Supervised ensemble learning methods towards automatically filtering Urdu fake news within social media

Muhammad Pervez Akhter [1] , **Jiangbin Zheng** [1] , **Farkhanda Afzal** [Corresp., 2] , **Hui Lin** [3] , **Saleem Riaz** [3] , **Atif Mehmood** [4]

[1] School of Software and Microelectronics,, Northwestern Polytechnical University, Xian, China

[2] Department of Humanities and Basic Sciences, MCS, National University of Sciences and Technology, Islamabad, Pakistan

[3] School of Automation, Northwestern Polytechnical University, Xian, China

[4] School of Computer Science and Technology, Xidian University, Xian, China

Corresponding Author: Farkhanda Afzal
Email address: farkhanda@mcs.edu.pk

The popularity of the internet, smartphones, and social networks has contributed to the proliferation of misleading information like fake news and fake reviews on news blogs, online newspapers, and e-commerce applications. Fake news has a worldwide impact and potential to change political scenarios, deceive people into increasing product sales, defaming politicians or celebrities, and misguiding visitors to stop visiting a place or country. Therefore, it is vital to find automatic methods to detect fake news online. In several past studies, the focus was the English language, but the resource-poor languages have been completely ignored because of the scarcity of labeled corpus. In this study, we investigate this issue in the Urdu language. Our contribution is threefold. First, we design an annotated corpus of Urdu news articles for the fake news detection tasks. Second, we explore three individual machine learning models to detect fake news. Third, we use five ensemble learning methods to ensemble the base-predictors' predictions to improve the fake news detection system's overall performance. Our experiment results on two Urdu news corpora show the superiority of ensemble models over individual machine learning models. Three performance metrics balanced accuracy, the area under the curve, and mean absolute error used to find that Ensemble Selection and Vote models outperform the other machine learning and ensemble learning models.

1 # Supervised ensemble learning methods towards
2 # automatically filtering Urdu fake news within social
3 # media

5 Muhammad Pervez Akhter[1], Zheng Jiangbin[1], Farkhanda Afzal[2], Lin Hui[3], Saleem Riaz[3], Atif
6 Mehmood[4]

8 [1] School of Software and Microelectronics, Northwestern Polytechnical University 710072,
9 Xian, P.R. China
10 [2] Department of Humanities and Basic Sciences, MCS, National University of Sciences and
11 Technology, Islamabad, Pakistan.
12 [3] School of Automation, Northwestern Polytechnical University 710072, Xian, P.R. China
13 [4] School of Computer Science and Technology, Xidian University 710071, Xian, P.R. China

15 Corresponding Author:
16 Farkhanda Afzal
17 MCS, NUST, Islamabad, Pakistan.
18 Email address: farkhanda@mcs.edu.pk

20 ## Abstract

21 The popularity of the internet, smartphones, and social networks has contributed to the
22 proliferation of misleading information like fake news and fake reviews on news blogs, online
23 newspapers, and e-commerce applications. Fake news has a worldwide impact and potential to
24 change political scenarios, deceive people into increasing product sales, defaming politicians or
25 celebrities, and misguiding visitors to stop visiting a place or country. Therefore, it is vital to find
26 automatic methods to detect fake news online. In several past studies, the focus was the English
27 language, but the resource-poor languages have been completely ignored because of the scarcity
28 of labeled corpus. In this study, we investigate this issue in the Urdu language. Our contribution
29 is threefold. First, we design an annotated corpus of Urdu news articles for the fake news
30 detection tasks. Second, we explore three individual machine learning models to detect fake
31 news. Third, we use five ensemble learning methods to ensemble the base-predictors' predictions
32 to improve the fake news detection system's overall performance. Our experiment results on two
33 Urdu news corpora show the superiority of ensemble models over individual machine learning
34 models. Three performance metrics balanced accuracy, the area under the curve, and mean
35 absolute error used to find that Ensemble Selection and Vote models outperform the other
36 machine learning and ensemble learning models.

37 ## Introduction

38 Fake news is also known as deceptive news or misinformation. A news story is a piece of fake
39 news if its authenticity is verifiable false, and it intends to mislead the reader. As compared to
40 fake news, the authenticity of legitimate news is verifiable real, and it plans to convey authentic
41 information to the users (Abonizio et al., 2020). Fake news can take on numerous structures
42 including, edited text stories, photoshopped pictures, and unordered video clips. Fake news is
43 similar in appearance to legitimate news, but the aims are different. The aims of spreading fake
44 news are multipurpose, including deceiving readers into benefiting the author, propaganda about
45 a politician to win the election, increased sale of a product by posting fake positive reviews to
46 benefit a businessman, and defame a showbiz star (Monteiro et al., 2018). There are numerous
47 hazardous impacts on our society of the proliferation of fake news. Fake news changes the
48 manner of the individual to interpret and reply to legitimate news. Besides, fake news makes
49 individuals skeptical by destroying consumers' trust in the media by posting fabricated and
50 biased news stories (Agarwal & Dixit, 2020).
51 Spreading fake news is not a new problem in our time. Before the advent of the internet, fake
52 news was transmitted through face-to-face (oral), radio, newspaper, and television. In recent
53 years with the advent of the computer, the internet, smartphones, websites, news blogs, and
54 social media applications have contributed to transmitting fake news. There are several reasons
55 for spreading fake news through the internet and social media. It requires less cost and time than
56 traditional news media. It is very easy to manipulate legitimate digital news and share the
57 fabricated news story rapidly. Since 2017, there is a 13% global increase in social media users
58 (Kaur, Kumar & Kumaraguru, 2020). Fake news influences different groups of people, products,
59 companies, politicians, showbiz, news agencies, and businessman.
60 It requires more energy, cost, and time to manually identify and remove fake news or fake
61 reviews from social media. Some previous studies conclude that humans perform poorly than
62 automated systems to separate legitimate news from fake news (Monteiro et al., 2018). For the
63 last few years, machine learning methods' focus is to differentiate between fake and legitimate
64 news automatically. After the U.S. presidential elections in 2015, few popular social media
65 applications like Twitter, Facebook, and Google started to pay attention to design machine
66 learning and natural language processing (NLP) based mechanisms to detect and combat fake
67 news. The remarkable development of supervised machine learning models paved the way for
68 designing expert systems to identify fake news for English, Portuguese (Monteiro et al., 2018;
69 Silva et al., 2020), Spanish (Posadas-Durán et al., 2019), Indonesian (Al-Ash et al., 2019),
70 German, Latin, and Slavic languages (Faustini & Covões, 2020). A major problem of machine
71 learning models is that different models perform differently on the same corpus. Their
72 performance is sensitive to corpus properties like corpus size, distribution of instances into
73 classes (Pham et al., 2021). For example, the performance of K-nearest neighbor (KNN) depends
74 on the number of nearest points (k) in the dataset. SVM suffers from numerical instability when
75 solving optimization problems (Xiao, 2019). Similarly, the performance of an artificial neural
76 network (ANN) is prone to optimal architecture and tuning its parameters (Pham et al., 2021).

77   Ensemble learning is considered an efficient technique that can boost the efficiency of individual
78   machine learning models, also called base-models, base-predictors, or base-learners, by
79   aggregating the predictions of these models in some way (Lee et al., 2020). Ensemble learning
80   aims to exploit the diversity of base-predictors to handle multiple types of errors to increase
81   overall performance. Ensemble learning techniques show superior performance in various recent
82   studies about fake news detection. In a recent study, the ensemble learning technique
83   outperformed the four deep learning models including the deep structured semantic model
84   (DSSM) with RNN, intentCapsNet, LSTM model, and capsule neural network (Hakak et al.,
85   2021). In another recent study, (Mahabub, 2020) applied eleven machine learning classifiers
86   including the neural network-based model MLP on a fake news detection corpus. After that three
87   out of eleven machine models were selected to ensemble a voting model. Ensemble voting with
88   soft voting outperformed the other models. (Gutierrez-Espinoza et al., 2020) applied two
89   ensemble methods bagging and boosting with SVM and MLP base-predictors to detect fake
90   reviews detection. Experiments show that boosting with MLP outperforms the other.
91   This can be achieved in numerous ways, including homogenous models with diverse parameters,
92   heterogeneous models, resampling the training corpus, or using different methods to combine
93   predictions of base-predictors (Gupta & Rani, 2020). Ensemble learning can be of two types:
94   parallel and sequential. In the parallel ensemble, base-predictors are trained independently in
95   parallel. In the sequential ensemble, base-predictors are trained sequentially, where a model
96   attempts to correct its predecessor (Pham et al., 2021). Ensemble learning methods have shown
97   good performance in various applications, including solar irradiance prediction (Lee et al., 2020),
98   slope stability analysis(Pham et al., 2021), natural language processing (Sangamnerkar et al.,
99   2020), malware detection (Gupta & Rani, 2020), traffic incident detection (Xiao, 2019). In the
100   past, several studies explored machine learning models for fake news detection task in a few
101   languages like Portuguese (Monteiro et al., 2018; Silva et al., 2020), Spanish (Posadas-Durán et
102   al., 2019; Abonizio et al., 2020), Urdu (Amjad et al., 2020; Amjad, Sidorov & Zhila, 2020),
103   Arabic (Alkhair et al., 2019), Slavic (Faustini & Covões, 2020; Kapusta & Obonya, 2020), and
104   English (Kaur, Kumar & Kumaraguru, 2020; Ozbay & Alatas, 2020). As compared to machine
105   learning, a few efforts have been made to explore ensemble learning for fake news detection like
106   Indonesian (Al-Ash & Wibowo, 2018; Al-Ash et al., 2019), English (Kaur, Kumar &
107   Kumaraguru, 2020; Sangamnerkar et al., 2020). Therefore, this study aims to investigate
108   ensemble learning methods for the fake news detection task.
109   Urdu is the national language of Pakistan and the 8th most spoken language globally, with more
110   than 100 million speakers (Akhter et al., 2020a). Urdu is the South Asian severely resource-poor
111   language. As compared to resource-rich languages like English, a few annotated corpus from
112   very few domains are available for research purposes. Besides, insufficient linguistic resources
113   like stemmers and annotated corpora make the research more challenging and inspired.
114   Particularly in Urdu, studying fake news detection has several challenges. First, unavailability of
115   some sufficient annotated corpus. A recent study (Amjad et al., 2020) proposed an annotated
116   fake news corpus with a few hundred news articles. Experiments on this corpus reveal the poor

