Journal: Contemporary Mathematics
Special Issue: Contemporary Contributions to Statistical Modelling of Futuristic Machine Learning Algorithms for IoT and Clustered Data Science

Cover Letter

Dear Editor,

Hereby, we submit a manuscript for consideration of publication in Contemporary Mathematics - Special Issue: Contemporary Contributions to Statistical Modelling of Futuristic Machine Learning Algorithms for IoT and Clustered Data Science. The manuscript is entitled "Identifying Fake Digital Information Using Machine Learning Algorithms: Performance Analysis and Recommendation System".

"Fake Digital Information" refers to things that aren't true but are written in a way that makes the reader think they are. Propaganda is subtly interwoven into these tales to sway readers. Many machine learning and deep learning algorithms were used in this study to detect bogus data, and a recommendation system was presented to boost readers' confidence in online articles.

The manuscript is original and it has not been previously published in whole or in part, and it is not being considered for publication elsewhere. All the authors have read the final manuscript, have approved the submission to the journal, and have accepted full responsibilities pertaining to the manuscript's delivery and contents. For any correspondence, please write to: rutvij.jhaveri@sot.pdpu.ac.in.

Yours Sincerely,
Ashish Patel, Yogesh Jadhav, Rutvij Jhaveri, Roshani Raut, Faisal Mohammed Alotaibi, Dhavalkumar Thakker

Identifying Fake Digital Information Using Machine Learning Algorithms: Performance Analysis and Recommendation System

Ashish Patel, Yogesh Jadhav, Rutvij Jhaveri*,
Roshani Raut, Faisal Mohammed Alotaibi+, Dhavalkumar Thakker

Abstract

This work focuses on the detection of fake digital information using various machine learning and deep learning algorithms to prevent its spread in IoT devices and systems. The research highlights the significance of detecting and preventing false or misleading information in critical areas such as healthcare, public safety, and emergency response. The study compares the performance of several supervised machine learning algorithms and identifies logistic regression as the most accurate (98.03%). The empirical analysis used data from The Indian Express, PolitiFact, and Kaggle and leveraged Natural Language Processing (NLP) to prepare, clean, and model the data. To detect fraudulent posts, the study employed Random Forest, a supervised machine learning algorithm, which achieved an impressive accuracy rate of 99.71% on a Kaggle dataset. The research also developed a model for detecting false reporting related to COVID-19, utilizing the Support Vector Machine technique, which achieved an accuracy rate of 78.69%. The presented work also determined authenticity of images through Convolutional Neural Networks (CNN). Lastly, a content-based recommendation system was developed to enhance people's security and confidence.

Keywords: COVID-19; Deep Learning; Google Colab; Image Processing; Kaggle

1 Introduction

From an IoT perspective, the proliferation of "fake news" can have severe consequences on connected devices and systems. Misinformation, defined as inaccurate or incorrect information unintentionally made or distributed, can spread rapidly through IoT devices, leading to false assumptions and misguided actions [1, 2]. Meanwhile, the intentional creation and dissemination of disinformation can manipulate public opinion and obstruct factual evidence, potentially causing damage to the IoT system's trust and the government's ability to make informed decisions. In such cases, it becomes crucial to employ

*Correspondence: rutvij.jhaveri@sot.pdpu.ac.in, +Co-correspondence: Fasiaal.sdn@gmail.com

techniques like fake digital information detection to prevent the spread of false or misleading information through IoT devices, ensuring that decisions are based on accurate and factual evidence [3, 4]. Misinformation and disinformation can be broken down into seven distinct categories, which are as follows:

1. Satire or Parody: The use of satire or parody is risky, even though the author does not intend to cause harm to anyone.
2. Misleading Content: Misleading Content is characterized by inappropriate use of facts to present a topic or a person in an unfavorable light.
3. Imposter Content: Imposter content is created when genuine sources are misrepresented as being something else.
4. Fabricated Content: Recently fabricated content is wholly untrue, and its creators intend it to mislead and harm people.
5. False Connection: This is what happens when the headlines, photos, or captions do not adequately support the content.
6. Context: This occurs when authentic content is distributed together with contextualized fake information.
7. Manipulated Content: It can be defined as changing truthful information or imagery to trick someone into believing something untrue.

Websites that specialize in creating or sensationalizing stories are frequently the ones responsible for the dissemination of fake news [5]. It often employs controversial and provocative headings. Its growth poses a significant obstacle to the functioning of democratic societies in the modern world [6]. Fact-checkers are people who investigate the claims made in the media to ensure that they are accurate. These professionals debunk false claims made in fake news by pointing out why they are false. The use of machine learning and natural language processing (NLP) techniques, as shown in [7, 8], can now be used to supplement the manual fact-checking procedure that has traditionally been used. An application of artificial intelligence known as machine learning allows a computer to learn independently without being specifically programmed to do so. It not only processes but also learns from the data. We provide machine learning with the dataset, and the software generates the algorithm independently. There are three different ways of acquiring knowledge: (i) Supervised Learning (ii) Unsupervised Learning (iii) Reinforcement Learning.

"What we see is what we believe". A person cannot readily think of a made-up truth without first validating it, which results in its dissemination to others. Individuals begin feeling something is proven only after it has been exposed several times, which explains the illusory truth effect. The repetition effect persuades us to think previously incorrect information is true; this is the impact's strength. The illusory truth effect is one of the factors that contribute to misleading news reports gaining traction and drawing an audience [9, 10]

A tremendous amount of image data has been generated by social networking platforms such as Facebook and Instagram. GANs (Generative Adversarial Networks) is a burgeoning topic of Machine Learning/Artificial Intelligence that generates a lot of fake images. Many people have fallen victim to picture counterfeiting using image and video processing software like GNU Gimp, Adobe Photoshop, etc. Fake news and mob provocation rely heavily on photos like this, which are excellent targets for malicious manipulation [11, 12, 13].

Several techniques for facial manipulation in videos, including deepfake and faceswap, have been developed in the last several years, allowing anyone to easily change faces in video sequences with amazingly realistic results and no effort [14, 15, 16, 17, 18]. Modern technology, powerful smartphone cameras, and the widespread availability of high-speed internet connections have enormously boosted social media's ever-growing reach. Technological advancements, increased network access, and improved peer-to-peer connectivity have simplified the creation and transmission of digital videos across media-sharing platforms. Deep learning has become extremely powerful as processing power has increased, which was thought to be unachievable only a few years ago [19]. However, this disruptive technology has introduced some new obstacles. Free deep learning-based software tools have made it possible to create credible face exchanges in films that leave minimal indications of manipulation, dubbed DeepFake (DF) videos or AI-generated media. The proliferation of DF on social media platforms has resulted in spamming and the propagation of false information [20]. DF detection is critical for resolving such a problem.

Supervised learning entails training the model with labeled samples, after which the machine performs the task on unseen data [21, 22, 23]. Our study will utilize a supervised learning algorithm to detect bogus news. Supervised learning applies to two distinct sorts of problems: classification and regression [24, 25, 26]. In this endeavor, we classify news as 'false' or 'real'. The optimal classification method for the false news detector is chosen based on the maximum accuracy achieved after comparing various classification algorithms [27, 28].

An approach based on deep learning can successfully distinguish artificial intelligence-generated false videos (DF Videos) from real videos [29, 30, 31, 32]. Convolutional Neural Networks and Recurrent Neural Networks are used to identify DF. The system uses a convolutional neural network at the frame level to extract features (CNN). These features are used to train a recurrent neural network (RNN) that learns to classify whether a video has been manipulated or not and whether detection of temporal inconsistencies between frames introduced by the DF creation tools is possible [33, 34, 35].

From Amazon to LinkedIn, Uber Eats to Spotify, and Netflix to Facebook, recommender systems are widely used to offer "similar things," "related employment," "preferred cuisines," "interesting movies or series," and "songs to try based on interest" to their customers. The number of press releases has expanded significantly, and it has become difficult for one to go through all online news resources in search of relevant news stories. Search engines assist consumers in navigating the massive amounts of information available online. Since then, recommendation systems have evolved to address various issues and provide users with information relevant to their requirements, either based on their preferences or on content similarity [36, 37, 38]. Each online news publisher controls its content and employs various strategies to propose stories to users based on shared interests. In a very dynamic environment, it becomes difficult to propose news articles due to various obstacles, including frequent changes in the set of news articles, the set of users, and rapid changes in user preferences. As a result, recommendation systems must be capable of continuously processing incoming news streams in real-time.

There are two kinds of recommendation methods:

- Content-based recommendation
- Collaborative filtering

A content-based recommendation system uses similarities between users or objects as determined by their qualities [39, 40, 41]. It uses additional information (metadata) about persons or products, i.e., it uses already-existing content. This metadata may include the user's age, gender, occupation, location, and skill sets. It comprises the item's name, specs, category, and registration date. In this article, we develop a content-based recommendation system to recommend news articles comparable to previously read articles based on article headline, category, author, and publishing date.

