Societal Transformation: AI and Big Data Journal
Vol: 2(2): 1-18, 2024
10.20547/aibd.242201

Fake News Detection with Urdu Data Set

Umer Bin Siddiq*
Muhammad Bilawal Raheem†

Abstract

This paper investigates how fake news in Urdu language can be identified on social media content. There has been increasing usage of social media and the news is ubiquitous and spread daily over the social media. It is very difficult to identify original news from a fake news. As we all know, billions of users are using social media from all over the world and they spread all types of news. It is very difficult to identify if a news is fake or real and this poses serious concerns as the ultimate objective of the original poster may be malicious. To resolve this issue, this paper used machine learning techniques for detection of fake news. Dataset of various types were analyzed, and then Fire2021 dataset from Kaggle was selected, and algorithms were trained to automatically detect a fake news. The proposed approach comprises multiple phases. The first phase comprises steps for preprocessing of the dataset. It attempts to validate the authenticity of news content. For this purpose, it employs several linguistic features. A number of various machine learning algorithms such as Multinomial Naive Bayes, Support Vector machine classifier (SVC), Bernoulli Naïve Bayes (BNB), Logistic Regression (LR), Decision Tree (DT), Random Forest(RF) and AdaBoost have been used over the fake news benchmark datasets. In addition, k-fold cross-validation is performed to obtain the best accuracy. The paper reports various performance metrics for these algorithms. Ada Boost is found to give the best results in terms of precision, recall, accuracy and F1-score.

Keywords: *fake news dataset, machine learning, decision tree, SVM, logistic regression*

1. Introduction

The task of automatically identifying false news is crucial and has found numerous applications. As these stories are frequently disseminated over social media platforms to a global audience, they may influence the perceptions of users. A political campaign, the induction of a conflict, or other goals can be achieved by the fake news and can drastically change public opinion. The spread of fake news has a severe effect on our society, political government, even it can lead to enthusiasm, anxiety, or action.

*Karachi Institute of Economics and Technology, Karachi
† Karachi Institute of Economics and Technology, Karachi

Verifying the veracity of news is crucial in stopping the spread of false information. Due to time restrictions, users tend to trust the captivating headline and image. Sensational headlines therefore lead to misinterpreted and fabricated facts. Identifying fake news is a difficult task. Fake news can be produced for various reasons, including manipulating public figures, or promoting skewed viewpoints to influence the course of significant political events. Another phenomenon observed in Covid-19 days is that people are battling with unparalleled anxiety and dependence on social media. However, this dependence has led to the rise of fake news further [5].

Before proceeding to the topic, let's discuss the recent interest in fake news detection. Several scholars have focused on the challenge of automatic fake news identification, particularly in the wake of the US 2016 presidential election. At that time, false news was extensively employed as a political tool. Recent overview of publications, for instance, discussed scientific concerns related to the identification of reporting false news [6-13]. Substantial collaborative projects were working in numerous languages, including for Arabic [12], Spanish [9], English [7, 14]. In terms of the quantity of datasets with annotations and the accessibility of NLP tools, research typically concentrate on languages with an abundance of resources, such as English [9], Spanish [10], German [15], and Chinese [8].

As the fake news is still spreading across the worldwide web and social media, all languages—including those with limited resources such as Urdu, need technical assistance for recognizing fake information. There are a number of challenges for fake news detection in Urdu. There are multiple meanings of a word in Urdu and the meaning is inferred from the context. This makes fake news detection very difficult. The second challenge is the shortage of reliable datasets in Urdu. In addition, the language has high misinformation susceptibility, political and religious sensitiveness, limited fact checking infrastructure, emotional appeal and virality.

Realizing this gap, this paper employs a dataset for Urdu language and performed various machine learning techniques for identification of fake news. The dataset comprises 1600 files that are labelled as genuine / fake. Each of the news is 100-500 words long. The major contributions of the paper is to compare various machine learning/ deep learning algorithms for detection of fake news for Urdu dataset. The novelty of the paper lies in employing these algorithms for Urdu language as very few work has been done on Urdu language. We only find FireFly as the reliable dataset [37]. A few studies are reported as discussed in literature review, but major work is lacking in this domain. There are a number of test cases where the proposed approach has application. For example, lot of contents are available in Urdu on social media. If a news is generated, it can be verified using the machine learning.

The rest of the sections of the paper have been organized as follows. The next section presents a literature review on the topic. This is followed by a proposed methodology. Then, we present the results. Finally, the paper concludes with limitations of proposed work and future directions.

2. Literature Review

There has been a lot of research done on fake news detection in English language. Primarily, it is observed that most of the fake news detection work is based on machine learning. An overview on various approaches for fake detection using machine learning can be found in [35]. In addition, deep learning has been used in various studies for fake news detection. We briefly discuss below various conventional and machine learning based approaches for fake news detection.

A number of conventional approaches have been proposed in literature. [22] employed features such as subjective nature, understanding, and knowledge traits to identify fake reviews on a dataset. Sentiment-encompassing words are used by [23] as a text feature. They discovered that tweets with higher sentiment words typically contain more unverified material. [24] employed topic modelling with Latent Dirichlet allocation (LDA) [25] to identify the authenticity of various topic over social media posts. The proposed approach by them, described web pages and profiles of the various news article distributors. Additionally, they looked at each news article's key phrases and word embeddings. The incorporation of words shows promising results in the identification of false news.