117 performance of machine learning models. Second, labeling a news article as 'fake' or
118 'legitimate' needs experts' opinions, which is time taking. Last, hiring experts in the relevant
119 domains is an expensive task. Therefore, in this study, we design a machine-translated corpus of
120 Urdu news articles translated from English news articles using Google Translate. We followed
121 the same procedure in the study (Amjad, Sidorov & Zhila, 2020). Experiments reveal that
122 machine learning models do not perform well on machine-translated corpus compared to the real
123 dataset (Amjad, Sidorov & Zhila, 2020). Because of the small size, the corpus is not sufficient to
124 make any conclusion about machine learning models' performance. Further, to the best of our
125 knowledge, no study explores ensemble learning models for Urdu fake news detection tasks.
126 Inspired by the work done in other languages, we are investigating the issue of fake news
127 detection in the Urdu language. The major aim of this study is to explore the capability of
128 ensemble learning models in improving fake news predictions in resource-poor language Urdu.
129 Our significant contributions to this study have been summarized below:

130 • We manually built an annotated news corpus composed of Urdu news articles distributed
131   into legitimate and fake categories.
132 • We perform several experiments using three diverse traditional machine learning
133   classifiers Naïve Bayes (NB), Decision Tree (DT), and Support Vector Machine (SVM),
134   and five ensemble models, including Stacking, Voting, Grading, Cascade Generalization,
135   and Ensemble Selection, to achieve improved prediction quality relative to conventional
136   individual machine learning models.
137 • We investigate the performance of our models using three feature sets generated through
138   character-level, word-level, and statistical-based feature selection methods.
139 • We report experiments of both machine learning and ensemble learning models on two
140   fake news corpora of the Urdu language.
141 • We comparatively analyze the performance of our models using four performance
142   measures, including balanced accuracy, the area under the curve, time and mean absolute
143   error.
144 Henceforward this article is organized as follows: Section 2 presents the essential related works.
145 Section 3 provides a brief overview of machine learning and ensemble learning models used in
146 this study. Section 4 will show the architecture of the adopted framework and corpus
147 characteristics. The results of the experiments are comparatively discussed in Section 5. Finally,
148 Section 6 ends the article with conclusions and future directions.

149 **Related Work**
150 Online social media and instant messaging applications like Facebook, Google, and Twitter are
151 popular these days in talking to your loved ones, expressing your opinion, sharing professional
152 information, or posting news about the subject of interest. Further, it is common to find some
153 information on the internet quickly. Unfortunately, all the information available on social media
154 sites is not accurate and reliable as it is straightforward to manipulate digital information and
155 quickly spread it in the world. Therefore, it is vital to design some accurate, efficient, and
156 reliable automated systems to detect fake news from a large corpus.

157   In the past, numerous machine learning methods have been used to combat fake news. (Monteiro
158   et al., 2018) showed that the multi-layer perceptron (MLP) model outperforms the NB and
159   random forest to identify fake news from a large news corpus. The study of (Faustini & Covões,
160   2020) concludes that SVM with bag-of-word (BoW) feature outperformed the other on five
161   corpora of three languages Germanic, Latin, and Slavic. A benchmarking study for fake news
162   detection concludes that SVM with linguistic-based word embedding features enables us to
163   classify fake news with high accuracy (Gravanis et al., 2019). A study about Portuguese fake
164   news detection reveals that random forest outperforms the other five machine learning models
165   (Silva et al., 2020). AdaBoost achieves the best performance on a small corpus than the other six
166   models to separate fake news from legitimate news (Amjad et al., 2020). A recent study of fake
167   news detection using supervised artificial intelligence methods shows that the DT is the best
168   model out of twenty-three models (Ozbay & Alatas, 2020). After analyzing the above studies, we
169   can conclude designing an effective and high-performing system with a careful selection of the
170   machine learning model and the feature selection method.
171   To overcome individual machine learning models' issues and increase the classification
172   performance, an ensemble of several diverse machine learning models has shown superior
173   performance than individual machine learning in several applications. Xiao applied ensemble
174   techniques with SVM and KNN base learners to detect traffic incidents (Xiao, 2019).
175   Experiments show the superiority of the ensemble model over individual machine learning
176   models. A recent study about detecting fraud in credit cards concludes that the ensemble
177   approach based on cost-sensitive meta-learning outperforms traditional machine learning models.
178   (Pham et al., 2021) used four ensemble models: boosted trees, bagged trees, random forest, and
179   generalized random forest for slope stability analysis. Experimental results prove the best
180   performance of extreme gradient boosting over other ensemble models and machine learning
181   models. For malware detection, (Gupta & Rani, 2020) used five base-predictors, and the output
182   of each base-predictor was ranked by calculating and aggregating the output weights. Then using
183   two ensemble techniques Voting and Stacking to rank the output. A higher value ranked by the
184   ensemble technique was the final prediction of the model. After extensive experiments, the study
185   reveals that weighted Voting shows the best performance than Stacking.
186   Compared to other applications, fake news detection using ensemble learning techniques has
187   very few studies in the past. (Kaur, Kumar & Kumaraguru, 2020) proposed a multi-level Voting
188   model for the fake news detection task. The study concludes that the proposed model
189   outperforms the other eleven individual machine learning and ensemble learning models. For
190   multiclass fake news detection, (Kaliyar, Goswami & Narang, 2019; Abonizio et al., 2020) used
191   Gradient Boosting ensemble techniques and compare their performance with several individual
192   machine learning models. Results on multiple corpora show that Gradient Boosting achieves the
193   best performance than any other individual models. A recent study (Huang & Chen, 2020)
194   proposed a self-adaptive harmony search algorithm to get optimized weights of ensemble
195   models. The proposed algorithm achieved outstanding performance with 99.4% accuracy. The

196   Bagging approach to detect fake news showed superior performance than SVM, Multinomial
197   Naïve Bayes, and Random Forest (Al-Ash et al., 2019).
198   English is a resource-rich language, and many linguistic resources are publically available for
199   research purposes. Therefore, several research studies have been performed for the fake news
200   detection task. A study gives a comparison of twenty three publically available datasets of
201   English (Sharma et al., 2019). A recent survey compares different techniques to identify fake
202   news, their credibility detection, and discusses fundamental theories and opportunities (Zhou &
203   Zafarani, 2020). There is a severe lack of research studies for fake news detection in languages
204   other than English. For the Indonesian language, a study by (Al-Ash et al., 2019) shows that the
205   bagging model outperforms the three individual models: SVM, Random Forest, and Multinomial
206   Naïve Bayes. In (Abonizio et al., 2020) applied three machine learning models (KNN, SVM, and
207   random Forest) and extreme gradient boosting (ensemble model) on five datasets of three
208   languages (English, Portuguese, and Spanish). In another study about fake news detection for
209   Portuguese, random forest shows high accuracy in most of the experience (Silva et al., 2020).
210   Extreme gradient boosting shows the best performance than other individual models. For the
211   Urdu language (Amjad et al., 2020), Adaboost outperforms the other seven machine learning
212   models on a very small corpus. DT improves the classification accuracy for fake news detection
213   for the Slovak language (Kapusta & Obonya, 2020).
214   The lack of availability of a benchmarked annotated corpora of resource-poor languages are the
215   major problem to investigate and compare the performance of numerous automated methods for
216   fake news detection. Therefore, in several other than English studies, authors designed their
217   corpus by collecting news articles from the internet and other web resources and manually
218   annotating these articles into fake and legitimate news. A list of corpora for several resource-
219   poor languages is given in Table 1. It can be noticed that all the corpora except 'Fake.Br' are
220   very small in size. Because corpus designing is a costly and time-consuming task, the annotation
221   process requires several experts from various fields to decide about the news article (Amjad,
222   Sidorov & Zhila, 2020). To date, our proposed corpus Urdu fake news (UFN) is the largest
223   corpus than others.
224

Machine Learning and Ensemble Learning Models

Machine Learning Models

227   This section gives a brief overview of three traditional machine learning models: Naïve Bayes,
228   Decision Tree, and Support Vector Machine. We also described their significant drawbacks,
229   which limit their performance on various tasks.

Naïve Bayes (NB)

231   Naïve Bayes used a probabilistic approach based on Bayesian Theorem with two assumptions: 1)
232   all the features are independent of each other in the dataset, and 2) all the features have equal
233   effects. It is simple, popular, and useful for the classification of a large corpus, although the
234   corpus does not hold independence. NB is challenging to interpret for a large corpus, and its
235   assumption about features independence makes its performance poor, especially when the data

236   distribution is very skewed (Komiya et al., 2011). Several studies have used NB for fake news
237   detection tasks like for Portuguese (Monteiro et al., 2018) and English (Gravanis et al., 2019).