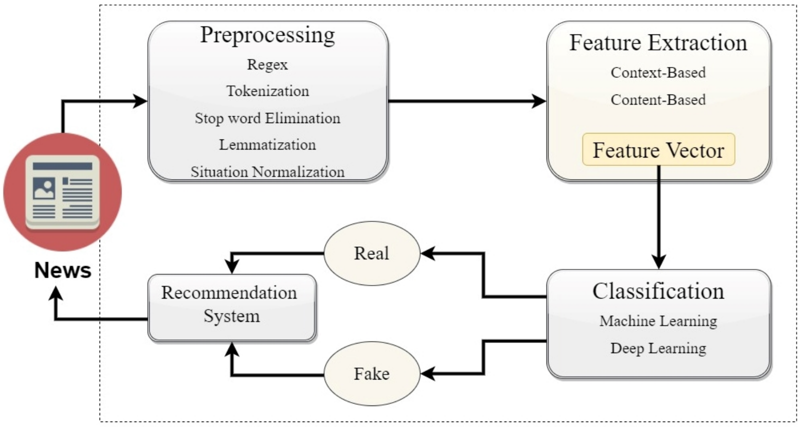


Figure 1: Fake news detection and recommendation approach

In this paper, we provide the results of an empirical study to identify fake news as shown in fig.1. The accuracy of the detection has been measured using a variety of machine learning and deep learning methodologies. Immediately after detecting fake information, we introduced the concept of a recommended system. This approach boosts the user's confidence when working with digital media, which is beneficial.

The key contributions of this work are:

- We propose an intelligent approach to deal with fake digital information.

- We conduct experiments to evaluate our approach to detect fabricated news stories, job listings, COVID-19 information, and manipulated images online.

- We propose a trustworthy recommendation system to increase users' confidence in consuming online digital media.

The organization of the paper is as follows. The first section overviews bogus digital information, its influence on society, and work requirements. The second section of the article describes the current condition of the field. Section 3 provides a detailed evaluation of the empirical inquiry. The fourth section addresses the analysis of the empirical study's results. Section 5 suggests a method for suggesting to the user that they maintain their faith while ingesting digital information. Section 6 concludes with a synopsis of the work and its potential scope.

2 Related Work

Kesarwani at el. [42] demonstrated a straightforward method for detecting fake news on social media using the K-Nearest Neighbor classifier, with a 79 percent accuracy rate. The K-Nearest Neighbors (KNN) algorithm is a supervised machine learning algorithm that addresses classification and regression issues. The article attempts to assess fake news using the K-Nearest Neighbor classification algorithm by collecting users' implicit and explicit n characteristics across many dimensions. According to the article, KNN becomes slower as the model size increases. As indicated in the paper, the dataset for the model was obtained from Buzz Feed News (BFN), and it was utilized to train and test the model in this research. BFN uses the social analytics platform BuzzSumo to determine the best-performing Facebook content among 167 websites that post articles consistently. This dataset collected data on Facebook posts, each representing a news article. The authors culled these news pieces from Politico, ABC News, and CNN. The dataset contained the following four categories: "mainly true," "primarily false," "combination of true and false," and "no factual material." Apart from this, information on user social activity is collected, including the number of shares, comments, and reactions to each post via the Facebook API. After testing the model with various values of K, the study determined that the value of K provided the greatest accuracy. On the test set, the proposal achieved an approximate 79 percent classification accuracy for this model. The model's average weight precision was 0.75, and its recall was 0.79. This article offered a framework for predicting bogus news on social media. The method of extracting features from datasets was deemed critical due to their application in the data mining algorithm K-Nearest Neighbor for classifying news articles on social media.

The purpose of [43] is to categorize news stories as genuine or fabricated. The authors identify fake news using various models and classifiers and then forecast the model's accuracy using the Kaggle dataset. They constructed models using Natural Language Processing, Machine Learning, and Deep Learning approaches and compared them to ensure optimum accuracy. The study analyzed results swiftly using an Nvidia DGX-1 supercomputer. This investigation examined several different models, including CNN, KNN, LSTM, Decision tree, Random forest, and Nave Bayes theorem, as well as Deep Learning networks, including Shallow Convolutional Neural Networks (SCNN), Very Deep Convolutional Neural Networks (VDCNN), Long Short-Term Memory Networks (LSTM), Gated Recurrent Unit Networks (GRU), Combination of Convolutional Neural Networks with Long Short-Term Memory (CNN-LS (CNN-GRU). We investigated feature extraction and features such as the n-gram. The model was constructed using TF-IDF features that were extracted. Additionally, they discovered features such as word embedding and word2vec in Deep Neural networks. For feature extraction in the machine learning model, the proposal used pick k best and chi2.

The study presented by [44] develops a model for classifying Instagram photographs in order to identify potential threats and fake images. The model is constructed utilizing deep learning methods, Convolutional Neural Networks (CNNs), the AlexNet network, and AlexNet transfer learning. According to the findings model, the AlexNet network had a higher accuracy rate of 97 percent than the other networks. This research presents a method for classifying images by taking them as input and classifying them using a practical system (the CNN model). The training phase will use a label or classification assigned to the input sample. Two tags are used: one for the original image class and another for the fictitious image class. The target photographs were collected from the Instagram program, and they represent the dataset, which is necessary to answer the research questions, test the hypothesis, and evaluate the results. The convolutional layer was developed with typical mathematical methods to extract visual features. These convolutional processes operate as two-dimensional digital filters. Following that, the activation function was created. Because picture data is nonlinear, a nonlinear activation function called the Rectified Linear Unit (ReLU) was applied. A technique called max pooling was used to minimize the array's size.

The authors of the [45] presented a solution for detecting misleading information about the covid-19. A model is developed using World Health Organization, UNICEF, and the United Nations data. The study constructed a voting ensemble classifier using seven feature extraction approaches and ten distinct machine-learning algorithms. To ensure the validity of the acquired data, the proposal used fivefold cross-validation and then produced an evaluation technique. The detection technique used ensemble learning to train ten machine learning classifiers on the acquired ground truth data. The author collected COVID-19 ground truth by scraping the websites of the World Health Organization and its regional branches, UNICEF, related organizations, and the United Nations. It gathered all information about the COVID-19 outbreak from these organizations' daily situation updates, the WHO Director-briefing General on COVID-19, and news published in their newsrooms on their websites. The Google Fact Check Tools API was used, which enables users to browse and search for facts from a variety of fact-checking websites worldwide, including opensecrets.org, snopes.com, factcheck.afp.com, washingtonpost.com, factcheck.org, and politifact.com. The results established the veracity of the gathered ground truth data and yielded favorable results. The Neural Networks, Decision Trees, and Logistic Regression classifiers produced the best results in the empirical study.

The article [46] described a new deep-learning method that can distinguish AI-generated fake videos from real videos. To detect the DF, it is necessary to grasp the Generative Adversarial Network (GAN). GAN is fed a video and an image of a specific target. It generates further films in which the target's face is replaced by another individual who serves as a source. GAN divides video into frames and outputs the input image at the end of each frame. Additionally, it reconstructs and utilizes additional autoencoders. Due to resource constraints and production time constraints, the DF method can only generate face images of a given size. They must undergo affinal warping in order to match the source's face configuration. The approach discussed here detects such abnormalities by dividing the movie into frames and comparing the resulting face regions to their surrounding areas. The features are then extracted using a ResNExt Convolutional Neural Network (CNN), a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM), and by recording the GAN-induced temporal discrepancies between frames. It uploaded and classified videos using a web-based platform; the primary purpose was to assess performance, security, user-friendliness, correctness, and reliability.

The primary goal of [47] was to determine whether or not a job posting was phony using the Kaggle dataset. There were 17880 job postings in the datasets. Prior to using this data in the classifier, it is pre-processed to eliminate stop words, superfluous attributes, and excessive spaces. The Decision Tree Classifier outperformed the Naive Bayes, Multi-Layer Perceptron, and K-Nearest Neighbor Classifiers. The technique evaluated the ensemble approach method in order to determine whether it might improve the model's performance or not. We implement and compare Random Tree classifiers, AdaBoost classifiers, and Gradient Boost classifiers. The investigation demonstrated that the Random Tree classifier outperformed the others. The Random Forest Classifier achieved an accuracy of 98.27 percent, a Cohen-kappa score of 0.74, an F1-score of 0.97, and an MSE of 0.02, indicating that it was the most accurate approach.

In its simplest form, a metadata analyzer is a tag search algorithm. Assume that terms such as Photoshop, Gimp, and Adobe are discovered in the text, maybe at a higher rate. The technique provided in [48] resulted in the creation of two distinct variables, dubbed fakeness and realness. Each variable describes the probability of an image being genuine or a forgery. After the tag is assigned, it is assessed, and the relevant variables are incremented by a specified weight. Metadata analysis revealed encouraging results in the field of non-shared photos. Under minimal processing, it can detect anomalies in all 'photoshopped' photographs. However, it did not work with pictures posted over WhatsApp, Google+, or other social media platforms. When the system provides photos with modified metadata, it generates an error. They trained the neural network successfully utilizing error level analysis on 4000 fake and 4000 true photos. The image was classified as false or real using a neural network trained to a maximum success rate of 83 percent. Using this program on mobile devices will almost certainly result in a decrease in the propagation of fraudulent photographs via social media. A reliable false picture identification algorithm is created and evaluated by combining metadata analysis (40%) and neural network output (40%) findings (60%).

The system they develop in [49] learns from sentences how to identify bogus news.

