Exploring the Impact of Machine Translation on Fake News Detection: A Case Study on Persian Tweets about COVID-19

Masood Hamed Saghayan
Department of Computer Science
Allameh Tabataba'i University
Tehran, Iran
saghayan_m@atu.ac.ir

Seyedeh Fatemeh Ebrahimi
Languages and Linguistics Center
Sharif University of Technology
Tehran, Iran
sfati.ebrahimi@student.sharif.edu

Mohammad Bahrani
Department of Computer Science
Allameh Tabataba'i University
Tehran, Iran
bahrani@atu.ac.ir

***Abstract*— Fake news detection has become an emerging and critical topic of research in recent years. One of the major complications of fake news detection lies in the fact that news in social networks is multilingual, and therefore developing methods for each and every language in the world is impossible, especially for low resource languages like Persian. In an effort to solve this problem, researchers use machine translation to uniform the data and develop a method for the uniformed data. In this paper, we aim to explore the impacts of machine translation on fake news detection. For this purpose, we extracted and labeled a dataset of Persian Tweets from Twitter on the subject of COVID-19 and developed a method for detecting fake news on the extracted Tweets based on the SVM classifier, then we machine translated the data and applied our proposed method to it. Finally, the result for binary class (only fake and legitimate) fake news detection was 87%, and for multiclass (satire, misinformation, neutral and legitimate) fake news detection was 62%, and our findings demonstrate that machine translation has a 4% negative impact on binary classification accuracy and a 23% negative impact on multiclass classification.**

Keywords*— Fake news detection, the impact of machine translation, classification, COVID-19, Persian language.

I. INTRODUCTION

The Internet has increasingly infused itself into daily life regardless of one's geographical location. This modern channel has played a significant role in transferring news and information throughout the web [1]. This point indicates that users have been expending enormous data and information every day on social networks such as Twitter, known as news sources. Therefore, misinformation or satire may disseminate through social networking and mislead users into believing and sharing with other users. Consequently, this influences people's views and attitudes, harms individuals and society, and reduces public trust [2].

Attention in developing effective techniques for fake news detection has been increased drastically due to the limited availability of language processing tools in a low resource language such as Persian [3]. Developing a unified, effective fake news detection system by proposing different models is critical since employing different approaches for such a low resource language and a lack of Persian fake news datasets is too expensive and time-consuming.

These hypotheses motivate us to investigate the machine translation impact on fake news detection. Employing a single method does not provide a solution for different languages. As we know, each language has certain complex features that are dependent on that specific language. Based on recent evidence, linguistic features play a significant role in discrimination types of fake news [4]. This means that most of those linguistic features that make each language unique and distinct may be removed or transformed by translation. This is the first study that undertakes to explore the machine translation impact of Persian Tweets to the best of our knowledge.

What exactly does fake news mean? There is no agreed definition of the term fake news in literature, but we discuss some of the recently reviewed literature's widespread definitions. Individuals may define different definitions for the "fake" term. According to, "news articles that are intentionally written to mislead or misinform readers for an invidious purpose but can be verified as false utilizing other sources [2]." In terms of existing content, fake news can either contain false information, distort the facts to some extent, or intentionally contain a portion of the truth. Fake content can also be published intentionally or unknowingly. Fake news can be categorized based on authentication and the intended features. As a result, the concepts commonly known as fake news are deceptive in [5], satire news in [6], disinformation, false news, rumor, clickbait, legitimate, etc.

To address this challenge, we investigate the fake news detection for Persian Tweets on the COVID-19 issue. This work's primary focus is on investigating the machine translation effect on fake news classifier efficiency. Based on this process, we employed a basic feature such as n-gram. The highest efficient performance in fake news detection compared to global weighting schemes such as TF-IDF decreases the classification performance according to [7, 8]. We also used different types of classifiers to achieve the best accuracy; the Support Vector Machine (SVM) classification model that has been one of the most effective models such as proposed in [9], achieved the best efficiency in fake news

2021 29th Iranian Conference on Electrical Engineering (ICEE) | 978-1-6654-3365-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICEE52715.2021.9544409

detection in our experiment and previous studies as in on the Spanish language [7].