Decision Tree (DT)

239   The decision tree algorithm learns a decision rule inferred from the training data to design a
240   decision tree model. Nodes of the tree represent the feature vectors taken from the text of the
241   news article. Leaf nodes represent the set of possible labels or classes in the corpus. In our case,
242   there are two possible labels: fake and legitimate. The DT predicts the article's label by learning
243   features from the tree's root to one of the leaf nodes. It is simple and easy to interpret because all
244   the information about the model behavior and influential variables is available. Therefore, the
245   DT is also known as a white-box model. Drawbacks of the DT include overfitting and instability,
246   a complex tree for a high-dimensional dataset that is not easy to interpret (Pham et al., 2021). For
247   the fake news detection task, DT has shown good performance for Slovak (Kapusta & Obonya,
248   2020), Portuguese (Silva et al., 2020), English (Gravanis et al., 2019; Ozbay & Alatas, 2020),
249   and Urdu (Amjad et al., 2020) languages.

Support Vector Machine (SVM)

251   Support vector machine is a non-parametric machine learning model. The performance of SVM
252   is usually considered suitable for binary classification tasks with high-dimensional data. SVM
253   maps all the features obtained from news articles' text into n-dimensional space where a feature
254   represents the particular coordinate. During training, SVM learns a hyper-plan that best
255   discriminates the features of one class to another. Therefore, SVM is also known as a
256   discriminative classifier. Although SVM performs well with high-dimensionality data, it is
257   difficult to interpret, requires significant computing resources, and faces numerical instability for
258   optimization problems (Pham et al., 2021). SVM shows excellent performance for fake news
259   detection task in several studies of various languages like English (Monteiro et al., 2018;
260   Gravanis et al., 2019), Urdu (Amjad, Sidorov & Zhila, 2020), Portuguese (Silva et al., 2020),
261   Dutch (Verhoeven & Daelemans, 2014), Germanic, Latin, and Slavic (Faustini & Covões, 2020).

Ensemble Learning Models

263   Ensemble learning aggregates the individual machine learning models (base-predictors) to design
264   a superior model to increase overall accuracy by handling the shortcomings of the base-
265   predictors. It is known as the most efficient technique for improving the performance of machine
266   learning models. Nowadays, ensemble learning methods are gaining more popularity than
267   traditional individual machine learning models in numerous classification tasks like fake news
268   detection (Kaur, Kumar & Kumaraguru, 2020), malware detection (Gupta & Rani, 2020).
269   Ensemble learning methods fall into two categories: parallel ensemble and sequential ensemble.
270   Both are shown in Figure 1(a) and Figure 1(b). In the parallel ensemble, the base-predictors are
271   trained in parallel on the input corpus. The parallel ensemble has the advantages of simultaneous
272   predictions, utilizing different CPU cores to execute the models simultaneously, and utilizing the
273   characteristics of independence among them. In the sequential ensemble, the base-predictors are
274   trained sequentially where the output of the one base-predictor plus the input corpus is the input
275   to the next base-predictor. In other words, the base-predictors are dependent on each other. The

276 next base-predictor challenge is to try to correct the errors of the previous base-predictor to
277 improve the overall prediction accuracy (Pham et al., 2021). Base-predictors can be homogenous
278 or heterogeneous. In homogenous, a single machine learning model (like DT or NB) is trained in
279 parallel or sequentially, while in heterogeneous different machine learning models (like DT and
280 NB) are trained in parallel or sequentially. The ensemble learning method is advantageous if the
281 heterogeneous machine learning models are used as base-predictor (Kittler, Hater & Duin, 1996).
282 Heterogeneous ensemble learning can be performed using different classifiers with different
283 feature sets, training sets, and evaluation methods. In this section, we provide a brief description
284 of the five ensemble models used in this study.

Stacking
286 Stacking model ensembles several base-predictors machine learning models using the stacking
287 method. It was initially proposed by (Ting & Witten, 1997) and used in several studies for
288 classification tasks like malware detection (Gupta & Rani, 2020), credit card fraud detection
289 (Olowookere & Adewale, 2020), and spam detection (Saeed, Rady & Gharib, 2019). It can
290 perform both classification and regression on data. Base-predictors are trained on input data, and
291 the output of these base-predictors is given to a meta-classifier which makes the final prediction
292 about the class of an input sample. Meta-classifier can be any classifier like Adaboost,
293 Regression, etc. The primary aim of meta-classifier is to learn the optimal weights to combine
294 the predictions of base-predictors and produce better prediction results than individual base-
295 predictor results. Therefore, Stacking reduces the variance and improve the quality of
296 classification. For unseen test articles from the test set, the article is passed to the base-predictor
297 to classify these test articles. Their classification is then passed to stacking-based ensemble
298 learners to make the final classification of that article as either fake or legitimate.

Voting
300 Voting is a meta-classifier that combines several base-predictors using different combination
301 rules. Base-predictor can be any machine learning model. Individual base-predictors are trained
302 on training data. The output of these base-predictors is combined using some combination rules
303 like majority voting, minimum or maximum probabilities, or product of probabilities (Kittler,
304 Hater & Duin, 1996). Majority voting is also known as hard-voting as the class with majority
305 votes is considered the input article's final class. In soft-voting, the final class is a class with the
306 highest probability averaged over the individual predictors (González et al., 2020). Voting
307 method have used in several classification tasks like fake news detection (Kaur, Kumar &
308 Kumaraguru, 2020), spam detection (Saeed, Rady & Gharib, 2019), and slope stability analysis
309 (Pham et al., 2021)

Grading
311 Grading is an ensemble model originally presented by (Seewald & Fürnkranz, 2001). Grading is
312 a type of meta-classifier that corrects the graded predictions of base-predictors at the base-level
313 assuming that different base-predictors make different errors. Graded predictions are those
314 predictions that are marked as incorrect or correct predictions by the base-predictor. A meta-
315 classifier is learned for each base-predictor whose basic job is to predict when the base-predictor

316   will error. These meta-classifiers are trained on a training set constructed from the graded
317   predictions of corresponding base-predictors as new class labels. Grading is different from
318   Stacking that uses the incorrect predictions of base-predictors as the attribute value for meta-
319   classifier. Several studies show that Grading outperforms the Voting and Stacking ensemble
320   models on classification tasks (Seewald & Fürnkranz, 2001; González et al., 2020).

321   ***Cascade Generalization***
322   Cascade Generalization belongs to a special Stacking generalization that uses a sequentially
323   layered architecture for combining multiple classifiers. The predictions of several base-predictors
324   are used in the next stage for final prediction (Gama & Brazdil, 2000). An extension of the
325   original data is achieved at each level by inserting new attributes derived from the probability
326   class distribution given by the base-predictors. Cascade Generalization is different from Stacking
327   generalization in that Stacking is parallel, while Cascade is sequential in nature. Because of its
328   sequential nature, intermediate models have access to the original attributes and the low-level
329   models' predictions. Cascade provides the rules to choose the high-level and low-level models
330   for classification. A major problem of the Cascade is that it transforms the feature space into a
331   new high-dimensional feature space (the curse of dimensionality) that sometimes leads to a more
332   difficult learning problem (Gama & Brazdil, 2000).

333   ***Ensemble Selection***
334   Ensemble selection is a method to construct an ensemble of several machine learning models. It
335   starts with an empty ensemble and adds a model into the ensemble if it increases the performance
336   of the ensemble. This process is repeated up to a specified number of iterations or until all the
337   models have been used (Caruana, Ksikes & Crew, 2014). Models are added into an ensemble
338   using numerous ways like forwarding selection, backward elimination, and the best model. It
339   stops adding models into the ensemble when the ensemble's performance starts to decrease after
340   achieving the best performance. Ensemble selection allows ensembles to be optimized to
341   performance metrics such as accuracy, cross-entropy, mean precision, or ROC Area (Ballard &
342   Wang, 2016; Nguyen et al., 2020). In a recent study, Ensemble Selection outperforms the other
343   ensemble models to classify 62 datasets (Nguyen et al., 2020).

344

345   ## Methodology and corpus construction
346   The proposed framework for fake news detection consists of four-phases. The first phase
347   describes the procedure to design a corpus of Urdu news articles. The second phase explains the
348   preprocess operations performed on the text of news articles. The third phase shows feature
349   selection or dimensionality reduction approaches. The fourth phase provides the description of
350   individual machine learning models or base-predictors for ensemble models. At last, the fifth
351   phase describes the numerous ensemble learning models used in this study. The architecture with
352   five layers is shown in Figure 2.

353   **Corpus Design**
354   In this study, we choose two corpora of text news articles of Urdu language for experiments. As
355   Urdu is a resource-poor language, there is no standard corpus available for fake news detection

356   task to the best of our knowledge. Because of the lack of linguistic resources, the collection of
357   news articles from multiple sources is a tough task. Besides, the annotation process of these news
358   articles based on the articles contents needs expert knowledge, a lot of time, and budget.
359   Therefore, augmented corpus design is the only way to perform research about fake news
360   detection for resource-poor languages. Our first corpus is Urdu Fake News (UFN) augmented
361   corpus. It contains two thousand news articles randomly selected and translated from English
362   language fake news corpus using online Google Translator. The original English corpus contains
363   nearly 8000 news articles. We picked a subset of two thousand articles because 1) manual
364   translation of all the articles in the original corpus is time taking and difficult to perform, 2)
365   English-Urdu translation using Google Translator is not hundred percent accurate and to the best
366   of our knowledge, no study investigates this problem. This is an open research question and can
367   be investigated in future studies. 3) we aim to explore the detection performance of ensemble
368   learning techniques. Several recent studies about fake news detection in Slovak (Kapusta &
369   Obonya, 2020), Italian (Fornaciari & Poesio, 2013), and Spanish (Posadas-Durán et al., 2019)
370   used corpora with even less than two thousand news articles (see Table 1). Our second corpus is
371   a small size Bend the Truth (BET) corpus designed and annotated by (Amjad et al., 2020). This
372   corpus contains only 900 original news articles in Urdu. A sample of the Urdu news articles is
373   shown in Table 2. After translation, the Urdu article label was the same as in the English corpus's
374   corresponding article. The final corpus is available online on GitHub in CSV file format. The
375   statistics of both corpora are shown in Table 3. It can be noticed that our designed corpus UFN is
376   larger than the BET corpus based on the total number of articles, size of vocabulary, and the
377   length of the article.
378

Corpus Preparation and Preprocessing

380   Articles in the corpus are in an unstructured format and cannot be processed directly by the
381   machine learning models. We must have to perform a series of operations on the corpus to
382   convert an unstructured corpus into a structured corpus. We have cleaned and processed both
383   corpora's news articles before generating the feature vectors for feature selection. We tokenized
384   the text using space characters. Special characters, email addresses, and website URLs were
385   removed from the text. After cleaning the text, we removed the most frequent and rare words of
386   the Urdu language (also known as stopwords) from the text. The cleaned and the preprocessed
387   articles were converted into numeric feature vectors using the term frequency-inverse document
388   frequency (TF-IDF) method as used in a recent study (Ozbay & Alatas, 2020). Both corpora
389   were passed through the same number of preprocessing steps.

