IJICTR
International Journal of Information & Communication Technology Research

Volume 15- Number 1 – 2023 (63 -71)

The Effect of Data Augmentation Techniques on Persian Stance Detection

Mojgan Farhoodi
Department of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran
mojgan.farhoodi@srbiau.ac.ir

Abbas Toloie Eshlaghy*
Department of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran
toloie@gmail.com

Mohamadreza Motadel
Central Tehran Branch, Islamic Azad University, Tehran, Iran
dr.motadel@gmail.com

Received: 5 May 2022 – Revised: 25 August 2022 - Accepted: 27 November 2022

Abstract—**The purpose of stance detection is to identify the author's stance toward a particular topic or claim. Stance detection has become a key component in applications such as fake news detection, claim validation, argument searching, and author profiling. Although significant progress has been made in stance detection in languages such as English, little attention has been paid in some other languages, including Persian. One of the main problems of research in Persian stance detection is the shortage of appropriate datasets. In this article, to address this problem, we consider data augmentation, the artificial creation of training data, which is used to conquer the shortage of datasets. In this research, we studied several methods of data augmentation such as EDA, back-translation, and merging source dataset with similar one in English language. The experimental results indicate that combining the primary data set with the translation of another dataset with similar content in another language (for example English) result in a significant improvement in the performance of the model.**

Keywords: stance detection; data augmentation; fake news; dataset.

Article type: Research Article

Publisher: ICT Research Institute

I. Introduction

Recently, the rapid spread of news in social media has led people to depend on these platforms as the main sources of information. Therefore, validating the content and information exchanged in these media has become a vital issue.

Sources such as Twitter, Facebook, online news sites, other social media and personal blogs of journalists have unconsciously become essential players in providing news content [1]. Therefore, governments, journalists and social media platforms are working hard to distinguish real news from fake news. The first helpful step in fake news detection is to find out what other news sources, posts, or comments have a stance towards this news. Therefore, stance detection is the first and most crucial step in detecting fake news [2], which is still in the early stages of research and in recent years it has attracted a lot of attention from many researchers [3, 4]. The stance detection process requires a large amount of labeled data. On the other hand, most of the articles have worked on stance detection in English [5, 6, 7], and many of the prepared datasets are also in the same language [8, 9, 10].

* Corresponding Author

Nevertheless, for stance detection in low data resource languages, it is necessary to use techniques that either does not depend on the data or can increase the amount of data without generating new labeled data, which is called data augmentation techniques. In this study, we investigate the impact of data augmentation methods on the accuracy of Persian stance detection in social media. The structure of this paper is as follows: Section II describes the concept of stance detection and data augmentation, and while examining their methods, points out the related activities carried out in these two areas. Section III explains the proposed method. Section IV, while describing the used datasets in this study, will discusses and analyzes the results of applying data augmentation methods on Persian stance detection. Finally, Section V expresses conclusions and future works.

II. Related Works

A. Stance Detection

Stance detection (which is also known as stance classification, stance analysis, or stance prediction) is usually considered as a subset of the sentiment analysis problem. Its aim is to automatically determine the position of an author towards a statement, a target or a subject that is explicitly stated in the text, or only implied [11]. A statement can be a claim, news, an idea, or a part of a text, and the target or subject can be a person, an organization, a government policy, a product, or an event [8].

In the existing literature, stance detection can be categorized into different types:

1) Target-specific stance detection: Most researches are based on this category [12, 13, 14], and its aim is the detection of the stance expressed in a text towards a specific target [15].

2) Multi-target stance detection: More recently, since people often comment on multiple target entities in the same text, multi-target stance detection was designed. Multi-target stance detection aims to detect social media users' opinions toward two or more targets [10, 16]. In [10] was state because in many applications, there are many natural dependencies among targets, target-specific models are not effective. Therefore, it focuses on the problem of multi-target stance detection.

3) claim-based stance detection: It is considered a suitable method to analyze the integrity of the news. For that reason, claim-based stance detection has been heavily used for rumor resolution studies [3, 17, 6].

In recent years, several research focus on posts and tweets on social networks. Still, regardless of the type of the content, the stance detection approaches can be divided from three main perspectives [18]: 1) Feature-based machine learning approaches that often use machine learning algorithms such as logistic regression, Support Vector Machine (SVM), decision tree, and so on for learning [19, 20]. 2) Deep learning approaches that usually use deep neural networks such as Recurrent Neural Network (RNN), or Long Short Term Memory (LSTM) [21, 22]. Some of the common features used in these approaches are vector representation of words, i.e., Word2Vec [23] and GloVe [24], phrase embedding, n-grams of words or letters. 3) Ensemble learning approaches that use more than one classifier to get the final result of the stance detection [14, 22]. The simplest method considered in these approaches is majority voting.

