Overview of the Shared Task on Fake News Detection in Urdu at FIRE 2020

Maaz Amjad[a], Grigori Sidorov[a], Alisa Zhila[b], Alexander Gelbukh[a] and Paolo Rosso[c]

[a]Center for Computing Research (CIC), Instituto Politécnico Nacional (IPN), Mexico
[b]Independent Researcher, United States
[c]Universitat Politècnica de València, Spain

Abstract

This overview paper describes the first shared task on fake news detection in Urdu language. The task was posed as a binary classification task, in which the goal is to differentiate between real and fake news. We provided a dataset divided into 900 annotated news articles for training and 400 news articles for testing. The dataset contained news in five domains: (i) Health, (ii) Sports, (iii) Showbiz, (iv) Technology, and (v) Business. 42 teams from 6 different countries (India, China, Egypt, Germany, Pakistan, and the UK) registered for the task. 9 teams submitted their experimental results. The participants used various machine learning methods ranging from feature-based traditional machine learning to neural networks techniques. The best performing system achieved an F-score value of 0.90, showing that the BERT-based approach outperforms other machine learning techniques.

Keywords

Natural Language Processing, Urdu language, fake news detection, low resource language,

1. Introduction

Fake news dissemination has been an important issue starting in the 15th[1] century. While meant to be objective, not all news articles follow the rigor of conveying fair facts chasing the fast-paced readers' attention by screaming headlines and sensational content. The spread of fake news brought many technical and social challenges. For example, it was a dispersion of fake news, which ignited the origin of antisemitism[2] in 1475, when a Franciscan preacher on the occasion of Easter Sunday claimed that Jewish community killed a toddler. Moreover, to celebrate the child's pass-over, some Jewish drained the child's blood and drank. The fake news spread fast and as a revenge, Trent's whole Jewish community was arrested, tortured, and fifteen Jewish were found guilty and burned at the stake. This story inspired surrounding local communities to commit similar atrocities. Thus, propagation of fake news brought terrifying results and inflamed social conflict.

Fire 20: Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad, India
maazamjad@phystech.edu (M. Amjad); sidorov@cic.ipn.mx (G. Sidorov); alisa.zhila@gmail.com (A. Zhila); gelbukh@gelbukh.com (A. Gelbukh); prosso@dsic.upv.es (P. Rosso)
https://nlp.cic.ipn.mx/maazamjad/ (M. Amjad)

[1]https://www.blackbird.ai/history-of-misinformation/
[2]https://libguides.ncl.ac.uk/fakenews/history

Automatic fake news identification is difficult, because we deal with very high level semantic phenomenon and at the first glance fake news look like real news. There are several types of news that are considered fake news. The researchers [1] classified fake news into six types: (i) fabrication, (ii) news satire, (iii) manipulation (e.g., editing pictures), (iv) advertising (e.g., ads are depicted as professional journalism), (v) propaganda, and (vi) news parody. Fabricated news can be defined as a deliberately fabricated news article manipulated to deliver a particular narrative, such as to create confusion, to be prominent in news headlines, or to make money.

The term "`fake news`" is not a simple concept. Publishers have been spreading false and misleading information even before the availability of the Internet. Different studies proposed various definitions of fake news [2, 1, 3]. For example, a recent study [3] defined fake news as a factually incorrect news article, which intentionally misleads a reader to believe that the conveyed information is true. There is a related term *clickbait*, which is defined as a snippet of a news article that is used to attract the reader's attention, and upon clicking, it redirects the reader to a different page. Clickbaits are used to generate revenues by online advertisements. Notably, the term *misinformation* – spreading untruths – emerged in the late 16th century[3]. Further, the more specific term *disinformation* came from Russian word[4] "dezinformacija". Disinformation is used when there is the intent to harm. The Guardian showed that this term was used during the cold war by all participants, being its meaning sowing falsehoods to confuse enemies.

