Detecting Fake News in URDU using Classical Supervised Machine Learning Methods and Word/Char N-grams

Yaakov HaCohen-Kerner, Natan Manor, Netanel Bashan, and Elyasaf Dimant

Computer Science Department, Jerusalem College of Technology, Jerusalem 9116001, Israel

Abstract

In this paper, we describe our submissions for the UrduFake 2021 track. We tackled the task entitled "Fake News Detection in the Urdu Language". We developed different models using three classical supervised machine learning methods: Support Vector Classifier, Random Forest, and Logistic Regression. Our machine learning models were applied to various sets of character or word n-gram features. Our best submission was an SVC model using 7,500 char trigrams. This model was ranked in 11th place out of 34 teams that participated in the discussed track.

Keywords

Fake news, supervised machine learning, word/char n-grams

1. Introduction

"Fake News is a term used to represent fabricated news or propaganda comprising misinformation communicated through traditional media channels like print, and television as well as non-traditional media channels like social media" [1]. In previous years, fake news has been used to influence politics and promote advertising. During the last two years, the phenomenon of fake news dramatically appeared in the field of coronavirus news.

There are various dangers in fake news such as incorrect (and sometimes even harmful) advice, social disorders, fear, panic, and hatred of population groups. Fake news in social networks (e.g., Facebook and Twitter) is spreading quickly and easily via various social media platforms. A large number of fake news in social media poses a huge challenge to the research community.

Therefore, there is a need for high-quality systems that can detect fake news in social media. Such systems will help to improve the protection and security of the people.

One of the recent results of this challenge was the organization of several fake news detection tournaments in different languages such as Constraint@AAAI2021 in English [2], FakeDeS 2021 in Spanish [3]; Author Profiling Task at PAN 2020 in English and Spanish [4]. In 2020, the first shared task on fake news detection in Urdu was arranged [5-6]. The current shared task is the second shared task on fake news detection in Urdu [7-8]. In these tournaments, researchers presented various models that combined natural language processing (NLP) and machine learning (ML) to detect fake news.

The structure of the rest of the paper is as follows. Section 2 introduces general background about fake news detection, natural language processing (NLP) in Urdu, and text preprocessing. Section 3 describes the UrduFake 2021 task and datasets. In Section 4, we present the applied models and their experimental results. Section 5 summarizes and suggests ideas for future research.

Forum for Information Retrieval Evaluation, December 13-17, 2021, India
EMAIL: kerner@jct.ac.il; natanmanor@gmail.com, netanelb56@gmail.com , elyasafdi@gmail.com
ORCID: 0000-0002-4834-1272

2. Related Work

2.1 Fake news detection

Posadas-Durán et al. [9] built a new fake news corpus for the Spanish language. This corpus contains 971 news collected from January to July of 2018. It is divided into 491 real news and 480 fake news. The corpus covers news from 9 different topics: Science, Sport, Economy, Education, Entertainment, Politics, Health, Security, and Society. The resource is freely available at https://github.com/jpposadas/FakeNewsCorpusSpanish. In addition, the authors trained four well-known classification methods on various lexical features BOW, POS tags, n-grams (with n varying from 3 to 5), and n-grams combinations. The highest accuracy result 0.7694 has been obtained by Rando Forest applied on BOW and POS features.

Shu et al. [10] explored the problem of exploiting social context for fake news detection. They propose a tri-relationship embedding framework TriFN, which models publisher-news relations and user-news interactions simultaneously for fake news classification. They conduct experiments on two real-world datasets, which demonstrate that the proposed approach significantly outperforms other baseline methods, e.g., RST, Castillo, and LIWC for fake news detection.

In another study, Shu et al. [11] described their tool called FakeNewsTracker that can automatically collect data for news pieces and social context, which benefits further research of understanding and predicting fake news with effective visualization techniques.

A systematic literature review on approaches to identify fake news is presented in [12]. The authors present the main approaches currently available to identify fake news and how these approaches can be applied in different situations.

2.2 NLP in Urdu

Amjad et al. [13] investigated whether machine translation from English to Urdu can be applied as a text data augmentation method to expand the limited annotated resources for Urdu. Yet the empirical results show that at its current stage, the machine translation quality for this language pair does not enable efficient automated data augmentation, in particular, for fake news detection which is regarded as a relatively high-level task.