Feature Selection

391   In our experiments, we have performed the experiments using three feature selection methods
392   character tri-grams, bag-of-word (BOW), and information gain (IG). A recent study shows the
393   superiority of the character n-gram method over word-level n-grams in short text classification
394   tasks (i.e. offensive language detection) in Urdu text comments (Akhter et al., 2020b). Character
395   n-gram is a contiguous sequence of characters in the text. In character n-grams, the value of n is

396   taken as three, which means the combination of three characters makes a tri-gram feature. From
397   the UFN corpus, 1084 character n-grams, and from the BET corpus 1091 n-grams were
398   extracted. BOW is a content-based feature representation in which a news article is represented
399   as a set of words that occur in it at least once. IG measures the goodness of the features in the
400   text. A comparative study concludes that IG is the best feature selection method for document-
401   level text classification of Urdu. In our experiments, we have selected the top one thousand IG
402   features from both corpora. 1225 and 1214 BoW features from BET and UFN, respectively.

403   **Heterogeneous Machine Learning Models**
404   In our experiments, for machine learning classification, we use three individual machine learning
405   models NB, SVM, and DT to detect fake news. All three models are heterogeneous. The working
406   of these models is entirely different from each other. Using character-level n-grams from text
407   articles, these models analyze the article's text and classify it into one of the categories legitimate
408   or fake. Detail description of these machine learning models is given in Section 3.1.

409   **Ensemble Learning Models**
410   Ensemble classification is usually based on two levels: base-level and ensemble-level. We use
411   three diverse machine learning models, SVM, DT, and NB, as base-predictors at the base-level.
412   Input to these base-predictors is the character-level n-grams extracted from the news articles.
413   Output predictions of these base-predictors are input to ensemble-level models. The basic aim of
414   using the ensemble model is to overcome the base-predictors' shortcomings and improve overall
415   prediction accuracy. We use five ensemble models for ensemble classification, known as Voting,
416   Grading, Stacking, Cascading Generalization, and Ensemble Selection. A brief description of our
417   ensemble models is given in Section 3.2.

418   **Performance Measures**
419   To compare the performance of individual machine learning models and ensemble learning
420   models, in this study, we employed the three well-known performance measures mean absolute
421   error (MAE), balanced accuracy (BA), and area under the curve (AUC).

422   ***Balanced Accuracy (BA)***
423   Separation of fake news from legitimate news is a binary classification task. A model has to
424   decide about an article, either a legitimate article or a fake article. As the focus of this study is to
425   detect both classes correctly, we used the balanced accuracy performance measure to compare
426   the performance of our models. Balanced accuracy calculates the average of the proportion of
427   corrects of each class individually. Balanced accuracy can be calculated as follows:

$$Balanced\ Accuracy\ (BA) = \left[\frac{TP}{TP + FP} + \frac{TN}{TN + FN}\right]/2 \tag{1}$$

428

429   ***Area Under the receiver operating characteristic Curve (AUC)***
430   Area Under the receiver operating characteristic curve, also known as area under the curve
431   (AUC), is used to estimate the performance of a machine learning model using a single value.
432   AUC provides the probability that the model will rank a randomly chosen positive sample higher
433   than a randomly chosen negative sample. AUC can be calculated by Equation 2. TP_{rate} is the

434   ratio of correctly predicted articles as fake articles. It is also known as recall and can be
435   calculated as above in Equation 3. FP_{rate} is the number of legitimate news articles that are
436   misclassified or incorrectly predicted as fake news articles.

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{2}$$

$$FP_{rate} = \frac{FP}{FP + TN} \tag{3}$$

437   General rules for categorizing the performance of the machine learning model using AUC are
438   given in Table 4. These rules are used and discussed in (Pham et al., 2021).
439
440   ***Mean Absolute Error (MAE)***
441   The error refers to the absolute difference between the actual values and the predicted values.
442   MAE measures the average magnitude of the error from a set of predictions made by the model.
443   MAE can be calculated as follows:

$$Mean\ Absolute\ Error\ (MAE) = \frac{1}{N}\sum_{j=1}^{n}|y_j - \hat{y}_j| \tag{4}$$

444   .

445   # Results
446   ## Experiment Setup
447   As mentioned earlier, in this study, three diverse machine learning models NB, DT, and SVM
448   have been used to classify news articles into legitimate and fake classes. We use a well-known
449   data mining tool, WEKA, for experiments. WEKA provides a list of supervised and
450   unsupervised machine learning models, data preprocessing techniques, and various performance
451   evaluation methods. Machine learning models have few parameters, called hyper-parameters, to
452   minimize the difference between training error and testing error. In our experiments, we do not
453   fine-tune the hyper-parameters of all these models. We use the default parameters given in the
454   WEKA as the default parameters have already the best values in most of the cases. We use the
455   J48 algorithm for DT implementation. LibLINEAR algorithm for SVM implementation. We use
456   the same DT, SVM, and NB models as base-predictors for all the ensemble models. For Voting
457   and Stacking, along with the three base-predictors, we use Adaboost as a meta-classifier.
458   ## Model training and testing
459   For training and validation of individual machine learning models and ensemble models, we use
460   k-fold cross-validation as mentioned in Section 4.1 that both corpora have not been divided into
461   a training subset and testing subset. k-fold cross-validation is a popular choice and used in many
462   past research studies. In our experiments, we use 10-fold cross-validation, where k-1 folds are
463   used for training, and one-fold is used to test the model's prediction performance. This process is
464   repeated ten times to achieve the final performance score.
465   ## Results and discussion of machine learning models

466   The experiment results achieved using 10-fold cross-validation from individual machine learning
467   models are shown in Table 5. We compare the performance using BA, AUC, MAE, and time. A
468   close observation of the results reveals that SVM outperforms the other for all corpora
469   performance metrics. A model is considered an accurate classifier if its balanced accuracy is
470   higher than the other models. The BA metric shows that SVM outperforms the others on the
471   UFN corpus. SVM achieves BA scores 81.6%, 86.7%, and 87.3% using tri-gram, BoW, and IG
472   features, respectively. SVM also outperforms the others on the BET corpus. It achieves 76.3,
473   62.7, and 62.4 using tri-gram, BoW, and IG features, respectively. IG feature outperforms the
474   others and achieves the maximum BA score 87.3% on large UFN corpus while tri-gram
475   approach achieves maximum BA scores of 76.3% on BET corpus. With the lowest balanced
476   accuracy scores, NB shows the worst performance. It is noticed that SVM has higher accuracy at
477   UFN than BET. The size of the UFN corpus, in terms of the number of articles and vocabulary
478   size, is almost double the of BET, and SVM is considered a good model for the classification of
479   high-dimensional feature space (Faustini & Covões, 2020).
480   Similarly, AUC scores of the SVM model are the highest score than DT and NB on both corpora.
481   SVM achieves 87.3% and 76.3% AUC metric values on UFN and BET corpora, respectively.
482   Here, again IG proves the best feature selection method for UFN while tri-gram on BET corpus
483   as SVM achieves the maximum AUC scores on IG and tri-gram features. Further, as per the rules
484   of Table 4, a comparison of AUC scores of all the models concludes that the performance of
485   SVM on UFN is excellent $(0.8 \leq AUC < 0.9)$ on all the features. On the BET corpus, SVM
486   performance is only acceptable $(0.7 \leq AUC < 0.8)$ on tri-gram features. The performance of DT
487   and NB is just acceptable.
488   From Table 5, it can be seen that in terms of MAE, the prediction error of SVM is the lowest
489   than others. SVM achieves the lowest MAE score 12.7% with IG on UFN while 23.5% with tri-
490   gram on BET corpus. The highest MAE values of NB proves its worst performance to detect
491   Urdu fake news. A model is considered efficient if it takes minimum time than other models to
492   build the model on some corpus. Again, SVM takes a minimum time of 0.15 on BOW and 0.17
493   seconds on IG to build the model on UFN and BET. DT takes the highest time on all features to
494   build the model for both corpora. Further, it is notable that all the models perform betters on our
495   designed machine-translated UFN corpus than the original news article's corpus BET. It shows
496   that Google API translated text, preprocessing methods, and feature selection methods all
497   together improve the classification accuracy of our models to detect fake news. Therefore, after
498   analyzing the results, we conclude that SVM is a more accurate, reliable, efficient, and robust
499   classification model among the three models to detect Urdu text's fake news articles.
500   **Results and discussion of ensemble models**
501   The values of four evaluation metrics balanced accuracy, the area under curve, mean absolute
502   error, and time achieved by five ensemble models on both corpora are given in Table 6 . Time
503   and MEA are the two metrics whose minimum values are required by the models. The other two
504   metrics balanced accuracy and AUC, whose maximum values are required to achieve by a
505   model. For the time metric, it is visible that Voting takes the minimum time than other models to