B. Data Augmentation

Since a significant amount of data is needed for automatic stance detection, in low data resources languages in which there is not enough labeled data, it is necessary to use methods to increase data, which these techniques are called data augmentation.

Data augmentation techniques refer to strategies that enable us to artificially increase training examples by generating different versions of real datasets without explicitly collecting new data [25]. In data augmentation, the optimal mode is to increase data and improve system performance. Data augmentation strategy is used in computer vision and Natural Language Processing (NLP) to deal with data scarcity and insufficient data diversity. It is relatively easy to create augmented images, but the same is not the case with Natural Language Processing due to the complexities inherent in the language. We cannot replace each word with a synonym and even if we do, the context will be different [26]. Data augmentation is usually done at different levels: letter level, word level, phrase level and document level [26]. On the other hand, data augmentation techniques usually take place in different ranges, from rule-based approaches [27] to model-based methods [28]. Rule-based methods are much easier to implement, but they may not provide significant improvement. Model-based methods greatly effect on performance but, they are more challenging to develop and use. On the other hand, the distribution of augmented data generated should neither be too similar nor too different from the original data because this may lead to overfitting or poor performance through practical data augmentation approaches should aim for a balance [26].

Common methods of data augmentation for NLP are as follows [29]:

1) Paraphrasing-based methods: They create appropriate and limited changes in the data, which have very little semantic difference from the original data. The most common method of this category is back-translation, which consists of three steps:

- Temporary translation: Each of the labeled sentences in the original dataset (source language) is translated into another language (destination language).

- Back-translation: Each of the translated sentences into the destination language is translated again into the source language.

- Removing duplicates: In which duplicate samples are removed from the combination of the two original data sets and the created data. This technique allows to produce of textual data of distinct wording to the original text while preserving the original context and meaning [30]. Several studies use this method to increase data [31, 32]. Table 1 shows a real example in our dataset.

TABLE I. HOW SOME EXAMPLES OF BACK TRANSLATED

Original data	تلف شدن مرغ ها با آنفولانزای پرندگان در یزد	Persian -> English	Loss of chickens due to bird flu in Yazd
Augmented data	تلف شدن جوجه‌ها بر اثر آنفولانزای پرندگان در یزد	Persian <- English	
Original data	وزیر جنگ داعش در حمله هوایی عراق به هلاکت رسید	Persian -> English	The war minister of ISIS was killed in an Iraqi airstrike
Augmented data	وزیر جنگ داعش در حمله هوایی عراق کشته شد	Persian <- English	

2) Noising-based methods: the focus of paraphrasing is to make the semantics of the augmented data as similar as possible to the original data. In contrast, the noising-based methods add weak noise that does not seriously affect the semantics so it appropriately deviates from the original data [29].

One of the most common methods in this category is Easy Data Augmentation (EDA), which consists of four simple but powerful operations. These operations include synonym replacement, random insertion, random swapping and random deletion of words [21].

The aim of synonym replacement is randomly choosing *n* words from the sentence that are not stop words and replacing each of them with one of its synonyms. This replacement is either based on word embedding that uses Pre-trained word embedding like GloVe, Word2Vec, and fast-text [33], or based on lexical that uses a network of word concepts such as WordNet [26]. In random insertion, a word is chosen randomly in the sentence and after choosing one of its synonyms, the synonym is added to an arbitrary position. In random swapping, randomly choose two words within the sentence and swap their positions. This method is usually not a good suggestion for languages that are very grammatically rich (such as Hindi Language) because it may change the meaning of the whole sentence. And finally, the aim of random deletion is randomly remove the words within the sentence. Many research has applied this method [27, 34, 35]. An example of the output of the four mentioned methods is given in Table 2.

TABLE II. AN EXAMPLE OF EDA METHOD

Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
Synonym replacement	A ***lamentable***, superior human comedy played out on the ***backward*** roads of life.
Random insertion	A sad, superior human comedy played out on ***funniness*** the back roads of life.
Random swap	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
Random deletion	A sad, superior human out on the roads of life.

One of the problems of this method is that in relatively short sentences, these changes may lead to noise and sometimes change the corresponding class. But long sentences, because they have more words than short sentences, can absorb more noise and retain their original class label. In this article, we ignored this issue and had the same treatment with all the sentences.

