Overview of the Shared Task on Fake News Detection in Urdu at FIRE 2021

Maaz Amjad[a], Sabur Butt[a], Hamza Imam Amjad[c], Alisa Zhila[b], Grigori Sidorov[a] and Alexander Gelbukh[a]

[a]*Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC), Mexico*
[b]*Ronin Institute for Independent Scholarship, United States*
[c]*Moscow Institute of Physics and Technology, Russia*

Abstract

Automatic detection of fake news is a highly important task in the contemporary world. This study reports the 2nd shared task called UrduFake@FIRE2021 on identifying fake news detection in Urdu language. The goal of the shared task is to motivate the community to come up with efficient methods for solving this vital problem, particularly for the Urdu language. The task is posed as a binary classification problem to label a given news article as a real or a fake news article. The organizers provide a dataset comprising news in five domains: (i) Health, (ii) Sports, (iii) Showbiz, (iv) Technology, and (v) Business, split into training and testing sets. The training set contains 1300 annotated news articles —750 real news, 550 fake news, while the testing set contains 300 news articles —200 real, 100 fake news. 34 teams from 7 different countries (China, Egypt, Israel, India, Mexico, Pakistan, and UAE) registered for participation in the UrduFake@FIRE2021 shared task. Out of those, 18 teams submitted their experimental results and 11 of those submitted their technical reports, which is substantially higher compared to the UrduFake shared task in 2020 when only 6 teams submitted their technical reports. The technical reports submitted by the participants demonstrated different data representation techniques ranging from count-based BoW features to word vector embeddings as well as the use of numerous machine learning algorithms ranging from traditional SVM to various neural network architectures including Transformers such as BERT and RoBERTa. In this year's competition, the best performing system obtained an F1-macro score of 0.679, which is lower than the past year's best result of 0.907 F1-macro. Admittedly, while training sets from the past and the current years overlap to a large extent, the testing set provided this year is completely different.

Keywords

Natural Language Processing, NLP, fake news detection, shared task, Urdu language, text classification, low resource language, medium resource language

1. Introduction

The proliferation of social media brought in various forms of cybercrime that urgently need automatic solution for the safety of people online and beyond [1, 2, 3]. Among these problems, fake news dissemination is a critical problem that spreads in the form of advertisements, posts,

FIRE 21: Forum for Information Retrieval Evaluation, December 13–17, 2021, India
maazamjad@phystech.edu (M. Amjad); sabur@nlp.cic.ipn.mx (S. Butt); hamzaimamamjad@phystech.edu (H. I. Amjad); alisa.zhila@ronininstitute.org (A. Zhila); sidorov@cic.ipn.mx (G. Sidorov); gelbukh@gelbukh.com (A. Gelbukh)
https://nlp.cic.ipn.mx/maazamjad/ (M. Amjad)

news articles and others. It is an outstanding threat to journalism, democracy, and freedom of expression that negatively affects trust between the media outlets and the users. The socio-political impact of fake news can be observed with the incidents such as 2016 United States presidential elections. Post election studies showed [4, 5] various occasions of fake news spiking on social media with content emphasising nonexistent cause–effect relationship aggravating the division between the political groups. Behavioural studies [6, 7] showed the effect that exposure to fake news has on political and social issues through randomized controlled experiments. The results established that fake news can cause a change in views and behaviour regarding topics of broad domain including politics. Hence, the status quo of fake news needs immediate attention and robust solutions.

Natural language processing (NLP) researchers formulated the problem into subcategories of fake news such as satire [8, 9], propaganda [10, 11], deception [12, 13], fact cherry picking [14, 15], clickbaits [16, 17, 18], hyperpartisanship [19, 20], and claim "check-worthiness" for potentially untruthful facts [21, 22, 23, 24, 25, 26]. Each subcategory has distinct features and solutions to achieve desirable results. Fake news becomes a very challenging problem to control because of the Velocity, Volume, Variety, and Time Latency of its spread [27]. The community behind the fake news content marches the spread at a pace which becomes higher than the real news dissemination itself.

This paper describes the UrduFake@FIRE2021 shared task and its results. The task invited the participants to tackle the problem of automatic fake news detection in Urdu in Nastalíq script . The problem is shaped into a binary classification problem in which news articles from various sources including such news outlets as BBC Urdu News, CNN Urdu, Express-News, Jung News, Naway Waqat, and others, are offered for classification as fake or real. During the active competition phase the ground truth annotations for the testing set were hidden from the participants, while the training set was provided with the corresponding ground truth annotations. After the end of the competition, the both parts of the dataset were made publicly available along with the corresponding ground truth annotations at the CICLing 2021 UrduFake track at FIRE 2021 shared task homesite [1]. This year's track is the continuation of CICLing 2020 UrduFake track at FIRE 2020 [28, 29] with the core difference being the size of the offered dataset. The training data has increased to facilitate a wider range of neural network and particularly deep learning studies and to get more insightful information from data analysis. In the shared task the participating teams were requested to submit only their top 3 different runs, among which the best run was considered for submission of the technical report paper describing the approach.

