SSNCSE_NLP@Fake news detection in the Urdu language (UrduFake) 2020

Nitin Nikamanth Appiah Balaji, B. Bharathi

Department of CSE, Sri Siva Subramaniya Nadar College of Engineering,Tamil Nadu, India

Abstract

The broadcasting of fake news always hammers out the truth with considerable growth. Fake news and false rumors are spreading further and faster, reaching more people, and penetrating deeper into social networks. Social media interaction is one of the major sources of spreading the news across the world nowadays. The fake news also spread among the people very faster using digital media. The objective of this proposed work to detect unreliable information from the news content in the Urdu language using digital media text collected from different sources. We have experimented with this task using the features namely TFIDF, fastText. We have achieved an accuracy of 90% for development data and 78.7% for test data respectively.

Keywords

TFIDF, fastText, Gradient boosting algorithm, Random forest classifier

1. Introduction

Fake news detection has recently fascinated a growing interest from the public and research community as the spread of unreliable information online increases, predominantly in media outlets such as social media feeds, news blogs, and online newspapers. In recent research, fake news detection is one of the dominant task using natural language processing. Urdu belongs to the Indo-Aryan language group, and it is the most commonly spoken language in the world with more than 100 million speakers. Urdu is an under-resourced language, only a small amount of data is publicly available. To automate the fake news detection process a corpus has been developed by [1]. The Urdu fake news dataset, named Bend-The-Truth [1], is composed of news articles in six different domains which are mentioned in Table 1. The fake news detection is considered to be a classification task. In this paper, we propose to develop a binary classification task to detect the given news belongs to fake news or real news using linguistic features present in the given text with different machine learning classifiers.

The organization of the paper is as follows: Section 2, lists the literature related to the fake news detection task. The data set used in the proposed work is tabulated in section 3. The proposed methodology is briefly explained in section 4. Results and discussions are presented in section 5. The section 6 concludes the paper.

FIRE 2020: Forum for Information Retrieval Evaluation, December 16-20, 2020, Hyderabad, India

nitinnikamanth17099@cse.ssn.edu.in (N. N. A. Balaji); bharathib@ssn.edu.in (B. Bharathi)

0000-0002-6105-0998 (N. N. A. Balaji); 0000-0001-7279-5357 (B. Bharathi)

2. Related work

The process of detecting fake news can be grouped in different ways. Based on the natural language approach used to extract the features from the document. Based on the machine learning or deep learning model used to classify the news. Based on the fact presented in the news, how the news is written, how the news spread in social media, and so on. Automated fake news detection is the task of assessing the truthfulness of claims in news. Automated fake news detection is one of the challenging tasks in natural language processing. One of the datasets available for detecting fake news is FAKENEWSNET [2], it is the project for collecting fake news research. In [3], a data augmentation method for fake news detection in Urdu language using machine translation is presented. In [4], a combination of linguistic and semantic features are used to discriminate real and fake news. Recurrent and convolution neural network is used to detect the fake news by the authors of [5]. Fake news detection is carried out using emotional content using LSTM by authors [6].

3. Data set

The Urdu fake news dataset, named Bend-The-Truth, in which the news articles are collected from different sources such as sports, social media, education sector, technology domain, business, and entertainment. The real news was collected by following a very rigorous procedure using a variety of mainstream news websites predominantly in Pakistan, India, UK, and the USA. These news channels are BBC Urdu News, CNN Urdu, Express-News, Jung News, Naway Waqat, and many other reliable news websites for the time frame from January 2018 to December 2018 [1]. The official description of the task is given in [7, 8]. The distribution of data from different categories is given in Table 1.

Table 1
Dataset distribution

Category	Real	Fake
Business	100	50
Health	100	100
Showbiz	100	100
Sports	100	50
Technology	100	100

4. Proposed work

For this task various feature extraction technique is studied for effective feature extraction. The investigated feature extraction strategies are explained in this section. These extracted features are then classified using machine learning models such as Multi-Layer Perceptron (MLP), AdaBoost (AB), ExtraTrees (ET), Random Forest(RF), Support Vector Machine (SVM),

gradient boosting (GB) algorithms. The scikit-learn implementation of machine learning models is used. The performance of the models is compared using the F1-scores.

4.1. TF-IDF

The Term Frequency Inverse Document Frequency gives a better normalized representation of the sentences by removing the impact of overly repeated banal words. A TFIDF vectorizer is trained from scratch using the given training data-set for this particular task. The char TFIDF is used to get character-level relationships as there are generally differences in the usage of words in different forms and tenses. So the word and the char TFIDF are considered for comparison. An n-gram range of 1-4 is used for the study of the two systems. These extracted feature vectors are fitted and trailed using different machine learning models such as Random Forest, Extra Trees, Gradient Boosting, Ada Boosting.