Automatic detection of fake news is crucial to prevent the devastating and havoc impact that the fake news phenomenon causes worldwide. Throughout history, humans have always been prejudiced and intolerant to different views. For example, in 1620, Francis Bacon[5] emphasized the consequences of inaccurate language in his book *Novum Organum*, "The ill and unfit choice of words wonderfully obstructs the understanding." To this point, fake news detection is a means to eliminate the vast and disastrous effects produced by misinformation. Further, the ever increasing pace and scale of fake news propagation can be mitigated only by automating the solutions and increasing their effectiveness.

Fortunately, fake news detection attracted many researchers, in particular after the US 2016 presidential election. The importance of the task visibly increased for the Natural language processing (NLP) community as well as the demand for the solutions from the industry. Given the topicality of fake news detection and the urgency of coming up with an effective solution, this competition aims to gain the attention of a larger research community and to incentivize development of different solutions to combat the propagation of fake news on the Web.

2. Importance of Fake News Detection in Urdu

A large number of existing studies in the literature examined automatic fake news detection in multiple languages, such as English, Spanish, German, Chinese, and Arabic. However, a very limited work was done in Urdu language to automatically identify fake news content on the web. Urdu is a widely spoken language having more than 100 million native speakers worldwide[6].

[3]https://www.theguardian.com/books/2019/nov/22/factitious-taradiddle-dictionary-real-history-fake-news

[4]https://www.theguardian.com/books/2019/jun/19/dominic-raab-disinformation

[5]https://www.theguardian.com/technology/2016/nov/29/fake-news-echo-chamber-ethics-infosphere-internet-digital

[6]http://www.bbc.co.uk/voices/multilingual/urdu.shtml

However, according to the best of our knowledge, there are no automatic web sources to verify the authenticity of news articles in Urdu language. This situation requires the attention of researchers working in NLP to develop tools and solutions in verifying the authenticity of news articles written in Urdu language. Note that Urdu is a low resource language, i.e., it does not have many NLP tools and corpora data.

Urdu is spoken mainly in Pakistan. The fake news phenomenon had bad effects in Pakistan's social, politics and economical situation. For example, a Pakistani TV anchor Dr. Shahid Masood[7], was sent to jail and barred from hosting the TV show due to deceptive claims and spreading fake news on the rape case of a teenage girl during a television show. Similarly, according to the Washington Post[8], fake news about child trafficking led to many deaths of innocent people in India.

In addition to this, BBC reported that some Indian sites claimed a civil war[9] had broken out in one of the cities of Pakistan. The report mentioned that some Indian websites described the situation in Pakistan as dangerous and the civil war resulted in the deaths of many city police officers. Moreover, the websites claimed that tanks had been seen on the streets, which eventually proved to be fake news. Therefore, this urge to conduct studies to combat the dissemination of fake content in Urdu language.

3. Literature Review and Task Overview

3.1. Related Work

A number of approaches for automatic fake news detection were proposed. These approaches are based on statistical text analysis to tackle fake news detection. Previous studies used various datasets, which comprised mainstream media news articles and news published on social media. Notably, the majority of work focused on English [4, 5], with some efforts in other languages such as Spanish [4, 6], German [7], Arabic [8, 9], Persian [10], Indonasian [11], Bangla [12], Portuguese [13], Dutch [14], Italian [15], and Hindi [16].

Correspondingly, challenges[10] and shared tasks on automatic fake news detection were proposed, such as SemEval 2017 task 8 RumourEval for English [17], SemEval-2019 task 7 RumourEval for English [18], and PAN 2020[4], the latter focused on the author profiling of Fake News Spreaders on Twitter. Moreover, previous studies [19, 20] have shown that emotional information can be helpful to identify fake news on social media effectively. To the best of our knowledge, this is the first shared task on fake news detection for the Urdu language. This task incentivizes the development of fake news detection in Urdu as well as provides an opportunity to compare the system performances with the recent shared tasks in other languages.