506   build the model on the input corpus. Voting takes 11.52 seconds and 3.12 seconds to build the
507   model on UFN and BET corpora, respectively. As the size of the BET model is very small, the
508   Voting takes the minimum time to build the model than all the other models. It can also be
509   noticed that the minimum time taken by Voting to build a model on both corpora is using tri-
510   gram. It shows the efficiency of the tri-gram method over IG and BoW to build a model.
511   For the MAE metric, again, the Voting model achieves the minimum values than others on both
512   corpora, which shows that magnitude of the error is significantly less in predicting the labels of
513   both types of articles. The average magnitude error of the Voting model is 18.41% on tri-gram
514   and 10.7% on IG for BET and UFN, respectively. It means that IG is a good feature selection
515   method over other methods on large size UFN corpus while tri-gram is good for small size BET
516   corpus.
517   To estimate an ensemble model's performance and decide whether a model's performance is
518   acceptable or not, we use a performance ranking metric AUC. On the BET corpus, only
519    $AUC \geq 90$ is achieved by the Ensemble Selection model over the tri-gram feature method. With
520   IG and BoW features the AUC scores of all the other models are $AUC < 75$ which means the
521   performance of these models is acceptable. On UNF corpus, Cascade Generalization achieves
522   outstanding performance to detect fake news with BoW and IG (see Table 4). It achieves 92.0%
523   and 92.7% AUC scores ($AUC \geq 90$) for BoW and IG methods. Cascade Generalization with
524   86.8% AUC score categorizes its performance ranking to excellent. Again, Ensemble Selection
525   achieves the best AUC score using IG on UFN while Cascade Generalization achieves the best
526   AUC using tri-gram features on BET.
527   As we are interested to know the performance of a model to predict both labels ('fake',
528   'legitimate) correctly in the corpus, we use balanced accuracy. The maximum BA achieved by a
529   model means that the model is performing more accurately than others to distinguish fake
530   articles from legitimate articles. Experiment results reveal that Ensemble Selection and Voting
531   outperform the other models on BET corpus and UFN corpora. Ensemble Selection achieves the
532   maximum 83.3% BA on BET corpus using the tri-gram feature On the UFN corpus, the Voting
533   model significantly outperforms the other four ensemble models and achieves an 89.3% BA
534   score using the IG feature. Again, it is noticed that IG outperforms the other methods on UFN
535   while tri-gram outperforms the other feature selection methods on the BET corpus.
536
537   Voting model ensemble the numerous base-predictors using some ensemble rule. The four
538   popular rules are majority voting, the product of probabilities, minimum probabilities, and
539   maximum probabilities. By impressive Voting performance on both corpora using balanced
540   accuracy, MAE, and Time metrics, as given in Table 6   and discussed above, we further
541   investigate its performance using different ensemble rules. The mean absolute error values
542   achieved by each ensemble rule is shown in Figure 3. We conclude that the minimum
543   probabilities rule is impressive to ensemble the predictions of the base-predictors at it achieves
544   the lowest error values on both corpora. The vote model achieves 16.74% and 18.41% MAE

545 scores on UFN and BET corpora. Hence, in our experiments, minimum probabilities, and
546 product of probabilities, both rules perform the same on both corpora.

547 **Performance comparison of machine learning and ensemble learning models**

548 It is important to know the difference in the performance of ensemble models and individual
549 machine learning models.

550 A summary of the results achieved by the best ML and EL model with the best feature selection
551 method is given in Table 7. Comparative analysis of the results shows that machine learning
552 models are efficient than EL models and take less time to build a model on both corpora. SVM
553 takes a minimum time of less than a second to build the model on both corpora. Among the EL
554 models, Voting is efficient and takes 11.52 and 3.12 seconds on UFN and BET. But Voting is
555 much costly than SVM. It is because of the multiple base-predictors in the EL model. EL model
556 combines the three heterogeneous ML models and then the final Voting model predicts the final
557 label based on the prediction of the base-models.

558 For error analysis, MAE values show that EL models have the lowest values of MAE than
559 individual ML models. Again SVM outperforms the NB and DT by achieving minimum MAE
560 scores on both corpora. On the other side, Voting outperforms the other EL models on both
561 corpora. The lowest score of MAE for EL models means that these models are more accurate in
562 fake news detection. EL models reduce MAE at two levels: at the base-predictors level and
563 ensemble-level. Voting takes the advantage of MAEs of its base-predictor. It reduces the MAE
564 scores of its three base-predictors using minimum probability to predict the final class.

565 SVM achieves maximum scores of AUC 87.3% and 76.3% on UFN and BET. AUC scores rank
566 SVM predictions to excellent on UFN and acceptable on BET. Cascade Learning and Ensemble
567 Learning achieve 92.7% and 91.0% AUC scores on UFN and BET. It categorizes the detection
568 performance of both models as outstanding. SVM outperforms the other ML models and it
569 achieves the maximum BA scores. SVM achieves 87.3% BA on UFN and 76.3% on BET. From
570 EL models, Voting achieves 89.3% BA and outperforms the other EL and ML models on the
571 UFN corpus. On BET corpus, Ensemble Selection models produce 83.3% BA that is the
572 maximum BA among all the models.

573 The comparison of EL and ML methods using three feature selection methods is interesting
574 valuable. SVM shows the best performance among the three ML models on small and large
575 corpora. SVM achieves the best scores in all the performance measures. Character tri-gram
576 works well on small size corpus BET while IG works well on large size corpus UFN to boost
577 SVM performance. Voting performance is the best performance among EL models using Time
578 and MAE performance measures on both corpora. Ensemble Selection is good at small corpus
579 BET on two performance measures. IG feature works well with Voting to predict the class of a
580 news article on UFN while tri-gram is the best with Voting and Ensemble Learning. Further, it
581 can be seen that the IG feature works well on large size corpus while character tri-gram is good
582 on small size corpus.

583

584 **Conclusions**

585   Fake news detection through ensemble models is the most prominent topic of machine learning.
586   If the traditional machine learning models are used for fake news detection task, the performance
587   is not encouraging because their performance is limited to corpus characteristics. In this study,
588   we deliberately choose ensemble methods to classify fake and legitimate news articles of the
589   Urdu language. First, we use three machine learning models to classify two fake news corpora.
590   Our experiments on two Urdu news corpora conclude that the individual machine learning model
591   SVM outperforms the DT and NB on both corpora. SVM achieves the best scores of balanced
592   accuracy and AUC and the minimum score of MAE. Second, we use five ensemble models for
593   the same task.
594   We find that ensemble models with three base-predictors DT, NB, and SVM, Ensemble
595   Selection, and Vote models outperform the other on BET and UFN corpora, respectively. After
596   the analysis of MAE, AUC, time, and BA values, we conclude that Voting with minimum
597   probability is the best EL model for the fake news detection task. IG feature works well with
598   large size corpus while character tri-gram works well on small size corpora.
599   This study has several limitations that need to be addressed in future studies. The proposed
600   corpus UFN still needs to grow by adding more news articles to it. We used online Google
601   translation API in English-to-Urdu translation and we believe that translation accuracy is not a
602   hundred percent. A study is vital in the future to explore the translation accuracy and quality of
603   various translation APIs like Baidu, Google, etc. The potential of deep learning models also can
604   be explored to detect fake news for Urdu. Further, we also hope to design another multilingual
605   fake news corpus of English and Urdu news articles.
606

607   **References**

608   Abonizio HQ, de Morais JI, Tavares GM, Junior SB. 2020. Language-independent fake news
609   detection: English, Portuguese, and Spanish mutual features. *Future Internet* 12:1–18. DOI:
610   10.3390/FI12050087.
611   Agarwal A, Dixit A. 2020. Fake News Detection: An Ensemble Learning Approach.
612   *Proceedings of the International Conference on Intelligent Computing and Control Systems,*
613   *ICICCS 2020*:1178–1183. DOI: 10.1109/ICICCS48265.2020.9121030.
614   Akhter MP, Jiangbin Z, Naqvi IR, Abdelmajeed M, Fayyaz M. 2020a. Exploring deep learning
615   approaches for Urdu text classification in product manufacturing. *Enterprise Information*
616   *Systems*:1–26. DOI: 10.1080/17517575.2020.1755455.
617   Akhter MP, Jiangbin Z, Naqvi IR, Abdelmajeed M, Sadiq MT. 2020b. Automatic Detection of
618   Offensive Language for Urdu and Roman Urdu. *IEEE Access* 8:91213–91226. DOI:
619   10.1109/ACCESS.2020.2994950.
620   Al-Ash HS, Putri MF, Mursanto P, Bustamam A. 2019. Ensemble Learning Approach on
621   Indonesian Fake News Classification. In: *2019 3rd International Conference on Informatics*
622   *and Computational Sciences (ICICoS)*. 1–6. DOI: 10.1109/ICICoS48119.2019.8982409.
623   Al-Ash HS, Wibowo WC. 2018. Fake News Identification Characteristics Using Named Entity
624   Recognition and Phrase Detection. In: *2018 10th International Conference on Information*
625   *Technology and Electrical Engineering (ICITEE)*. 12–17. DOI:
626   10.1109/ICITEED.2018.8534898.
627   Alkhair M, Meftouh K, Smaïli K, Othman N. 2019. An Arabic Corpus of Fake News: Collection,

628   Analysis and Classification. In: 292–302. DOI: 10.1007/978-3-030-32959-4_21.

629   Amjad M, Sidorov G, Zhila A. 2020. Data Augmentation using Machine Translation for Fake

630   News Detection in the Urdu Language. *Proceedings of The 12th Language Resources and*

631   *Evaluation Conference*:2530–2535.