Detection of threatening language and target identification in Tweeter messages written in Urdu is described in Amjad et al. [14] In this paper, the authors introduced a dataset that contains 3,564 Tweeter messages manually annotated by human experts as either threatening or non-threatening. The threatening tweets are further classified by the target into one of two types: threatening to a person or threatening to a group. Extensive experiments using various machine learning (ML) methods including deep learning classifiers showed that the best threatening language detection was achieved using an MLP classifier with a combination of word n-grams and the best target identification was achieved using an SVM classifier using fastText pre-trained word embedding.

2.3 Text preprocessing

An important component for the success of the text classification (TC) process is the preprocessing component. In many cases, preprocessing can “clean” the data and improve its quality. There are various basic types of preprocessing methods e.g., conversion of uppercase letters into lowercase letters, HTML tag removal, punctuation mark removal, and stop-word removal.

HaCohen-Kerner et al. [15] investigated the impact of all possible combinations of six preprocessing methods (spelling correction, HTML tag removal, converting uppercase letters into lowercase letters, punctuation mark removal, reduction of repeated characters, and stopword removal) on TC in three benchmark mental disorder datasets. In one dataset, the best result showed a significant improvement

over the baseline result using all six preprocessing methods. In the other two datasets, several combinations of preprocessing methods showed minimal improvements over the baseline results.

In another study, HaCohen-Kerner et al. [16] explored the influence of various combinations of the same six basic preprocessing methods (mentioned in the previous paragraph) on TC in four general benchmark text corpora using a bag-of-words representation. The general conclusion was that it is always advisable to perform an extensive and systematic variety of preprocessing methods, combined with TC experiments because this contributes to improving TC accuracy.

3. Task and Dataset Description

The 2021 shared task on fake news detection in Urdu [7-8] addresses the problem of "Fake News Detection in the Urdu Language". This task is coarse-grained binary classification in which participating systems are required to classify tweets into two classes: Real and Fake.

The Urdu fake news dataset [17] is composed of news articles in six different domains: business, education, entertainment, politics, sports, and technology. The real news was collected from several mainstream Urdu news websites in Pakistan, India, the UK, and the USA. The fake news was intentionally written by a group of professional journalists, each proficient in corresponding topics. The fake news is in the same domains and of the approximately same length as the real news.

General statistics about the training dataset[2] that we used are provided in Table 1. This training dataset is divided into training sub-dataset and test sub-dataset where each sub-dataset contains real and fake news.

Table 1
General statistics about the training dataset

	Training sub-dataset	Test sub-dataset	Total
Real news	600	150	750
Fake news	438	112	550
Total	1038	262	1300

4. Applied Models and their Experimental Results

We used the training dataset, which is described in the previous section, according to its given split. Due to time limitations, we applied only one preprocessing method - converting uppercase letters into lowercase letters and only three classical supervised ML methods: Support Vector Classifier (SVC), Random Forest (RF), and Logistic Regression (LR) using classical features such as character n-gram features and word n-gram features.

SVC is a variant of the support vector machine (SVM) ML method [18] implemented in SciKit-Learn. SVC uses LibSVM [19], which is a fast implementation of the SVM method. SVM is a supervised ML method that classifies vectors in a feature space into one of two sets, given training data. It operates by constructing the optimal hyperplane dividing the two sets, either in the original feature space or in higher dimensional kernel space.

Random forest (RF) is an ensemble learning method for classification and regression [20]. Ensemble methods use multiple learning algorithms to obtain improved predictive performance compared to what can be obtained from any of the constituent learning algorithms. RF operates by constructing a multitude of decision trees at training time and outputting classification for the case at hand. RF combines Breiman's

[2] https://github.com/MaazAmjad/Urdu-Fake-news-detection-FIRE2021/blob/main/Training%20Dataset%40FIRE2021.zip

"bagging" (Bootstrap aggregating) idea in [21] and a random selection of features introduced by Ho [22] to construct a forest of decision trees.

Logistic Regression (LR) [23-24] is a linear model for classification. It is known also as maximum entropy regression (MaxEnt), logit regression, and the log-linear classifier. In this model, the probabilities describing the possible outcome of a single trial are modeled using a logistic function.