The paper is structured as follows. An overview of previous relevant research can be found in Section 3. We provide the task description in Section 4 and explain in detail the data collection and annotation procedure in Section 5. Training and testing set splits and statistics are outlined in Section 5.2. Sections 6 and 7 describe the choice of evaluation metrics and baselines correspondingly. A high level overview and comparison of the solutions and approaches submitted by the participants is provided in Section 8 along with the final results summarized in Section Sections 9. A brief summary of the UrduFake@FIRE2021 track can be found in a separate publication [30].

[1] https://www.urdufake2021.cicling.org/home

2. Importance of Fake News Detection in Urdu

Urdu is the national language of Pakistan and has more than 230 million [2] speakers worldwide. Many of these speakers carry out their written communication in the Nastalíq script. Urdu is commonly written in the Nastalíq script, while the Devanagari script is commonly used for Hindi. However, due to cultural and geographical proximity, Devanagari may be also used for writing in Urdu. This creates a situation of *digraphia* for the Urdu language when two scripts are used for writing in a language. Apart from this commonality, Urdu has other structural similarities with Hindi and other South Asian languages [31]. The emergence of Urdu came in the form of tribal movement which resulted in the merging of morphological and syntactic structures of Arabic, Persian, Turkish, Sanskrit, and recently English in the conversational usage. Due to the mixture of various languages, Urdu has more complexity than the other existing languages and, consequently, requires more careful processing.

South Asia has been suffering from numerous instances of fake news affecting its political, social, and economic situation. For example, Dr. Shahid Masood [3] who works as a TV anchor in Pakistan, was exiled and tortured for spreading false information about a child rape case. Another case of fake news in India was reported in the Washington Post [4], where many innocent people died because of a child trafficking report.

These severe consequences of fake news reporting surge the urge for high quality automation of fake news detection in Urdu. Given that despite the numerous speakers Urdu is still a low/medium resourced language, we strive for providing larger annotated datasets and incentivize the community to develop state-of-the-art solutions for early detection of fake news.

3. Literature Review

Contemporary fake news is not solely produced by humans, but can also be generated through bots [32]. These bots replicate human behaviour and are created for the purpose of spamming, spreading rumours and misinformation on various social media platforms. Social context [27] has been one of the key indicators to differentiate between fake and real news patterns. Researcher have dealt with the fake news problem with the aid of a wide range of feature based approaches [33, 34, 35, 36, 37] including features such as engagement, user attributes, stylistic features, linguistic features, and personality based features.

Earlier solutions [27] in fake news detection used fact checking with the aid of experts, however, the solution was time consuming and labor-cost intensive. Hence, NLP experts moved on to finding automatic solutions based on machine learning and deep learning algorithms [38, 39, 40, 41]. Studies have found unique emotional language cues [42] and emotional pattern [43] differences between real and fake news. Among the supervised machine learning techniques [38, 44, 45], we have seen Random Forest (RF), Support Vector Machine (SVM), and Decision Trees repeatedly used for fake news detection. Other research have used neural network ensembles combining various neural network architectures. Thus, Roy et al. [41] fed article representations

[2] https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/
[3] https://www.globalvillagespace.com/dr-shahid-masoods-claims-about-zainabs-murderer-prove-false/
[4] https://tinyurl.com/ynhsudnx

provided by CNN and Bi-LSTM models into MLP for the final classification which allowed for considering more contextual information. Yet another approach towards identifying fake news is looking at the news sources instead of the text content in the article, as news sources can provide valuable insights [46].

The dataset created for fake news identification mostly rely on social media platforms and news outlets. The majority of the existing datasets are available in English [29, 27]. Recently, datasets and studies on various subcategories of fake news appeared in other languages: Persian [47], Spanish [46, 48], Arabic [49, 50], German [51], Bangla [11], Dutch [19], Italian [52], Portuguese [53], Urdu [54], and Hindi [55].

Some of the online challenges to improve automatic fake news systems include Fake News Challenge [5], multiple fake news detection competitions on Kaggle [6] as well as shared task tracks organized by the academic community: PAN 2020 [46], RumourEval Task 8 of SemEval 2017 for English [56], RumourEval Task 7 of SemEval-2019 for English [57], and others.

4. Task Description

This task is aimed to motivate the community to come up with methods and systems for automatic fake news detection in Urdu language by providing an annotated dataset with a train/test split and competitive settings. The challenge is posed as a binary classification task where participants are to train their classifiers on the provided training part of the dataset and to submit the labels, either fake or real, for each news article from the testing set, the ground truth annotations for the latter being hidden from the participants. Organizers compute the evaluation metrics for each submission by comparing the submitted labels to the ground truth annotations.

The motivations of this shared task is to investigate whether and to which extent the textual content alone can be grounds for fake news detection and examine the efficiency of machine learning algorithms in identifying fake news articles written in Urdu in the Nastalíq script.