3) Sampling-based methods: In this method, new samples are added based on the data distribution. For example, it is possible to create a larger dataset by merging the original dataset and a similar dataset in another language. The similarity between two dataset means the both datasets have been prepared with the same purpose (here, to stance detection of a text, a news or a tweet reply toward a claim or tweet). Also, both datasets have the same labels or, in case of differentiation, the labels can be mapped to each other.

III. PROPOSED METHODOLOGY

To investigate the effect of data augmentation methods on stance detection performance, used the methods mentioned in Section II. Our methodology is composed of two big steps which have been shown in pseudo code in detail in Figure 1 and Figure 2.

The first step in natural language processing is data preprocessing, which helps to improve the quality of data and extract better meaning from it. In this step, sequences such as numbers, punctuation marks, extra spaces, stop words, and unwanted characters were cleared from the text, after tokenizing it. For this purpose, we used Hazm library for pre-processing the Persian dataset. Hazm is an open-source library to perform the necessary processing for the Persian NLTK[1] library. The NLTK library is also used for the preprocessing of the English dataset.

In the next step, each data augmentation methods were applied to the data and the augmented data were produced. Then, the features should be extracted. We used Bag of Words (BOW) representation and Term Frequency-Inverted Document Frequency (TFIDF) to extract features from our texts. Bag of words is a vector space model used to extract features from textual data simply and flexibly [36]. TFIDF is a vector that shows the importance of a word to a document in a set of documents [37].

In the modeling step, we first split dataset into train and test, then we considered 80% of the datasets as training data and the rest as testing data. Also, we used k-fold cross-validation and set k=10. On the other hand, because the samples in the data are not balanced, that is, the number of samples in each class is not equal, the stratified-KFold library in Python was used to shuffle the data in a balanced way.

Since our goal is to investigate the impact of data augmentation methods on the model's performance, we tried to use only the simple but practical SVM algorithm that shows promising results in text classification. After training the model and evaluating it, we applied the model to test data for predicting their labels. The results of using this algorithm on the augmented data are given in Section IV.

IV. EXPERIMENTAL RESULTS

In this section, we will first describe the dataset used in the article, and after introducing the criteria used, we will present the results of the experiments and compare them with existing similar researches.

[1] Natural Language Toolkit

A. Dataset

A.1. Persian dataset[2]

This dataset contains 534 claims collected from Shayeaat[3] and Fakenews[4]. It consists of two parts [38]: the first includes claims with news headlines and the second includes claims with the article's body text. Each news headline or article's body has one of the following labels:

- Agree: The article states that the claim is true without any hedging or quotation.
- Disagree: The article states that the claim is false, without any kind of hedging and quotation.
- Discuss: The claim is reported in the article without evaluating its truth.
- Unrelated: The claim is not reported in the article.

The first part of this data set which contains the pairs of news headlines and claim has 2029 samples, and the second part, which contains the pairs of article's bodies and claim has 1997 samples. Table 3 shows the distribution of labels in each part of dataset.

Algorithm: Pseudo-code for the **DataAugmentation**.

Input: *PersianDS, EnglishDS* [*input Datasets as global variables*]
Global variables:
Aug_EDA_Data, [Augmented data generated using EDA method]
Aug_BT_Data_Eng, [Augmented data generated using back-translation method by considering English as destination language]
Aug_BT_Data_Ar, [Augmented data generated using back-translation method by considering Arabic as destination language]
Aug_Trans_Data_Eng, [Augmented data generated using merging method in English language]
Aug_Trans_Data_Pers [Augmented data generated using merging method in Persian language]

Function *DataAugmentation*
PreProcPersDS = PreProcess (*PersianDS*) [*After tokenizing each sentences in input dataset, punctuation marks, numbers, additional spaces, stop words, and undesirable characters were removed in the text.]*
PreProcEngDS = PreProcess (*EnglishDS*)
EDA_Data = EDA (*PreProcPersDS) [EDA refer to Easy Data Augmentation method]*
BT_Data_Eng = BT (*PreProcPersDS, Persian, English*) *[BT refer to Back Translation method]*
BT_Data_Ar = BT (*PreProcPersDS, Persian, Arabic*)
Trans_Data_Eng = TR (*PreProcPersDS, English*) *[Translate PreProcPersDS to English*]
Trans_Data_Pers = TR (*PreProcEngDS, Persian*) *[Translate PreProcEngDS to Persian*]
Aug_EDA_Data = Merge (*EDA_Data, PreProcPersDS*)
Aug_BT_Data_Eng = Merge (*BT_Data_Eng, PreProcPersDS*)
Aug_BT_Data_Ar = Merge (*BT_Data_Ar, PreProcPersDS*)
Aug_Trans_Data_Eng = Merge (*Trans_Data_Eng, PreProcPersDS*)
Aug_Trans_Data_Pers = Merge (*Trans_Data_Pers, PreProcEngDS*)
Return
End function