Several machine learning based approach for fake news identification are available in literature. In [36], a three pronged solution for fake news identification has been proposed. They used various machine learning classifiers for English news. The ML based fake news detection are specifically tuned for certain scenarios and only functions optimally on those specific situations[28]. It can be said that no single classifier could guarantee the optimal outcomes across all datasets. Researchers began employing deep-learning approaches to cover complex datasets. [29] proposed a deep convolutional neural network known as the fake news detection network (FNDNet) model. The data set from Kaggle was utilized. The accuracy of the suggested method was 98.36%. However, one drawback of the approach is that generalized text was overlooked [30]. [31] suggested a method for identifying fake news on Facebook that relies on Chrome extensions. They used classifiers from both deep learning and machine learning. Comparing the proposed method against machine learning classifiers, the maximum accuracy of 99.4% was obtained using long short term memory (LSTM) model. The method looks at shared posts and user profiles to identify fake news in real time on the user's Chrome environment.

A number of hybrid approaches have also been proposed in literature. In this direction, [32] presented TraceMiner, a method for identifying fake news. After obtaining a message trace, TraceMiner categorizes the message as genuine of fake. They classified the news using the hybrid LSTM-RNN model. Trueman et al. [33] proposed a novel approach in which they focused on making the model more widely applicable. Convolutional and recurrent neural networks were combined to create the hybrid model. The ISOT and FA-KES datasets are used in the experiments. Their accuracy in generalization was 50%.

For identifying fake news, Nasir et al. [34] combined convolutional neural network (CNN) and bi-directional LSTM (Bi-LSTM) model with attention mechanism. Vectorized text is generated by using GLOVE words embedding. They implemented their approach using the LIAR dataset.

Let's briefly talk about the work in languages besides the English for fake news identification. To detect fake reviews, [26] provided 234 variables that took linguistic and syntactic aspects into account. Leveraging the Italian social network dataset, [3] employed text, user-specific, and message-specific variables for hoax identification. They have used several models based on machine learning, such as linear and logistic regression to analyze data. Employing linear regression, the suggested method produced results with approximately 90% efficiency. Similar features i.e. text, consumer, and message are used in another study by [27] to assess the integrity of 489,330 twitter profiles. They employed a 10-fold cross-validation technique with random forest, decision trees, Naive Bayes, and feature-rank Naïve Bayes techniques. In [38], fake news detection of human and machine authored text was performed on 4 datasets in Urdu. Hierarchical approach to fake news detection was also done. In [39], transfer learning was used in the context of COVID 19 for detecting fake news in Urdu.

Table 1: Summary of Selected Literature for Fake News Detection

Reference	Paper Title	Novelty
[22]	Linguistic characteristics of shill reviews.	Employed features such as subjective nature, understanding, and knowledge traits
[23]	Information credibility on twitter.	Sentiment-encompassing words are used as a text feature
[24]	A first step towards combating fake news over online social media	Employed topic modelling with Latent Dirichlet allocation
[36]	Ternion: An autonomous model for fake news detection	A three pronged solution for fake news detection
[31]	Multiple features based approach for automatic fake news detection on social networks using deep learning	Chrome extension was used for fake news detection
[32]	Tracing fake-news footprints: Characterizing social media messages by how they propagate	LSTM-RNN was used for fake news detection
[33]	Attention-based C-BiLSTM for fake news detection	Convolutional and recurrent neural network was used for fake news detection
[34]	Fake news detection: A hybrid CNN-RNN based deep learning approach	Convolutional neural network and bi-directional LSM was used

Having reviewed the literature, it has been found that work on fake news detection over the Urdu dataset is very shallow. We don't find any significant work over this topic. We only find [37] as the reliable dataset for Urdu fake news detection. Additionally, dedicated fact check

platforms are very few for Urdu language. Hence, this work research paper compares fake news detection in Urdu language for various machine learning algorithms.

3. Methodology

This section presents the proposed methodology. We start with a brief discussion on the Urdu language.

3.1. Urdu language

Urdu is a popular language in sub-continent. The vast South Asian population has made Urdu an increasingly prevalent language in South Asia and around the world [16]. With a modified Perso-Arabic alphabet, Urdu is written in a context-sensitive Nastalitic writing style that is cursive. Urdu is distinct because it borrows colloquial terminology from South Asian local languages while maintaining a literary vocabulary derived from Persian and Arabic [17].

Lack of capitalization, the use of diacritical markings optionally, and space's unreliability as a word boundary marker are a few of the difficulties faced by Urdu computers [18-20]. When trying to infer a word's pronunciation in the absence of diacritical marks, context is essential. A general sentence in Urdu has the following arrangement: Subject-Object-Verb dialect [21].

3.2. Processing pipeline

For fake news detection, we employed a proper machine learning pipeline as shown in Figure 1.

Figure 1: Processing pipeline/ methodology for fake-news detection

Following are the major steps employed for the pipeline:

a) **Data acquisition:** The very first step is the acquisition of the data. We acquired the data from Kaggle. Fire2021 [37] Urdu dataset is to be used for training and evaluation. Table 1 shows the sample text from dataset.