Here, a fake news article and fake news detection are defined as follows:

- `Fake News`: A news article that contains factually incorrect information with the intention to deceive a reader and to make the reader believe that it is factually correct.
- `Fake News Detection`: Suppose that n is a news article (without annotation) and $n \in N$, where N is the total number of news articles. A fake news detection is a process in which an algorithm calculates the likelihood of whether a given news article n is a fake news article by assigning a value between 0 and 1. In mathematical terms, this can be described as $S(n) \in [0, 1]$. In other words, if $S(\hat{n}) > S(n)$, this indicates that the $\hat{n}$ new article has a higher chances to be fake news than the n news article. Also, it is important to define a threshold. The threshold β is a hyperparameter cut value selected by the algorithm developers such that if the algorithm assigns an equal or higher value to a news article as compared to the threshold, then the news article will be tagged as fake. A threshold β can be defined so that the prediction function $F(n)$: $n \rightarrow \{not\ fake, fake\}$ is:

[5] http://www.fakenewschallenge.org/

[6] https://www.kaggle.com/c/fake-news/data, https://www.kaggle.com/c/fakenewskdd2020

$$F(N) = \begin{cases} fake, & if \quad S(n) >= \beta), \\ not\, fake, & \text{otherwise.} \end{cases}$$

More elaborated definition of fake news is provided in our previous work [54].

5. Dataset Collection and Annotation

This section gives an outline of the dataset created for the UrduFake shared task at FIRE 2021. Our previous research [54] reported the first version of this dataset, called "Bend The Truth" that contained 500 real news and 400 corresponding fake news. A new training dataset and test dataset data was acquired using the dataset collection and annotation guidelines presented in our previous research [54]. The dataset presented in this shared task is publicly available and can be used for research objectives [7].

The training dataset was released on April 30, 2021 [8]. It is important to mention that the training dataset used in 2021 UrduFake task comprised 1300 news article. This dataset was made up by combining the training dataset, which we presented in our previous research [54] "Bend The Truth" and testing dataset collected for UrduFake 2020 shared task. The training dataset contained 750 real news articles and 550 fake news articles. we presented a new test dataset that contained 200 real news and 100 fake news articles collected from January 2021 to August 2021 to test the proposed systems.

A crowdsourcing technique was used to collect the fake news articles. In other words, the fake news were composed by hiring professional journalists who deliberately wrote fake news of the corresponding real news. The journalists were provided a set of instructions to follow while writing fake news articles. This dataset contains five domains of the news: (i) Business, (ii) Health, (iii) Sports, (iv) Showbiz (entertainment), and (v) Technology.

5.1. Procedure for Dataset Annotation

All the news articles were labelled into two two types of news: (i) real news article, and (ii) fake news article. Different techniques were used to annotate and assemble real and fake news. This dataset can be used for future research using supervised machine learning and deep learning techniques. Figure 1 shows the list of news organizations used to crawl news articles.

5.1.1. Real News Collection and Annotation

To assemble real news articles, various traditional news media mainstream were used to crawl news manually. Manual procedures were followed to annotate a news article using the underline guidelines, the news would label as real news. The news organizations used to gather news items for annotation are presented in Figure 1 and all the news were manually crawled. The following guidelines were used to annotate a news item as a real news:

[7] https://github.com/MaazAmjad/Urdu-Fake-news-detection-FIRE2021
[8] https://www.urdufake2021.cicling.org/home

Name	URL	Origin
BBC News	www.bbc.com/urdu	England
CNN Urdu	cnnurdu.us	USA
Dawn news	www.dawnnews.tv	Pakistan
Daily Pakistan	dailypakistan.com.pk	Pakistan
Eteemad News	www.etemaaddaily.com	India
Express-News	www.express.pk	Pakistan
Hamariweb	hamariweb.com	Pakistan
Jung News	jang.com.pk	Pakistan
Mashriq News	www.mashriqtv.pk	Pakistan
Nawaiwaqt News	www.nawaiwaqt.com.pk	Pakistan
Roznama Dunya	dunya.com.pk	Pakistan
The daily siasat	urdu.siasat.com	India
Urdu news room	www.urdunewsroom.com	USA
Urdupoint	www.urdupoint.com	Pakistan
Voice of America	www.urduvoa.com	USA
Waqt news	waqtnews.tv	Pakistan

Figure 1: Legitimate websites

1. The news article was labeled as real news if the news meets the following criteria:
    - That news article is published by a credible newspaper or a prominent news media agency.
    - The integrity of that news article can be verified by other credible newspaper agencies. This was an important point to do fact-checking. For example, manual source verification was performed to check place of the event, image, date of the news and whether the provided information in the news article matched with the same news article but published by other newspaper or news agency as well.
    - Incongruity between news titles and its content was also confirmed to ensure that a news article has a correlation between the news headline and the body text. We read the complete news articles to check the incongruity between news titles and the body text.

It is important to highlight that a news article was removed If it did not fulfil one of the aforementioned criteria. Different news articles contained different words length. For example, CNN publish news articles that contains between 200-300 words. On the other hand, a news article published by BBC Urdu news typically contains on average 1500 words. Therefore, the real news articles contains heterogeneous length of words. This is how all the real news articles were collected and annotated.

5.1.2. Professional Crowdsourcing of Fake News

To obtain fake news, the services of professional journalist were used who work in different news organizations in Pakistan. We hired professional columnist because they are expertise in writing news articles, and use different journalists techniques to make the news interesting to hook and and their written fake news can easily trick the reader. The real news articles were provided to the journalists and they were asked to write fake news corresponding to the real

news. In other words, if a real news contains story about football, the correspond fake news article should also contain similar story but with fabricated information.