4.2. Text Embedding

As the training data-set for the Bend-The-Truth data-set is only 900 samples, which is very low, fine-tuning pre-trained models are considered as a better alternative. So pre-trained sentence to vector conversion techniques such as Word2Vec [9], FastText [10] and BERT [11] trained on CommonCrawl and Wikipedia data is used.

The Word2Vec and FastText are CBOW or Skip-gram based models that are trained in an unsupervised manner with large amounts of data. The Urdu specific pre-trained models are used for the task. These models generate a fixed-length representation of vector from variable length sentences. The fixed-length representations are then used to train machine learning models such as Multi-Layer Perceptron, Random Forest, and Support Vector Machine. The various hyper-parameters for the models are shown in Table 2.

BERT is a transformer-based neural network architecture that can be trained in an unsupervised manner and fine-tuned for particular tasks. BERT model is considered as it has shown excellent performance in the case of Twitter Data-set classification and other sentence classification tasks. In our experiment the model is trained end-to-end for 50 epochs keeping the *BERT-base multilingual cased* pre-trained weights as the initial starting weights.

5. Results and discussions

The performance of the Urdu Fake News detection task is analyzed in this section. Out of the TFIDF models, the char TFIDF method showed better results than the word TFIDF method as expected, even though the improvement was very small. Out of all the machine learning models the Random Forest model gave the best output for the TFIDF model. For the embedding techniques, FastText along with the Multi-Layer Perceptron model proved to be the better performing model than the word2vec and the BERT model. The results of the performance of models on dev-set and the test-set are tabulated in Table 2 and Table 3 respectively. Even though the performance of the TFIDF was better for dev-set, the FastText model produced a better performance as expected as it is pre-trained on a larger collection of unseen data.

Table 2
Performance of the fake news detection using dev-set

Feature	Classifier	Parameter	F1-score
Char TFIDF	Random forest	N range 1-4	0.9024
Char TFIDF	Extra trees	N range 1-4	0.8875
TFIDF	Gradient boosting	N range 1-4	0.8930
TFIDF	Adaboost	N range 1-4	0.87
fastText	MLP	1024,512,128 layers	0.8228
word2vec	Random forest	N_estimators =10000	0.7425
word2vec	SVM	urdu_web_news_vector300	0.73
Bert	Cola processor	50 epochs	0.7214

Table 3
Performance of the fake news detection using test-set

Feature	Classifier	Parameter	F1-score	Accuracy
TF-IDF	Random forest	N range 1-4	0.7872	0.8050
TF-IDF	Gradient boosting	N range 1-4	0.6997	0.7400
fastText	MLP	N range 1-4	**0.7881**	0.7875

6. Conclusion

In the current scenario, fake news is the biggest problem in our society. The fake news is spreading in society through social media very faster which will cause different problems. Natural language processing plays a major role in automatically classifying the given news into real or fake news. In the proposed system, fake news detection in the Urdu language is studied using the "Bend the truth" benchmark dataset. Our system showed an accuracy of 90% for development data and 78.7% for test data respectively.

References

[1] M. Amjad, G. Sidorov, A. Zhila, H. Gomez 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 (2020) 1–13. doi:10.3233/JIFS-179905.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, CoRR abs/1708.01967 (2017). URL: http://arxiv.org/abs/1708.01967. arXiv:1708.01967.

[3] 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, European Language Resources Association, Marseille, France, 2020, pp. 2537–2542. URL: https://www.aclweb.org/anthology/2020.lrec-1.309.

[4] M. Hardalov, I. Koychev, P. Nakov, In search of credible news (2019). arXiv:1911.08125.

[5] Y. Liu, Y. Wu, Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks, in: AAAI, 2018.

[6] P. R. Bilal Ghanem, F. Rangel, An emotional analysis of false information in socialmedia and news articles, in: ACM Trans. Internet Technology, 2020.
[7] M. Amjad, G. Sidorov, A. Zhila, A. Gelbukh, P. Rosso, Urdufake@fire2020: Shared track on fake news detection in urdu (2020). Proceedings of the 12th Forum for Information Retrieval Evaluation (FIRE 2020), Hyderabad, India.
[8] M. Amjad, G. Sidorov, A. Zhila, A. Gelbukh, P. Rosso, 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.
[9] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013. arXiv:1301.3781.
[10] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.
[11] 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).