# Ensembled Feature Selection for Urdu Fake News Detection

Fazlourrahman Balouchzahi[1], Hosahalli Lakshmaiah Shashirekha[2] and Grigori Sidorov[1]

[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, CDMX, Mexico
[2]Department of Computer Science, Mangalore University, Mangalore, India

**Abstract**

Identifying fake news shared on social media is a vital task due to its immense effects in a negative way on the society, community, an individual or whoever is the target. Controlling and managing the fake news shared on social media manually is an impractical task due to the increasing number of social media users, increasing volume of fake news and the speed in which the fake news spreads on social media. Hence, there is a great demand for the automatic identification of fake news quickly and efficiently. Most of the fake news detection works carried out focus on resource rich languages like English and Spanish leaving the under-resourced languages like Urdu and many Indian languages less attended or unattended. UrduFake 2021 - a shared task in Forum for Information Retrieval Evaluation (FIRE) 2021 promotes detecting fake news in Urdu - an under-resourced language. This paper presents the description of the model proposed and submitted by our team MUCIC to UrduFake 2021 which aims to classify Urdu news article into one of the two categories, namely: Fake and Real. The major focus of this work is on feature engineering part to enhance the performance of traditional Machine Learning (ML) classifiers using very simple features such as word and char n-grams. Three Feature Selection (FS) algorithms, namely: Chi-square, Mutual Information Gain (MIG), and f_classif are ensembled to select the top informative features for the classification of Urdu news articles. The proposed methodology using an ensemble of five popular ML classifiers with soft voting obtained 8$^{th}$ rank in the shared task with an average macro F1-score of 0.592.

**Keywords**

UrduFake, Feature Engineering, Feature Selection, Machine Learning

## 1. Introduction

Fake news on social media is a common issue nowadays with the recent events in the world such as the outbreak of Covid-19, return of Taliban to Afghanistan, etc. These situations and the attributes of social media in being anonymous provide a lot of opportunities to content polluters to share false information on social media networks [1, 2]. The task of fake news detection is to identify the news articles that contain legit contents and distinguish them from false information. Moreover, fake news detection is a significant issue since fake news might be

*Forum for Information Retrieval Evaluation, December 13-17, 2021, India*
frs_b@yahoo.com (F. Balouchzahi); hlsrekha@gmail.com (H. L. Shashirekha); sidorov@cic.ipn.mx (G. Sidorov)
https://sites.google.com/view/fazlfrs/home (F. Balouchzahi);
https://mangaloreuniversity.ac.in/dr-h-l-shashirekha (H. L. Shashirekha); https://www.cic.ipn.mx/~sidorov/ (G. Sidorov)
0000-0003-1937-3475 (F. Balouchzahi)

a use-case to manipulate peoples' mind towards a country, community, religion or any other objective for political purposes or induction to a war, etc. [3, 4]. Detecting fake news manually is totally ruled out due to the increase in the number of social media users, increasing volume of fake news and the speed at which the fake news content is being generated and spread on social media [5]. Therefore, developing automatic tools and models to identify the real news from fake ones have become the primary requirements of the day to prevent the dissemination of false information quickly and efficiently [6].

Fake news detection is a remarkable issue in many languages other than English [7]. Several studies, shared tasks and workshops are being conducted to tackle the fake news detection and profiling in high-resource languages such as English, Spanish, and German. However, very few works have explored issues related to fake news detection in low-resource languages. One among them is UrduFake 2021[1] [8, 9] - a shared task at Forum for Information Retrieval Evaluation (FIRE) 2021 organized in continuation with the previous shared task - UrduFake 2020 [6] to promote fake news detection in Urdu language. The goal of this shared task is to classify Urdu news articles into one of the two classes: Fake and Real.

To detect fake news in Urdu language, we - team MUCIC, present a feature engineering-based ML model to investigate the effectiveness of ensembled FS algorithms on traditional ML classifiers using char and word n-grams features. The objective of the methodology is to keep the features and model simple and focus on the FS part.

The rest of paper is organized as follows: Section 2 gives a summary of the works submitted to UrduFake 2020 followed by the description of Methodology in Section 3. The Experiments and results are mentioned in Section 4 and the paper concludes in Section 5.