The rest of the paper is categorized as follows: section 2 represents the review literature of similar research on fake news detection by using diverse approaches. Section 3 describes the methods which include dataset, proposes different models and considers the impact of machine translation on detecting fake news. Sections 4, 5, 6 include discussion, result, and conclusion. In short, future works will be presented.

II. Related Works

Based on the recent research papers, there is a fairly large body of machine learning methods for automatically detecting diverse fake news on social media. So far, fake news detection has been developed to a larger extent for the English language mentioned below, according to [10, 11].

Reference [12] investigates fake news detection by employing machine learning approaches. Data was collected from Twitter, and the utilized methods, i.e., Naive Bayes, Neural Network, and Support Vector Machine, perform very well in detecting the fake news. Two classes, believable and unbelievable, were used in the dataset. The Neural Network and Support Vector Machine have equivalently measured precision, recall, and F-measure, and they reported 99.9% accuracy.

Reference [13] also explored automated fake news detection. They used various models such as *TF-IDF*, CountVectorizer, Word2Vec to convert text to numeric representation, and various algorithms such as Naive Bayes, LSTM, and Support Vector Machine. Their dataset is publicly available [14]. The results of the classification are evaluated by precision, recall, f-measure, and accuracy. A combination of CountVectorize*r* and LSTM model is the best among all the models. However, these approaches perform very well in detecting fake news as a binary classification using different types of machine learning methods depending on the amount of data available.

Reference [15] proposed detecting fake news by exploring the application of Natural Language Processing methods to machine identification of misleading news sources. Dataset was obtained from Signal Media. They applied the TF-IDF of bi-grams and probabilistic context-free grammar (PCFG). The TF-IDF of bi-grams fed into the SGD model outperforms quite well at detecting fake news.

Reference [10] investigated data augmentation using machine translation for fake news detection in the Urdu language. They used machine translation for annotated text data augmentation to fake news detection in the Urdu language. The result shows that the classifier trained on the original Urdu dataset shows better results than the purely MT-translated and the augmented (the combination of the two) datasets.

We reviewed the research provided to detect fake news in Persian. To the best of our knowledge, only one valid research has been done, which classifies rumors in Persian Tweets. Reference [16] achieved accuracy of about 80% by using two sets of structural and content-based features. This automatic detection system classifies the Tweets into two categories, rumor and non-rumor. Nevertheless, there are two major drawbacks to this study. First and foremost, the dataset is derived from two separate sources, a reliable news source and a source of rumors posted on two unofficial sites. Accordingly, this is a classification between popular literature and official literature rather than between rumor and non-rumor. Second, the basis for selecting the Tweets, and their period is not specified as the Tweet's time period is very influential. If the rumors and non-rumors are from two different time periods, the separation of the two will be greatly simplified due to the diverse topics of the day, and in fact, the classification has been done on the topic rather than on fake and legitimate news. Most of these recent studies used modern and traditional machine learning approaches, which require an available dataset.

III. Methods

In this section, we review the process of detecting fake news and machine translation impacts to detect fake news. In the first part, we introduce a new dataset, preprocessing, and feature extraction. We then examine models and evaluation, and in the next section, we report the classification results and analysis on both the original dataset and machine-translated versions.

This research's potential contribution primarily focuses on analyzing the impact of machine translation on fake news detection. The fundamental insight into this idea arises from fake news detection for low-resource languages such as the Persian language and is expended mainly with less expense and time. Early studies in this field mostly provided an English language dataset publicly available to investigate the new fake news detection approaches [17]. Basically, there is no existing Persian dataset publicly accessible to propose a new approach for detecting fake news since it needs lots of effort, time, and expenses. To explore the result of detecting fake news, we used a machine translation tool to transform Persian Tweets into the English language. In summary, transforming Persian Tweets to the machine-translated version (English language) allows us to unify a low resource language to the prevalent language that has been provided lots of Natural Language Processing tools and available data as in [10].