632   Amjad M, Sidorov G, Zhila A, Gómez-Adorno H, Voronkov I, Gelbukh A. 2020. "Bend the

633   truth": Benchmark dataset for fake news detection in urdu language and its evaluation.

634   *Journal of Intelligent & Fuzzy Systems*:1–13. DOI: 10.3233/jifs-179905.

635   Ballard C, Wang W. 2016. Dynamic ensemble selection methods for heterogeneous data mining.

636   In: *2016 12th World Congress on Intelligent Control and Automation (WCICA)*. IEEE,

637   1021–1026. DOI: 10.1109/WCICA.2016.7578244.

638   Caruana R, Ksikes A, Crew G. 2014. Ensemble selection from libraries of models. In:

639   *Proceedings of the twenty-first international conference on Machine learning.*

640   Faustini PHA, Covões TF. 2020. Fake news detection in multiple platforms and languages.

641   *Expert Systems with Applications* 158:113503. DOI: 10.1016/j.eswa.2020.113503.

642   Fornaciari T, Poesio M. 2013. Automatic deception detection in Italian court cases. *Artificial*

643   *Intelligence and Law* 21:303–340. DOI: 10.1007/s10506-013-9140-4.

644   Gama J, Brazdil P. 2000. Cascade Generalization. *Machine Learning* 41:315–343. DOI:

645   10.1023/A:1007652114878.

646   González S, García S, Del Ser J, Rokach L, Herrera F. 2020. A practical tutorial on bagging and

647   boosting based ensembles for machine learning: Algorithms, software tools, performance

648   study, practical perspectives and opportunities. *Information Fusion* 64:205–237. DOI:

649   10.1016/j.inffus.2020.07.007.

650   Gravanis G, Vakali A, Diamantaras K, Karadais P. 2019. Behind the cues : A benchmarking

651   study for fake news detection. *Expert Systems With Applications* 128:201–213. DOI:

652   10.1016/j.eswa.2019.03.036.

653   Gupta D, Rani R. 2020. Improving malware detection using big data and ensemble learning.

654   *Computers and Electrical Engineering* 86:106729. DOI:

655   10.1016/j.compeleceng.2020.106729.

656   Gutierrez-Espinoza L, Abri F, Namin AS, Jones KS, Sears DRW. 2020. Fake Reviews Detection

657   through Ensemble Learning. *arXiv*:1–8.

658   Hakak S, Alazab M, Khan S, Gadekallu TR, Maddikunta PKR, Khan WZ. 2021. An ensemble

659   machine learning approach through effective feature extraction to classify fake news.

660   *Future Generation Computer Systems* 117:47–58. DOI: 10.1016/j.future.2020.11.022.

661   Huang YF, Chen PH. 2020. Fake news detection using an ensemble learning model based on

662   Self-Adaptive Harmony Search algorithms. *Expert Systems with Applications* 159:113584.

663   DOI: 10.1016/j.eswa.2020.113584.

664   Kaliyar RK, Goswami A, Narang P. 2019. Multiclass Fake News Detection using Ensemble

665   Machine Learning. In: *Proceedings of the 2019 IEEE 9th International Conference on*

666   *Advanced Computing, IACC 2019*. 103–107. DOI: 10.1109/IACC48062.2019.8971579.

667   Kapusta J, Obonya J. 2020. Improvement of misleading and fake news classification for flective

668   languages by morphological group analysis. *Informatics* 7. DOI:

669   10.3390/informatics7010004.

670   Kaur S, Kumar P, Kumaraguru P. 2020. Automating fake news detection system using multi-

671   level voting model. *Soft Computing* 24:9049–9069. DOI: 10.1007/s00500-019-04436-y.

672   Kittler J, Hater M, Duin RPW. 1996. Combining classifiers. *Proceedings - International*

673   *Conference on Pattern Recognition* 2:897–901. DOI: 10.1109/ICPR.1996.547205.

674 Komiya K, Sato N, Fujimoto K, Kotani Y. 2011. Negation naive bayes for categorization of
675   product pages on the web. *International Conference Recent Advances in Natural Language*
676   *Processing, RANLP*:586–591.
677 Lee J, Wang W, Harrou F, Sun Y. 2020. Reliable solar irradiance prediction using ensemble
678   learning-based models: A comparative study. *Energy Conversion and Management*
679   208:112582. DOI: 10.1016/j.enconman.2020.112582.
680 Mahabub A. 2020. A robust technique of fake news detection using Ensemble Voting Classifier
681   and comparison with other classifiers. *SN Applied Sciences* 2:525. DOI: 10.1007/s42452-
682   020-2326-y.
683 Monteiro RA, Santos RLS, Pardo TAS, de Almeida TA, Ruiz EES, Vale OA. 2018.
684   Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic
685   Detection Results. In: *Lecture Notes in Computer Science (including subseries Lecture*
686   *Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer International
687   Publishing, 324–334. DOI: 10.1007/978-3-319-99722-3_33.
688 Nguyen TT, Luong AV, Dang MT, Liew AW-C, McCall J. 2020. Ensemble Selection based on
689   Classifier Prediction Confidence. *Pattern Recognition* 100:107104. DOI:
690   10.1016/j.patcog.2019.107104.
691 Olowookere TA, Adewale OS. 2020. A framework for detecting credit card fraud with cost-
692   sensitive meta-learning ensemble approach. *Scientific African* 8. DOI:
693   10.1016/j.sciaf.2020.e00464.
694 Ozbay FA, Alatas B. 2020. Fake news detection within online social media using supervised
695   artificial intelligence algorithms. *Physica A: Statistical Mechanics and its Applications*
696   540:123174. DOI: 10.1016/j.physa.2019.123174.
697 Pham K, Kim D, Park S, Choi H. 2021. Ensemble learning-based classification models for slope
698   stability analysis. *Catena* 196:104886. DOI: 10.1016/j.catena.2020.104886.
699 Posadas-Durán J-P, Gómez-Adorno H, Sidorov G, Escobar JJM. 2019. Detection of fake news in
700   a new corpus for the Spanish language. *Journal of Intelligent & Fuzzy Systems* 36:4869–
701   4876. DOI: 10.3233/JIFS-179034.
702 Saeed RMK, Rady S, Gharib TF. 2019. An ensemble approach for spam detection in Arabic
703   opinion texts. *Journal of King Saud University - Computer and Information Sciences*. DOI:
704   10.1016/j.jksuci.2019.10.002.
705 Sangamnerkar S, Srinivasan R, Christhuraj MR, Sukumaran R. 2020. An ensemble technique to
706   detect fabricated news article using machine learning and natural language processing
707   techniques. *2020 International Conference for Emerging Technology, INCET 2020*:1–7.
708   DOI: 10.1109/INCET49848.2020.9154053.
709 Seewald AK, Fürnkranz J. 2001. An evaluation of grading classifiers. *Lecture Notes in Computer*
710   *Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*
711   *Bioinformatics)* 2189:115–124. DOI: 10.1007/3-540-44816-0_12.
712 Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y. 2019. Combating Fake News: A
713   Survey on Identification and Mitigation Techniques. *ACM Trans. Intell. Syst. Technol.* 10.
714   DOI: 10.1145/3305260.
715 Silva RM, Santos RLS, Almeida TA, Pardo TAS. 2020. Towards automatically filtering fake
716   news in Portuguese. *Expert Systems with Applications* 146:113199. DOI:
717   10.1016/j.eswa.2020.113199.
718 Ting KM, Witten IH. 1997. Stacking bagged and dagged models. In: Fisher DH ed. *Proc. of*
719   *ICML '97*. 367–375.

720   Verhoeven B, Daelemans W. 2014. CLiPS stylometry investigation (CSI) corpus: A Dutch
721   corpus for the detection of age, gender, personality, sentiment and deception in text. In:
722   *Proceedings of the 9th International Conference on Language Resources and Evaluation,*
723   *LREC 2014*. 3081–3085.
724   Xiao J. 2019. SVM and KNN ensemble learning for traffic incident detection. *Physica A:*
725   *Statistical Mechanics and its Applications* 517:29–35. DOI: 10.1016/j.physa.2018.10.060.
726   Zhou X, Zafarani R. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods,
727   and Opportunities. *ACM Comput. Surv.* 53. DOI: 10.1145/3395046.
728