## 2. Related Work

Fake news detection in a resource-poor language like Urdu is a real challenge due to lack of resources. UrduFake 2020[2][6] is a shared task in FIRE 2020[3] that aims at identifying fake news from real ones which can be modeled as binary Text Classification (TC). The shared task participants were provided with a Train and Test set and the statistics of the datasets is given in Table 1. Further, the models submitted by the participants were evaluated and ranked based on macro F1-score. Several researchers had submitted their models to this shared task and the description of the models which exhibited good performance are given below:

Amjad et al. [6] developed "Bend the truth" - a benchmarked dataset for fake news detection in Urdu language and its evaluation [10]. The news items collected from Business, Health, Showbiz, Sports, and Technology domains consisted of 500 real and 400 fake news. They collected real news from different reliable sources such as BBC Urdu News, CNN Urdu, Roznama Dunya, Voice of America, Mashriq News, etc. and the fake news were manually generated by hired journalists from various news agencies such as Express news and Dawn news from Pakistan. The authors also experimented several baselines with various weighting schemes. LR classifier trained on Term Frequency–Inverse Document Frequency (TF-IDF) vectors generated from char bi-grams

[1] https://www.urdufake2021.cicling.org/
[2] https://www.urdufake2020.cicling.org/
[3] http://fire.irsi.res.in/fire/2020/home

**Table 1**
Statistics of the Train and Test sets in UrduFake 2020

| Dataset | Categories | | Total |
|---|---|---|---|
| | Fake | Real | |
| **Train set** | 400 | 500 | 900 |
| **Test set** | 150 | 250 | 400 |

exhibited the best performance obtaining 2nd rank with a macro F1-score of 0.889. Lin et al. [11] represented text as word and char level units and then obtained sentence embedding in terms of word and char level vectors using Robustly optimized Bidirectional Encoder Representations from Transformers approach (RoBERTa) and Char Convolutional Neural Network (CharCNN) respectively and fed them to softmax as a connecting layer to predict the authenticity of the text. Utilizing label smoothing to enhance the performance of the model based on a smoothing factor and the number of categories, this model secured 1st rank with a macro F1-score of 0.900.

The XLNet architecture benefits the generalized autoregressive method that utilizes autoencoding and autoregressive language modeling schemes [12]. Additionally, integration of the recurrence mechanism and relative encoding scheme are the other advantages of XLNet. Khilji et al. [13] fine-tuned XLNet [12] model with the maximum length of sequences to 512, and used an architecture of six layers with four attention heads and an embedding dimension of 256. With this settings they obtained a macro F1-score of 0.823 and 2nd rank in the shared task. Kumar et al. [14] proposed two models based on char n-grams in the range (1, 3) that were transformed to Term Frequency — Inverse Document Frequency (TF-IDF) vectors. While the first model is an ensemble of Random Forest (RF), Decision Tree (DT), and AdaBoost (AB) with majority voting, the other model is a multi-layer dense Neural Network (NN) architecture comprised of four dense layers containing 2,048, 512, 128, and 2-neurons with a dropout layer after each dense layer. Using Rectified Linear unit (ReLu) activation function for the first three dense layers and softmax activation function for the fourth layer, the multi-layer dense NN model secured 3rd rank with a macro F1-score of 0.791.

Out of the three models experimented by Reddy et al. [15] the first two models are based on bi-directional Gated Recurrent Unit (GRU) that employed Skipgram word embedding with similar architecture and difference in only pooling layer. While max pooling and average pooling were concatenated in one model, the other model employed only average pooling for the spatial connections. However, in both the models the output of the pooling layer is fed to the sigmoid layer for classification. The authors also experimented a multi-head transformer containing a pipeline of vector generation along with a global max pooling layer and a hidden layer with a dropout of 10% after each layer followed by softmax as final layer for classification. Among the three models, the model using max pooling and average pooling exhibited the best performance with a macro F1-score of 0.807. Balaji et al. [16] transformed word and char n-grams into TF-IDF vectors along with pre-trained word embeddings, namely: Word2Vec, fastText, and Bidirectional Encoder Representations from Transformers (BERT) trained on the traditional ML classifiers, namely: Multilayer Perceptron (MLP), AB, ExtraTrees (ET), RF, Support Vector Machine (SVM), and Gradient Boosting (GB). MLP classifier trained on fastText vectors obtained macro F1-score of 0.788 and outperformed the other classifiers.