Table 2: Sample Text from the dataset

سٹٹگارٹ (اُردو پوائنٹ اخبارتازہ ترین 03 دسمبر2018ء) جرمنی کی موٹر ساز کمپنی ڈیملر نے آئندہ سال سے چین میں امرکین ٹیرف کی وجہ سے الیکٹرک کاریں بند کرنے کا اعلان کیا ہے ڈیملر کے چین کے حوالے سے سُعبے کے سربراہ ہوبرٹس ٹروسکا نے بتایا کہ کمپنی آئندہ سال 2019 کے آخر تک چین میں الیکٹرک کاروں کی پیداوار بند کر دے گی جن میں مرسٹیز بینز ای کیو سی ماڈل بھی شامل ہے انکا مزید کہنا تھا کہ چین میں امرکین ٹیرف کی وجہ سے الیکٹرک کاریں جیسا کہ مرسٹیز بینز مے بیچ ماڈل کی طلب میں کمی آرہی ہے اس لیے یہ فیصلہ کیا گیا ہے	ماسکو (اُردو پوائنٹ اخبارتازہ ترین 17 دسمبر2018ء) روس کی ایروسپیس کمپنی یونائٹڈ ایئرکرافٹ کارپوریشن نے نئے مسافر بردار طیارہ بنانے کا اعلان کردیا جس پر 10 ارب روبل 150.5 ملین ڈالر لاگت آئے گی یہ طیارہ سابقہ طیاروں کی نسبت چوڑا ہوگا اس کی تجرباتی بنیادوں پر تیاری دسمبر 2019ء تک ہوسکے گی سرکاری ذرائع کے مطابق نئے مسافر بردار جیٹ طیارے کی چوڑائی پہلے سے موجود طیاروں کی نسبت زیادہ II-96-400 ہوگی اور یہ طویل فاصلے تک پرواز کرسکے گا تجرباتی بنیادوں پر تیار ہونے والے اس طیارے پر خرچ آنے والے 10 ارب روبل میں سے 7.6 ارب روبل طیارے پر جبکہ باقی اس کی انجینئرنگ ڈرائنگ اور دیگر شعبوں پر خرچ کیے جائیں گے طیارے کی تجرباتی بنیادوں پر تیاری دسمبر 2019 ء تک مکمل ہوسکے گی جبکہ تجرباتی پرواز نومبر 2019ء سے جنوری 2020ء کے درمیانی عرصے میں کی جاسکے گی طیارے کی کمرشل بنیادوں پر پیداوار 2020ء کے آغاز میں شروع ہوگی
لاہور(کامرس رپورٹر) لاہور چیمبر کے صدر الماس حیدر نے حکومت پر زور دیتے ہوئے کہا ہے کہ وہ کاروباری برادری کے لئے جدید ٹیکنالوجی اور ون ونڈو آپریشن سسٹم کے ذریعے پراپرٹی کی رجسٹریشن کے عمل میں آسانی پیدا کرنے کے حکومتی اقدام کو بہت سراہتے ہیں اور وہ حکومت پاکستان کے مشکور ہیں لاہور چیمبر کے صدر نے کہا کہ کاروباری برادری کے لئے پراپرٹی کی رجسٹریشن ایک بڑا مسئلہ تھا جیسے حکومت نے حل کر دیا ہے انھو ں نے کہا کہ انفراسٹرکچر کی مضبوطی زمینی حدود معلوم کرنے کے لئے الیکٹرونک ڈیٹا بیس معلومات میں شفافیت جغرافیائی کوریج زمینی تنازعات کا حل اور پراپرٹی رائٹس کے حصول تک سب کی یکساں رسائی جیسے مسائل کرنے سے کاروبار میں بہت اضافے کا امکان ہے لاہور چیمبر کے سینئر نائب صدر خواجہ شہزاد ناصر نے کہا ہے کہ پراپرٹی کی رجسٹریشن کے عمل میں حکومتی ٹھوس اور فوری اقدامات پر انتہائی شکر گزار ہیں	نیویارک (اُردو پوائنٹ اخبارتازہ ترین 08 دسمبر2018ء) امریکا میں سویابین کے نرخ پانچ ہفتے کی بلند ترین سطح پر رہے جس کی وجہ چین کے ساتھ تجارتی کشیدگی میں کمی کے باعث جنس کی برآمدات میں اضافے کی توقع ہے گندم اور مکئی کے نرخوں میں بھی اضافہ ریکارڈ کیا گیا شکاگو بورڈ آف ٹریڈمیں سویابین کے وقفے سے قبل جنوری کیلئے سودے 6.50 سینٹ بڑھ کر 9.16 ڈالر فی بشل طے پائے گندم کے مارچ کیلئے سودے 15.25 سینٹ بڑھ کر 5.31 ڈالر فی بشل کے قریب طے پائے مکئی کے مارچ کیلئے سودے 2.25 سینٹ اضافے کے ساتھ 3.85 ڈالر فی بشل طے پائے

b) **Pre-processing:** In the next step, we performed pre-processing of the dataset. The following are the major steps performed for pre-processing:
   - The noise in the data such as URL, websites, and other handles were removed from the dataset. Listing 1 shows the data cleaning step performed over the data.

```
def preprocess_text(text):
    cleantext = re.sub('\d','0',text)
    return cleantext
```