We used professional "crowdsourcing" for collecting fake news and the reasons are described as follows:

1. The news articles analysis with manual procedures for verification through web scraping approach was unfeasible. This is due to the facet that it is extremely challenging task to find the corresponding fake news of a real news article.
2. No online service in Urdu language is available for news fact-checking. Unlike English, the news fact-checking is manually performed in Urdu.

This dataset contains news of five domains: (i) business, (ii) education, (iii) sports, (iv) showbiz (entertainment), and (v) technology. The journalists expertise was taken into account to ensure that the fake news corresponding to the real news is written by the domain expert. The journalists were asked to keep the same length of the news (fake news article should have the same words length as real news). In addition, we also instructed journalists to mitigate defined patterns so that the undesirable clues should not be induced to classify news articles. Therefore, journalists' expertise were used to collected all the fake news articles.

5.2. Training and Testing Split

5.2.1. Training and Validation Set

The training set contained 1300 news articles, in which 750 news articles were annotated as real, and 550 news articles were annotated as fake news article. The training set and the testing set contained five types of news: (i) Business, (ii) Health, (iii) Showbiz (entertainment), (iv) Sports, and (v) Technology. Participants were allowed to use of the training set for validation, development, and parameter tuning. The training dataset made up by combining the training dataset, which we presented in our previous research [54] "Bend The Truth" and the testing dataset collected for UrduFake 2020 shared task.

5.2.2. Test dataset

The new test set was introduced that contained 200 real news and 100 fake news articles collected from January 2021 to August 2021. The test set was presented without the ground truth labels so that all the participants could evaluate and test the performance of their proposed systems. To evaluate and compare the performance of the classifiers submitted by the participants, the organizers used the truth labels of the test set. It is worth mentioning that the participants were unaware of the distributions of real and fake news in the test set.

5.3. Dataset Statistics

In this shared task, we divided the dataset into two parts: (i) training set, and (ii) testing set. Initially, the training set was released so that the participants can train their classification models. Then, the test set was released so that the participants can predict the labels of whether

a given news is real or fake. Table 1 describes the corpus distribution of the news articles by topics for the training and testing sets.

Table 1
Domain Distribution in Train and Test subsets

Domain	Train		Test	
	real	fake	real	fake
Business	150	80	40	20
Health	150	130	40	20
Showbiz	150	130	40	20
Sports	150	80	40	20
Technology	150	130	40	20
Totals	**750**	**550**	**200**	**100**

6. Evaluation Metrics

This is a binary classification task in which the task is to classify a news article as fake or real. All the participating teams were allowed to submit up to 3 different runs, i.e., labels for the testing set generated by their proposed classifiers. The ground truth annotations were used to compare the labels predicted by the participants' classifiers. We used the evaluation metrics commonly used to measure the performance of binary classification on imbalanced datasets: two sets of *Precision* (P), *Recall* (R), and *F1* score, one for the "real" class treated as a target class and the other for the "fake" class; the inter-class metrics *Accuracy* and *F1-macro*. The macro-averaged F1-macro, which is the average of $F1_{real}$ and $F1_{fake}$, was also calculated to accommodate the dataset skew towards the real class. As detection of both classes (real and fake) is equally important, this is why we evaluated performance against both classes.

7. Baselines

To introduce a baseline, we used the bag of words (BoW) approach. We used a combination of character, word, and function word bi-grams with TF-IDF weighting scheme for text representation. Function words are similar to stopwords, for more elaborated definition and list we suggest to refer to [58]. Decision Tree was selected as a classifier, which achieved surprisingly good results compared to other traditional ML classifiers on our trial runs. d In the trial runs, five weighting schemes (tf-idf, log-ent, norm, binary, relative frequency) [54] were used for the experiments along with different machine learning classifiers such Decision Tree, Random Forest, Logistic Regression, AdaBoost, SVM, and Naive Bayes. We tried different *n*-grams, $n = \{1, ..., 7\}$. We noticed that the classifiers started to obtain insignificant results when $n = 5$ or higher. Finally, the Decision Tree algorithm outperformed other classifiers in identifying fake news. The baseline code is publicaly available [9].

[9] https://github.com/MaazAmjad/Urdu-Fake-news-detection-FIRE2021

8. Overview of the Submitted Approaches

This section briefly overviews the methods applied in the competition by the teams. In total 34 teams registered for the competition, and 18 teams submitted experimental results on a test dataset. We report the findings of 11 teams who submitted their methodologies in the form of technical report papers. The registered participants were from the countries where Urdu language has presence or cause interest: Pakistan, India, United Arab Emirates, Israel, and Egypt. Table 2 shows the approaches used by the teams and table 3 tells the best run scores achieved through those methods.