Function EDA (*Dateset*)
Part1, Part2, Part3, Part4 = SplitDS (*Dataset*) *[Divide the Dataset into four* parts *for our methods]*
Aug_Part1 = SynReplace (Part1) [*randomly choosing a word from each sentence in Part1 and replacing it with one of its synonyms]*
Aug_Part2 = RandInsert (Part2) [*randomly choosing a word from each sentence in Part2 and its synonyms is added to an arbitrary position]*
Aug_Part3 = RandSwap (Part3) [*randomly choosing two words within each sentence in Part3 and swap their positions]*
Aug_Part4 = RandDelete (Part4) [*randomly removing the one or some words within each sentence in Part4]*
EDA_Data = Merge (Merg(Aur_Part1, Aug_Part2*)*, Merge (ug_Part3, Aug_Part4*))) [Merging all Aug_part*$_i$ *datasets]*
Return *EDA_Data*
End Function

Function BT (*Dateset, SourceLang, DestLang)*
TransDt_DS = TR *(Dataset, DestLang) [Translate Dataset to DestLang Language]*
TransSr_DS = TR (TransDt_DS, SourceLang) *[Translate TransDt_DS to SourceLang Language]*
BT_Aug = Merge (Dataset, *TransSr_Ds*)
Return *BT_Aug*
End Function

Function Merg (*DS1, DS2)*
MergedDS = Union (DS1, DS2*)*
Return *MergedDS*
End Function

Figure 1. Pseudo code for data augmentation

[2] https://github.com/majidzarharan/persian-stanceclassification
[3] Shayeaat.ir
[4] Fakenews.ir

Algorithm: Pseudo-code for the **StanceDetection**

Input: *AugData* [*input Datasets as global variables*]
Function StanceDetection *(AugDataDS)*
    Vectorized_DS = FeatureExtr (*AugDataDS*) [*Extracting features of each document in AugDataDS and Present them in Matrix format]*
    Train_Data, Test_data = Split (Vectorized_DS, n) *[split Vectorized_DS to n, 1-n; which 0<n<1]*
    Trained_Model = TrainModel (*Train_Data)*
    Predicted_Out = Predict *(Trained_Model, Test_Data)*
    Return
End Function

Function TrainModel (*Train_Data, k) [k is the number of folds]*
    Divide Train_Data into *k* folds with approximately equal distribution of cases
    For *fold* k_i in *the k folds:*
        Set fold k_i *as the test set*
        Perform automated feature selection on the remaining k-1 folds
        Trained_Model = Train model on k-1 folds using hyper-parameter combination [train with SVM algorithm]
        Evaluate model performance on fold k_i
    Calculate average performance over k folds
Return *Trained_Model*
End Function

Function Predict (*Trained_Model, Test_Data)*
    Divide DS into *k* folds with approximately equal distribution of cases
    For S_i in *Test_Data:*
        out_i *= Trained_model* (S_i) *[for each instance in Test_Data create a lable as output]*
Return *[out is a list of all* $out_{i]}$
End Function

Figure 2. Pseudo code for stance detection

TABLE III. DISTRIBUTION OF LABELS IN PERSIAN DATASET

Label / Type	Agree	Disagree	Discuss	Unrelated	Total
Part 1 (Headline-Claim Stance)	405	164	802	658	2029
Part 2 (Article-Claim Stance)	137	206	1068	586	1997

A.2. English dataset

This dataset is related to SemEval-2017-Task8[5] and contains 297 rumors - which are gathered around eight events taken from the urgent news- along with 5271 response tweets, which is a total of 5568 pairs (tweets and tweet responses), which are divided into two parts, training data and test data. Table 4 shows the distribution of tags in this data set:

TABLE IV. DISTRIBUTION OF LABELS IN ENGLISH DATASET

Label / Type	Support	Deny	Query	Comment
Train	910	344	358	2907
Test	94	71	106	778

This dataset used the tree-structure of tweets [3]. It composed of tweets that respond to the rumor tweet. The labels of this dataset are Support, Deny, Query and Comment. Therefore, this dataset is developed to determine the stance of response tweets toward a rumor tweet (that can be direct or nested responses).