Several studies [5, 6] reported that linguistic features can be helpful to identify fake news. For example, a recent study [5] demonstrated that linguistic features proved to be helpful to capture the differences of writing styles between fake and real news. Another study [6] exploited some

[7]https://www.globalvillagespace.com/dr-shahid-masoods-claims-about-zainabs-murderer-prove-false/

[8]https://www.washingtonpost.com/world/asia_pacific/as-mob-lynchings-fueled-by-whatsapp-sweep-india-authorities-struggle-to-combat-fake-news/2018/07/02/683a1578-7bba-11e8-ac4e-421ef7165923_story.html

[9]https://www.bbc.com/news/world-asia-54649302

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

linguistic features and showed that these features provided cues for differentiating between fake and real news.

3.2. Task Description

The task of fake news detection is to solve a binary classification problem, in which the input is a news article and the output is the assigned label (real or fake). Built around the idea reported in recent studies that the textual content can be helpful to identify fake news [5, 6], this shared task is aimed to explore the efficiency of models to detect fake news, in particular, for the news articles written in Urdu language.

The task was made publically available[11] on June 30, 2020, and the same day the training dataset was released as well. The training dataset was splitted into two subfolders, (i) real news subfolder, and (ii) fake news subfolder. The real news subfolder contained 500 real news articles, and the fake news subfolder contained 400 fake news articles. We released the test dataset on August 31, 2020. Like for the training dataset, the testing dataset was also splitted into two subfolders, (i) real news, and (ii) fake news. The real news subfolder contained 250 real news articles, and the fake news subfolder contained 150 fake news articles. The participating teams submitted their system until September 10, 2020. Each team could submit up to 3 different runs.

4. Dataset Collection and Annotation

This section provides an overview of the dataset developed for the shared task. A smaller version of the dataset, named "Bend The Truth", along with the detailed information about the collection and annotation description was presented in the recent study [3]. For this task, we performed additional data collection and annotation following the exactly same procedure. As a result, we obtained the final dataset of the annotated fake and real news in Urdu that was 1.5 times the size of the original "Bend-The-Truth" dataset. It is publicly available for academic research[12].

The fake news articles were intentionally written by hired professional journalists under specific instructions. The domains of the news present in our dataset are: (i) Business, (ii) Health, (iii) Showbiz (entertainment), (iv) Sports, and (v) Technology. They are similar to the dataset [5] used to identify fake news in English language. Nonetheless, only one domain of news, namely, related to education, is not available in our dataset because it was difficult to find enough verifiable news in Urdu related to the education domain.

For the training and development purposes, we offered 900 news articles from the previously published "Bend The Truth" dataset. The 400 news articles from the previously unseen and unpublished newly collected part were held out as a test dataset.

4.1. Dataset Annotation Procedure

For the dataset annotation, rigorous guidelines and annotation procedures were defined. The news articles were annotated into two types of news: (i) real news article, and (ii) fake news

[11] https://www.urdufake2020.cicling.org/home
[12] https://github.com/UrduFake/urdufake2020eval.git

article. The dataset can be used as a corpus for supervised machine learning. It is possible to use the knowledge of the dataset annotation procedure for applying to the underlying characteristics of fake news in addition to linguistics features, but we do not recommend it, because in real life there is no such information. We followed different strategies for real and fake news annotation.

4.1.1. Real News Annotation

For the annotation of real news articles, initially numerous news articles from different mainstream were collected. Table 1 shows news agencies used to collect news articles for annotation. To annotate a news article as a real news, the major points in the real news data collection and annotation were:

1. The data collection and annotation procedure were performed manually.
2. The news article was labeled as real news if the news meets the following criteria:
    - A reliable newspaper or a prominent news agency published that news article.
    - Other authentic and credible newspaper agencies published the same news article and the veracity of the news article can be easily verified using information such as place of the event, image, date, etc. We also performed manual source verification from where the news are originated. We further compared and cross checked different sources (mainstream news agencies) to verify the information present in the news article.
    - We also confirmed that a news article has a correlation between its title and its content. We read the complete news articles to find out the correlation between the title and the content.

If a news article does not follow one of these criteria, we simply discard that news article.

Note that the length of all the news articles is heterogeneous. The reason is that each news agency has a different style of news articles. For example, BBC Urdu contained on average more than 1,500 words in a news article. Thus, we selected real news articles carefully following the described procedure.