1. **Nayel:** The best performing model used the linear classifier function from the scikit-learn package `sklearn.linear_model.SGDClassifier` that by default fits a linear SVM classifier with Stochastic Gradient Descent (SGD) optimization algorithm. The team trained it on word token tri-gram features weighted with TF-IDF scheme. The model uses tokens without any preprocessing, which increases the number of features.
2. **Abdullah-Khurem:** The team experimented with neural network techniques: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and textCNN, for fake news detection in Urdu. The final submission used textCNN with TF-IDF features which and ranked second in the competition.
3. **Hammad-Khurem:** The methodology used no pro-processing and proposed a voting-based approach with a majority voting ensemble of boosting-based ML classifiers: AdaBoost, LightGBM and XGBoost. The proposed approach employed BoW features.
4. **Muhammad Homayoun:** The participant reported results using Convolution Neural Network (CNN) with four input channels. Before classification, the data was pre-processed by removing diacritic, normalization, stopword removal and lemmatization. The best results submitted used character level sequences (n-grams) for text representation.
5. **Snehaan Bhawal:** The transformer methods (MuRIL, BERT) gave the best results with no pro-processing. Multilingual Representations for Indian Languages (MuRIL) was submitted to the competition as the final submission and slightly outranked the non-specialized BERT.
6. **MUCIC:** The participants used three feature selection algorithms (Chi-square, Mutual Information Gain (MIG), and f_classif) to choose the best features from the word and character n-grams. The intersection of selected features was passed into an ensemble of ML classifiers (Linear SVM (LSVM), LR, MLP, XGB, and RF) with soft voting and feature selection to achieve the best results.
7. **SOA NLP:** The submitted method used character level uni, bi and tri-gram TF-IDF features as an input to dense neural network (DNN). The best results used a learning rate of 0.001, a dropout rate of 0.3, a batch size of 16, Adam as an optimizer and binary cross-entropy as a loss function with 100 epoch training.
8. **Dinamore&Elyasafdi_SVC:** The team used classical machine learning algorithms: SVM, Random Forest (RF), and Logistic Regression (LR). They used character tri-gram features with only one pre-processing step of lowercasing all letters.
9. **MUCS:** In the pre-processing stage the participants removed non-relevant characters, stopwords and punctuation. They used pre-trained Urdu word embeddings from fastText

and TF-IDF of words as well as character n-grams as features. Similar to team MUCIC, an ensemble of ML classifiers (RF, MLP, AdaBoost, and GraidentBoost) were used with soft voting to achieve the highest F1 macro.

10. **Iqra Ameer:** This is another study that used BERT-base model. The best results were reported using both the training and validation set for training of the model.
11. **Sakshi Kalra:** The best team runs used an ensemble of various transformer methods (RoBERTa, XLM-RoBERTa and Multilingual BERT) as well as a single specialized transformer RoBERTa-urdu-small. The text input was normalized. Interestingly, the best performing method on the test set turned out to be RoBERTa-urdu-small which exceeded the three-transformer ensemble method (XLM-RoBERTa+Multilingual BERT+RoBERta).

Table 2
Approaches used by the participating teams

System/Team Name	Text Representation	Feature Weighting Scheme	Classifying Algorithm	is NN-based?
Nayel	tri-gram	TF-IDF	linear SVM with SGD	No
Abdullah-Khurem	Word2Vec, GloVe, fastText	TF-IDF	textCNN	Yes
Hammad-Khurem	BoW	count *(?)*	ensemble XGBoost+LightGBM+AdaBoost	No
Muhammad Homayoun	char 2, 6-gram	*N/A*	CNN	Yes
Snehaan Bhawal	transformer embeddings	*N/A*	MuRIL	Yes
MUCIC	word- & char 1, 2-grams	TF-IDF	ensemble linSVM+LR+MLP+XGB+RF	Yes (MLP) & No
SOA NLP	char 1, 3-grams	TF-IDF	DNN	Yes
Dinamore&Elyasafdi_SVC	char 3-grams	TF-IDF	SVM	No
MUCS	word fastText emb & char 2, 3-grams	TF-IDF for char-grams	ensemble MLP+AdaBoost+GraidentBoost+RF	Yes (MLP) & No
Iqra Ameer	transformer emb	*N/A*	BERT-base	Yes
Sakshi Kalra	transformer emb	*N/A*	RoBERTa-urdu-small	Yes

9. Results and Discussion

Each team submitted three runs (proposed three different systems), and only the best run was considered for comparison. We calculated the results of all the submitted runs by each teams individually and only reported the results obtained by the best run. Table 3 shows the the results of the best run (among up to three submitted runs) submitted by the participating teams. We used F1-macro score to rank the participants systems. The aggregated statistics about the performance is presented in Table 4.

It can be observed that only two systems outperformed the baseline, and all the other systems did not beat the F1-macro score of the baseline. The team Nayel obtained the the best results in terms of F1-macro, Accuracy, P_{fake} (precision) scores. The team Abdullah-Khurem obtained the the second best results in terms of F1-macro, Accuracy, R_{fake} (recall), P_{real} (precision), and $F1_{fake}$ scores. Moreover, the baseline approach with the combination of char-word-function words bi-gram with tf-idf weighting scheme using Decision Tree classifier the third position in the shared task with the difference of 2.8% from Nayel system and 1.2% from Abdullah-Khurem system in F1-macro score.

Table 3 presents the best results of the submitted systems.

Table 4 presents aggregated statistics of the submitted systems.

Table 3
Participants' best run scores.