The initial step was to amass datasets. They gathered data from a variety of online sites. The acquired data was then labeled as training data. 30% of that data was used for validation purposes. The data was then processed using the processes detailed in the flowchart below. Following the implementation of the method, the data were further processed using the Bidirectional LSTM model. Two stages are required for the bi-LSTM architecture. In the first stage, encoding headlines and bodies into input before passing them to the classifier based on terms that have been sub-worded, 28 headlines will be entered based on the largest number of words in the real headline. Simultaneously, 1000 words will be fed into the body to ensure that it does not consume an excessive amount of memory. It retains the article's text, with a vector dimension of 300 for each word entered in the headline or news body. They then utilized a Softmax activation function to determine class opportunities in the second stage.

According to the observed findings in [50], KNN is the weakest algorithm in terms of predictive capacity, with a mean accuracy of 75% (EN-English), 89% (PT-Portuguese), and 75% (EN-Portuguese) (ES-Spanish). The ensemble algorithms RF and XGB produced comparable and steady results, with RF achieving 79.9 percent (EN), 93.9 percent (PT), and 82.3 percent (ES), while XGB achieved 80.3 percent (EN), 94.7 percent (PT), and 82.3 percent (ES) (ES). SVMs showed the second-worst prediction performance on EN (79%) but the highest on PT (95%) and tied with ensemble algorithms on ES (82%). According to the results, the PT and ES collections are more linearly separable than the EN collection, with PT exhibiting slightly more linear behavior. Identifying false news is crucial in light of the increased news consumption via OSM, which allows for unregulated content streaming. Due to the rapid expansion of OSM, detection requires an automated method. The concept proposed a comparative examination of language-independent characteristics, stylometric, complexity, and psychological types in a multilingual setting.

3 Empirical Study: Fake Information Detection

The following sections detail the procedures we took to identify fake news. The step-by-step explanation will assist readers in comprehending the use of machine learning algorithms.

3.1 Data Collection

A web scraper was used for the collection of data. The dataset for fake news detection was made by scraping data from websites like PolitiFact and The Indian Express. GitHub was also used for obtaining data.

3.2 Assembling Data in CSV Files

Data collected from various sources were arranged together depending upon the four columns for fake news detection, which were, title, author, text, and label.

3.3 Platform Used

Google Colaboratory (Colab) Notebooks were used for the implementation of our model. For performing deep learning tasks, Google Colab is proven to be an excellent tool. It is a hosted Jupyter notebook that requires no setup, and has an excellent free version, giving free access to Google computing resources such as GPUs and TPUs.

3.4 Importing Libraries

Our model made use of the Nltk, Pandas, NumPy, Matplotlib, Wordcloud, Regex, and Sklearn packages. NLTK, Natural Language Toolkit, provides a systematic foundation for developing systematic Python programs for working with human language data. The interfaces include over 50 corpora and lexical resources, such as WordNet, as well as text-processing libraries for classification, tokenization, stemming, lemmatization, tagging, parsing, and semantic reasoning wrappers. Pandas is a tool for manipulating large amounts of data at a high level. Constructed using the Numpy package and utilizing the DataFrame as the primary data structure.

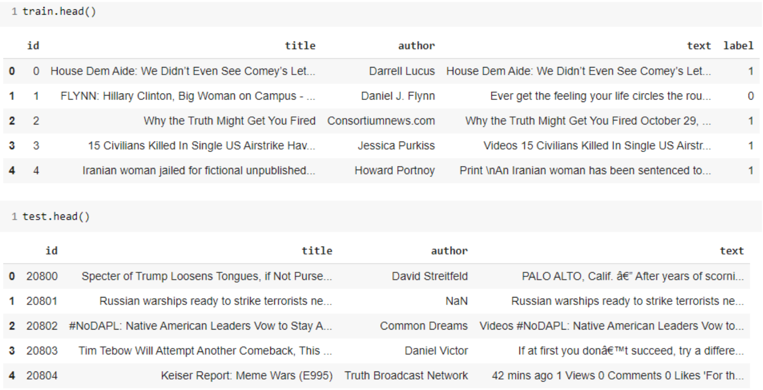
The NumPy library contains multidimensional array objects and a plethora of algorithms for processing them (Numerical Python). NumPy is used to conduct mathematical and logical operations on arrays. Matplotlib is a powerful Python package that enables the creation of static, interactive, and animated displays. We created a data visualization using WordCloud to represent textual data. This library aids in visualizing the significance of each word based on its size, which corresponds to its frequency in the dataset. Regular expressions (abbreviated REs, regexes, or regex patterns) are a small, highly specialized computer language that is integrated into Python and imported via the re-module. It is a string that has a sequence of characters that defines a search pattern. It can be used to determine whether a string contains the defined search pattern. Scikit-learn (Sklearn) is the most efficient library, providing a well-organized collection of methods for machine learning and statistical modeling, including classification, regression, clustering, and dimensionality reduction, via a logical Python interface.

3.5 Information about the datasets

The training dataset has 20800 rows and five columns, whereas the testing dataset contains 5988 rows and four columns. In training data, a label of 1 indicates that the news is genuine, whereas a label of 0 indicates that the news is fraudulent, as illustrated in fig.2.

3.6 Cleaning and Pre-processing

Typically, material scraped from websites is in the raw textual format. Prior to analyzing or fitting a model to this data, it must be cleaned. Cleaning (or pre-processing) is the process of converting data to a format that a computer can interpret. This stage entails the removal of superfluous data. Stop words are meaningless words (data) in natural language processing.



```
train.head()
```

	id	title	author	text	label
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

```
test.head()
```

	id	title	author	text
0	20800	Specter of Trump Loosens Tongues, if Not Purse...	David Streitfeld	PALO ALTO, Calif. â€” After years of scorni...
1	20801	Russian warships ready to strike terrorists ne...	NaN	Russian warships ready to strike terrorists ne...
2	20802	#NoDAPL: Native American Leaders Vow to Stay A...	Common Dreams	Videos #NoDAPL: Native American Leaders Vow to...
3	20803	Tim Tebow Will Attempt Another Comeback, This ...	Daniel Victor	If at first you donâ€™t succeed, try a differe...
4	20804	Keiser Report: Meme Wars (E995)	Truth Broadcast Network	42 mins ago 1 Views 0 Comments 0 Likes 'For th...

Figure 2: Top five rows of training & testing data

- Regex, which is based on a context-free grammar approach, is beneficial for reducing superfluous punctuation.

- Tokenization: For English, we utilized the software Punkt tokenizer. An unsupervised algorithm divides a text into a list of sentences in order to create a model for abbreviated words, collocations, and words that begin sentences.

- Stop Words Elimination: These are frequently used terms (such as "the," "a," "an," and "in") that a search engine has been trained to ignore, both when indexing entries for searching and when retrieving them in response to a search query. To avoid clogging up our database or wasting precious processing time, stop words must be removed. This is accomplished by keeping a list of words deemed to stop words. Subsequently, the dataset will be cleaned of stop words.

- Lemmatization is the process of gathering together the various versions of a word so that they can be studied as a single item. This method grasps the term in its base form, called Lemma, i.e., in its dictionary form. This procedure preserves contextual information. Pre-processing of text entails both stemming and lemmatization. Wordnet is a freely accessible lexical database containing over 200 languages that was created primarily for natural language processing. It contains semantic associations (e.g. synonyms) between its terms. It is a frequently used lemmatizer approach. Synsets are a way for WordNet to organise synonyms. Synsets are a collection of semantically identical data components. After removing stop-words from our dataset, we employed a WordNet lemmatizer.

- Situation Normalization of Items: In this case, we are just turning all characters in the text to lowercase. Because Python is a case-sensitive language, it distinguishes between NLP and NLP. The .lower() method can be used to transform a string to lowercase.

3.7 Applying Feature Extraction Techniques

Conversion of document corpora into a numerical structure is required to make them more intelligible to computers. To do this, a Vector Space model, commonly known as a Bag-of-Words (BoW) model, is used. Vectorization is the process of transforming a collection of textual texts into numerical feature vectors. The Bag of Words or "Bag of n-grams" representation is the methodological approach for tokenization, counting, and normalizing. The Bag-of-Words technique entails the following:

- Splitting of the documents into tokens.

- Giving weights to each token proportional to the frequency with which it arises up in the document and/or corpora.

- Creating a document-term matrix with every row showcasing a document and each column addressing a token.

This approach involves describing documents in terms of word occurrences while fully disregarding the document's relative location information for the words. The most frequently used feature extraction algorithms from the SKLearn.feature extraction. text classes are CountVectorizer and TfidfVectorizer. CountVectorizer converts text documents to a matrix of token counts, taking into account the occurrences of tokens in each document. As a result, a sparse representation of the counts is obtained.

The location of tokens or words is completely ignored when this technique is used in corpora. If a word is the entirety of a phrase, it will still be assigned a frequency of one; this is a significant disadvantage of the count vectorizer. As a result, the introduction of tfidf is necessary. A count matrix is transformed into a normalized tf: term-frequency or tf-idf: term-frequency times matrix. The inverse document-frequency representation is accomplished using the TfidfTransformer. Equation 1 specifies the formula for calculating the tf-idf for a term 't' in a document 'd' in a document collection.

$$tf\text{-}id(t,d) = tf(t,d) * log\left(\frac{n}{df(t)+1}\right) \tag{1}$$

3.8 Data Modelling

The present work makes use of supervised machine-learning classification techniques. The accuracy, f1-score, f beta-score, precision, specificity, AUCROC, and recall of several models are compared.