4.1.2. Professional Crowdsourcing of Fake News

To obtain fake news in this dataset, we used professional journalist services from various news agencies in Pakistan: Express news, Dawn news, etc., who were asked to write fake news stories that correspond to the original real news articles. This is a peculiar attribute of this dataset, because it ensured that the fake news articles realistically imitated real life approach to fake news creation. In real life, it is journalists who are responsible for writing fake news articles. Obviously, not all journalists do it. Still, people of this profession have a better understanding of how to write an article (real or fake) and make it interesting to hook and, in case of fake news, trick the reader.

The reasons to use professional “crowdsourcing” to collect fake news are the following:

1. Finding and verifying the falsehood of fake news in the same domain as the available real news articles is a challenging task that requires a huge amount of time and resources unavailable to a small group of organizers. Thus, manual analysis of hundreds of thousands of news articles for verification through web scraping approach was unfeasible.
2. Unlike the case of the English language, most of the news verification in Urdu language is done manually due to the absence of web services that offer news validation.

We should mention that the news articles style and language characteristics vary depending on the news domain. Our dataset contains news in five major domains: sports, business, education, technology, and showbiz. Thus, we assigned news articles according to the journalists expertise in the corresponding domain. Moreover, all the journalists were given instructions to minimize the possibilities of introducing defined patterns that can provide undesirable clues in the classification task. Also, some technical guidelines, such as the requirement that the lengths of fake news should be in the range of those of the original news, were provided. Finally, all fake news articles were prepared using journalists' expertise.

Table 1
Legitimate websites

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

4.2. Training and Testing Datasets

4.2.1. Training and Validation Dataset

The training set was made available to the participants to develop their approaches to identify fake news. It contained 900 news articles, annotated in a binary manner as real or fake. 500 news articles were annotated as real, and 400 articles were annotated as fake. The real news part of the dataset was retrieved from January 2018 to December 2018, which is different from the test set.

All five topic domains, i.e., (i) Business, (ii) Health, (iii) Showbiz (entertainment), (iv) Sports, and (v) Technology were present in the training dataset. That is, we did not hold out any domain from the training set to make it "unseen" for the participants.

The use of the training set for validation, development, and parameter tuning purposes was at the participants' discretion.

4.2.2. Test dataset

The test dataset was used to evaluate the performance of the submitted classifiers. It was provided to all the participants without the ground truth labels. The truth labels were only used by the organizers to evaluate and compare the performance of participants' approaches.

To create the testing dataset, news articles were retrieved from January 2019 to June 2020. It also has all five types of news as the training set. The test dataset is composed of 400 news articles. The ground truth distribution among these 400 news articles was 250 real news articles and 150 fake news articles. We emphasize again that this information, along with labels, was not made available to the participants.

4.3. Dataset Statistics

To prepare the data for the experiments, the corpus was split into train and test sets. In the first stage of the shared task, the training dataset was released which contained 900 news samples (500 real and 400 fake). In the second stage, we released the test dataset which contained 400 news articles (250 real and 150 fake). Table 2 describes the corpus distribution of the news articles by topics for the training and testing sets.

Table 2
Domain Distribution in Train and Test subsets

Domain	Train		Test	
	real	fake	real	fake
Business	100	50	50	30
Health	100	100	50	30
Showbiz	100	100	50	30
Sports	100	50	50	30
Technology	100	100	50	30
Totals	**500**	**400**	**250**	**150**

5. Evaluation Metrics

The task consists in classifying a news article as fake or real news. First, the training dataset was released for the participants to develop and train their systems, and subsequently, the test dataset was released. Each participant team had an option to submit only 3 different runs. The participants' submissions were evaluated by comparing the labels predicted by the participants' classifiers and the ground truth labels. To quantify the classification performance, we employed

the commonly used evaluation metrics: Precision (P), Recall (R), Accuracy, and two F1-scores, namely, $F1_{real}$ for the prediction of label "real" and $F1_{fake}$ for the prediction of label "fake" out of all news. Additionally, to accommodate the skew towards the real class, which dominates (it has more samples than the fake news class), we used the macro-averaged F1-macro which is the average of $F1_{real}$ and $F1_{fake}$.