These ML methods were applied using the following tools and information sources: The Python 3.7.3 programming language and Scikit-learn – a Python library for ML methods.

In our experiments, we test dozens of TC models. As mentioned above, we applied three different supervised ML methods for various combinations of character and/or word n-gram features. Under the user called Elyasafdi, we submitted the three models described in Table 2.

The models in Table 2 are sorted according to their accuracy results. The best model was SVC applied on 7,500 char trigrams (colored in gray). This model was ranked in 11th place out of 34 teams. Our main results were F-Measure of 0.550 (while the F-Measure results of the teams that were ranked at the 9th and 10th place were 0.592 and 0.590, respectively) and Accuracy of 0.703 (while the Accuracy results of the teams that were ranked at the 9th and 10th place were much lower than our Accuracy result, 0.65 and 0.590, respectively). Table 2 provides detailed results for the three submitted models on the test dataset[3] (nine leftmost columns) and the training dataset (two rightmost columns).

Table 2

Detailed results for the three submitted models on the test and training sub-datasets

Model	Results on the Competition Test Dataset								Results on the Training Dataset	
	Fake class			Real class			Average F1 Macro	Accuracy	Average F1 Macro	Accuracy
	Precision	Recall	F1 Macro	Precision	Recall	F1 Macro				
SVC - 7500 char trigrams	0.720	0.180	0.288	0.701	0.965	0.812	0.550	0.703	0.832	0.793
SVC - 4000 char trigrams	0.633	0.190	0.292	0.700	0.945	0.804	0.548	0.693	0.806	0.759
SVC - 2533 char bigrams	0.667	0.100	0.174	0.684	0.975	0.804	0.489	0.683	0.834	0.793

As can be seen from Table 2, our results on the training dataset (F-Measure of 0.832 and Accuracy of 0.793) were significantly higher than our results on the competition test dataset (F-Measure of 0.550 and Accuracy of 0.703). Possible explanations for these significant differences might be: (1) The training dataset is more balanced (550 fake news and 750 real news) than the competition test dataset (100 fake news and 200 real news) and (2) the content of a relatively high number of news items in the competition test dataset is fundamentally different from the content of the news in the training dataset.

[3] https://github.com/MaazAmjad/Urdu-Fake-news-detection-FIRE2021/blob/main/Test%20Dataset%20%40%20FIRE%202021.zip

5. Conclusions and Future Work

In this paper, we described our submitted models for the UrduFake 2021 track, which addresses the detection of fake news in the Urdu language. We applied three classical ML methods (SVC, RF, and LR) on various sets of character and/or word n-gram features. The best-submitted model was an SVC model applied on 7,500 char trigrams. This model obtained an F-Measure result of 0.550 and an accuracy result of 0.703 and it was ranked in 11th place out of 34 teams.

Potential future ideas are application of: various deep learning models; acronym disambiguation [25-26]; skip character n-grams that can serve as generalized n-grams [27]; stylistic feature sets [28]; key phrases [29]; and summaries [30].

Acknowledgments

We are grateful to the anonymous reviewers and the organizers for their fruitful comments and suggestions.

6. References

[1] A. Thota, P. Tilak, S. Ahluwalia, N. Lohia, Fake news detection: a deep learning approach, SMU Data Science Review, 1(3) (2018), Article 10.

[2] P. Patwa, M. Bhardwaj, V. Guptha, G. Kumari, S. Sharma, S. Pykl, ..., T. Chakraborty, Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts, In International Workshop on Combating On line Hostile Posts in Regional Languages during Emergency Situation, 2021, pp. 42-53, Springer, Cham.

[3] H. Gómez-Adorno, J. P. Posadas-Durán, G. Bel-Enguix, C. Porto, Overview of fakedes task at iberlef 2020: Fake news detection in Spanish, Procesamiento del Lenguaje Natural, 67(0) (2021).

[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, In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org, 2020.

[5] 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 FIRE (Working Notes) , 2020 ,pp. 434-446.

[6] M. Amjad, G. Sidorov, A. Zhila, A. F. Gelbukh, P. Rosso, UrduFake@ FIRE2020: Shared Track on Fake News Identification in Urdu, In Forum for Information Retrieval Evaluation, 2020, pp. 37-40.