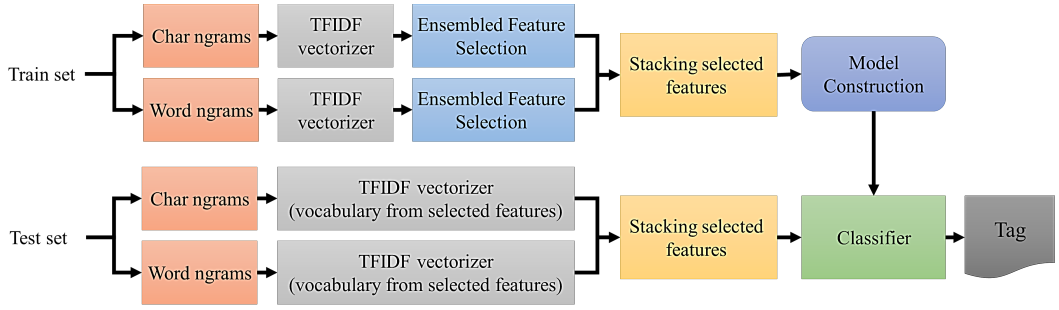


**Figure 1:** Overview of the proposed methodology

Balouchzahi et al. [17] experimented three learning approaches, namely: ML, Deep Learning (DL), and Transfer Learning (TL). The ML model was an ensemble of LR, SVM, and Multinomial Naïve Bayes (MNB) trained on the count vectors of the combination of word and char n-grams. A BiLSTM architecture employing Skipgram word embedding generated from the training set was used as DL model and for TL model Universal Language Model Fine-Tuning (ULMFiT) architecture borrowed from Howard et al. [18] was used. ULMFiT model which makes use of a pre-trained and publicly available Language Model[4] fine-tunes the training set for classification. The authors also submitted a majority voting of all the learning models to the shared task. However, the best performance was exhibited by the ensemble of traditional ML classifiers with a macro F1-score of 0.789.

## 3. Methodology

The proposed methodology contains two main steps: Feature Engineering (FE) and Model construction and the strength of the methodology lies in the FE part. As the data is already pre-processed only the punctuation if any are removed. Overview of the proposed methodology is shown in Figure 1.

### 3.1. Feature Engineering

As the main objective of this work is to investigate the effectiveness of FS algorithms on simple features, char and word n-grams are considered as features and are extracted separately. Selecting the range of n-grams in an intelligent way could improve the performance of the system. However, in this work, the range of $n$ for char and word n-grams are set to (1, 3) and (2, 5) respectively, based on the results obtained by conducting various experiments.

The aim of FS algorithms is to reduce the number of features by selecting the relevant features which may improve the performance of the model. An ensemble of two or more FS algorithms will result in better features as compared to a single FS algorithm. However, ensembling may be done in various ways. One way of ensembling is to consider the intersection of all the feature sets obtained by the FS algorithms in the ensemble, ie., a feature is selected by an ensembled FS

[4]https://github.com/anuragshas/nlp-for-urdu

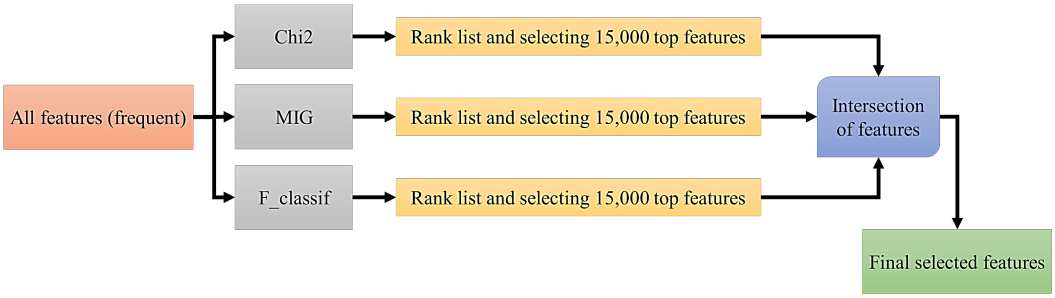


**Figure 2:** Steps involved in selecting the features using ensembled FS algorithm

algorithm if and only if it is selected by all the FS algorithms. Algorithm 1 gives the ensembled FS algorithm based on the intersection property of sets. The pictorial representation of the steps involved in selecting the features using ensembled FS algorithm is shown in Figure 2.