- Logistic Regression: When the dependent variable or target is categorical, logistic regression is utilized. Here, we're forecasting if our news is phony, or whether it's real, or whether it's a 1. Because there are only two possible outcomes, binary logistic regression will be employed. The logistic regression value must be between 0 and 1 and cannot exceed this number to produce a curve in the "S" shape referred to as the Sigmoid function or logistic function. The logistic sigmoid function transforms the output, producing a probability value that may then be mapped to two or more discrete classifications. It will be beneficial to investigate a model's nonlinearity. A logit function is used in logistic regression. The model becomes non-linear as a result of this logic function. In logistic regression, the concept of a threshold value is used to determine the probability of a value is either 0 or 1. Values greater than the threshold value are assigned a value of 1, while values less than the threshold value are assigned a value of 0.

- KNN is a supervised machine learning technique that may be used to handle classification and regression problems. Due to the fact that it does not immediately learn from the training set, it is often referred to as a lazy learner algorithm. Rather than that, a dataset is stored, and action is performed on it during categorization. KNN can estimate which of two categories, Category A, Real, and Category B, Fake, a new data point will fall into.

- Random Forest: This is a commonly used machine learning algorithm that falls under the category of supervised learning. It is utilized in both classification and regression issues in Machine Learning. Random Forest is a machine learning algorithm that is based on the concept of ensemble learning. It solves a complex problem by combining numerous classifiers and optimizing the model's performance. The ultimate forecast is formed by integrating all of the trees' projections. Ensemble approaches are a method for collecting data in order to get a final conclusion.

- Decision Tree: A decision tree is a subset of supervised machine learning in which data is continually partitioned by a defined parameter. A decision tree's internal nodes reflect dataset properties; branches indicate decision rules and a leaf node represents the outcome. The tree is divided into two sections, which we refer to as decision nodes and leaves. The leaves denote the final outcomes, and the decision nodes denote the data splitting points. The tests are run in accordance with the data's properties. After posing a question, the tree is separated into yes or no responses. It is a graphical depiction that is used to determine all viable solutions to a problem/decision given certain conditions. The CART algorithm is utilized. CART stands for Classification and Regression Tree algorithm.

- Support Vector Machine (SVM): SVM is a supervised machine learning technique that is used to solve classification and regression problems. Its objective is to determine the ideal line or decision boundary (hyperplane) for classifying n-dimensional space. A hyperplane is a mathematical term for this optimal choice boundary. The SVM algorithm determines the hyperplane's extreme points/vectors. This type of vector is referred to as a support vector. The figure below depicts how a decision boundary or hyperplane is used to classify two distinct groups. Our dataset is referred to be linearly separable data since it can be divided into two groups using a single straight line, and the classifier utilized is a Linear SVM classifier.

4 Result Analysis

Binary classification has four possible types of results, given by a confusion matrix.

1. True Negatives (TN): correctly predicted negatives (zeros)

2. True Positives (TP): correctly predicted positives (ones)

3. False Negatives (FN): incorrectly predicted negatives (zeros)

4. False Positives (FP): incorrectly predicted positives (ones)

For the dataset we employed, the following result is obtained:

1. TP: Out of 5200 samples, 2502 samples were real and were predicted real.

2. TN: 2496 samples were fake and were predicted fake.

3. FP: 62 samples were fake and were predicted real.

4. FN: 40 samples were real and were predicted fake.

To measure the performance of various machine learning algorithms, we used the following indicators of binary classifiers:

- Accuracy (eq. 2): It is the most often used categorization metric, and it is straightforward to grasp.

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \tag{2}$$

- Precision (eq. 3): The precision of a forecast is an acceptable option of evaluation metric when we want to be certain about our prediction.

$$Precision = (TP)/(TP + FP) \tag{3}$$

- Recall (eq. 4): When it comes to diagnostic accuracy, recall, also known as sensitivity or true positive rate, is defined as the ratio of true positives to actual positives.

$$Recall = (TP)/(TP + FN) \tag{4}$$

- F1 Score (eq. 5): It is a number between 0 and 1, and it is also referred to as the harmonic mean of precision and recall, among other things. The F1 score assigns equal importance to precision and recall, resulting in significant difficulty.

$$F1\ Score = (TP)/(TP + 1/2(FP + FN)) \tag{5}$$

- F-Beta Score (eq. 6): It creates a weighted F1 metric as below where beta manages the trade-off between precision and recall.

$$F_\beta\ Score = (1 + \beta 2) * (precision * recall)/((\beta 2 * precision) + recall) \tag{6}$$

- Specificity (eq. 7): The specificity (or rate of true negatives) is the ratio of true negatives to actual negatives.

$$Specificity = TN/FN + TN \tag{7}$$

- AUC-ROC Curve: AUC (Area Under The Curve) - Receiver Operating Characteristics (ROC) curve is a performance statistic for classifying problems over several thresholds. The receiver operating characteristic curve (ROC) represents the probability curve, whereas the area under the curve (AUC) represents the degree or measure of separability. It demonstrates the model's capacity to distinguish between classes. The greater the AUC, the more accurate predictions for identifying 0s as 0s and 1s as 1s are made.

For the purpose of automating machine learning workflows, a machine learning pipeline is utilized. Their business model incorporates the process of transforming a sequence of data and correlating it with one another in a model that is capable of being tested and evaluated in order to arrive at a result, which may be positive or negative. The process of training a model typically takes place in a series of stages known as a pipeline. They are carried out over and over again in order to accomplish developing an effective algorithm and continually improve the model's precision.

Joblib is a SciPy module that assists with pipelining. Additionally, it facilitates the saving and loading of objects that make use of Numpy data. The joblib API serializes Python objects to NumPy arrays efficiently. joblib.dump() and joblib.load() acts as a replacement for pickle when dealing with large amounts of data, most notably large Numpy arrays. Front-end web development was carried out using HTML and CSS, while back-end web development was carried out using Python. Flask was used to deploy the website.

4.1 Fake News Detection

As shown in Table 1, Logistic Regression achieves the highest accuracy, 98.03 percent. SVM came in second place with an accuracy of 97.92 percent.

4.2 Fraudulent Job-Postings Detection

Datasets from Kaggle's datasets were used to predict fraudulent job posts. In the CSV file, the following categories were used to group the data for job postings: job id, title, location, department, salary range, company profile, description, requirements, benefits,

Table 1: Fake news performance parameters

Model	Accuracy (%)	F1-Score	F-beta Score	AUC	Specificity	Recall	Precision	Log Loss
Logistic Regression	98.03	81.90	91.68	84.63	69.53	100	76	5.38
K-Nearest Neighbors	80.84	80.78	81.61	85.71	79.43	95.24	93.65	6.61
Random Forest	82.82	82.83	83.51	81.93	81.71	95.24	93.65	5.93
Decision Trees	95.67	95.75	95.54	95.66	96.09	95.24	93.65	1.49
Support Vector Machine	97.92	97.95	97.71	97.91	98.36	97.46	96.79	0.71

telecommuting, company logo (if present), questions (if present), employment type, required experience, required education, industry, function, and fraudulent. Our concept was put into practice using Google Colaboratory (Colab) Notebooks. The dataset has 18 columns and 17880 rows. All fraudulent job profiles have a value of 1, while all legitimate job profiles have a value of 0.

Table 2: Fraudulent job-postings performance parameters

Model	Accuracy (%)	F1-Score	F-beta Score	AUC	Specificity	Recall	Precision	Log Loss
Logistic Regression	69.08	69.21	68.57	69.62	70.31	67.88	62.59	10.07
K-Nearest Neighbors	93.30	93.65	90.22	92.52	100	86.76	88.07	23.24
Random Forest	99.71	99.71	99.54	100	100	99.43	99.42	9.84
Decision Trees	97.72	99.71	99.54	53.56	100	99.43	99.42	9.82
Support Vector Machine	73.50	73.57	71.98	76.53	76.40	70.79	65.60	9.15

From Table 2, it is noted that the Random Forest Supervised Classification Algorithm has the greatest accuracy of 99.71%. In comparison to other algorithms, several evaluation criteria were also higher. With other criteria like f1-score, f-beta score, specificity, recall, precision, and log loss having the same value, decision trees provided the next-best accuracy of 97.72%.

4.3 Fraudulent Covid-19 Information Detection

Datasets from Kaggle were gathered for the Covid-19 and fraudulent job post prediction. The data for the Covid-19 dataset was organized into the following categories in the CSV file: title, text, source, and label. Our concept was put into practice using Google Colaboratory (Colab) Notebooks. There are 4 columns and 1164 rows in the dataset. Title, text, source, and label are the many columns of data. The website data that has been scraped is available in raw text form. Prior to analysis or model fitting, this data needs to be cleaned. As was done earlier in "Fake News Detection", this entails using the regex library, performing tokenization, deleting stop words, lemmatization, and case normalization.

From Table 3, it is noted that the Support Vector Machine (SVM) Supervised Classification Algorithm has the greatest accuracy of 78.69%. F1-score and F-beta score were two additional evaluation measures that outperformed existing methods. With other measures, such as recall and precision, having a greater value than SVM, KNN provided the next-best accuracy of 73.88%.

