ABSTRACT
This study reports the second shared task named as UrduFake@FIRE2021 on identifying fake news detection in Urdu language. This is a binary classification problem in which the task is to classify a given news article into two classes: (i) real news, or (ii) fake news. In this shared task, 34 teams from 7 different countries (China, Egypt, Israel, India, Mexico, Pakistan, and UAE) registered to participate in the shared task, 18 teams submitted their experimental results and 11 teams submitted their technical reports. The proposed systems were based on various count-based features and used different classifiers as well as neural network architectures. The stochastic gradient descent (SGD) algorithm outperformed other classifiers and achieved 0.679 F-score.

CCS CONCEPTS
• Artificial Intelligence; • Natural Language Processing;

KEYWORDS
Fake news detection, low resource languages, Urdu language

1 INTRODUCTION
Fake news is used as a tool to sway the opinion of people that has a direct impact on society, economy, politics, health and various other domains of life. The COVID-19 pandemic brought a sea of fake news generated through social media, for example, influencing people from getting vaccinated [9]. Similarly, in the political domain, we saw Indian media reporting fabricated news of the outbreak of civil war in Karachi, Pakistan [1]. The incidents became frequent and need timely action to prevent social unrest. As much as fake news affects the rest of the world, the South Asian people had their fair share of fake news exposure. Hence, there exists a vacuum for fake news study in low resource languages, for example, Urdu, which is spoken by more than 230 million people in South Asia and worldwide.

Automatic detection of fake news in English has many textual approaches to classify fake news in various forms. NLP experts used supervised machine learning algorithms (Random Forest (RF) [2], Support Vector Machines (SVM) [2], Decision Trees, etc.) [2, 6, 10] and deep learning algorithms (Long short-term memory (LSTM), Recurrent Neural networks (RNN), Gated Recurrent Units (GRU), etc.) [7, 8] with the combination of various embeddings, linguistic and user-level features. Though many of these approaches are successful in the English language, Urdu has a different morphological and syntactic structure and lacks resources for many deep learning approaches. Hence, it is important to continue developing resources and methods to reach real-life deployable solutions.

This paper gives a summary of the UrduFake track at FIRE 2021, which is in the continuation of its 2020 version. The dataset, methodologies and baselines changed with the expansion of the dataset between these two tasks. A detailed explanation of the UrduFake track at FIRE 2021 [1] and 2020 [3] is available for the research community. The goal is this study is to generate more resources for fake news classification in the Urdu language. We encouraged participation from all over the world and the submitted methodologies are compiled and explained in Section 6. We summarized the challenges, differences and future direction for combating fake news distribution on digital media platforms.

2 TASK DESCRIPTION
The binary classification task remained unchanged since the last edition [3, 4]. The teams were required to assign a label (real or
fake news) for given news articles written in Urdu Nastaliq script. The used definition of fake news is presented in the overview and dataset articles [3, 5].

Fake News Detection: For an unannotated news article, denoted as $\alpha$, where $\alpha \in N$ ($N$ represents the total articles), an automatic fake news detection algorithm assigns a score $S(\alpha) \in [0, 1]$. Given that $S(\tilde{\alpha}) > S(\alpha)$, we predict that $\tilde{\alpha}$ has higher probability to be a fake news article. A threshold $\gamma$ can be defined, such that the prediction function $F : N \rightarrow \{\text{not fake, fake}\}$ is:

$$F(N) = \begin{cases} \text{fake}, & \text{if } S(\alpha) \in \gamma, \\ \text{not fake}, & \text{otherwise}. \end{cases}$$

3 DATA COLLECTION AND ANNOTATION

The dataset used in this shared task contained five types of news articles: (i) Business, (ii) Health, (iii) Showbiz (entertainment), (iv) Sports, and (v) Technology. To assemble the dataset, numerous news articles were crawled from well-known traditional media news streams (national and international) using Python library Newspaper3. The Newspaper library automatically removes unimportant information, such as author’s name, date, advertisements, location of the publisher, HTML tags and images etc. The dataset is publicly available for research purposes4.

- **Real News Collection for Training and Testing Sets:**
  All the real news were retrieved from news agencies [5] and were manually annotated using a set of guidelines. For example, if the information mentioned in a news article can be verified from other news sources, then the news article can be annotated as a real news article. The detailed instruction of how the real news were annotated are reported in our earlier research [5].

- **Fake News Collection by Professional Crowdsourcing:**
  The fake news used in this shared task were written by professional journalists. The real news were provided to the professional journalists and they were asked to write corresponding fake news articles. Professional crowdsourcing was used because collecting corresponding fake news is a challenging task. The detailed instruction of how the fake news were collected are reported in our earlier research [5].