Figure 1

Parallel vs. Sequential Ensemble

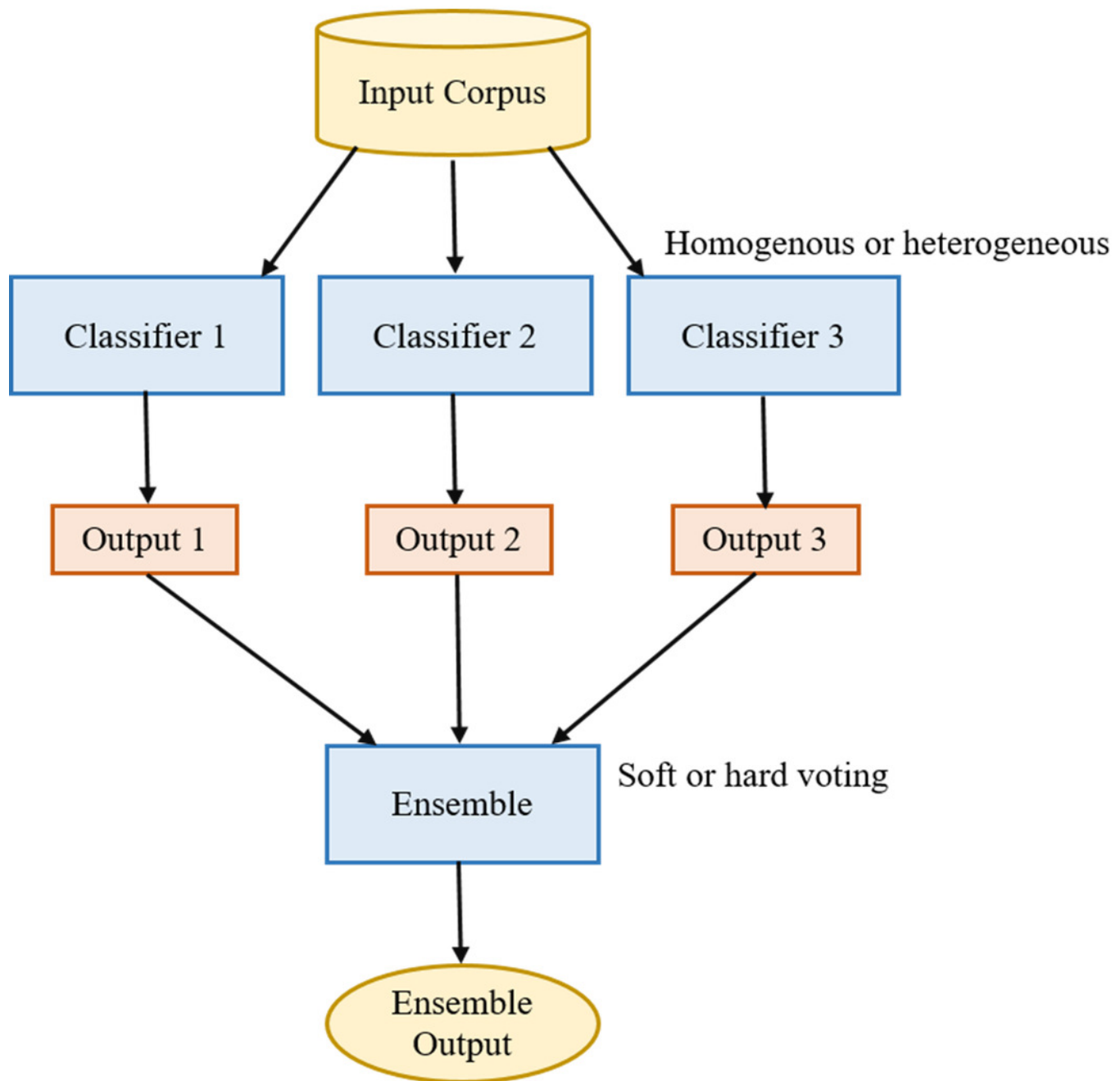



Figure 2

The architecture of the proposed framework used for fake news detection

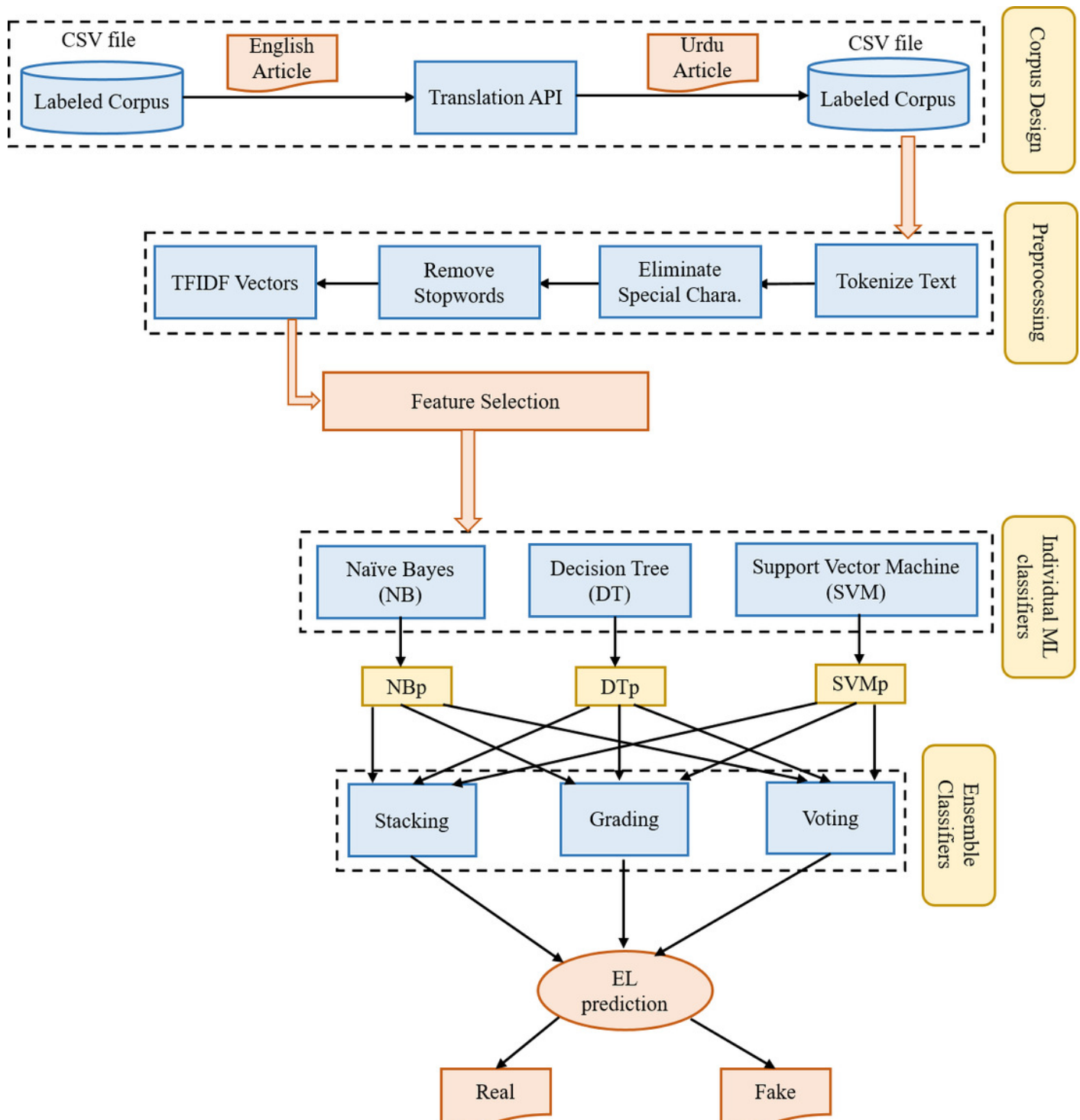
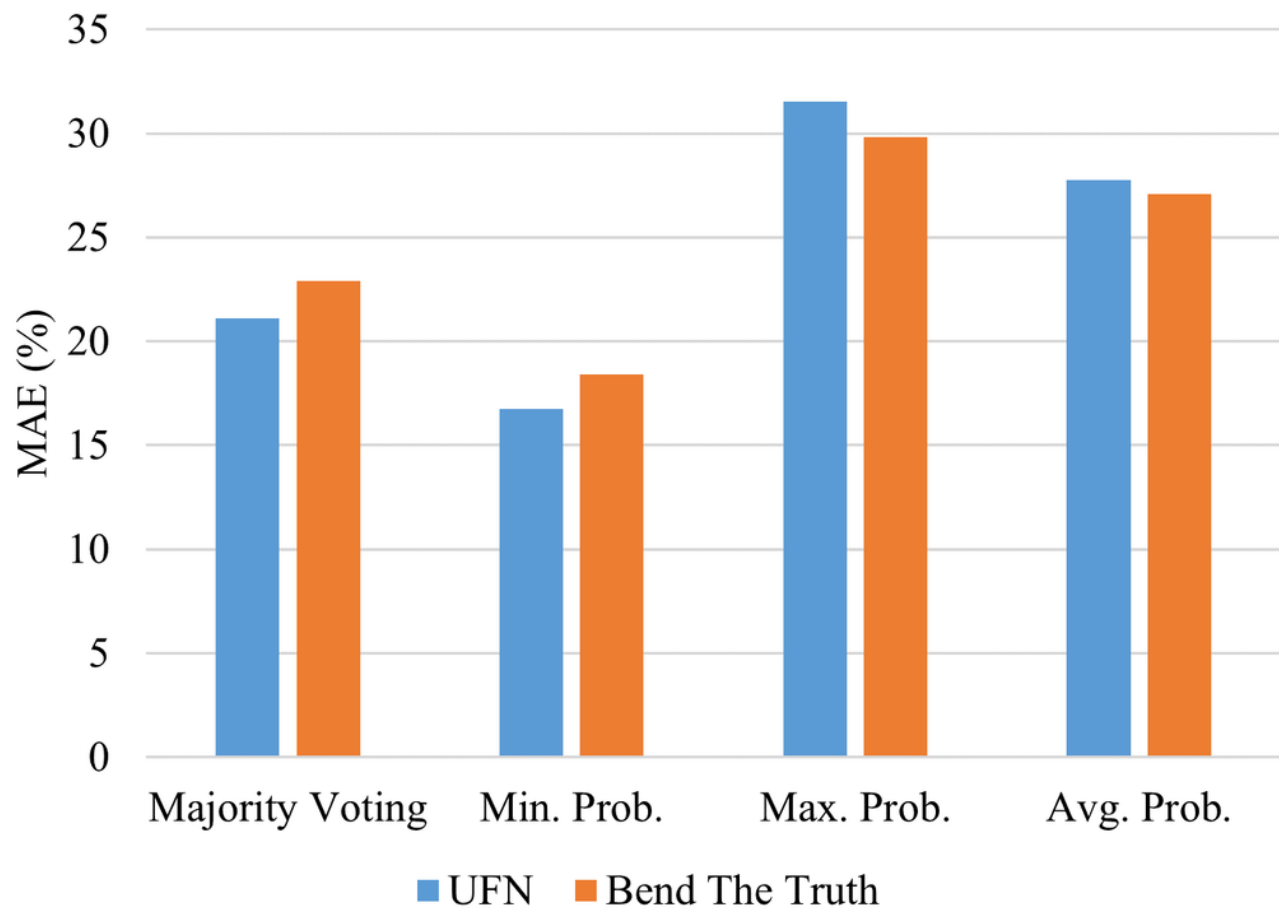