Table 3: Fraudulent Covid-19 information performance parameters

Model	Accuracy (%)	F1-Score	F-beta Score	Specificity	Recall	Precision
Logistic Regression	54.98	54.98	54.98	100	100	53
K-Nearest Neighbors	73.88	73.88	73.88	100	95	83
Random Forest	63.57	63.57	63.57	100	90	71
Decision Trees	60.82	60.82	60.82	100	90	71
Support Vector Machine	78.69	78.69	78.69	100	70.79	65.60

4.4 Forged Image Detection

Datasets from GitHub and Kaggle were used to gather information for faked image detection. Data were separated into training and testing folders. Two folders—one containing phony photographs and the other real images—were contained inside the training folder. Comparably, the testing folder contained phony images and a separate folder of real images.

Convolutional Neural Networks (CNN or ConvNet) were used in this study to classify images as fake or real. Convolutional Neural Network classifications of images start with an input image, process it, and assign it to one of several categories. It is a deep learning system that recognizes photos, groups them based on shared characteristics, and then identifies objects in the images. CNN employs a distinctive aspect of an image to identify it. CNN is superior to older neural networks for image processing. CNN and ANN are comparable. CNN first gathers information using layers before using artificial neural networks to extract some of the image's properties. A high-level neural network API called Keras runs on top of TensorFlow. The implementation of that algorithm is provided by the Python module MTCNN. We will utilize Multi-task cascade CNN (MTCNN), which is based on deep learning because it is thought to be the best method for detecting faces in images.

First, we go through the folder iteratively, find and crop faces, and then save them in the appropriate folder. In this case, we'll be using the image-processing library Pillow. The most crucial module of PIL (Python Image Library) is a library named Pillow that operates on top of PIL. But after 2011, PIL module has not been used. Pillow offers a variety of features that are compatible with all popular operating systems and even support Python 3. It is compatible with a wide range of image formats, including "bmp", "gif", "ppm", "tiff", "jpeg", and "png". The pillow library utilizes an image class to display the image. The image module in Pillow comes with built-in features including the ability to load or generate new images, among others.

The code that begins with "from PIL import Image" imports the image module from the pillow before calling Image.open() and supplying the image filename as a parameter. Fake photos can be identified by executing the detect faces() function once the model has been loaded and configured. This gives a list of dictionary objects that each have a number of keys for the information about each face that was detected. For example, the dictionary object "box" has the x and y coordinates of the bottom left corner of the bounding box, along with its width and height. and the image will then be cropped.

A function in the pillow named crop() is used to crop a picture's rectangular section. It takes a four-element tuple as input and returns the cropped rectangle portion of the image. We are transforming the created image from an array back to a NumPy array and saving it.

Three convolution layers, three max-pool layers, one flattening layer, and ultimately an output layer with sigmoid activation are included in this classifier. With the sequential API, a model is built. Sequential is one of the most used Keras models. The sequential API allows us to stack layers one on top of the other. A Sequential object is made, then layers are added using the add function to form a convolutional neural network. The method above builds a Sequential object first, then adds some convolutional, max-pooling, and dropout layers. It then flattens the output and sends it to a final dense layer before sending it to our output layer. The CONV layer is used to compute the result of neurons connected to local regions in the input. Each neuron computes a dot product between its weights and the local area in the input volume that it is connected to. The downsampling operation for the POOL layer is carried out along the spatial dimensions (width, height). Breaking up linearity is done with an activation function. We will employ the activation functions of sigmoid and ReLU.

- Padding: Convolutional operations as mentioned above decrease the size of the image that's why we apply Padding to preserve our input size.

- Pooling: This layer is used for reducing parameters and the computation process. Here, in this pooling process, max pooling will be used.

- Flattening: Flattening is taking a matrix coming from convolutional and pooling processes and turning it into a one-dimensional array. This is important because the input of a fully-connected layer or Artificial Neural Network consists of a one-dimensional array.

We set up the learning process before we begin training the model. The optimizer, a loss function, and, optionally, certain metrics like accuracy, must all be specified. Because optimizers employ gradients to update the weights, they can reduce the loss function, which assesses how effectively our model meets the specified objective. Adam Optimizer has been utilized. Adam is an optimization technique that dynamically modifies the learning rate.

By combining multiple orientations of existing data, augmentation produces new information; overfitting is avoided. Before supplying the photos to the model as input during training or model evaluation, the pixel values in the image should be scaled. Image augmentation is possible using the imageDataGenerator function. Before modeling, the ImageDataGenerator class in Keras provides a variety of methods for scaling variables in the image dataset. The class encapsulates the picture collection before returning photos in batches and performing the necessary scaling operations as needed.

Our model is now prepared for training. In order to train the model, we used a fit generator and gave train data, the number of epochs, and the batch size to our generator.

To track the loss and accuracy on both sets, we also handed it a validation set, which is currently test data. It will also pass if Steps per epoch, a requirement when using a generator, is set to the length of the training set and Validation steps is set to the length of test data. Epoch refers to the overall quantity of times the algorithm examines the full data set. We can split up an epoch into smaller pieces if it is too big to run all at once. A collection of these pieces is referred to as a batch.

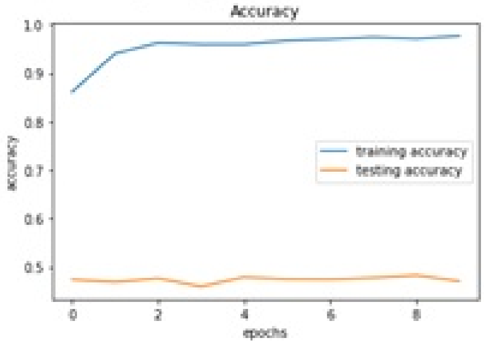


Figure 3: MTCNN: Variation of accuracy over no. of epochs

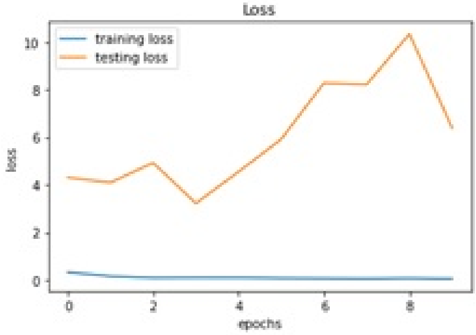
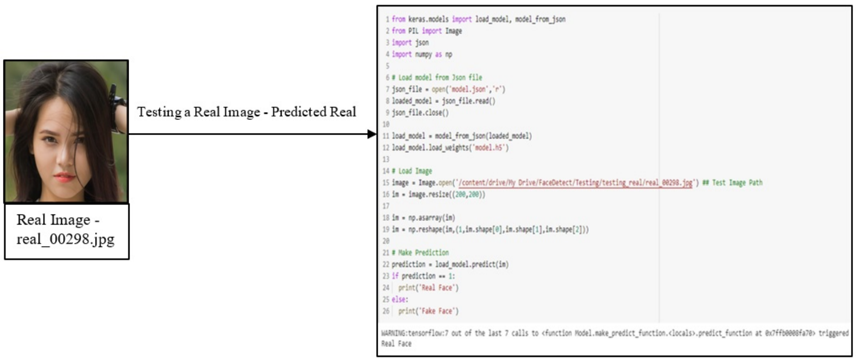


Figure 4: MTCNN: Variation of loss over no. of epochs

For describing any model in JSON format, Keras provides the to json() function. A function named model from json() is used to load this information once it has been saved to a file and creates a new model from the JSON file. The weights are then immediately saved using the load weights() symmetrical function and saved directly using the save weights() method. The complete model is then written to model.json in JSON format. A new model is built when the model and weight information are loaded from the saved files. Later, the model can be loaded by executing the load model() function with the filename as an argument. The model with the same architecture and weights is returned by the function. By using the Image.open() function, which returns a value of the picture object data type, the image is loaded. It recognizes the file type automatically.

Figure 3 shows the accuracy over a number of epochs and Figure 4 shows the loss over



```
from keras.models import load_model, model_from_json
from PIL import Image
import json
import numpy as np

# Load model from Json file
json_file = open('model.json','r')
loaded_model = json_file.read()
json_file.close()

load_model = model_from_json(loaded_model)
load_model.load_weights('model.h5')

# Load Image
image = Image.open('/content/drive/My Drive/FaceDetect/Testing/testing_real/real_00298.jpg') ## Test Image Path
im = image.resize((200,200))

im = np.asarray(im)
im = np.reshape(im,(1,im.shape[0],im.shape[1],im.shape[2]))

# Make Prediction
prediction = load_model.predict(im)
if prediction == 1:
  print('Real Face')
else:
  print('Fake Face')
```

```
WARNING:tensorflow:7 out of the last 7 calls to <function Model.make_predict_function.<locals>.predict_function at 0x7ffb0008fa70> triggered
Real Face
```

Figure 5: MTCNN:Testing a real image - Predicated real

a number of epochs. The dataset comprises real and fake images as shown in Figure 5. We achieved an accuracy of 97.69% with the use of MTCNN.

5 Recommendation System

Kaggle datasets were used for the data collection for the recommendation algorithm. The link, text, title, date, keywords, summary, and title summary columns of the recommendation were used to group the data together. The application of our concept was carried out using Google Colaboratory (Colab) Notebooks. Imported packages included Sklearn, Numpy, Pandas, Matplotlib, NLTK, Regex, Random, and String. By adding a new column id, eliminating duplicates, and deleting any null values that were previously included in the dataset, the dataset is transformed. Id, date, title, text, and link are some of the columns in the modified dataset, which has 1496 rows and 5 columns.