Listing 1: Data cleaning performed over the dataset

- Stop words were removed from the dataset. Table II shows the sample stop words in Urdu.
- As the dataset is in Urdu, it is irrelevant to perform case normalization.
- Outliers present in the dataset was removed

Table 3: Sample stop words used during preprocessing

آ	اس
آئی	اجنبی
آئیں	از
آئے	استعمال
آتا	اسی
آتی	اسے
آتے	البتہ
آخری	الف
آس	ان
آنا	اندر
آنی	انہوں
آنے	انہی
آپ	انہیں
	اور
آگے	اوپر
آیا	اپ
ابھی	اپنا
	اپنی

c) **Feature Engineering:** As the next step, we performed feature engineering before employing any model for classification. The objective is to have a usable representation for better results.

d) **Data Augmentation**: Data augmentation is an approach employed for generating synthetic data for a relatively smaller dataset.

e) **Data Split:** The data has been split into train-test data so that training is performed on 80% of the data and rest of 20% of the data is used for evaluation.

f) **Classification:** In the next step, various machine learning models have been applied for classification of the news as fake or genuine

3.3. Feature Engineering

The feature engineering is a very important step in our processing pipeline. So, we will discuss it in detail. After the data has been acquired from Kaggle, the dataset files contain textual information that we can use to create features to train the machine learning model. However, there is a challenge here. The patterns of the file don't contain sufficient discriminating information. Therefore, one needs to convert the news data into a better form. Listing 2 shows the feature extraction step. We extracted three types of n-gram: word n-grams, character n- grams and function n-grams from the dataset.

For example, the phrase: "Fake news spreads quickly" can yield following
bigrams: ["Fake news", "news spreads", "spreads quickly"]. The N-grams capture context in a

morphologically rich language. In addition, it handles complex word order and script. It helps preserve syntactic structure. It can handle exclamation, exaggeration as it is normal in a Urdu sentence.

```
def extract_features(text,cn,wn,fn):
    text = text.lower()
    #text=clean_text(text)
    features = []
    for n in wn:
        if n != 0:
            features.extend(wordNgrams(text,n))
    for n in cn:
        if n != 0:
            features.extend(charNgrams(text,n))
    for n in fn:
            if n != 0:
                features.extend(funcNgrams(text,n))
    return features
```

Listing 2: Feature extraction step

In some instances, the dataset may contain enough information to train the model to give better results. However, the input data should be converted. This allows us to determine the suitable model for features, which is primarily better and differs from a model that works on purely raw input data. Before feature engineering, we also clean the dataset as discussed earlier. This also includes removal of the outliers (carefully), removing irrelevant input data according to business logic.

3.4. Dataset

In this section, we describe the dataset used to evaluate a news as legitimate or fake. It is very important to understand the dataset structure which is used to train the model. The used dataset Fire2021 is categorized into 2 basic classes i.e. genuine or fake news. The dataset was selected because it has been used for research challenges / competition. The dataset comprises diverse data sources, and has accurate ground truth for model training. The dataset is well structured as well. It contains news articles from diverse sources and have various topics. Figure 2 shows the snapshot of the dataset. There are two folders corresponding to each label and there are around 800 files on each folder. There are two classes with the name from fold1 to fold2. This aids in the comparison and training of the machine learning classifier. In addition, for the

پاکستان کے وزیراعظم عمران خان سعودی عرب کے دارالحکومت ریاض میں ملک میں سرمایہ کاری کے حوالے سے سالانہ کانفرنس میں شرکت کر رہے ہیں حکومت پاکستان کا کہنا ہے کہ سعودی عرب نے پاکستان کو معاشی بحران سے نمٹنے میں مدد کے لیے ایک سال کے لیے تین ارب ڈالر دینے پر اتفاق کیا ہے دفتر خارجہ کی جانب سے منگل کی شب جاری ہونے والے اعلامیے میں بتایا گیا ہے کہ یہ فیصلہ وزیر اعظم عمران خان کی سعودی قیادت سے ملاقات کے بعد کیا گیا

عمران خان سعودی عرب میں سرمایہ کاری کے حوالے سے سالانہ کانفرنس میں شرکت کے لیے ریاض میں موجود ہیں