As this is a binary classification problem and the dataset contains two equally important classes, we measured both of them, evaluating the quality of predicting whether a news article is real, i.e., treating the "real" label as a positive, or target, class, and the quality of predicting whether the news article is fake, i.e., considering the label "fake" as a positive class. It is important to mention that since the dataset is not balanced, this is why these metrics are used.

The final ranking is based on the F1-macro score. It can be observed that F1-macro and accuracy are correlated, as it is expected.

6. Baselines

We provided three baseline systems with the goal that their performance could serve as reference points for qualitative evaluation of the submissions' placement in the ranking. First, we provided the Random Baseline as the most basic and trivial baseline, which is expected to be ranked at the bottom with a more massive gap from the participating systems. Second, we provided the most traditional baseline: bag of words (BoW) model. It uses words as features and then apply a machine learning classifier. In this baseline, we used binary weighting scheme (i.e., a feature is present or not) with Logistic Regression classifier. For the third baseline, we provided the results of character bi-gram with tf-idf weighting scheme using Logistic Regression classifier, which achieved surprisingly good results. Overall, we tried five weighting schemes (tf-idf, logent, norm, binary, relative frequency) [11] along with various classifiers such as Logistic Regression, SVM, Adaboost, Decision Tree, Random Forest, and Naive Bayes, but we got the best results with Logistic Regression, which we are reporting.

7. Overview of the Submitted Approaches

This section gives a brief overview of the systems submitted to this competition. 42 teams registered for participation, 9 teams submitted their runs. Registered participants were from 6 different countries (India, Pakistan, China, Egypt, Germany, and the UK). This wide range of the regions where the interested participants were located confirms the importance of this task. The team members came from various types of organizations: universities, research centers, and industry.

As the initial step of the experimental setup, the majority of participating teams performed data cleaning and preprocessing such as stop words elimination. In particular, the system submitted by the team MUCS removed the stopwords, whereas the systems submitted by teams BERT 4EVER, Chanchal_Suman, CNLP-NITS, SSNCSE_NLP and NITP_AI_NLP decided to leave them.

Further, all the news articles were represented with different text representation techniques. The team MUCS used traditional bags of words representation. Similarly, the teams MUCS,

NITP_AI_NLP, and SSNCSE_NLP used the n-gram (words or characters) representation weighted with tf-idf. The teams BERT 4EVER, CNLP-NITS, and Chanchal_Suman represented news article texts using word embeddings. In particular, only one team, SSNCSE_NLP, represented texts using Word2Vec embeddings, while the team SSNCSE_NLP employed FastText embeddings. Furthermore, the team Chanchal_Suman used Urdu word embedding and only one team, BERT 4EVER, used the contextual representation using BERT [21], which is one of the most recent and advanced manners of text representation.

To implement their classifiers, some participating teams used the traditional, i.e., non-neural machine learning algorithms, while some teams submissions were based on various neural network architectures. The team MUCS used two classical machine learning classifiers such as Multinomial Naive Bayes and Logistic Regression with default parameters, and the same team also used LSTM in the experimental setup. Similarly, the team SSNCSE_NLP used machine learning classifiers such as Multi-Layer Perceptron (MLP), AdaBoost (AB), ExtraTrees (ET), Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (GB). Another team, NITP_AI_NLP, used ensemble models by combining Random Forest, Decision Tree and AdaBoost classifiers. The team NITP_AI_NLP also used a multi-layer dense neural network. In contrast, one team, Chanchal_Suman, used Gated Recurrent Unit (GRU). All the participating teams, except one team (NITP_AI_NLP), used Transformers. Description of the approaches is presented in Table 3.

Table 3
Approaches used by the participating systems.