By invoking the modules in the NLTK library, the cleaning and pre-processing stage entails changing the data of the "text" column to lowercase and eliminating stop words and punctuation. By first compiling the HTML tag patterns using the regex library's compile function and then replacing any occurrences with spaces, HTML tags are eliminated from the dataset.

We used the Term frequency-inverse document frequency (TF-IDF) encoding method for our dataset because it gives each term a weight based on how important it is to the document. The more often a term appears, the bigger its weight. It also provides each item with a weighted inverse of how often that particular term appears in the dataset. So, it draws attention to rare words in the entire dataset but is essential to the text at hand.

The parameters for Tfidf vectorizer are:

- Analyser: The analyzer is used to decide if the feature should be made up of n-grams of individual words and characters. In our case, the thing is words.

- ngram_rangetuple (min_n, max_n): The lower and upper limits of the range of n-values to be extracted for different n-grams. All n values between min_n and max_n are used. In our case, it's in the ballpark (1,3).

- max_dffloat or int: When building the vocabulary, don't include words that appear in a document more than the given threshold. In this case, the threshold is 0.8.

- min_dffloat or int: When building the vocabulary, don't include words that appear in fewer than the given number of documents. This number is also known as the cut-off. Here, its value is 0.0.

- use_idfbool, default = True: Turn on inverse-document-frequency reweighting.

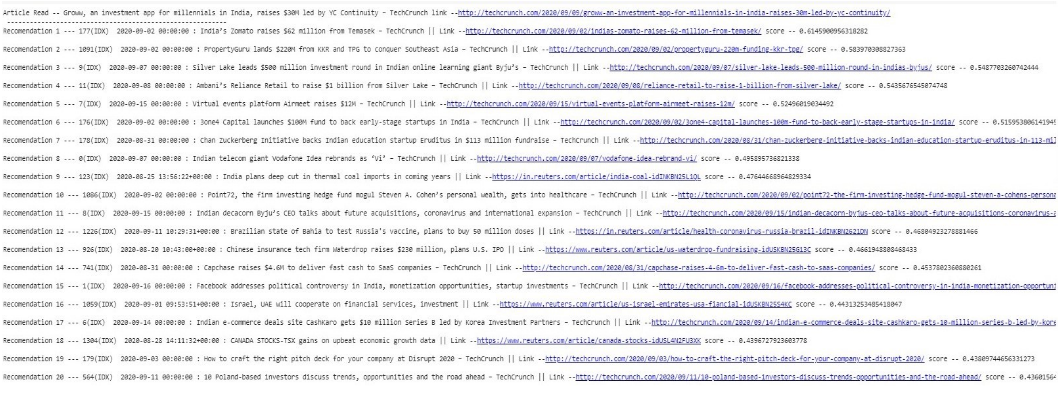
Now, we calculated how relevant or similar one document was to another. Here, each item is stored as a vector of its attributes (also vectors) in an n-dimensional space, and the angles between the vectors are used to figure out how similar they are. The user's likes and dislikes are measured by taking the cosine of the angle between the user profile vector (Ui) and the document vector. In our case, it's the angle between two document vectors. The reason for using cosine is that the cosine value increases as the angle between two vectors decreases, which means that the vectors are more alike.

When making suggestions for a user, a recommender system has to choose between two ways to send information:

- Exploitation: The system picks documents that are similar to the ones the user has already chosen.

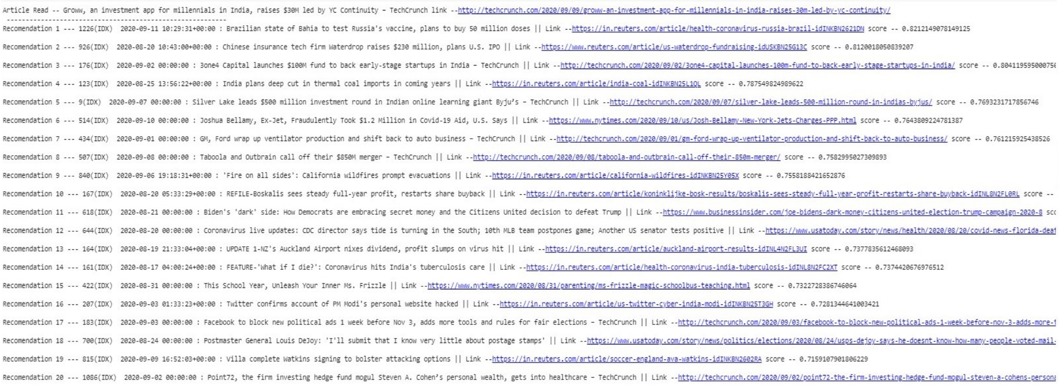
- Exploration: The system picks documents for which the user profile doesn't give enough information to predict how the user will react.

We used the exploitation technique as shown in fig. 6 and fig. 7.



```
Article Read -- Groww, an investment app for millennials in India, raises $30M led by YC Continuity - TechCrunch link --http://techcrunch.com/2020/09/09/groww-an-investment-app-for-millennials-in-india-raises-30m-led-by-yc-continuity/
 ------------------------------------------------------------
Recomendation 1 --- 177(IDX)  2020-09-02 00:00:00 : India's Zomato raises $62 million from Temasek - TechCrunch || Link --http://techcrunch.com/2020/09/02/indias-zomato-raises-62-million-from-temasek/ score -- 0.6145900956318282
Recomendation 2 --- 1091(IDX)  2020-09-02 00:00:00 : PropertyGuru lands $220M from KKR and TPG to conquer Southeast Asia - TechCrunch || Link --http://techcrunch.com/2020/09/02/propertyguru-220m-funding-kkr-tpg/ score -- 0.583970308827363
Recomendation 3 --- 9(IDX)  2020-09-07 00:00:00 : Silver Lake leads $500 million investment round in Indian online learning giant Byju's - TechCrunch || Link --http://techcrunch.com/2020/09/07/silver-lake-leads-500-million-round-in-indias-byjus/ score -- 0.5487703260742444
Recomendation 4 --- 11(IDX)  2020-09-08 00:00:00 : Ambani's Reliance Retail to raise $1 billion from Silver Lake - TechCrunch || Link --http://techcrunch.com/2020/09/08/reliance-retail-to-raise-1-billion-from-silver-lake/ score -- 0.5435676545074748
Recomendation 5 --- 7(IDX)  2020-09-15 00:00:00 : Virtual events platform Airmeet raises $12M - TechCrunch || Link --http://techcrunch.com/2020/09/15/virtual-events-platform-airmeet-raises-12m/ score -- 0.52496019034492
Recomendation 6 --- 176(IDX)  2020-09-02 00:00:00 : 3one4 Capital launches $100M fund to back early-stage startups in India - TechCrunch || Link --http://techcrunch.com/2020/09/02/3one4-capital-launches-100m-fund-to-back-early-stage-startups-in-india/ score -- 0.51595380614194
Recomendation 7 --- 178(IDX)  2020-08-31 00:00:00 : Chan Zuckerberg Initiative backs Indian education startup Eruditus in $113 million fundraise - TechCrunch || Link --http://techcrunch.com/2020/08/31/chan-zuckerberg-initiative-backs-indian-education-startup-eruditus-in-113-mil
Recomendation 8 --- 0(IDX)  2020-09-07 00:00:00 : Indian telecom giant Vodafone Idea rebrands as 'Vi' - TechCrunch || Link --http://techcrunch.com/2020/09/07/vodafone-idea-rebrand-vi/ score -- 0.495895736821338
Recomendation 9 --- 123(IDX)  2020-08-25 13:56:22+00:00 : India plans deep cut in thermal coal imports in coming years || Link --https://in.reuters.com/article/india-coal-idINKBN25L1QL score -- 0.47644668964829334
Recomendation 10 --- 1086(IDX)  2020-09-02 00:00:00 : Point72, the firm investing hedge fund mogul Steven A. Cohen's personal wealth, gets into healthcare - TechCrunch || Link --http://techcrunch.com/2020/09/02/point72-the-firm-investing-hedge-fund-mogul-steven-a-cohens-persona
Recomendation 11 --- 8(IDX)  2020-09-15 00:00:00 : Indian decacorn Byju's CEO talks about future acquisitions, coronavirus and international expansion - TechCrunch || Link --http://techcrunch.com/2020/09/15/indian-decacorn-byjus-ceo-talks-about-future-acquisitions-coronavirus-a
Recomendation 12 --- 1226(IDX)  2020-09-11 10:29:31+00:00 : Brazilian state of Bahia to test Russia's vaccine, plans to buy 50 million doses || Link --https://in.reuters.com/article/health-coronavirus-russia-brazil-idINKBN2621DN score -- 0.46804923278881466
Recomendation 13 --- 926(IDX)  2020-08-20 10:43:00+00:00 : Chinese insurance tech firm Waterdrop raises $230 million, plans U.S. IPO || Link --https://www.reuters.com/article/us-waterdrop-fundraising-idUSKBN25G13C score -- 0.4661948808468433
Recomendation 14 --- 741(IDX)  2020-08-31 00:00:00 : Capchase raises $4.6M to deliver fast cash to SaaS companies - TechCrunch || Link --http://techcrunch.com/2020/08/31/capchase-raises-4-6m-to-deliver-fast-cash-to-saas-companies/ score -- 0.4537802360880261
Recomendation 15 --- 1(IDX)  2020-09-16 00:00:00 : Facebook addresses political controversy in India, monetization opportunities, startup investments - TechCrunch || Link --http://techcrunch.com/2020/09/16/facebook-addresses-political-controversy-in-india-monetization-opportuni
Recomendation 16 --- 1059(IDX)  2020-09-01 09:53:51+00:00 : Israel, UAE will cooperate on financial services, investment || Link --https://www.reuters.com/article/us-israel-emirates-usa-fiancial-idUSKBN25S4KC score -- 0.44313253485418047
Recomendation 17 --- 6(IDX)  2020-09-14 00:00:00 : Indian e-commerce deals site CashKaro gets $10 million Series B led by Korea Investment Partners - TechCrunch || Link --http://techcrunch.com/2020/09/14/indian-e-commerce-deals-site-cashkaro-gets-10-million-series-b-led-by-kore
Recomendation 18 --- 1304(IDX)  2020-08-28 14:11:32+00:00 : CANADA STOCKS-TSX gains on upbeat economic growth data || Link --https://www.reuters.com/article/canada-stocks-idUSL4N2FU3XK score -- 0.4396727923603778
Recomendation 19 --- 179(IDX)  2020-09-03 00:00:00 : How to craft the right pitch deck for your company at Disrupt 2020 - TechCrunch || Link --http://techcrunch.com/2020/09/03/how-to-craft-the-right-pitch-deck-for-your-company-at-disrupt-2020/ score -- 0.43809744656331273
Recomendation 20 --- 564(IDX)  2020-09-11 00:00:00 : 10 Poland-based investors discuss trends, opportunities and the road ahead - TechCrunch || Link --http://techcrunch.com/2020/09/11/10-poland-based-investors-discuss-trends-opportunities-and-the-road-ahead/ score -- 0.436015
```