No	Team Names	Fake Class			Real Class				
		Prec	Recall	F1_Fake	Prec	Recall	F1_Real	**F1_Macro**	Accuracy
1	Nayel	0.754	0.400	0.522	0.757	0.935	0.836	0.679	0.756
2	Abdullah-Khurem	0.592	0.480	0.530	0.762	0.835	0.797	0.663	0.716
3	Baseline	0.584	0.450	0.508	0.753	0.840	0.794	0.651	0.710
4	Hammad-Khurem	0.634	0.330	0.434	0.729	0.905	0.808	0.621	0.713
5	Muhammad Homayoun	0.480	0.490	0.485	0.742	0.735	0.738	0.611	0.653
6	Snehaan bhawal	0.960	0.240	0.384	0.723	0.995	0.837	0.610	0.743
7	MUCIC	0.821	0.230	0.359	0.716	0.975	0.826	0.592	0.726
8	SOA NLP	0.793	0.230	0.356	0.356	0.715	0.823	0.590	0.590
9	Dinamore & Elyasafdi _SVC	0.720	0.180	0.288	0.701	0.965	0.812	0.550	0.703
10	MUCS	0.850	0.170	0.283	0.703	0.985	0.820	0.552	0.713
11	Iqra Ameer	0.454	0.100	0.163	0.676	0.940	0.786	0.475	0.660
12	Sakshi kalra	0.266	0.120	0.165	0.654	0.835	0.734	0.449	0.596

Table 4
Aggregated statistics of the submitted systems and the baseline.

Stat. metric	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$	F1-macro	Acc.
mean	0.596	0.294	0.348	0.679	0.837	0.757	0.553	0.658
std	0.201	0.171	0.136	0.105	0.201	0.137	0.093	0.106
min	0.262	0.070	0.115	0.356	0.155	0.228	0.296	0.303
percentil 10%	0.319	0.100	0.165	0.610	0.684	0.713	0.448	0.584
percentil 25%	0.462	0.145	0.229	0.680	0.802	0.761	0.490	0.646
percentil 50%	0.592	0.240	0.364	0.716	0.905	0.797	0.590	0.686
percentil 75%	0.737	0.430	0.451	0.726	0.970	0.816	0.61	0.713
percentil 80%	0.769	0.462	0.473	0.734	0.975	0.821	0.615	0.714
percentil 90%	0.826	0.506	0.510	0.753	0.981	0.828	0.653	0.729
max	0.960	0.600	0.530	0.762	0.995	0.837	0.679	0.756

10. Conclusion

Automatic fake news detection is an important task, especially in low resource languages. This research presents the second shared task (the first task was organized in 2020) in identifying fake news in Urdu namely the UrduFake 2021 track at FIRE 2021. A training and testing dataset was presented so that the participants could train and test their proposed systems. The dataset contained news in five domains (business, health, sports, showbiz, and technology). All the real news were crawled from credible sources and manually annotated while the fake news were written by the professional journalists.

In this shared task, thirty four teams from seven different countries registered and eighteen teams submitted their proposed systems (runs). The participants used different techniques ranging from the traditional feature-crafting and application of traditional ML algorithms to word representation through pre-trained embeddings to contextual representation and end-to-end neural network based methods. The approaches used included ensemble methods, CNN, and non-Urdu specialized Transformers (BERT, RoBERTa) as well as Urdu-specialized (MuRIL, RoBERTa-urdu-small) .

Team Nayel outperformed all the proposed systems by using the linear SVM optimized with Stochastic Gradient Descent and obtained F1-macro score of 0.67. This result reveals that classical feature-based models perform better compared to the contextual representation and large neural network algorithms. The characteristics of the dataset require further investigation to better explain this observation.

This shared task aims to attract and encourage researchers working in different NLP domains to address the automatic fake news detection task and help to mitigate the proliferation of fake content on the web. Moreover, this also offers a unique opportunity to explore the sufficiency of textual content modality alone and effectiveness of fusion methods. In addition, an annotated news dataset in Urdu is also provided to encourage more research to address the automatic fake news detection in Urdu language.

Acknowledgments

This competition was organized with the support from the Mexican Government through the grant A1-S- 47854 of the CONACYT, Mexico and grants 20211784, 20211884, and 20211178 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico.

References

[1] S. Butt, N. Ashraf, G. Sidorov, A. Gelbukh, Sexism identification using BERT and data augmentation - EXIST2021, in: International Conference of the Spanish Society for Natural Language Processing SEPLN 2021, IberLEF 2021, Spain, 2021.

[2] N. Ashraf, R. Mustafa, G. Sidorov, A. Gelbukh, Individual vs. group violent threats classification in online discussions, in: Companion Proceedings of the Web Conference 2020, WWW '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 629–633.

[3] R. U. Mustafa, N. Ashraf, F. S. Ahmed, J. Ferzund, B. Shahzad, A. Gelbukh, A multiclass depression detection in social media based on sentiment analysis, in: 17th International Conference on Information Technology–New Generations (ITNG 2020), Springer International Publishing, Cham, 2020, pp. 659–662.

[4] K. Ali, K. Zain-ul abdin, Post-truth propaganda: heuristic processing of political fake news on facebook during the 2016 us presidential election, Journal of Applied Communication Research 49 (2021) 109–128.