Table 1 describes the distribution of the news articles in the dataset.

|          | Real News | Fake News | Total |
|----------|-----------|-----------|-------|
| Train    | 750       | 550       | 1,300 |
| Test     | 200       | 100       | 300   |
| Total    | 950       | 650       | 1,600 |

1https://newspaper.readthedocs.io/en/latest/
2https://github.com/MaazAmjad/Urdu-Fake-news-location FIRE2021.git

4 EVALUATION METRICS

The shared task presented a challenge to classify a news article as fake or real. The participants were provided with the train and development dataset at the beginning of the competition. The test dataset was later revealed for the teams to evaluate their performances. All the participants were only allowed three attempts to submit their best performing models.

The test labels provided by the participants were compared with the ground truth labels. The standard evaluation metrics for fake news: Recall ($R$), Precision ($P$), Accuracy, and two F1-scores (F1-score for each class and F1-macro) were used. The F1-score has multiple variants like F1-macro, weighted F1, or F1-micro. To calculate the label of the “real” class $F_{\text{real}}$ was used and the label of the “fake” class out of all news was calculated through $F_{\text{fake}}$.

The macro-averaged F1-macro, which is the average of $F_{\text{real}}$ and $F_{\text{fake}}$ was used to tackle the imbalance of the dataset. F1-macro penalizes when a system does not perform well for the minority classes and does not use weights for the aggregation. It should be noted, that we only report F1-macro. This is because when the weighted F1 of both classes is summed up, it gives more weight to the majority class.

5 BASELINE SYSTEMS

A baseline was proposed to serve as reference points so that the proposed systems can be evaluated and ranked them accordingly. We used bag of words (BoW) model and n-gram (char, word) features and trained different machine learning classifiers. Overall, Decision Tree classifier provided the best results using bi-grams of the combination of char-word-function words with tf-idf.

6 OVERVIEW OF THE SUBMITTED APPROACHES

The submitted approaches spanned from machine learning methods to deep learning and transformer methods. 34 teams from 6 different countries (India, Pakistan, China, Egypt, Germany, and the UK) registered for participation, 18 teams submitted their experimental results and from which 11 teams submitted their technical reports. Registered participants were from 6 different countries (India, Pakistan, China, Egypt, Germany, and the UK). The results of the proposed approaches did not prove to be better than the baseline methods, except for two participant teams. Table 3 shows the approaches used by the teams and table 2 presents the best run scores achieved using those methods.

7 CONCLUSION

This shared task aims to attract researcher to address the fake news detection in Urdu language. A first news dataset in Urdu language has been proposed for fake news detection task which contains 1,600 news articles. All the real news articles were crawled from national and international reliable sources and the information in these articles was manually verified. On the other hand, the professional journalists were hired for the corresponding fake news generation. In this shared task, the Stochastic Gradient Descent algorithm outperformed all the classifiers and obtained the highest score in identifying fake news in Urdu. We aim to increase the size
Table 2: Participants' best run scores.

| Team Names     | Fake Class |          |          |          |          |          |          |          |
|----------------|------------|----------|----------|----------|----------|----------|----------|----------|
|                | Prec       | Recall   | F1       | Prec     | Recall   | F1       | Macro    | Accuracy |
| Nayel          | 0.754      | 0.400    | 0.522    | 0.757    | 0.935    | 0.836    | 0.679    | 0.756    |
| Abdullah-Khurem | 0.592      | 0.450    | 0.530    | 0.761    | 0.835    | 0.797    | 0.663    | 0.716    |
| Baseline       | 0.584      | 0.400    | 0.508    | 0.751    | 0.840    | 0.784    | 0.651    | 0.718    |
| Hammed-Khurem   | 0.634      | 0.350    | 0.470    | 0.772    | 0.905    | 0.853    | 0.631    | 0.753    |
| Muhammad Homayun| 0.480      | 0.400    | 0.485    | 0.742    | 0.873    | 0.778    | 0.611    | 0.653    |
| Suhair Bhawal   | 0.560      | 0.450    | 0.508    | 0.729    | 0.905    | 0.808    | 0.621    | 0.713    |
| Snehaan Bhawal  | 0.960      | 0.240    | 0.384    | 0.723    | 0.995    | 0.837    | 0.610    | 0.743    |
| MUCIC           | 0.821      | 0.250    | 0.356    | 0.716    | 0.975    | 0.826    | 0.592    | 0.726    |
| Baseline        | 0.793      | 0.230    | 0.356    | 0.716    | 0.975    | 0.826    | 0.592    | 0.726    |
| Dinamore & Elyasafdi _SVC | 0.720    | 0.180    | 0.288    | 0.701    | 0.965    | 0.812    | 0.550    | 0.703    |
| MUCS            | 0.850      | 0.170    | 0.356    | 0.716    | 0.985    | 0.820    | 0.592    | 0.726    |
| Iqra Ameer      | 0.454      | 0.100    | 0.163    | 0.676    | 0.940    | 0.786    | 0.475    | 0.660    |
| Sakshi Kalra    | 0.266      | 0.120    | 0.165    | 0.654    | 0.835    | 0.734    | 0.449    | 0.596    |