Figure 6: Recommended articles in case of min_df=0.1



```
Article Read -- Groww, an investment app for millennials in India, raises $30M led by YC Continuity – TechCrunch link --http://techcrunch.com/2020/09/09/groww-an-investment-app-for-millennials-in-india-raises-30m-led-by-yc-continuity/
 ----------------------------------------------------------
Recomendation 1 --- 1226(IDX)  2020-09-11 10:29:31+00:00 : Brazilian state of Bahia to test Russia's vaccine, plans to buy 50 million doses || Link --https://in.reuters.com/article/health-coronavirus-russia-brazil-idINKBN2621DN score -- 0.8212149078149125
Recomendation 2 --- 926(IDX)  2020-08-20 10:43:00+00:00 : Chinese insurance tech firm Waterdrop raises $230 million, plans U.S. IPO || Link --https://www.reuters.com/article/us-waterdrop-fundraising-idUSKBN25G13C score -- 0.8120018050839207
Recomendation 3 --- 176(IDX)  2020-09-02 00:00:00 : 3one4 Capital launches $100M fund to back early-stage startups in India – TechCrunch || Link --http://techcrunch.com/2020/09/02/3one4-capital-launches-100m-fund-to-back-early-stage-startups-in-india/ score -- 0.80411959500075
Recomendation 4 --- 123(IDX)  2020-08-25 13:56:22+00:00 : India plans deep cut in thermal coal imports in coming years || Link --https://in.reuters.com/article/india-coal-idINKBN25L1QL score -- 0.787549824989622
Recomendation 5 --- 9(IDX)  2020-09-07 00:00:00 : Silver Lake leads $500 million investment round in Indian online learning giant Byju’s – TechCrunch || Link --http://techcrunch.com/2020/09/07/silver-lake-leads-500-million-round-in-indias-byjus/ score -- 0.769323171785674б
Recomendation 6 --- 514(IDX)  2020-09-10 00:00:00 : Joshua Bellamy, Ex-Jet, Fraudulently Took $1.2 Million in Covid-19 Aid, U.S. Says || Link --https://www.nytimes.com/2020/09/10/us/Josh-Bellamy-New-York-Jets-Charges-PPP.html score -- 0.7643809224781387
Recomendation 7 --- 434(IDX)  2020-09-01 00:00:00 : GM, Ford wrap up ventilator production and shift back to auto business – TechCrunch || Link --http://techcrunch.com/2020/09/01/gm-ford-wrap-up-ventilator-production-and-shift-back-to-auto-business/ score -- 0.761215925438526
Recomendation 8 --- 507(IDX)  2020-09-08 00:00:00 : Taboola and Outbrain call off their $850M merger – TechCrunch || Link --http://techcrunch.com/2020/09/08/taboola-and-outbrain-call-off-their-850m-merger/ score -- 0.7582995027309893
Recomendation 9 --- 840(IDX)  2020-09-06 19:18:31+00:00 : 'Fire on all sides': California wildfires prompt evacuations || Link --https://in.reuters.com/article/california-wildfires-idINKBN25Y05X score -- 0.7558188421652876
Recomendation 10 --- 167(IDX)  2020-08-20 05:33:29+00:00 : REFILE-Boskalis sees steady full-year profit, restarts share buyback || Link --https://in.reuters.com/article/koninklijke-bosk-results/boskalis-sees-steady-full-year-profit-restarts-share-buyback-idINL8N2FL0RL score --
Recomendation 11 --- 618(IDX)  2020-08-21 00:00:00 : Biden's 'dark' side: How Democrats are embracing secret money and the Citizens United decision to defeat Trump || Link --https://www.businessinsider.com/joe-bidens-dark-money-citizens-united-election-trump-campaign-2020-8 sco
Recomendation 12 --- 644(IDX)  2020-08-20 00:00:00 : Coronavirus live updates: CDC director says tide is turning in the South; 10th MLB team postpones game; Another US senator tests positive || Link --https://www.usatoday.com/story/news/health/2020/08/20/covid-news-florida-deat
Recomendation 13 --- 164(IDX)  2020-08-19 21:33:04+00:00 : UPDATE 1-NZ's Auckland Airport nixes dividend, profit slumps on virus hit || Link --https://in.reuters.com/article/auckland-airport-results-idINL4N2FL3UI score -- 0.7377835612468093
Recomendation 14 --- 161(IDX)  2020-08-17 04:00:24+00:00 : FEATURE-'What if I die?': Coronavirus hits India's tuberculosis care || Link --https://in.reuters.com/article/health-coronavirus-india-tuberculosis-idINL8N2FC2XT score -- 0.7374420676976512
Recomendation 15 --- 422(IDX)  2020-08-31 00:00:00 : This School Year, Unleash Your Inner Ms. Frizzle || Link --https://www.nytimes.com/2020/08/31/parenting/ms-frizzle-magic-schoolbus-teaching.html score -- 0.7322728386746064
Recomendation 16 --- 207(IDX)  2020-09-03 01:33:23+00:00 : Twitter confirms account of PM Modi's personal website hacked || Link --https://in.reuters.com/article/us-twitter-cyber-india-modi-idINKBN25T3GH score -- 0.7281344641003421
Recomendation 17 --- 183(IDX)  2020-09-03 00:00:00 : Facebook to block new political ads 1 week before Nov 3, adds more tools and rules for fair elections – TechCrunch || Link --http://techcrunch.com/2020/09/03/facebook-to-block-new-political-ads-1-week-before-nov-3-adds-more-t
Recomendation 18 --- 700(IDX)  2020-08-24 00:00:00 : Postmaster General Louis DeJoy: 'I'll submit that I know very little about postage stamps' || Link --https://www.usatoday.com/story/news/politics/elections/2020/08/24/usps-dejoy-says-he-doesnt-know-how-many-people-voted-mail-
Recomendation 19 --- 815(IDX)  2020-09-09 16:52:03+00:00 : Villa complete Watkins signing to bolster attacking options || Link --https://in.reuters.com/article/soccer-england-ava-watkins-idINKBN2602RA score -- 0.7159107901806229
Recomendation 20 --- 1086(IDX)  2020-09-02 00:00:00 : Point72, the firm investing hedge fund mogul Steven A. Cohen’s personal wealth, gets into healthcare – TechCrunch || Link --http://techcrunch.com/2020/09/02/point72-the-firm-investing-hedge-fund-mogul-steven-a-cohens-persona
```

Figure 7: Recommended articles in case of min_df=0.2

A machine learning method for spotting bogus news was discussed here. We presented a suggestion mechanism as soon as we discovered the bogus news. Increased confidence in online digital media consumption can be achieved through the deployment of a user-friendly recommendation system.

6 Conclusion and Future Scope

It is crucial to follow basic rules and regulations when consuming information through digital media to ensure the safety and security of IoT systems. The affordability, convenience, and rapid transmission of information through social media have made it a popular source for news consumption. However, this convenience also makes spreading " fake news easier," which includes deliberately misleading and deceptive material. The widespread dissemination of false information can be detrimental to individuals and society, potentially leading to incorrect decisions and actions by IoT systems and devices. Consequently, detecting false news on social media has emerged as a critical research topic, requiring unique machine learning and deep learning techniques that can identify and mitigate these threats. In this article, we have discussed various methods for identifying fake news on social media, focusing on enhancing trust in digital media sources. Additionally, we have suggested a recommendation system to promote informed decision-making and enhance the security of IoT devices and systems.