اعلامیے کے مطابق سعودی عرب پاکستان کو نہ صرف ایک سال کے لیے تین ارب ڈالر دے گا بلکہ تین سال تک مؤخر ادائیگیوں پر تقریباً نو ارب ڈالر مالیت تک کا تیل دینے پر بھی رضامند ہو گیا ہے دفتر خارجہ کا کہنا ہے کہ سعودی عرب کی جانب سے دیے جانے والے تین ارب ڈالر ادائیگیوں میں توازن لانے میں مدد دیں گے وزیراعظم ہاؤس کا کہنا ہے کہ سعودی عرب نے پاکستان میں آئل ریفائنری پراجیکٹ میں بھی دلچسپی کی تصدیق کی ہے اس کے علاوہ اعلامیہ میں بتایا گیا کہ سعودی حکومت بلوچستان میں معدنیات کی کانوں میں بھی دلچسپی رکھتی ہے اور اس سلسلے میں ان کی حکومت کا اعلیٰ وفد پاکستان کے دورے پر بھی آئے گا وزیراعظم عمران خان نے دورہ سعودی عرب میں ان کے ہمراہ وزیر خارجہ شاہ محمود قریشی وزیر خزانہ اسد عمر وزیر اطلاعات فواد چوہدری اور وزیراعظم کے مشیر برائے تجارت عبدالرزاق داؤد کے علاوہ چیئرمین بورڈ آف انویسٹمنٹ ہارون شریف بھی شامل ہیں کانفرنس میں شرکت سے پہلے عمران خان کہہ چکے ہیں کہ پاکستان کے معاشی بحران سے نمٹنے کے لیے وہ آئی ایم ایف سے قرض لینے کی بجائے دوست ملکوں کا دروازہ کھٹکھٹانے کو ترجیح دیں گے عمران خان نے وزارتِ عظمیٰ کا عہدہ سنبھالنے کے بعد اپنے پہلے دورے کے لیے سعودی عرب کا ہی انتخاب کیا تھا اور دورے کے بعد وزیر اطلاعات فواد چوہدری نے اسلام آباد میں پریس کانفرنس سے خطاب کرتے ہوئے کہا تھا کہ سعودی عرب وہ پہلا ملک ہے جسے پاکستان نے چین اور پاکستان کے درمیان جاری منصوبے سی پیک میں تیسرا پارٹنر بننے دعوت دی ہے تاہم بعد میں حکومت نے اپنے موقف کو تبدیل کیا تھا لیکن سعودی عرب کے ایک اعلیٰ سطح نے پاکستان کا دورہ کیا تھا ناکہ سرمایہ کاری کے منصوبوں پر بات ہو سکے اور اس وفد نے گوادر کا دورہ بھی کیا تھا لیکن اس دورے کے بعد دونوں ممالک کے ردعمیان سرمایہ کاری یا امدادی پیکج کے بارے میں کوئی واضح اعلان نہیں کیا گیا تھا

Figure 2: Snapshot of dataset

fake news, a CSV file is available that comprises metadata. This includes every single section and this file contain columns. These columns include filed, label, classid corresponding to label, salience, etc, and the format of the files is(txt format)

3.5. Implementation details

We will now discuss the implementation details. Table 4 provides the summary of implementation. We have used different machine learning models such as support vector machine, Naïve Bayes, logistic regression, decision tree, and Ada Boost. Training /test split was 80-20. I7 has been used for implementation.

Table 4: Implementation details

Models	Multinomial Naive Bayes SVC Bernoulli Naïve Bayes Logistic Regression Decision Tree Random Forest AdaBoost
Train/test split	80-20
Real News	600
Fake News	438
Libraries	Sk-learn, scipy, nltk
Hardware	I7 6-cores

3.5.1. Software details

For the experiment and implementation, we have used NLTK and SciPy for data pre-processing and feature extraction. Sometimes pre-processing becomes necessary to increase the accuracy of the model. The SciPy library has many methods already built-in to extract features. These methods are generalized enough to accept a numy array and returns an array of transformed features. In this research, we have used various combinations of those features. All the implementations have been performed on Jupyter Notebook.

The selected dataset is split with a ratio of 0.8 i.e. 80% is used for training and 20% is used for testing. On the dataset (Fire2021), the performance of classification is tested with respect to a range of parameters. Following are the main criteria for evaluation: accuracy, specificity, sensitivity, precision and F1score.

3.5.2. Hardware details

For the experiments we used Lenovo Legion 81FV Laptop. Complete specs of the system are as follows:

- Intel Core-i7 with 6 cores
- 12 logical processors were used with the graphics card of Nvidia GeForce GTX 1080
- Each has 8GB GDDR5X VRAM
- The total RAM of this system is 16GB

3.6. Introduction to the models used for fake news detection

We now briefly discuss the models used for evaluating the performance of fake news dataset.

- **Naïve Bayes classifier:** It is based on Bayesian theorem. Here, the posterior probability depends on prior probability and conditional probability. The assumption is that features are conditionally dependent given the target class

- **Logistic Regression:** In this type of machine learning models, the sigmoid function is used to separate the target classes and to find the decision boundaries. It can only be used for binary classification.

- **Adaboost:** AdaBoost is a boosting algorithm. They are not exactly classifiers. They can be employed with an array of any classifier i.e. random forest or support vector machine (SVM). AdaBoost only helps in base learning calculations multiple times. This works in a set of rounds i.e. t= 1..... T. One of the main objective of the calculation is to have improvement over the preparation set.

- **SVM**: The main objective of SVM algorithm is to identify a hyperplane in an n-dimensional space (where n is the number of features) that distinguishes between data points. There are a number of hyperplanes. It actually split the two groups of data points. The goal of SVM is to find a plane with the greatest objective function. In other words it is the greatest distance between data points from both classes. It attempts in maximizing the marginal distance. This makes it more easier to discriminate future data points.