**Algorithm 1** Ensembled FS algorithm

```
procedure ENSEMBLED_FS(features, FS)          ▷ FS: feature selection algorithms
    Selected_features = { }
    for each FS algorithm do :
        Compute the feature importance scores for all the features
        Rank the features based on scores
        Select the features with top 15,000 ranks and store in Feat_i
                                                  ▷ i for i^th FS algorithm
    for each feature f do:
        if  f in Feat_i and f in Feat_i+1 and ... and f in Feat_n then
            Selected_features.add(f)              ▷ intersection of all feature sets
        else
            Continue;

    return Selected_features
```

The FS algorithm explored by Shashirekha et al. [19] is an ensemble of three FS algorithms, namely: Chi-square, Mutual Information Gain (MIG), and f_classif from sklearn library. This ensembling algorithm is used in the proposed methodology to select the relevant features. The char and word n-grams features are selected in two levels as follows:

- 30,000 top frequent features are selected from the given dataset and converted into TF-IDF vectors (the number 30,000 is fixed based on the results obtained by conducting various experiments)
- out of 30,000 top frequent features, the relevant features are selected by applying the ensemble of FS algorithms explicitly

The number of n-grams features of both types at various stages are given in Table 2. Features selected by the ensembled FS algorithm for the Train set is used to construct the models and for that of Development (Dev.)/Test set are used to evaluate the models.

**Table 2**
Number of features in each stage

| Feature | All features | Frequent features | Selected features |
|---|---|---|---|
| **Char n-grams** | 262,137 | 30,000 | 22,477 |
| **Word n-grams** | 355,030 | 30,000 | 22,616 |

**Table 3**
Classifiers and their parameter values

| Classifier | Parameters |
|---|---|
| **LR** | Default parameters |
| **Linear SVM** | Kernel="linear", probability= "True" |
| **RFC** | N_estimators= 1000 |
| **MLP** | hidden_layer_sizes= (150,100,50), max_iter=300, activation = "relu", solver= "adam", random_state=1 |
| **XGB** | max_depth= 20, n_estimators= 80, learning_rate= 0.1, colsample_bytree= 0.7, gamma=.01, reg_alpha=4, objective= "multi:softmax", num_class= 2 |

### 3.2. Model Construction

To evaluate the effectiveness of the proposed FE algorithm, five popularly used traditional ML classifiers, namely: Linear SVM (LSVM), LR, MLP, XGB, and RF are trained individually and also as ensembled model with soft and hard voting. The individual and ensemble classifiers are trained with and without FS to compare the effectiveness of FE module on various classifiers. The classifiers and their parameters values used in this work are presented in Table 3.

## 4. Experiments and Results

UrduFake 2021 shared task organizers provided a dataset consisting of Train and Dev. set to enable the participants to build and evaluate the models locally. However, for final evaluation, Test set was provided without labels and the participating teams were supposed to predict the labels of the Test set and submit the predictions to the organizers for evaluation and ranking which was done based on average macro F1-score. The statistics of the datasets given in Table 4 shows that the datasets are highly imbalanced and it may negatively effect the performance of the system.

As the number of runs that can be submitted by a participant was restricted to 4, four sets of predictions were obtained on the Test set by applying the ensemble of five ML classifiers with hard voting, soft voting, with FS and without FS (only frequent features) and were submitted to the shared task organizers for evaluation and ranking.

A comparison of the performances of all the classifiers with and without FS (only frequent features) for Dev. and Test set in terms of average macro F1-score is shown in Figures 3 and 4 respectively. The performances of many individual and ensembled classifiers with FS are good compared to the ones without FS which justifies the effectiveness of FS algorithms. The results also reveals that the LR classifiers with FS has given the best performance among all the

**Table 4**
Statistics of the datasets in UrduFake 2021

| Dataset | Fake | Real | Total |
|---|---|---|---|
| **Train set** | 438 | 600 | 1,038 |
| **Dev. set** | 112 | 150 | 262 |
| **Test set** | 100 | 200 | 300 |