In the future, we shall focus on the following aspects:

- The models of the work are based solely on English, which is one of its significant constraints. Other languages are not taken into account. Therefore, our strategy will be to create a model to identify fake news using multilingual data. We’ll strive to use a variety of tongues, such as French, Greek, Sanskrit, Hindi, German, and Spanish.

- Since predetermined datasets are typically used for detection, we built a dynamic dataset that would gather all real-time news and refresh the dataset with the most recent information. In the future, accuracy prediction will be based on a different algorithm.

- We will create a cross-platform, hybrid application based on our website (Android and iOS).

- Plans include creating fake video detection models and the online deployment of video and picture detection models.

- We aim to implement the recommendation algorithm online as a dashboard to display various news categories based on the user's interests.

- Additionally, a methodology for spotting fake news produced by social media platforms (WhatsApp, Twitter, Facebook, and Instagram) will be developed.

References

[1] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[2] Akshay Jain and Amey Kasbe. Fake news detection. In *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1–5. IEEE, 2018.

[3] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 312–320, 2019.

[4] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

[5] Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 836–837, 2019.

[6] Xichen Zhang and Ali A Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025, 2020.

[7] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

[8] Ray Oshikawa, Jing Qian, and William Yang Wang. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*, 2018.

[9] Roseline Oluwaseun Ogundokun, Micheal Olaolu Arowolo, Sanjay Misra, and Idowu Dauda Oladipo. Early detection of fake news from social media networks using computational intelligence approaches. *Combating Fake News with Computational Intelligence Techniques*, pages 71–89, 2022.

[10] Rajesh Prasad, Akpan Uyime Udeme, Sanjay Misra, and Hashim Bisallah. Identification and classification of transportation disaster tweets using improved bidirectional encoder representations from transformers. *International Journal of Information Management Data Insights*, 3(1):100154, 2023.

[11] Adebayo Abayomi-Alli, Olusola Abayomi-Alli, Sanjay Misra, and Luis Fernandez-Sanz. Study of the yahoo-yahoo hash-tag tweets using sentiment analysis and opinion mining algorithms. *Information*, 13(3):152, 2022.

[12] Taiwo Olaleye, Adebayo Abayomi-Alli, Kayode Adesemowo, Oluwasefunmi Tale Arogundade, Sanjay Misra, and Utku Kose. Sclavoem: hyper parameter optimization approach to predictive modelling of covid-19 infodemic tweets using smote and classifier vote ensemble. *Soft Computing*, pages 1–20, 2022.

[13] Ada Peter, Rose Omole, Sanjay Misra, Lalit Garg, and Jonathan Oluranti. Machine learning approaches for classifying the peace-war orientations of global news organizations' social media posts. In *Information Systems and Management Science: Conference Proceedings of 4th International Conference on Information Systems and Management Science (ISMS) 2021*, pages 301–317. Springer, 2022.

[14] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405, 2019.

[15] Shivam B Parikh and Pradeep K Atrey. Media-rich fake news detection: A survey. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 436–441. IEEE, 2018.

[16] Mykhailo Granik and Volodymyr Mesyura. Fake news detection using naive bayes classifier. In *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*, pages 900–903. IEEE, 2017.

[17] Aswini Thota, Priyanka Tilak, Simrat Ahluwalia, and Nibrat Lohia. Fake news detection: a deep learning approach. *SMU Data Science Review*, 1(3):10, 2018.

[18] Esther Omolara Abiodun, Abdulatif Alabdulatif, Oludare Isaac Abiodun, Moatsum Alawida, Abdullah Alabdulatif, and Rami S Alkhawaldeh. A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities. *Neural Computing and Applications*, 33(22):15091–15118, 2021.

[19] Olayiwola Tokunbo Taofeek, Moatsum Alawida, Abdulatif Alabdulatif, Abiodun Esther Omolara, and Oludare Isaac Abiodun. A cognitive deception model for generating fake documents to curb data exfiltration in networks during cyber-attacks. *IEEE Access*, 10:41457–41476, 2022.

[20] Usman Ahmed, Rutvij H Jhaveri, Gautam Srivastava, and Jerry Chun-Wei Lin. Explainable deep attention active learning for sentimental analytics of mental disorder. *Transactions on Asian and Low-Resource Language Information Processing*, 2022.

[21] Ashish Patel and Jigarkumar Shah. Sensor-based activity recognition in the context of ambient assisted living systems: A review. *Journal of Ambient Intelligence and Smart Environments*, 11(4):301–322, 2019.

[22] Ashish Patel and Jigarkumar Shah. Real-time human behaviour monitoring using hybrid ambient assisted living framework. *Journal of Reliable Intelligent Environments*, 6(2):95–106, 2020.

[23] Ashish D Patel and Jigarkumar H Shah. Performance analysis of supervised machine learning algorithms to recognize human activity in ambient assisted living environment. In *2019 IEEE 16th India Council International Conference (INDICON)*, pages 1–4. IEEE, 2019.

[24] Yogesh Jadhav and Harsh Mathur. Detection of breast cancer by using various machine learning and deep learning algorithms. In *Handbook of Machine Learning for Computational Optimization*, pages 51–70. CRC Press, 2021.

[25] Yogesh Jadhav, Vishal Patil, and Deepa Parasar. Machine learning approach to classify birds on the basis of their sound. In *2020 International Conference on Inventive Computation Technologies (ICICT)*, pages 69–73. IEEE, 2020.

[26] Aasim Khan, Gautam Worah, Mehul Kothari, Yogesh H Jadhav, and Anant V Nimkar. News popularity prediction with ensemble methods of classification. In *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2018.

[27] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, 2017.

[28] Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. Fake news detection using machine learning ensemble methods. *Complexity*, 2020, 2020.

[29] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921, 2019.

[30] Georgios Gravanis, Athena Vakali, Konstantinos Diamantaras, and Panagiotis Karadais. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128:201–213, 2019.

[31] Sneha Singhania, Nigel Fernandez, and Shrisha Rao. 3han: A deep neural network for fake news detection. In *International conference on neural information processing*, pages 572–581. Springer, 2017.

[32] Reem K Alkhodhairi, Shahad R Aljalhami, Norah K Rusayni, Jowharah F Alshobaili, Amal A Al-Shargabi, and Abdulatif Alabdulatif. Bitcoin candlestick prediction with deep neural networks based on real time data. *CMCCOMPUTERS MATERIALS & CONTINUA*, 68(3):3215–3233, 2021.

[33] Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S Yu. Ti-cnn: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749*, 2018.

[34] Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, and Raviraj Joshi. Evaluating deep learning approaches for covid19 fake news detection. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*, pages 153–163. Springer, 2021.

[35] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5644–5651, 2019.

[36] Xiaojun Chen, Shengbin Jia, and Yang Xiang. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141:112948, 2020.

[37] Saptarshi Sengupta, Sanchita Basak, Pallabi Saikia, Sayak Paul, Vasilios Tsalavoutis, Frederick Atiah, Vadlamani Ravi, and Alan Peters. A review of deep learning with special emphasis on architectures, applications and recent trends. *Knowledge-Based Systems*, 194:105596, 2020.

[38] Kunni Han. Personalized news recommendation and simulation based on improved collaborative filtering algorithm. *Complexity*, 2020, 2020.

[39] Umair Javed, Kamran Shaukat, Ibrahim A Hameed, Farhat Iqbal, Talha Mahboob Alam, and Suhuai Luo. A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET)*, 16(3):274–306, 2021.

[40] Shristi Shakya Khanal, PWC Prasad, Abeer Alsadoon, and Angelika Maag. A systematic review: machine learning based recommendation systems for e-learning. *Education and Information Technologies*, 25(4):2635–2664, 2020.

[41] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. *Recommender systems handbook*, pages 73–105, 2011.

[42] Ankit Kesarwani, Sudakar Singh Chauhan, and Anil Ramachandran Nair. Fake news detection on social media using k-nearest neighbor classifier. In *2020 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pages 1–4. IEEE, 2020.

[43] Rohit Kumar Kaliyar. Fake news detection using a deep neural network. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pages 1–7. IEEE, 2018.

[44] Njood Mohammed AlShariah, A Khader, and J Saudagar. Detecting fake images on social media using machine learning. *International Journal of Advanced Computer Science and Applications*, 10(12):170–176, 2019.

[45] Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. Detecting misleading information on covid-19. *Ieee Access*, 8:165201–165215, 2020.

[46] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[47] Shawni Dutta and Samir Kumar Bandyopadhyay. Fake job recruitment detection using machine learning approach. *International Journal of Engineering Trends and Technology*, 68(4):48–53, 2020.

[48] Muhammed Afsal Villan, A Kuruvilla, Johns Paul, and Eldo P Elias. Fake image detection using machine learning. *IRACST-International Journal of Computer Science and Information Technology & Security (IJCSITS)*, 2017.

[49] Suyanto Suyanto et al. Synonyms-based augmentation to improve fake news detection using bidirectional lstm. In *2020 8th International Conference on Information and Communication Technology (ICoICT)*, pages 1–5. IEEE, 2020.

[50] Hugo Queiroz Abonizio, Janaina Ignacio de Morais, Gabriel Marques Tavares, and Sylvio Barbon Junior. Language-independent fake news detection: English, portuguese, and spanish mutual features. *Future Internet*, 12(5):87, 2020.