- **Decision Tree**: The decision tree has the ability to model both regression and classification problems. The primary purpose of this methodology is to predict the value of a target class variable based on the input model. For this purpose, a decision tree mimics the representation in the form of multiple feature comparison at each node. The class label can be represented through the leaf node and the attributes are represented by interior nodes of the tree.

4. Results

4.1. Performance of the models

Table 5-11 shows the confusion matrix for various machine learning classifiers for the Urdu fake news dataset. The precision, recall, accuracy and F1-scores are provided for the models. Figure 3 shows the accuracy of the various classification models used. To summarize them, Table I2 shows the summary of results.

Table 5: Confusion matrix for Multinomial Naive Bayes

	precision	recall	f1-score	support
0.0	0.64	1.00	0.78	150
1.0	1.00	0.25	0.40	112
accuracy			0.68	262
macro avg	0.82	0.62	0.59	262
weighted avg	0.79	0.68	0.62	262

Table 6: Confusion matrix for Support Vector Classifier

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	150
1.0	0.43	1.00	0.60	112
accuracy			0.43	262
macro avg	0.21	0.50	0.30	262
weighted avg	0.18	0.43	0.26	262

Table 7: Confusion matrix for Logistic Regression

	precision	recall	f1-score	support
0.0	0.77	0.99	0.87	150
1.0	0.97	0.61	0.75	112
accuracy			0.82	262
macro avg	0.87	0.80	0.81	262
weighted avg	0.86	0.82	0.81	262

Table 8: Confusion matrix for Decision Tree

	precision	recall	f1-score	support
0.0	0.79	0.90	0.84	150
1.0	0.84	0.68	0.75	112
accuracy			0.81	262
macro avg	0.81	0.79	0.79	262
weighted avg	0.81	0.81	0.80	262

Table 9: Confusion matrix for Random Forest

	precision	recall	f1-score	support
0.0	0.63	1.00	0.78	150
1.0	1.00	0.22	0.36	112
accuracy			0.67	262
macro avg	0.82	0.61	0.57	262
weighted avg	0.79	0.67	0.60	262

Table 10: Confusion matrix for Bernoulli Naïve Bayes

	precision	recall	f1-score	support
0.0	0.75	0.63	0.68	150
1.0	0.59	0.71	0.65	112
accuracy			0.66	262
macro avg	0.67	0.67	0.66	262
weighted avg	0.68	0.66	0.67	262

Table 11: Confusion matrix for AdaBoost

	precision	recall	f1-score	support
0.0	0.81	0.95	0.88	150
1.0	0.92	0.71	0.80	112
accuracy			0.85	262
macro avg	0.87	0.83	0.84	262
weighted avg	0.86	0.85	0.84	262

Table 12: Summary of Accuracy and F1-score for the machine learning models

Model Name	Accuracy	F1-Score
Multinomial Naive Bayes	67.9	40
SVC	42.7	59.8
Bernoulli Naïve Bayes	66.4	64.5
Logistic Regression	82.4	74.7
Decision Tree	80.5	74.8
Random Forest	66.79	36.4
AdaBoost	84.7	79.7

It can be seen that AdaBoost provides the best performance (84.7% accuracy and 79.7% F1- score). Logistic Regression also provides good accuracy in the same range. Decision tree achieved 80.5% accuracy while random forest achieved 66.79% accuracy. Figure 10 compares the performance of

various machine learning classifiers in terms of Bar diagram.

Figure 3: Comparison of performance for various machine learning classifiers

4.2. Cross validation of the model

We have also performed the cross validation on the machine learning models. Table III shows the contribution in the proposed model. Figure 4 shows the accuracy of the models after 10- fold cross validation has been performed. It can be seen that AdaBoost still provides the best accuracy. Figure 5 shows the comparison of accuracy and F1-score using line graph.