Figure 3

Performance of Voting model using numerous ensemble rules

Resource-poor language corpora for fake news detection

1   **Table 1**:

2   **Resource-poor language corpora for fake news detection**

Corpus	Language	Legitimate	Fake	References
--	Chinese	131	187	(Zhang et al., 2009)
Slovak National Corpus	Slovak	80	80	(Kapusta & Obonya, 2020)
DECOUR	Italian	1202	945	(Fornaciari & Poesio, 2013)
--	English and Spanish	100	100	(Pérez-Rosas et al., 2018)
CSI	Dutch	270	270	(Verhoeven & Daelemans, 2014)
Fake.Br	Portuguese	3600	3600	(Silva et al., 2020)
TwitterBR	Brazilian	4589	4392	(Faustini & Covoes, 2019)
Btvlifestyle	Bulgarian	69	69	(Hardalov, Koychev & Nakov, 2016)
--	Bangla	1548	993	(Hussain et al., 2020)
Spanish Fake News	Spanish	491	480	(Posadas-Durán et al., 2019)
Bend the Truth	Urdu	500	400	(Amjad et al., 2020)
MT	Urdu	400	400	(Amjad, Sidorov & Zhila, 2020)
UFN	Urdu	1032	968	Our corpus

Samples of the legitimate and fake news

1 **Table 2:**

2 **Samples of the legitimate and fake news**

Legitimate	Fake
، امریکہ عالمی جنگ 3 کی تیاری کر رہا ہے - 31 اکتوبر بجے شام 4:52 بجے انتخابی خدمت میں شامل 2016 خواتین کے علاوہ خوبصورت حقیقت پسندی۔ امریکی فوج ابھی بھی صرف رضاکارانہ ہے اور ویتنام کی جنگ کے بعد سے اس کا مسودہ نہیں رہا ہے۔ خواتین کے مسودے کے بارے میں فوج کے ایک 4 اسٹار جنرل نے یہ تبصرہ کیا تھا اور انہوں نے کہا تھا کہ لبرل یابو کو بند کردیں۔	جمہوری امیدواروں کے لئے مباحثے کے اہداف لاس ویگاس میں 2016 کی دوڑ کے پہلے - جمہوری صدارتی مباحثے کے دوران منگل کو پانچ امیدوار اسٹیج پر ہوں گے۔ شام کے اختتام پر بہتر پوزیشن میں دور آنے کے لئے ہر ایک کو کیا کرنا ہے یہ یہاں ہے:
اصلی انکشاف! سیکریٹ ایلین بیس مون کے ٹائکو کرٹر میں ملا - اصلی انکشاف! چاند کے ٹائکو کریکٹر میں ملنے والا خفیہ ایلین بیس # گرے 52 اصلی انکشاف وہ جگہ ہے جہاں آپ کو قمری سطح پر کچھ مل جاتا ہے جو ممکنہ طور پر موجود نہیں ہوسکتا جب تک کہ کوئی اسے نہ بنائے۔ کوئی راستہ نہیں یہ قدرتی تشکیل ہے --- کچھ زاویہ اجنبی 1es Âایسی تعمیر کی گئی ہے جو ---- 90 انسان ساختہ ساختہ باہمی تعامل کے بغیر ممکن بھی نہیں ہے۔ / مزید 'سگریٹ نوشی بندوق' بیرون ملک سے ذہانت کا ناقابل	نیو یارک ڈیموکریٹک پرائمری محسوس کرنے کی پانچ وجوہات - نیو یارک ، نیو یارک؟ € "اگر ہلیری کلنٹن جمہوری صدارتی دوڑ جیت رہی ہیں تو اسے ایسا کیوں محسوس ہوا جیسے وہ ہار رہی ہیں؟ ہاں ، مسز کلنٹن نے منگل کو نیویارک میں ایک اہم فتح حاصل کی ، اور انہوں نے ابتدائی طور پر اپنایا ہوا آبائی ریاست جیت لیا۔ لیکن کیا اس کا نتیجہ کبھی بھی شک میں رہنا چاہئے تھا؟ نیویارک کی سابقہ سینیٹر کی حیثیت سے وہ

The statistical description of the Urdu language corpora for fake news detection

1 **Table 3:**

2 **The statistical description of the Urdu language corpora for fake news detection**

Corpus	Bend the Truth		Urdu Fake News	
Labels	Fake	Legitimate	Fake	Legitimate
Vocabulary	1,84,023	1,20,394	11,47,547	9,54,254
Minimum Length Article	57	59	25	25
Maximum Length Article	2159	1153	7045	6068
Total Article	400	500	968	1032

Rules for classifying the discrimination using AUC

1 **Table 4:**

2 **Rules for classifying the discrimination using AUC**

AUC values	Classifier categories
$AUC < 0.5$	No Discrimination
$0.7 \leq AUC < 0.8$	Acceptable
$0.8 \leq AUC < 0.9$	Excellent
$0.9 \leq AUC$	Outstanding

The evaluation metrics values of supervised individual machine learning models

1 **Table 5:**

2 **The evaluation metrics values of supervised individual machine learning models**

Ensemble Models	BA		AUC		MAE		Time	
	UFN	BET	UFN	BET	UFN	BET	UFN	BET
Naïve Bayes	64.8	67.4	69.4	75.5	35.5	34.2	**0.94**	**0.32**
Support Vector Machine	**81.6**	**76.3**	**81.5**	**76.3**	**18.5**	**23.5**	2.56	0.67
Decision Tree	71.2	76.1	69.8	73.1	29.2	23.5	8.02	2.19
	BoW							
Naïve Bayes	75.3	56.3	80.1	61.9	24.6	43.4	0.73	**0.41**
Support Vector Machine	**86.7**	**62.7**	**86.7**	**62.7**	**13.3**	**37.1**	**0.15**	0.8
Decision Tree	76.6	55.6	76.6	58.4	23.8	43.5	10.96	4.39
	IG							
Naïve Bayes	75.3	56.7	80.5	**63.5**	24.4	42.2	0.59	0.24
Support Vector Machine	**87.3**	**62.4**	**87.3**	62.4	**12.7**	**37.7**	**0.19**	**0.17**
Decision Tree	76.8	57.0	76.8	61.5	23.6	41.6	10.28	3.56

Evaluation metrics for supervised ensemble models using both corpora

1 **Table 6:**

2 **Evaluation metrics for supervised ensemble models using both corpora**

Ensemble Models	BA		AUC		MAE		Time	
	UFN	BET	UFN	BET	UFN	BET	UFN	BET
Ensemble Selection	77.8	**83.3**	85.9	**91.0**	32.2	24.0	40.13	12.01
Cascade Generalization	80.8	83.1	**86.8**	90.5	27.0	22.0	72.89	40.43
Voting	**83.2**	81.1	77.3	76.2	**16.7**	**18.41**	**11.52**	**3.12**
Grading	77.4	79.5	77.4	79.8	22.65	20.2	195.59	51.11
Stacking	81.4	80.9	80.8	87.3	30.1	25.1	125.82	36.8
	BoW							
Ensemble Selection	78.9	**66.1**	87.9	**74.0**	31.7	40.4	77.65	17.75
Cascade Generalization	86.1	56.1	**92.0**	62.2	21.94	44.3	94.39	37.3
Voting	**86.7**	62.4	85.8	62.2	**13.3**	**37.3**	**14.87**	**6.14**
Grading	83.4	59.1	83.4	59.1	16.5	39.9	346.02	64.86
Stacking	86.6	55.9	85.8	63.9	23.6	56.3	205.3	50.3
	IG							
Ensemble Selection	80.6	**66.7**	88.4	**74.9**	31.0	40.1	47.18	32.88
Cascade Generalization	87.2	61.0	**92.7**	67.9	21	43.3	146.19	108.55
Voting	**89.3**	64.2	84.1	61.9	**10.67**	**35.4**	**24.63**	**18.38**
Grading	85.8	62.2	85.8	60.8	14.1	39.3	262.4	101.1
Stacking	87.1	54.5	92.5	60.6	19.47	46.9	232.62	61.05

Summary of the overall results obtained from the best ML model, EL model, and features

1   **Table 7:**

2   **Summary of the overall results obtained from the best ML model, EL model, and features**

	Classifier	**UFN**	**Feature**	**Classifier**	**BET**	**Feature**
Time	SVM	0.15	BoW	SVM	0.17	IG
	Voting	11.52	Tri-gram	Voting	3.12	Tri-gram
MAE	SVM	12.2	IG	SVM	23.5	Tri-gram
	Voting	**10.67**	IG	**Voting**	18.41	**Tri-gram**
AUC	SVM	87.3	IG	SVM	76.3	Tri-gram
	Cascade Learning	92.7	IG	Ensemble Learning	91.0	Tri-gram
BA	SVM	87.3	IG	SVM	76.3	Tri-gram
	Voting	89.3	IG	Voting	83.3	Tri-gram