deep Model for Fake News Detection." *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17).* doi:https://doi.org/10.1145/3132847.3132877.

Sammut, C., Webb, G.I. (eds). 2017. "Decision Stump. Encyclopedia of Machine Learning." In *Encyclopedia of Machine Learning*, by C., Webb, G.I. (eds) Sammut, 262–263. Boston, MA.: Springer. doi:10.1007/978-0-387-30164-8_202.

Thaher, T., M. Saheb, H. Turabieh, and H. Chantar. 2021. "Intelligent Detection of False Information in Arabic Tweets Utilizing Hybrid Harris Hawks Based Feature Selection and Machine Learning Models." *Symmetry 13* 556. doi:https://doi.org/10.3390/sym13040556.

Tuyen, T.T., A. Jaafari, H.P.H. Yen, T. Nguyen-Thoi, T. Van Phong, H.D. Nguyen, and B.T. Pham. 2021. "Mapping forest fire susceptibility using spatially explicit ensemble models based on the locally weighted learning algorithm." *Ecological Informatics.* doi:https://doi.org/10.1016/j.ecoinf.2021.101292.

University of Victoria. 2017. „ISOT Fake News dataset." https://www.uvic.ca/ecs/ece/isot/datasets/fake-news/index.php.

Varma, Sudhir, and Richard Simon. 2006. "Bias in error estimation when using cross-validation for model selection." *BMC bioinformatics 7.1.*

Yang, Z., D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. 2016. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies.*

Young, T., D. Hazarika, S. Poria, and E. Cambria. 2018. "Recent trends in deep learning based natural language processing." *ieee Computational intelligenCe magazine* 13 (3): 55-75. doi:10.1109/MCI.2018.2840738.

Yuliani, S.Y., M.F.B. Abdollah, S. Sahib, and Y.S. Wijaya. 2019. "A framework for hoax news detection and analyzer used rule-based methods." *International Journal of Advanced Computer Science and Applications.*

Zhu, J., and T. Hastie. 2005. "Kernel logistic regression and the import vector machine." *Journal of Computational and Graphical Statistics* 14 (1): 185-205.