System/Team Name	Feature Type	Feature Weighting Scheme	Classifying algorithm	NN-based
BERT 4EVER	context embedding BERT	BERT	CharCNN-Roberta	Yes
Character bi-gram (baseline)	char bi-grams	TF-IDF	Logistic Regression	No
BoW (baseline)	word uni-grams	Binary	Logistic Regression	No
CNLP-NITS	N/A	embedding	XLNet pre-trained model	Yes
NITP_AI_NLP	char 1-3 grams	TF-IDF	Dense Neural Network	Yes
Chanchal_Suman	N/A	embeddings	Bi-directional GRU model	Yes
MUCS	mix of char and word n-grams	embeddings	ULMFiT model	Yes
SSNCSE_NLP	char n-gram	TFIDF, fastText, word2vec	RF, Adaboost, MLP, SVM	Yes

8. Results and Discussion

Among all the submitted runs, the results of the best run (among up to three submitted runs) are presented in Table 4. The systems are ranked by the F1-macro score. Table 5 provides the aggregated statistics about the performance of all non-trivial systems, including the baselines, that is, all of the systems apart from the random baseline.

We observe that except one system, all the other participating teams' systems outperformed the random baseline in terms of F1-macro score. The BERT 4EVER system achieved the best F1-macro, Accuracy, as well as R_{fake} (recall), $F1_{fake}$, and P_{real} (precision) scores. However, the baseline approach with character bi-grams and Logistic Regression achieved the second position in the shared task with just 1.1% difference in F1-macro from BERT 4EVER, which is quite an unexpected result. The explanation of this fact is a question for further research.

Table 4 presents the best results of the submitted systems.

Table 4
Participants' best run scores.

Team names	Fake Class			Real Class			F1-Macro	Accuracy
	Precision	Recall	$F1_{Fake}$	Precision	Recall	$F1_{Real}$		
BERT 4EVER	0.890	0.860	0.874	0.918	0.936	0.926	0.900	0.908
Character bi-gram (baseline)	0.833	0.900	0.863	0.936	0.892	0.913	0.889	0.895
CNLP-NITS	0.836	0.713	0.769	0.842	0.916	0.877	0.823	0.840
NITP_AI_NLP	0.890	0.593	0.712	0.797	0.956	0.869	0.791	0.820
Chanchal_Suman	0.881	0.593	0.709	0.796	0.952	0.867	0.788	0.818
BoW (baseline)	0.722	0.746	0.734	0.845	0.828	0.836	0.785	0.798
SSNCSE_NLP	0.709	0.733	0.721	0.837	0.820	0.828	0.774	0.787
MUCS	0.783	0.627	0.696	0.800	0.896	0.845	0.770	0.795
CoDTeEM, NUST	0.771	0.607	0.679	0.791	0.892	0.838	0.758	0.785
Rana Abdul Rehman	0.422	0.433	0.427	0.654	0.644	0.649	0.538	0.565
Random (baseline)	0.373	0.420	0.395	0.623	0.576	0.599	0.497	0.517
Cyber Pilots	0.377	0.533	0.441	0.628	0.472	0.538	0.490	0.495

Over 50% of the systems obtained F1-macro and Accuracy of more than 0.8 (Table 4), which is a reasonably high result. In Table 3, we observe that most of these high performing systems achieve better Recall for fake news, and better Precision for the real news detection. This tells us that these systems tend to "mistrust" the news articles and tag more news as fake than there are in reality. This can be considered as an overly secure approach.

At this moment, it is hard to judge whether any of these approaches is ready to be applied "in the wild". While the results of $F1_{real}$ and $F1_{fake}$ over 0.9 shown by the winning BERT 4EVER system are impressively high, the modest size of the provided training and testing datasets cannot guarantee the same performance on an arbitrary text input. To ensure the scalability of the presented approaches, more multifaceted research at a larger scale is needed. We see that one of the paths is a community-driven effort towards the increase of available resources and datasets in the Urdu language.

Table 5 presents statistics of the submitted systems.

9. Conclusion

This paper describes the first competition on automatic fake news detection in Urdu, the UrduFake 2020 track at FIRE 2020. We provided the training and testing parts of the dataset that included news articles in five domains (business, health, sports, showbiz, and technology). The news articles in the dataset were manually annotated with labels "fake" and "real" with a slightly imbalanced ratio of approximately 60% real news and 40% fake news.