[7] M. Amjad, S. Butt, H. I. Amjad, A. Zhila, G. Sidorov, A. Gelbukh, UrduFake@ FIRE2021: Shared Track on Fake News Identification in Urdu, In Forum for Information Retrieval Evaluation, 2021.

[8] M. Amjad, S. Butt, H. I. Amjad, A. Zhila, G. Sidorov, A. Gelbukh. Overview of the shared task on fake news detection in Urdu at Fire 2021, In CEUR Workshop Proceedings, 2021.

[9] 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(5) (2019) 4869-4876.

[10] K. Shu, S. Wang, H. Liu, Beyond News Contents: The Role of Social Context for Fake News Detection, In The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), 2019.

[11] K. Shu, D. Mahudeswaran, H. Liu, FakeNewsTracker: a tool for fake news collection, detection, and visualization, Computational and Mathematical Organization Theory, 25(1) (2019) 60-71.

[12] D. De Beer, M. Matthee, Approaches to identify fake news: a systematic literature review, In International Conference on Integrated Science, pp. 13-22, Springer, Cham, 2020

[13] 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.

[14] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, A. Gelbukh, Threatening Language Detecting and Threatening Target Identification in Urdu Tweets, accepted for publication in IEEE Access, 2021.
[15] Y. HaCohen-Kerner, Y. Yigal, D. Miller, The impact of Preprocessing on Classification of Mental Disorders, in Proc. of the 19th Industrial Conference on Data Mining, (ICDM 2019), New York, 2019.
[16] Y. HaCohen-Kerner, D. Miller, Y. Yigal, The influence of preprocessing on text classification using a bag-of-words representation, PloS one, vol. 15, p. e0232525, 2020.
[17] 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(2) (2020) 2457-2469.
[18] C. Cortes, V. Vapnik, Support-vector networks, Machine learning, 20 (1995) 273–297.
[19] C.-C., Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM transactions on intelligent systems and technology (TIST), 2 (2011) 1–27.
[20] L. Breiman, Random forest, Machine Learning, 45(1) 2001 5-32.
[21] L. Breiman, Bagging predictors, Machine Learning, 24(2) (1996) 123-140.
[22] T. K. Ho, Random decision forests, In Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, Vol. 1, pp. 278-282, IEEE.
[23] D. R. Cox, The regression analysis of binary sequences, Journal of the Royal Statistical Society: Series B (Methodological), 20 (1958) 215–232.
[24] D. W. Hosmer Jr, S. Lemeshow, R. X. Sturdivant, Applied logistic regression, Vol. 398, John Wiley & Sons. Applied logistic regression (Vol. 398). John Wiley & Sons, 2013.
[25] Y. HaCohen-Kerner, A. Kass, A. Peretz, Combined one sense disambiguation of abbreviations. In Proceedings of ACL-08: HLT, Short Papers, Association for Computational Linguistics, Columbus, Ohio, 2008, pp. 61-64, URL: https://aclanthology.org/P08-2.
[26] Y. HaCohen-Kerner, A. Kass, A. Peretz, Haads: A hebrew aramaic abbreviation disambiguation system, Journal of the American Society for Information Science and Technology, 61(9) (2010) 1923–1932.
[27] Y. HaCohen-Kerner, Z. Ido, R. Ya'akobov, Stance classification of tweets using skip char Ngrams, In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2017, pp. 266-278, Springer, Cham.
[28] Y. HaCohen-Kerner, H. Beck, E. Yehudai, M. Rosenstein, D. Mughaz, Cuisine: Classification using stylistic feature sets and/or name-based feature sets, Journal of the American Society for Information Science and Technology, 61(8) (2010) 1644-1657.
[29] Y. HaCohen-Kerner, I. Stern, D. Korkus, E. Fredj, Automatic machine learning of keyphrase extraction from short html documents written in hebrew, Cybernetics and Systems: An International Journal, 38(1) (2007) 1–21.
[30] Y. HaCohen-Kerner, E. Malin, I. Chasson, Summarization of jewish law articles in hebrew, Proceedings of the 16th International Conference on Computer Applications in Industry and Engineering (CAINE), November 11-13, 2003, Imperial Palace Hotel, Las Vegas, Nevada, USA, 2003, pp. 172–177.