A. Dataset

A.1 **The Original Dataset**

The dataset used in this research was crawled from Persian Tweets on Twitter. This collection includes 2M Tweets related to COVID-19 disease that have had the highest

popularity (the most popular means the most liked Tweets), published in June and July 2020 during the peak period of COVID-19. Out of 2M Persian Tweets, 500 Tweets were annotated manually based on each individuals' veracity of Tweets. Each annotator labeled 250 Tweets individually, and each Tweet was voted for to ensure veracity.

The dataset was annotated with four class labels: legitimate: 96, misinformation: 90, neutral: 132, and satire: 182. The definition of each label is as follows; however, the definition and meaning of each type of news is generally different in many sources; we clarify them as shown below:

- Misinformation; False information that is published in the form of news or non-news, with malicious intent [18].
- Neutral; This type of fake news contains false information, and it does not matter that it was published knowingly and unknowingly in [6].
- Satire; This category includes inaccurate and distorted content published for entertainment purposes and may contain inappropriate content that people are reluctant to engage in, such as racist content [11].
- Legitimate; corresponds to actual events and presents those events with facts that have been checked and re-checked against other reports and found to be true, i.e., real.

The 2M crawled dataset includes many features: the name of the user, date of the Tweets, number of the retweets, number of following and followers, and likes, profiles, etc. However, we just used the text and their label in our experiment.

In fact, our dataset has an advantage that outperforms the previous datasets. Most of the data set is collected based on the user and its authenticity; therefore, it makes the detecting system classifies the news based on the difference between the formal Tweets and informal ones that are not fake or legitimate. Accordingly, to improve this problem's solution, we collected our dataset based on its popularity and manually annotated each class to provide a qualified dataset. The dataset used in this research has the following advantages:

- The most popular Tweets are used in this dataset because, in addition to the greater impact of these Tweets (COVID-19) on the community, our dataset's selection has caused our dataset not to be biased and have higher veracity.
- Secondly, selecting the most popular Tweets has caused the Tweets to be randomly selected in terms of time and source of publication so that these two features have not affected the results. In regard to the Tweets concerning COVID-19, our findings have not affected the topic's significance.

A.2 **The Machine-Translated Dataset**

This dataset is a machine-translated dataset created from the original dataset using a Google Translation API. It contains 500 translated Tweets that have the same labels as the original ones. Basically, this dataset is just the translated version of the original dataset with the same types of labels and number of Tweets.

B. Data Preprocessing

Data preprocessing is an important phase before applying machine learning approaches. At this stage of the process, our dataset should be ready for the training phase. First of all, we review the preprocessing that has been done on the Tweets. For preprocessing Tweets, we removed characters that were not textual, including;

- punctuations, signs, and delimiters,
- English letters, images, and videos
- Short Tweets of less than three words are filtered.
- Emojis or emoticons; in this case study, emojis may impact our findings regarding classification and provide a biased result; therefore, we filtered all emojis or emoticons to obtain a reasonable comparison of both datasets admissible results.

Considering that the Persian form of the word "Corona" is also used in Arabic and Urdu, some of the extracted Tweets are Arabic and Urdu Tweets. For filtering these Tweets, the stop words of Arabic and Urdu have also been used. We have done the same process for the machine-translated dataset as well to obtain a preprocessed text. In the end, we provided two datasets that contained preprocessed text separately, the original and machine-translated versions.

C. FeatureExtraction

In Natural Language Processing, the feature-based approach, which involves the feature extraction, feature engineering, and analysis of linguistic cues to identify specific target phenomena (e.g., fake product reviews from legitimate ones), has been a compelling model with relatively interpretable results [19]. In this work, to investigate the machine translation effect on detecting fake news, the counts of n-grams were extracted from two separate datasets as the lexical features. After preprocessing, which has been done on datasets, each token is vectorized, and the number of repetitions of each word in the text constitutes the values of the feature vector. This is done for two-word and three-word expressions, resulting in unigrams, bigrams, and trigrams. One of the most promising feature extraction techniques for fake news detection for investigating cross-linguistic studies is n-grams that as in [10, 9] proposed improvements in fake news detection. Based on these shreds of evidence proposed in [9], we have not used *TF_IDF* since, as proposed in [8], *TF_IDF* failed to improve classification performance. As

indicated in Table 1, we extracted n-grams featured on both the original Persian Tweets dataset and machine-translated version, as shown below in Table 1.