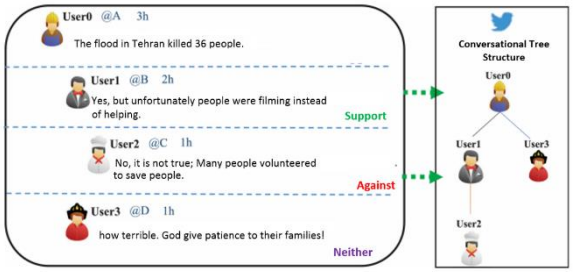

Figure 3. Tree structure of tweets in social media

An example of the tree structure of tweets is shown in Figure 3. In this figure, user1 and user3 directly respond to user0's tweet, but user2 has expressed his opinion in response to user1's post.

Since our Persian dataset only includes the first level (users who directly reply to the source tweet), we only consider the first level in the English dataset and ignore the rest. A real example of the dataset is shown in Figure 4 which L refers to the level.

L0: Unconfirmed reports claim that Michael Essien has contracted Ebola
    L1: his is not funny, people are dying from this. You should be ashamed of yourself.
        L2: Who mentioned it being funny? 'UNCONFIRMED reports'.... to translate: 'it might not be true, but...'
    L1: Glad to hear this is just a rumor. Disgusting whoever made this up. Apologies to @MichaelEssien and for any offence caused.
    L1: no he hasn't. The man himself confirmed not true

Figure 4. A real example in English dataset

[5] https://alt.qcri.org/semeval2017/task8

B. Performance Metrics

To demonstrate the performance of our proposed method, we calculated the accuracy and F-Measure as follows:

F-Measure determines the harmonic mean of precision and recall by giving information about the test's accuracy. It is expressed mathematically as follow:

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (1)$$

Accuracy measures the percentage of correct predictions relative to the total number of samples. It can be expressed as follow:

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} = \frac{TP + FN}{Tp + FN + TN + FP} \quad (2)$$

Recall measures the proportion of data predicted in its class. Mathematically, it can be expressed as follow:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Precision: measures the likelihood of a detected instance of data to its real occurrence. It can be calculated as follow:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

TP, FN, TN, FP stand for true positive, false negative, true negative and false positive respectively.

C. Results

In this section, first we present the results of experiments on the original dataset (without any data augmentation) in Table 5. Then we will examine the effects of applying each of the data augmentation methods on the performance of the used algorithm.

TABLE V. RESULTS ON THE ORIGINAL DATASET

Data	Size	Feature	Accuracy	Precision	Recall	F-Measure
Part 1 (Headline-Claim)	2029	BOW	0.41	0.40	0.41	0.40
		TFIDF	0.53	0.53	0.53	0.53
Part 2 (Article-Claim)	1997	BOW	0.55	0.50	0.55	0.51
		TFIDF	0.83	0.83	0.83	0.83

Table 3 shows that BOW does not lead to good accuracy in the original data set. But we will see later that when the data set increases, this feature can also have a positive effect on improving performance. Below are the methods used for data augmentation in this article, along with the results of each one:

1) Easy Data Augmentation (EDA)

In [32] all operations of EDA technique on the Persian dataset used in the current research have been examined and showed that combining these four operations would be more suitable for this data. Therefore, in this study, we used the combination of these operations on the Persian dataset. The results of this experiment are given in Table 6.

TABLE VI. THE RESULTS OF EDA ON PERSIAN DATASET

Type	Size	Feature	Accuracy	Precision	Recall	F-Measure
Part 1 (Headline-Claim)	4058	BOW	0.52	0.52	0.51	0.52
		TFIDF	0.77	0.77	0.77	0.76
Part 2 (Article-Claim)	3994	BOW	0.63	0.61	0.63	0.61
		TFIDF	0.80	0.80	0.80	0.79

2) Back-translation

In the current research, we considered English and Arabic as the target languages to examine the effect of the relevant method on the results and the effect of the target language. We used googletrans for translating the dataset. Googletrans is a free and unlimited python library that implemented Google Translate API. Tables 7 and 8 show the results of this method by considering English and Arabic, respectively.