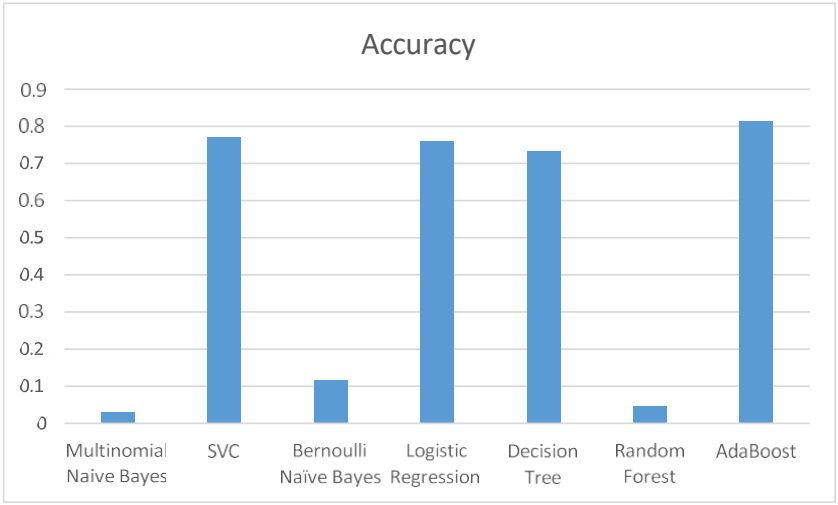

Figure 4: Accuracy of the models after 10-fold cross-validation

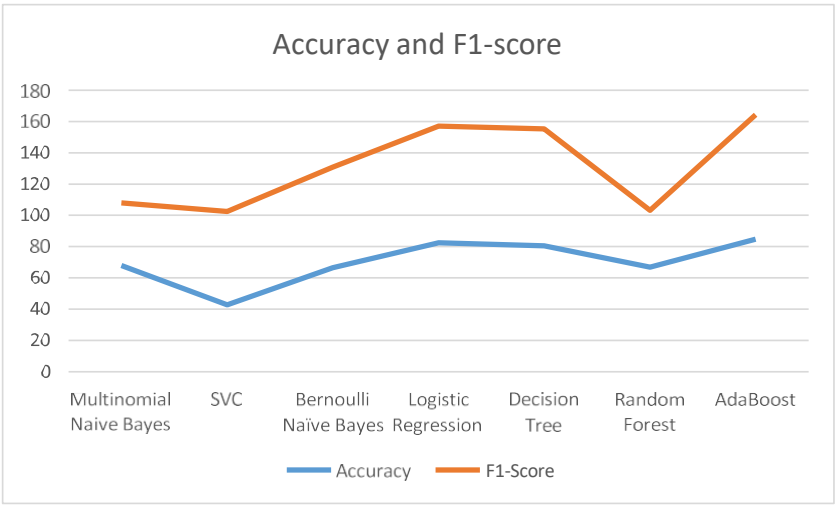

Figure 5: Accuracy and F1 score comparison using line graph

5. Conclusion

With the rapid growth of population, internet usage has increased a lot. As a result, it is becoming quite difficult to identify a genuine news from the fake news spreading over the internet. In this paper, we have used machine learning algorithms to identify fake news. We have used Fire2021 dataset to find out the best models for fake news detection in Urdu. The AdaBoost is found to give best result (84.7 accuracy and 79.7 F1 score). This is because it uses ensemble learning approach, combining several weak classifiers to form a stronger model. It is based on weighted learning from errors, handling noisy and complex data. It assigns higher weights to misclassified instances. As fake news involve noisy data, AdaBoost is robust against such complexities. It also assigns weights to features.

Different software and libraries were used for the implementation of the experiment i.e. to build and train ML and ANN model. Similar results have been reported in [40]. In [40], RF achieved 90.38% accuracy, LR achieved 89.21 accuracy, kNN 78.12% accuracy and SVM achieved 91.43% accuracy. The novelty of the paper lies in providing a comparison of various machine learning and deep learning algorithms specifically for Urdu news for detection of fake news. This work advances the state-of-the-art as no major work on using deep learning/machine learning algorithms for fake news detection. Even though, there are studies available for English and German languages realizing the gap, we used Fire21 and applied various machine learning model. This work serves as a foundation for further research on fake news detection, development of new datasets, improved model and fact checking initiatives. The proposed work can be extended to test other deep learning models such as LSTM, RNN, CNN. In addition, emerging transformer -based models such as BERT can be tested for classification of fake news in Urdu dataset. Similar results have been reported in literature for Arabic, Persian and Sindhi [42]. These languages have complex morphology similar to Urdu language, and high rate of code-mixing. A large number of datasets are also available for these languages. These languages have very high ambiguity as there are multiple meanings for the same words.

We conclude with the comment that in fake news detection, hardware constraints play a critical role, specifically in the context of training time, inference speed, and overall model feasibility. With the lack of sufficient hardware, we have slow text processing and feature extraction, limited training of large models and slow down training. It also slows dataset loading affecting model training efficiency.

References

[1] X. Zhang, et al., "Artificial intelligence in cognitive psychology—Influence of literature based on artificial intelligence on children's mental disorders," *Aggression and Violent Behavior*, p. 101590, 2021.

[2] L. Ru, et al., "A detailed research on human health monitoring system based on internet of things," *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1-9, 2021.

[3] A. Sharma and N. Kumar, "Third eye: an intelligent and secure route planning scheme for critical services provisions in internet of vehicles environment," *IEEE Systems Journal*, vol. 16, no. 1, pp. 1217-1227, 2021.

[4] Y. Liu, et al., "Line monitoring and identification based on roadmap towards edge computing," *Wireless Personal Communications*, pp. 1-24, 2021.

[5] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11765-11788, 2021.

[6] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Information Sciences*, vol. 497, pp. 38-55, 2019.

[7] G. Gorrell, et al., "SemEval-2019 Task 7: RumourEval 2019: Determining Rumour Veracity and Support for Rumours," in *Proc. of the 13th Int. Workshop on Semantic Evaluation: NAACL HLT 2019*, Association for Computational Linguistics, 2019.

[8] X. Zhang, et al., "Mining dual emotion for fake news detection," in *Proc. of the Web Conf. 2021*, 2021.

[9] R. Hou, et al., "Towards automatic detection of misinformation in online medical videos," in *Proc. of the 2019 Int. Conf. on Multimodal Interaction*, 2019.

[10] J.-P. Posadas-Durán, et al., "Detection of fake news in a new corpus for the Spanish language," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 4869-4876, 2019.