Forty two teams from six different countries registered for this task. Nine teams submitted their experimental results (runs). The approaches employed by the submitted systems varied from the traditional feature-crafting and application of traditional ML algorithms to word

Table 5
Statistics of the submitted systems and the baselines (excluding the random baseline).

Stat. metrics	P_{real}	R_{real}	$F1_{real}$	P_{fake}	R_{fake}	$F1_{fake}$	F1-macro	Acc.
mean	0.746	0.649	0.748	0.795	0.843	0.755	0.767	0.770
std	0.193	0.127	0.131	0.091	0.161	0.138	0.130	0.134
min	0.377	0.433	0.490	0.628	0.472	0.490	0.502	0.495
percentile 10%	0.418	0.523	0.534	0.652	0.627	0.534	0.560	0.558
percentile 25%	0.725	0.593	0.762	0.792	0.838	0.762	0.781	0.786
percentile 50%	0.810	0.617	0.781	0.798	0.906	0.781	0.799	0.806
percentile 75%	0.887	0.728	0.815	0.841	0.945	0.815	0.830	0.835
percentile 80%	0.890	0.745	0.828	0.850	0.949	0.839	0.846	0.850s
percentile 90%	0.891	0.800	0.850	0.888	0.952	0.902	0.882	0.892
max	0.890	0.860	0.874	0.918	0.936	0.926	0.900	0.908

representation through pre-trained embeddings to contextual representation and end-to-end neural network based methods. In particular, ensemble methods were used in the traditional ML case. LSTM, and Transformers (BERT) were used in neural network based solutions.

Among all the submissions, only the best submitted model, BERT 4EVER, outperformed the character bi-grams with Logistic Regression baseline achieving F1-macro score of 0.90. This confirms that contextual representation and large neural network techniques outperform classical features-based models, which has been shown in many recent studies in all branches of natural language processing.

This competition aimed to encourage researchers working in different NLP domains to attempt to tackle the proliferation of fake content on the web. It also provided an opportunity to fully explore the sufficiency of textual content modality and effectiveness of fusion methods. And last but not the least, this track provides a useful resource in the form of an annotated dataset for other researchers working in automatic fake news detection in Urdu.

Acknowledgments

This competition was organized with partial support of National Council for Science and Technology (CONACYT) A1-S-47854, SIP-IPN 20200797 and 20200859 and CICLing conference. The work of the last author was partially funded by MICINN under the research project MISMIS-FAKEnHATE on MISinformation and MIScommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

References

[1] E. C. Tandoc Jr, Z. W. Lim, R. Ling, Defining “fake news” a typology of scholarly definitions, Digital journalism 6 (2018) 137–153.

[2] V. L. Rubin, Y. Chen, N. K. Conroy, Deception detection for news: three types of fakes, Proceedings of the Association for Information Science and Technology 52 (2015) 1–4.
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] M. Zarharan, S. Ahangar, F. S. Rezvaninejad, M. L. Bidhendi, S. S. Jalali, S. Eetemadi, M. T. Pilehvar, B. Minaei-Bidgoli, Persian stance classification dataset (????).
[11] I. Y. R. Pratiwi, R. A. Asmara, F. Rahutomo, Study of hoax news detection using naïve bayes classifier in indonesian language, in: 2017 11th International Conference on Information & Communication Technology and System (ICTS), IEEE, 2017, pp. 73–78.
[12] M. Z. Hossain, M. A. Rahman, M. S. Islam, S. Kar, Banfakenews: A dataset for detecting fake news in bangla, arXiv preprint arXiv:2004.08789 (2020).
[13] 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.
[14] M. S. Looijenga, The Detection of Fake Messages using Machine Learning, B.S. thesis, University of Twente, 2018.
[15] F. Pierri, A. Artoni, S. Ceri, Investigating italian disinformation spreading on twitter in the context of 2019 european elections, PloS one 15 (2020) e0227821.
[16] A. Kumar, S. Singh, G. Kaur, Fake news detection of indian and united states election data using machine learning algorithm (2019).
[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).