TABLE VII. THE RESULTS OF BACK-TRANSLATION ON PERSIAN DATASET (TARGET LANGUAGE IS ENGLISH))

Type	Size	Feature	Accuracy	Precision	Recall	F-Measure
Part 1 (Headline-Claim)	3998	BOW	0.49	0.48	0.49	0.48
		TFIDF	0.67	0.67	0.67	0.66
Part 2 (Article-Claim)	4058	BOW	0.46	0.45	0.46	0.45
		TFIDF	0.62	0.62	0.62	0.62

TABLE VIII. THE RESULTS OF BACK-TRANSLATION ON PERSIAN DATASET (TARGET LANGUAGE IS ARABIC))

Data	Size	Feature	Accuracy	Precision	Recall	F-Measure
Part 1 (Headline-Claim)	3985	BOW	0.60	0.57	0.60	0.57
		TFIDF	0.72	0.71	0.72	0.70
Part 2 (Article-Claim)	3994	BOW	0.57	0.57	0.56	0.56
		TFIDF	0.66	0.65	0.66	0.65

TABLE IX. THE RESULTS OF DATA AUGMENTATION METHODS

Data Augmentation Method	Dataset	(Headline, Claim)				(Body, Claim)			
		BOW		TFIDF		BOW		TFIDF	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
	Persian Dataset	0.41	0.40	0.53	0.53	0.55	0.51	0.83	0.83
	Translated Persian Dataset	0.41	0.39	0.47	0.46	0.55	0.52	0.62	0.59
EDA	*Aug_EDA_Data*	0.52	0.52	0.77	0.76	0.63	0.61	0.80	0.79
Back-translation	*Aug_BT_Data_Eng*	0.49	0.48	0.67	0.66	0.60	0.57	0.72	0.70
	Aug_BT_Data_Ar	0.46	0.45	0.62	0.62	0.57	0.56	0.66	0.65
Merging datasets	*Aug_Trans_Data_Eng*	0.77	0.78	0.81	0.81	0.66	0.63	0.86	0.85
	Aug_Trans_Data_Pers	**0.77**	**0.78**	**0.82**	**0.82**	**0.83**	**0.83**	**0.86**	**0.86**

3) Merging the Persian dataset and English dataset

Most of the work done in Persian is in rumor detection, and there is no specific study on stance detection [38]. The only study has been done in this filed was explained in the beginning of Section IV [38]. One of the English datasets that can be considered equivalent to this is the dataset that was published in Semeval 2017-task 8, because its type is based on claims (claim-based), and on the other hand, its labels can be mapped to each other. The explanations of English dataset are given before.

Since the Persian dataset does not consider the tree structure, we also tried to select only the first level of English dataset tree structure, and thus the size of our used English dataset was reduced to 3272 samples.

At the next step, the English dataset labels were mapped to the Persian dataset as follows:

$$Agree \approx Support$$
$$Disagree \approx Deny$$
$$Discuss \approx Query$$
$$Comment \approx Unrelated$$

Finally, once the English dataset is translated into Farsi and added to the Persian dataset, and once again the Persian dataset is translated to English and added to the English dataset. Tables 9 and 10 show the test results on these two new datasets, respectively.

TABLE X. THE RESULTS OF MERGING DATASETS IN PERSIAN LANGUAGE

Data	Size	Feature	Accuracy	Precision	Recall	F-Measure
Translated English dataset + part1	5301	BOW	0.77	0.79	0.77	0.78
		TFIDF	0.81	0.81	0.81	0.81
Translated English dataset + part2	5269	BOW	0.66	0.64	0.66	0.63
		TFIDF	0.87	0.88	0.87	0.87

TABLE XI. THE RESULTS OF MERGING DATASETS IN ENGLISH LANGUAGE

Data	Size	Feature	Accuracy	Precision	Recall	F-Measure
Translated Part1 + English dataset	5301	BOW	0.77	0.79	0.77	0.78
		TFIDF	0.82	0.82	0.82	0.82
Translated Part2 + English dataset	5269	BOW	0.83	0.83	0.83	0.83
		TFIDF	0.85	0.86	0.85	0.85

D. Comparison and Analysis of results

Table 11 shows the summary of the results on each of datasets by the used features, the applied method and the evaluation criteria. The first two lines of this table show the results on the original data, the first line on the original data and the second line on the translation of the same original data. The other lines of the table show the results on the augmented data.

The results indicate that the best method to increase the quality of stance detection performance is the method of merging the original dataset with a similar dataset in other language, which leads to increase diversity in data and also there is no need to spend time and money to prepare the many samples. If such dataset is not found in other languages or it is not possible to access it, the next optimal method is the easy data augmentation (EDA) method, which also shows a good improvement in the performance of the algorithm.