[11] F. Rangel, et al., "Overview of the 8th author profiling task at PAN 2020: Profiling fake news spreaders on Twitter," in *CEUR Workshop Proc.*, Sun SITE Central Europe, 2020.

[12] F. Rangel, et al., "On the author profiling and deception detection in Arabic shared task at FIRE," in *Proc. of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, 2019.

[13] K. Shu, et al., "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22-36, 2017.

[14] C. Guo, et al., "Dean: Learning dual emotion for fake news detection on social media," *arXiv e-prints*, p. arXiv:1903.01728, 2019.

[15] I. Vogel and P. Jiang, "Fake news detection with the new German dataset 'GermanFakeNC'," in *Int. Conf. on Theory and Practice of Digital Libraries*, Springer, 2019.

[16] M. Humayoun, H. Hammarström, and A. Ranta, "Urdu morphology, orthography and lexicon extraction," *arXiv preprint arXiv:2204.03071*, 2022.

[17] M. Humayoun, H. Hammarström, and A. Ranta, "Implementing Urdu Grammar as Open Source Software," *Corpus*, vol. 1, pp. 23-696, 2007.

[18] M. Humayoun and H. Yu, "Analyzing pre-processing settings for Urdu single-document extractive summarization," in *Proc. of the 10th Int. Conf. on Language Resources and Evaluation (LREC'16)*, 2016.

[19] A. Nawaz, et al., "Extractive text summarization models for Urdu language," *Information Processing & Management*, vol. 57, no. 6, p. 102383, 2020.

[20] K. I. Malik, "Urdu news content classification using machine learning algorithms," *Lahore Garrison Univ. Res. J. of Computer Sci. and Information Technology*, vol. 6, no. 1, pp. 22-31, 2022.

[21] S. M. Virk, M. Humayoun, and A. Ranta, "An open-source Urdu resource grammar," in *Proc. of the 8th Workshop on Asian Language Resources*, 2010.

[22] T. Ong, M. Mannino, and D. Gregg, "Linguistic characteristics of shill reviews," *Electronic Commerce Research and Applications*, vol. 13, no. 2, pp. 69-78, 2014.

[23] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. of the 20th Int. Conf. on World Wide Web*, 2011.

[24] K. Xu, et al., "A first step towards combating fake news over online social media," in *Wireless Algorithms, Systems, and Applications: 13th Int. Conf., WASA 2018*, Springer, 2018.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[26] S. Shojaee, et al., "Detecting deceptive reviews using lexical and syntactic features," in *Proc. of the 13th Int. Conf. on Intelligent Systems Design and Applications*, IEEE, 2013.

[27] M. Alrubaian, et al., "A credibility analysis system for assessing information on Twitter," *IEEE Trans. on Dependable and Secure Computing*, vol. 15, no. 4, pp. 661-674, 2016.

[28] M. Fernández-Delgado, et al., "Do we need hundreds of classifiers to solve real-world classification problems?" *J. of Machine Learning Research*, vol. 15, no. 1, pp. 3133-3181, 2014.

[29] R. K. Kaliyar, et al., "FNDNet–a deep convolutional neural network for fake news detection," *Cognitive Systems Research*, vol. 61, pp. 32-44, 2020.

[30] G. Gravanis, et al., "Behind the cues: A benchmarking study for fake news detection," *Expert Systems with Applications*, vol. 128, pp. 201-213, 2019.

[31] S. R. Sahoo and B. B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using deep learning," *Applied Soft Computing*, vol. 100, p. 106983, 2021.

[32] L. Wu and H. Liu, "Tracing fake-news footprints: Characterizing social media messages by how they propagate," in *Proc. of the 11th ACM Int. Conf. on Web Search and Data Mining*, 2018.

[33] T. E. Trueman, et al., "Attention-based C-BiLSTM for fake news detection," *Applied Soft Computing*, vol. 110, p. 107600, 2021.

[34] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach," *Int. J. of Information Management Data Insights*, vol. 1, no. 1, p. 100007, 2021.

[35] K. Jabeen, et al., "Machine Learning Models to Detect Online Fake News: a Systematic Literature Review," *Telematique*, vol. 22, no. 01, pp. 552-561, 2023.

[36] N. Islam, et al., "Ternion: An autonomous model for fake news detection," *Applied Sciences*, vol. 11, no. 19, p. 9292, 2021.

[37] M. Amjad, "Urdu Fake News Detection," *GitHub*, Available: https://github.com/MaazAmjad/Urdu-Fake-news-detection-FIRE2021, Last accessed: Dec. 2024.

[38] M. Z. Ali, et al., "Detection of Human and Machine-Authored Fake News in Urdu," *arXiv preprint arXiv:2410.19517*, 2024.

[39] A. Hussain, et al., "Detecting Urdu COVID-19 misinformation using transfer learning," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 143, 2024.

[40] M. S. Farooq, et al., "Fake news detection in Urdu language using machine learning," *PeerJ Computer Science*, vol. 9, p. e1353, 2023

[41] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H., 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data, 8(3),* pp.171-188.

[42] Samadi, M., Mousavian, M., and Momtazi, S., 2021. Persian fake news detection: Neural representation and classification at word and text levels. *Transactions on Asian and Low- Resource Language Information Processing, 21(1),* pp.1-11.