Table 1. number of extracted n-grams

Types of the feature extraction	Unigram	Bigram	Trigram	Total extracted features
The original dataset	4551	8460	8256	21267
The machine-translated dataset	2148	5752	6807	14707

IV. Models & Evaluation

This work's primary goal is to investigate machine translation's impact on fake news detection, mainly through a special issue (COVID-19). Consequently, we selected some of these classifiers with high performance on fake news detection and many NLP tasks in earlier studies. As a part of our research, Naive Bayes (NB), Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) algorithms were fitted to our data. These algorithms are the most common and basic classification algorithms used in many tasks and research. Therefore, these well-known algorithms don't need to be explained.

The extracted feature vectors feed into various models to perform supervised classification. We investigated multiclass and binary classification; however, our dataset was annotated with four types of news (labels). Therefore, we transform both datasets for binary classification, so that satire, neutral, and misinformation are categorized as fake labels and legitimate as real. This downstream task was accomplished in two phases; first, on the original dataset and then on the machine-translated dataset. Each phase involves classification algorithms to detect which level of validity each news item is associated with. Furthermore, we use F1-measure and k-fold cross-validation with K=10 on each model to estimate the prediction error separately. Data has been splitted into ten parts, with one part used as test data and the other part as training data. The test set is selected in proportion to the training data.

V. Results & Discussion

The classification results on the original and machine-translated datasets (Tweets) are shown in Tables 2 and 3. These tables demonstrate the impacts of machine translation in detecting fake news. As can be seen, the best accuracy achieved by applying SVM, Multilayer Perceptron, and Logistic Regression algorithms. For the four classes fake news detection, the best accuracy is 62%, and for the binary classification is 87. In machine-translated dataset, the maximum accuracy we achieved for the four classes is 39%, and the highest accuracy for the binary classes is 83%. Comparing the classifiers' performance on the described datasets demonstrated that the classification accuracy in multiclass fake news detection had been significantly reduced for the machine-translated dataset. The accuracy score of detecting fake news in the multiclass classification decreased by about 23%, and in binary classification, by about 4%. One reason for decreasing the accuracy score on machine-translated datasets might be due to each language's differences, in the way each language has its unique linguistics features, such as the tone of speech. The other reason might be the poor quality of translations and unnatural sentences produced by machine translation tools. Also, when we translate a language to another, some of its features may be removed or changed. Our dataset includes Tweets on social networks, which contains mostly the informal styles. Generally, it may change to a formal style language after translation. Our findings indicate that machine translation quality does not improve fake news detection in the Persian language, but it has not a significant impact on detecting fake news, especially when we want to detect only the legitimacy of the news.

Table 2. Multiclass classification result on both datasets

Models	Multiclass (legitimate, satire, misinformation, neutral)		
	Original Dataset (%)	*Machine Translated Dataset (%)*	*Impact of Translation on the Accuracy (%)*
SVM	62	39	-23
DT	48	31	-17
NB	56	31	-25
MLP	62	33	-29
LR	62	39	-23

Table 3. Binary-classification result on both datasets

Models	Binary class (fake, legitimate)		
	Original Dataset (%)	*Machine Translated Dataset (%)*	*Impact of Translation on the Accuracy (%)*
SVM	87	83	-4
DT	84	69	-15
NB	86	69	-17
MLP	85	79	-6
LR	83	83	0