[5] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on twitter during the 2016 us presidential election, Science 363 (2019) 374–378.

[6] M. Anderson, Social media causes some users to rethink their views on an issue, Pew Research Center (2016).

[7] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, J. H. Fowler, A 61-million-person experiment in social influence and political mobilization, Nature 489 (2012) 295–298.

[8] C. Burfoot, T. Baldwin, Automatic satire detection: Are you having a laugh?, in: Proceedings of the ACL-IJCNLP 2009 conference short papers, 2009, pp. 161–164.

[9] A. N. Reganti, T. Maheshwari, U. Kumar, A. Das, R. Bajpai, Modeling satire in English text for automatic detection, in: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), IEEE, 2016, pp. 970–977.
[10] A. Zubiaga, M. Liakata, R. Procter, Exploiting context for rumour detection in social media, in: International Conference on Social Informatics, Springer, 2017, pp. 109–123.
[11] M. Z. Hossain, M. A. Rahman, M. S. Islam, S. Kar, Banfakenews: A dataset for detecting fake news in Bangla, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 2862–2871.
[12] Rubin, V. L, Chen, Yimin, Conroy, N. K, Deception detection for news: Three types of fakes, Proceedings of the Association for Information Science and Technology 52 (2015) 1–4.
[13] S. Feng, R. Banerjee, Y. Choi, Syntactic stylometry for deception detection, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2012, pp. 171–175.
[14] A. Asudeh, H. V. Jagadish, Y. Wu, C. Yu, On detecting cherry-picked trendlines, Proceedings of the VLDB Endowment 13 (2020) 939–952.
[15] V. F. Hendricks, M. Vestergaard, Alternative facts, misinformation, and fake news, in: Reality Lost, Springer, 2019, pp. 49–77.
[16] A. Chakraborty, B. Paranjape, S. Kakarla, N. Ganguly, Stop clickbait: Detecting and preventing clickbaits in online news media, in: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), IEEE, 2016, pp. 9–16.
[17] Y. Chen, N. J. Conroy, V. L. Rubin, Misleading online content: recognizing clickbait as "false news", in: Proceedings of the 2015 ACM on workshop on multimodal deception detection, 2015, pp. 15–19.
[18] M. Potthast, S. Köpsel, B. Stein, M. Hagen, Clickbait detection, in: European Conference on Information Retrieval, Springer, 2016, pp. 810–817.
[19] M. S. Looijenga, The Detection of Fake Messages using Machine Learning, B.S. thesis, University of Twente, 2018.
[20] Y. Jiang, J. Petrak, X. Song, K. Bontcheva, D. Maynard, Team bertha von suttner at SemEval-2019 task 4: Hyperpartisan news detection using elmo sentence representation convolutional network, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 840–844.
[21] N. Hassan, F. Arslan, C. Li, M. Tremayne, Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 1803–1812.
[22] C. Hansen, C. Hansen, S. Alstrup, J. Grue Simonsen, C. Lioma, Neural check-worthiness ranking with weak supervision: Finding sentences for fact-checking, in: Companion Proceedings of the 2019 World Wide Web Conference, 2019, pp. 994–1000.
[23] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: Proceedings of the 43rd European Conference on Information Retrieval, ECIR '21, Lucca, Italy, 2021, pp. 639–649.
[24] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam,

F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, S. Modha, M. Kutlu, Y. S. Kartal, Overview of the CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News, in: Proceedings of the 12th International Conference of the CLEF Association: Information Access Evaluation Meets Multiliguality, Multimodality, and Visualization, CLEF '2021, Bucharest, Romania (online), 2021.

[25] G. K. Shahi, J. M. Struß, T. Mandl, Overview of the CLEF-2021 CheckThat! Lab task 3 on fake news detection, in: Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum, CLEF '2021, Bucharest, Romania (online), 2021.

[26] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: A large-scale dataset for fact extraction and verification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 809–819.

[27] X. Zhou, R. Zafarani, Fake news: A survey of research, detection methods, and opportunities, arXiv preprint arXiv:1812.00315 (2018).

[28] M. Amjad, G. Sidorov, A. Zhila, A. Gelbukh, P. Rosso, Urdufake@ fire2020: Shared track on fake news detection in Urdu (2020), in: Proceedings of the 12th Forum for Information Retrieval Evaluation (FIRE 2020), Hyderabad, India, 2020.

[29] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3391–3401. URL: https://www.aclweb.org/anthology/C18-1287.

[30] M. Amjad, S. Butt, H. I. Amjad, A. Zhila, G. Sidorov, A. Gelbukh, Urdufake@ fire2021: Shared track on fake news detection in urdu (2021), in: Proceedings of the 13th Forum for Information Retrieval Evaluation (FIRE 2021), Hyderabad, India, 2021.

[31] F. Adeeba, S. Hussain, Experiences in building urdu wordnet, in: Proceedings of the 9th workshop on Asian language resources, 2011, pp. 31–35.

[32] A. Hall, L. Terveen, A. Halfaker, Bot detection in wikidata using behavioral and other informal cues, Proc. ACM Hum.-Comput. Interact. 2 (2018).