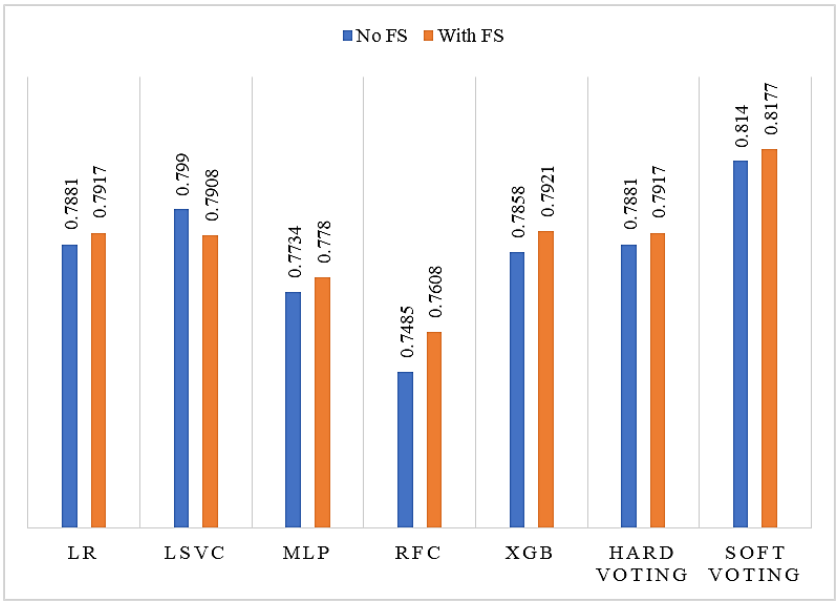


**Figure 3:** Performances on the Dev. set

experiments on the Test set with a macro F1-score of 0.617. However, the ensembled classifier with soft voting and FS have obtained 8$^{th}$ rank with a macro F1 score of 0.592.

The comparison of the performances of the proposed model (along with the best results) with the best performing teams of the shared task is given in Table 5. It can be observed that the LR classifier with FS could have obtain 5$^{th}$ rank in the shared task if it was submitted to the shared task (including baseline). Generally, the difference of less than 0.2 macro F1-score between 1$^{st}$ and 10$^{th}$ ranks in the shared task shows very competitive performances among the participating teams.

## 5. Conclusion and Future Work

This paper describes the methodology proposed by our team MUCIC for Urdu fake news detection in UrduFake 2021 shared task. The main objective of the methodology lies in FE part where three FS algorithms are ensembled to select the most informative char and word n-grams. Effectiveness of the ensembled FS algorithm studied by using individual ML classifiers and their ensemble as voting classifiers have shown promising results. Further, it is expected that other feature sets such as sub-words, syllables and morphological features may significantly improve

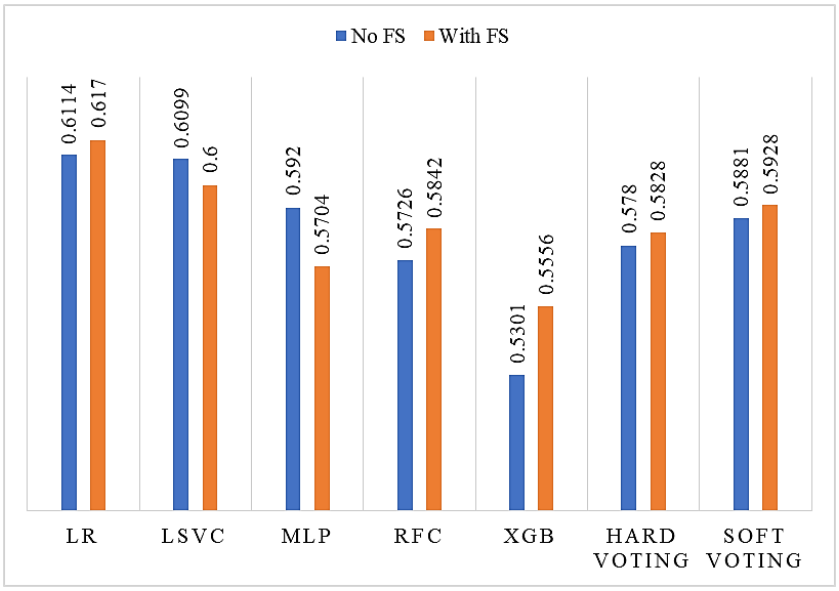


**Figure 4:** Performances on the Test set

**Table 5**
Comparison of the best performing teams in UrduFake 2021

| Team Name | Fake macro F1-score | Real macro F1-score | Average macro F1-score | Accuracy |
|---|---|---|---|---|
| Nayel | 0.522 | 0.836 | 0.679 | 0.756 |
| Abdullah-Khurem | 0.530 | 0.797 | 0.663 | 0.716 |
| Baseline | 0.508 | 0.794 | 0.651 | 0.710 |
| Hamed-Khurem | 0.434 | 0.808 | 0.621 | 0.713 |
| **LR with FS (out of competition)** | **0.400** | **0.834** | **0.617** | **0.7400** |
| Muhammad Homayoun | 0.485 | 0.738 | 0.611 | 0.653 |
| Snehaan bhawal | 0.384 | 0.837 | 0.610 | 0.743 |
| Junaid | 0.420 | 0.794 | 0.607 | 0.696 |
| **MUCIC (soft voting with FS)** | **0.359** | **0.826** | **0.592** | **0.726** |
| BERT4ever | 0.438 | 0.746 | 0.592 | 0.650 |

the performance of the models. The future plans include exploring robust FE-based models with various feature types and FS algorithms for TC in general in low resource languages such as Urdu, Persian and Dravidian languages.