Table 3: Approaches used by the participating systems.

| System/Team Name | Feature Type       | Feature Weighting Scheme | Classifying algorithm | NN-based |
|------------------|--------------------|--------------------------|-----------------------|----------|
| Nayel            | tri-gram          | TF-IDF                   | Stochastic Gradient Descent | No       |
| Abdullah-Khurem  | N/A                | TF-IDF, Word2Vec, GloVe, fastText | textCNN | Yes     |
| Hammad-Khurem    | n-gram             | N/A                      | Assemble (XG-Boost, Light GBM, Adaboost) | No       |
| Muhammad Homayun | word n-grams (1-4), char-gram (2-6) | CNN ( 4-channels) | CNN | Yes     |
| Snehaan Bhawal   | embeddings         | MuRIL                    | Yes                   |          |
| MUCIC            | char-gram (1-3)    | TF-IDF                   | Assemble (LSVM, LR, MLP, XGB, RF) | No       |
| SOA NLP          | char-gram (1-3)    | TF-IDF                   | Dense neural networks | Yes      |
| Dinamore & Elyasafdi _SVC | char tri-grams | TF-IDF                   | SVC | No       |
| MUCS             | word uni-grams, char-grams (2-3) | TF-IDF, fastText | Assemble (MLP, ADB, GB, RF) | No       |
| Iqra Ameer       | embeddings         | BERT-base                | Yes                   |          |
| Sakshi Kalra     | embeddings         | RoBERTa-urdu-small       | Yes                   |          |

of the news dataset in Urdu and use deep learning learning techniques to identify fake news 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] Maaz Amjad, Sabur Butt, Hamza Imam Amjad, Alisa Zhila, Grigori Sidorov, and Alexander Gelbukh. 2021. Overview of the shared Task on Fake News Detection in Urdu at FIRE 2021. CEUR Workshop Proceedings.

[2] Maaz Amjad, Grigori Sidorov, and Alisa Zhila. 2020. Data augmentation using machine translation for fake news detection in the Urdu language. In Proceedings of The 12th Language Resources and Evaluation Conference. 2537–2542.

[3] Maaz Amjad, Grigori Sidorov, Alisa Zhila, Alexander Gelbukh, and Paolo Rosso. 2020. Overview of the shared Task on Fake News Detection in Urdu at FIRE 2020. CEUR Workshop Proceedings (2020). Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020), Hyderabad, India.

[4] Maaz Amjad, Grigori Sidorov, Alisa Zhila, Alexander Gelbukh, and Paolo Rosso. 2020. UrduFake@FIRE2020: Shared Track on Fake News Detection in Urdu. In Proceedings of the 12th Forum for Information Retrieval Evaluation (FIRE 2020), Hyderabad, India.

[5] Maaz Amjad, Grigori Sidorov, Alisa Zhila, Helena Gómez-Adorno, Ilia Voronkov, and Alexander Gelbukh. 2020. Bend the Truth: A benchmark dataset for fake news detection in Urdu and its evaluation. Journal of Intelligent & Fuzzy Systems 39, 2 (2020), 2457–2469. https://doi.org/10.3233/JIFS-179905

[6] Noman Ashraf, Sabur Butt, Grigori Sidorov, and Alexander Gelbukh. 2021. CIC at CheckThat! 2021: Fake News detection Using Machine Learning And Data Augmentation. CLEF.

[7] Arjun Roy, Kingshuk Basak, Anil Ekbal, and Puspak Bhattacharyya. 2018. A deep ensemble framework for fake news detection and classification. arXiv preprint arXiv:1811.04670 (2018).

[8] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. 1201–1211.

[9] Sander van Der Linden, Jon Roszenbeek, and Josh Compton. 2020. Inoculating against fake news about COVID-19. Frontiers in psychology 11 (2020), 2928.
This figure "sample-franklin.png" is available in "png" format from:

http://arxiv.org/ps/2207.05144v1