[33] M. Zuckerman, B. M. DePaulo, R. Rosenthal, Verbal and nonverbal communication of deception, in: Advances in experimental social psychology, volume 14, Elsevier, 1981, pp. 1–59.

[34] R. S. Nickerson, Confirmation bias: A ubiquitous phenomenon in many guises, Review of general psychology 2 (1998) 175–220.

[35] M. Deutsch, H. B. Gerard, A study of normative and informational social influences upon individual judgment., The journal of abnormal and social psychology 51 (1955) 629.

[36] G. Loewenstein, The psychology of curiosity: A review and reinterpretation., Psychological bulletin 116 (1994) 75.

[37] B. E. Ashforth, F. Mael, Social identity theory and the organization, Academy of management review 14 (1989) 20–39.

[38] N. Ashraf, S. Butt, G. Sidorov, A. Gelbukh, CIC at checkthat! 2021: Fake news detection using machine learning and data augmentation, CLEF, 2021.

[39] C. Baziotis, N. Pelekis, C. Doulkeridis, Datastories at SemEval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis, in: Proceedings of

the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 747–754.

[40] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[41] A. Roy, K. Basak, A. Ekbal, P. Bhattacharyya, A deep ensemble framework for fake news detection and multi-class classification of short political statements, in: Proceedings of the 16th International Conference on Natural Language Processing, NLP Association of India, International Institute of Information Technology, Hyderabad, India, 2019, pp. 9–17.

[42] A. Giachanou, P. Rosso, F. Crestani, Leveraging emotional signals for credibility detection, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 877–880.

[43] B. Ghanem, P. Rosso, F. Rangel, An emotional analysis of false information in social media and news articles, ACM Transactions on Internet Technology (TOIT) 20 (2020) 1–18.

[44] P. Biyani, K. Tsioutsiouliklis, J. Blackmer, "8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[45] K. Wu, S. Yang, K. Q. Zhu, False rumors detection on sina weibo by propagation structures, in: 2015 IEEE 31st international conference on data engineering, IEEE, 2015, pp. 651–662.

[46] F. Rangel, A. Giachanou, B. Ghanem, P. Rosso, Overview of the 8th author profiling task at PAN 2020: Profiling fake news spreaders on twitter, volume 2696, CEUR-WS.org, 2020.

[47] M. Zarharan, S. Ahangar, F. S. Rezvaninejad, M. L. Bidhendi, M. T. Pilevar, B. Minaei, S. Eetemadi, Persian stance classification data set, in: M. Liakata, A. Vlachos (Eds.), Proceedings of the 2019 Truth and Trust Online Conference (TTO 2019), London, UK, October 4-5, 2019, 2019.

[48] J.-P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, J. J. M. Escobar, Detection of fake news in a new corpus for the Spanish language, Journal of Intelligent & Fuzzy Systems 36 (2019) 4869–4876.

[49] F. Rangel, P. Rosso, A. Charfi, W. Zaghouani, B. Ghanem, J. Sánchez-Junquera, On the author profiling and deception detection in Arabic shared task at FIRE, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, volume 2517, CEUR-WS.org, Kolkata, India, 2019, pp. 70–83.

[50] M. Alkhair, K. Meftouh, K. Smaïli, N. Othman, An Arabic Corpus of Fake news: Collection, Analysis and Classification, in: International Conference on Arabic Language Processing, Springer, 2019, pp. 292–302.

[51] I. Vogel, P. Jiang, Fake news detection with the new German dataset "GermanFakeNC", in: International Conference on Theory and Practice of Digital Libraries, Springer, 2019, pp. 288–295.

[52] F. Pierri, A. Artoni, S. Ceri, Investigating italian disinformation spreading on twitter in the context of 2019 european elections, PloS one 15 (2020) e0227821.

[53] R. A. Monteiro, R. L. Santos, T. A. Pardo, T. A. De Almeida, E. E. Ruiz, O. A. Vale, Contributions to the study of fake news in portuguese: New corpus and automatic detection results, in: International Conference on Computational Processing of the Portuguese Language, Springer, 2018, pp. 324–334.

[54] M. Amjad, G. Sidorov, A. Zhila, H. Gomez-Adorno, I. Voronkov, A. Gelbukh, Bend the Truth: A benchmark dataset for fake news detection in Urdu and its evaluation, Journal of Intelligent & Fuzzy Systems 39 (2020) 2457–2469. doi:10.3233/JIFS-179905.

[55] A. Kumar, S. Singh, G. Kaur, Fake news detection of indian and united states election data using machine learning algorithm (2019).

[56] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, A. Zubiaga, SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 69–76. URL: https://www.aclweb.org/anthology/S17-2006. doi:10.18653/v1/S17-2006.

[57] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, L. Derczynski, SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 845–854. URL: https://www.aclweb.org/anthology/S19-2147. doi:10.18653/v1/S19-2147.

[58] M. Amjad, G. Sidorov, A. Zhila, H. Gomez-Adorno, I. Voronkov, A. Gelbukh, Bend the Truth: A benchmark dataset for fake news detection in Urdu and its evaluation, Journal of Intelligent & Fuzzy Systems 39 (2020) 2457–2469. doi:10.3233/JIFS-179905.