Back translation method although has increased the amount of evaluation criteria, but compared to the other two methods, it makes less improvement in the algorithm. On the other hand, we tried to check the effect of the language on the results obtained in back translation method. Therefore, we chose two languages for target: English and Arabic. The results show that although the Persian language is more similar to the Arabic language in terms of alphabet and writing form but when the destination language is English, the results are better than what is achieved when it is Arabic. Because the Persian language and the English language do not have the grammar and syntax as the Arabic language.

Finally, we compared our model with the best models presented in [34] and [38] which are the only research that have been done in the field of Persian

stance detection. [38] used LSTM and [38] applied transfer learning and EDA data augmentation on Persian dataset which described in this paper. Table 12 shows the comparison results.

TABLE XII. COMPARISON OF THE PROPOSED MODEL WITH OTHER MODELS PRESENTED FOR PERSIAN STANCE DETECTION

Model	Headline-Claim		Body-Claim	
	Accuracy	F1	Accuracy	F1
[38]	0.67	0.67	72	71
[34]	0.75	0.75	0.76	0.76
The proposed model	**0.82**	**0.82**	**0.86**	**0.86**

V. CONCLUSION

In this paper, in order to resolving the problem of lack of data in Persian stance detection, we used data augmentation techniques. Then we analyzed the effect of each one in improving the performance of the SVM algorithm. In this regard, we investigated the Persian claim based stance detection and used different data augmentation methods (such as easy data augmentation, back translation, and merging similar datasets). The test results show that if we can merge the source dataset with similar dataset in other languages and create a larger dataset, we will get significant improvement without spending time, cost and human resources in collecting data and labeling them. If such a data set is not available or does not exist, a good improvement can be achieved by using the EDA technique.

Also, deep learning methods and transformers such as BERT can be used to improve the stance detection model and apply them to augmented data. Also, another thing that can be done to reduce dependence on dataset is to use transfer learning methods, so that the model is trained on English data and then the trained model is used on Persian data. It is also possible to examine the effect of other features such as the polarity of feeling and so on the results.

REFERENCES

[1] Bhatt, S., Goenka, N., Kalra, S., and Sharma, Y. (2022). Fake News Detection: Experiments and Approaches beyond Linguistic Features. In Data Management, Analytics and Innovation (pp. 113-128). Springer, Singapore.

[2] Pomerleau, D., Delip R, (2017). The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news

[3] Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G. W. S., & Zubiaga, A. (2017). SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).

[4] Allcott, H., and Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of economic perspectives, 31(2), 211-36.

[5] Kochkina, E., Liakata, M., and Augenstein, I. (2017). Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm. In proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)

[6] Yuan, C., Qian, W., Ma, Q., Zhou, W., and Hu, S. (2021, July). SRLF: a stance-aware reinforcement learning framework for content-based rumor detection on social media. In 2021 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[7] Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., and Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. PloS one, 11(3), e0150989.

[8] Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016, June). Semeval-2016 task 6: Detecting stance in tweets. In Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016) (pp. 31-41).

[9] Conforti, C., Berndt, J., Pilehvar, M. T., Giannitsarou, C., Toxvaerd, F., and Collier, N. (2020). Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter. In proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

[10] Sobhani, P., Inkpen, D., and Zhu, X. (2017, April). A dataset for multi-target stance detection. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (pp. 551-557).

[11] Kucuk, D., and Can, F. (2022, February). A Tutorial on Stance Detection. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (pp. 1626-1628).

[12] Lai, M., Cignarella, A. T., Farías, D. I. H., Bosco, C., Patti, V., and Rosso, P. (2020). Multilingual stance detection in social media political debates. Computer Speech and Language, 63, 101075.

[13] Zotova, E., Agerri, R., Nuñez, M., and Rigau, G. (2020, May). Multilingual stance detection in tweets: The Catalonia independence corpus. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 1368-1375).

[14] Swami, S., Khandelwal, A., Singh, V., Akhtar, S. S., and Shrivastava, M. (2018). An english-hindi code-mixed corpus: Stance annotation and baseline system. arXiv preprint arXiv:1805.11868.

[15] Du, J., Xu, R., He, Y., and Gui, L. (2017, August). Stance classification with target-specific neural attention networks. International Joint Conferences on Artificial Intelligence.