## Acknowledgments

Team MUCIC sincerely appreciate the organizers for their efforts to conduct this shared task.

## References


[1] H. L. Shashirekha, F. Balouchzahi, ULMFiT for Twitter Fake News Spreader Profiling, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_126.pdf.

[2] F. Balouchzahi, H. L. Shashirekha, G. Sidorov, MUCIC at CheckThat! 2021: FaDo-Fake News Detection and Domain Identification using Transformers Ensembling, in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 455–464. URL: http://ceur-ws.org/Vol-2936/paper-35.pdf.

[3] M. Amjad, G. Sidorov, A. Zhila, A. Gelbukh, P. Rosso, UrduFake@ FIRE2020: Shared Track on Fake News Identification in Urdu, in: FIRE 2020-Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation, Association for Computing Machinery, 2020, pp. 37–40.

[4] 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.

[5] N. Ashraf, S. Butt, G. Sidorov, A. F. Gelbukh, CIC at CheckThat! 2021: Fake News Detection Using Machine Learning and Data Augmentation, in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 446–454. URL: http://ceur-ws.org/Vol-2936/paper-34.pdf.

[6] M. Amjad, G. Sidorov, A. Zhila, A. F. Gelbukh, P. Rosso, Overview of the Shared Task on Fake News Detection in Urdu at FIRE 2020, in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 434–446. URL: http://ceur-ws.org/Vol-2826/T3-1.pdf.

[7] M. Amjad, G. Sidorov, A. Zhila, Data Augmentation using Machine Translation for Fake News Detection in the Urdu Language, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 2537–2542.

[8] M. Amjad, B. Sabur, I. A. Hamza, Z. Alisa, S. Grigori, G. Alexander, Overview of the Shared Task on Fake News Detection in Urdu at Fire 2021, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, India, December, 2021, CEUR Workshop Proceedings, CEUR-WS.org, 2021.

[9] M. Amjad, B. Sabur, I. A. Hamza, Z. Alisa, S. Grigori, G. Alexander, UrduFake@ FIRE2021: Shared Track on Fake News Identification in Urdu., in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, India, December, 2021, In Forum for Information Retrieval Evaluation, CEUR-WS.org, 2021.

[10] M. Amjad, G. Sidorov, A. Zhila, H. Gómez-Adorno, I. Voronkov, A. Gelbukh, "Bend the Truth": Benchmark Dataset for Fake News Detection in Urdu Language and its Evaluation, Journal of Intelligent & Fuzzy Systems 39 (2020) 2457–2469.

[11] N. Lin, S. Fu, S. Jiang, Fake News Detection in the Urdu Language using CharCNN-RoBERTa, in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation,

Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 447–451. URL: http://ceur-ws.org/Vol-2826/T3-2.pdf.

[12] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, Q. V. Le, XLNet: Generalized Autoregressive Pretraining for Language Understanding, in: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 5754–5764. URL: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

[13] A. F. U. R. Khilji, S. R. Laskar, P. Pakray, S. Bandyopadhyay, Urdu Fake News Detection using Generalized Autoregressors, in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 452–457. URL: http://ceur-ws.org/Vol-2826/T3-3.pdf.

[14] A. Kumar, S. Saumya, J. P. Singh, NITP-AI-NLP@ UrduFake-FIRE2020: Multi-layer Dense Neural Network for Fake News Detection in Urdu News Articles., in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation,Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 458–463.

[15] S. M. Reddy, C. Suman, S. Saha, P. Bhattacharyya, A GRU-based Fake News Prediction System: Working Notes for UrduFake-FIRE 2020, in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 464–468. URL: http://ceur-ws.org/Vol-2826/T3-5.pdf.

[16] N. N. A. Balaji, B. Bharathi, SSNCSE_NLP@ Fake News Detection in the Urdu Language (UrduFake) 2020, in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 469–473.

[17] F. Balouchzahi, H. L. Shashirekha, Learning Models for Urdu Fake News Detection, in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 474–479. URL: http://ceur-ws.org/Vol-2826/T3-7.pdf.

[18] J. Howard, S. Ruder, Universal Language Model Fine-tuning for Text Classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 328–339.

[19] H. L. Shashirekha, M. D. Anusha, N. S. Prakash, Ensemble Model for Profiling Fake News Spreaders on Twitter, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_136.pdf.