VI. Conclusion

This research study investigated machine translation's impact on fake news detection in the Persian language. A content-based method based on lexical features has been proposed and applied to a collection of the most popular Persian Tweets revolving around the subject of Covid-19. Results indicate that our classification accuracy for the multiclass fake news detection was 62%, and for the binary class fake news was 87%. Exploring the machine translation impact showed us that in multiclass fake news detection, machine translation has a significant negative impact of 23% on the classification accuracy, but on the binary class fake news detection, it only has a negligible 4% impact. This difference in the magnitude of impacts comes from the fact that many of the deeper features and meanings of sentences are lost in the translation which in turn causes a decrease in the efficiency of the procedure of classifying fake news into more than one category. It is also worth noting that to the best of our knowledge only one other study has been conducted on fake news detection in the Persian language until now, whose results have a lower accuracy than ours which we consider an innovation of ours alongside the explore of the impact of machine translation. Further research also needs to be done about the impact of machine translation as there is currently much room for improvement. One aspect of the field that perhaps is worth exploring, can be the impact of machine translation on the approaches which are built upon deep learning.

VII. Bibliography

[1] K. Shu, A. Silva, S. Wang, J. Tang, and H. Liu, "fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter,* vol. 19, no. 1, pp. 22-36, 2017.

[2] A. Bondielli, F. Marcelloni, "A survey on fake news and rumour detection techniques," *Information Sciences,* pp. 38-55, 2019.

[3] A. Abdella, A. Al-Sadi, and M. Abdullah, "A Closer Look at Fake News Detection: A Deep Learning Perspective," in *ICAAI 2019: Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence*, 2019.

[4] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any atretch of the imagination," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, 2011.

[5] H. Allcott, M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives,* vol. 31, no. 2, pp. 211-236, 2017.

[6] A. Kucharski, "Study epidemiology of fake news," *Nature 540,* vol. 7634, p. 525, 2016.

[7] J. P. F. P. Durán, H. G. Adorno, G. Sidorov, Grigori Sidorov, J. Moreno, "Detection of fake news in a new corpus for the Spanish language," *Journal of Intelligent and Fuzzy Systems,* vol. 36, no. 5, pp. 4869-4876, 2019.

[8] M. Amjad, G. Sidorov, A. Zhilla, H. G. Adorno, I. Voronkov, A. Gelbukh , ""Bend the truth": Benchmark dataset for fake news detection in Urdu language and its evaluation," *Journal of Intelligent and Fuzzy Systems,* vol. 39, no. 1, pp. 1-13, 2020.

[9] S. K. W. Chu, R Xie, and Y. Wang, "Cross-Language fake news detection," in *ASIS&T Asia-Pacific Regional Conference (Virtual Conference)*, Wuhan, China, 2020.

[10] M. Amjad, G. Sidorov, and A. Zhilla, "Data Augmentation using Machine Translation for Fake News Detection in the Urdu Language," in *LREC*, 2020.

[11] V. L. Rubin, Y. Chen, and N. J. Conroy, "Deception detection for news: three types of fakes," in *ASIST '15: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, 2015.

[12] S. Aphiwongsophon, P. Chongstitvatana, "Detecting Fake News with Machine Learning Method," in *15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Chiang Rai, Thailand, Thailand, 2018.

[13] S. Vijayaraghavan, Y. Wang, Z. Guo, J. Voong, W. Xu, A. Nasseri, J. Cai, L. Li, K. Vuong, and E. Wadhwa, "Fake News Detection with Different Models," in *Arxiv Prepreint*, 2020.

[14] Sairamvinay, "Github," [Online]. Available: https://github.com/Sairamvinay/Fake-News-Dataset.

[15] M. Granik, and V. Mesyura, "Fake news detection using naive Bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, Kiev, Ukraine, 2017.

[16] S. Zamani, M. Asadpour, and D. Moazzemi, "Rumor detection for Persian Tweets," in *Iranian Conference on Electrical Engineering (ICEE)*, Tehran, Iran, 2017.

[17] W. Y. Wang, ""Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada, 2017.

[18] N. Kshetri,and J. Voas, "The economics of "fake news"," *IT Professional,* vol. 6, pp. 8-12, 2017.

[19] F. T. Asr, and M. Tabaoda, "Big Data and quality data for fake news and misinformation detection," *Big Data & Society,* vol. 6, no. 1, 2019.