[16] Darwish, K., Magdy, W., and Zanouda, T. (2017, September). Trump vs. Hillary: What went viral during the 2016 US presidential election. In International conference on social informatics (pp. 143-161). Springer, Cham.

[17] Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., and Slonim, N. (2017, April). Stance classification of context-dependent claims. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (pp. 251-261).

[18] Kucuk, D., and Can, F. (2020). Stance detection: A survey. ACM Computing Surveys (CSUR), 53(1), 1-37.

[19] Wojatzki, M., and Zesch, T. (2016, June). ltl. uni-due at semeval-2016 task 6: Stance detection in social media using stacked classifiers. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 428-433).

[20] Cignarella, A. T., Lai, M., Bosco, C., Patti, V., and Paolo, R. (2020). Sardistance@ evalita2020: Overview of the task on stance detection in italian tweets. EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, 1-10.

[21] Wei, P., Lin, J., and Mao, W. (2018, June). Multi-target stance detection via a dynamic memory-augmented network. In The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1229-1232).

[22] Tutek, M., Sekulić, I., Gombar, P., Paljak, I., Čulinović, F., Boltužić, F., ... and Šnajder, J. (2016, June). Takelab at semeval-2016 task 6: Stance classification in tweets using a genetic algorithm based ensemble.

In Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016) (pp. 464-468).

[23] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301. 3781.

[24] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[25] Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., and Hovy, E. (2021). A survey of data augmentation approaches for NLP. In ACL 2021.

[26] Tidke, P. (2022, February). Text Data Augmentation in Natural Language Processing with Texattack

[27] Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in neural information processing systems, 28.

[28] Liu, R., Xu, G., Jia, C., Ma, W., Wang, L., and Vosoughi, S. (2020). Data boost: Text data augmentation through reinforcement learning guided conditional generation. In proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)

[29] Li, B., Hou, Y., and Che, W. (2022). Data augmentation approaches in natural language processing: A survey. AI Open.

[30] Beddiar, D. R., Jahan, M. S., and Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. Online Social Networks and Media, 24, 100153.

[31] Sennrich, R., Haddow, B., and Birch, A. (2015). Improving neural machine translation models with monolingual data. In 54th ACl 2016.

[32] Yu, A. W., Dohan, D., Le, Q., Luong, T., Zhao, R., and Chen, K. (2018, May). Fast and accurate reading comprehension by combining self-attention and convolution. In International Conference on Learning Representations (Vol. 2, No. 1).

[33] d'Sa, A. G., Illina, I., & Fohr, D. (2020, February). Bert and fasttext embeddings for automatic detection of toxic speech. In 2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA) (pp. 1-5). IEEE.

[34] Nasiri, H., and Analoui, M. (2022, February). Persian Stance Detection with Transfer Learning and Data Augmentation. In 2022 27th International Computer Conference, Computer Society of Iran (CSICC) (pp. 1-5). IEEE.

[35] Huang, W., and Wang, J. (2016). Character-level convolutional network for text classification applied to chinese corpus. The 3rd international conference on machine learning and machine intelligence (pp. 83-87)

[36] Zhang, Y., Jin, R., and Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. International journal of machine learning and cybernetics, 1(1), 43-52.

[37] Qaiser, S., and Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181(1), 25-29.

[38] Zarharan, M., Ahangar, S., Rezvaninejad, F. S., Bidhendi, M. L., Pilevar, M. T., Minaei, B., and Eetemadi, S. (2019). Persian Stance Classification Data Set. In Conference on Truth and Trust Online.

Mojgan Farhoodi received her B.Sc. degree in Software Engineering and her M.Sc. degree in IT Engineering from Amirkabir University of Technology. She has been working as a researcher at the ICT Research Institute since 2010. Currently, he is working as a faculty member of the research institute in the fields of Information Retrieval, Natural Language Processing, and Artificial Intelligence.

Abbas Toloie Eshlaghi received his Ph.D. degree in Industrial Management at Islamic Azad University. He is head of Industrial Management, Faculty of Management and Economics, Islamic Azad University, Science and Research Branch. He has published about 60 papers in Iranian journals, 70 international journals (indexed in web of science, web of knowledge and Scopus) and 17 books (Original and translated), and 46 conference papers.

Mohammadreza Motadel received his Ph.D. degree in Operations Management at Islamic Azad University. He is member of board of Azad University Central Tehran Branch. He has published about 12 paper in Iranian journals, 29 international journals (indexed in web of science, web of knowledge and Scopus), 4 books (Original and translated), and 7 conference papers.