An Ensemble Deep Model for Deceptive Opinion Detection Based on Opinion Text: For English and Persian Languages

Mahmoud Ali-Arab[1], Kazim Fouladi-Ghaleh[2]

[1] *M.Sc. of Information Technology Engineering, Deep Learning Research Lab, Faculty of Engineering, College of Farabi, University of Tehran, Iran*
aliarab.m@gmail.com
[2] *Assistant Professor, Department of Computer Engineering, Faculty of Engineering, College of Farabi, University of Tehran, Iran; Head of Cyberspace Research Lab, University of Tehran, Iran*
kfouladi@ut.ac.ir

Abstract

Spam reviews, written primarily to promote or demote a product or brand, mislead people for making purchases and make decisions difficult for customers. Much research has been done to detect spam reviews, and different methods have been developed, but these methods often use metadata to detect spam review, and because of the use of metadata, singleton reviews (reviews whose author has submitted only one comment) are dropped from the dataset because these types of reviews do not give much information to the model. In addition, in existing methods, comment text is considered any other text in text classification issues, while comment text contains many features that can be extracted and used to detect spam reviews. In this research, a hybrid model using 4 BiLSTM networks is presented, trained on the comment's text and the comments' polarity. Due to the lack of polarity of opinions in different datasets, a sentiment analysis model has been used that extracts the polarity of opinions from the comments text and adds it to the dataset. Since the model depends only on the comment's text and does not use metadata, there will be no problem in detecting singleton spam reviews using this model. The proposed model is evaluated for English and Persian languages. The performance of the proposed model is comparable for both Persian and English. For English, the accuracy was 89.4% on the OpSpam dataset and 87.7% on the Hotel domain (Doctor, Restaurant (HDR)) dataset. Also, 87.7% accuracy was obtained for the Persian language on the Digikala dataset.

Keywords: *Review Spam Detection, Opinion Spam, Deep Learning, Ensemble Model, Long Short-Term Memory (LSTM), Persian, English.*

1 Introduction

Customer reviews are critical because they significantly impact other customers, and the information it provides makes a user decide to buy a product. Nearly 95% of people read the reviews written about products before buying online and then decide to purchase [10]. The impact of these comments is not only on customers, but businesses use these reviews to improve the quality of their services or marketing decisions, etc. Due to the importance of these reviews and their impact on product sales, spam reviews have also spread. Spam review is an opinion that is not the result of a person's experience and is written to promote or demote a brand. Spam reviews can lead other customers to make wrong decisions. On many websites, people can post any review, and it is difficult for humans to tell if a comment is spam, which is why spam review is becoming more and more challenging to detect. Therefore, there is a need for a model that can recognize these spam reviews.

In recent years, much research has been done to identify spam reviews, and various businesses are looking for a way to deal with spam reviews. The number of researches in this field is increasing exponentially [1], and due to the increase of unrealistic information on the Internet, the research about spam detection is increasing every day.

There is a problem called singleton spam reviews in existing methods of detecting spam reviews. These comments are written by people who have written only one comment. If metadata such as username, IP, etc. are used to train the model for detecting spam reviews, singleton reviews do not give any information to the model, and therefore in many of existing methods, these reviews are dropped from the dataset, and these methods cannot detect singleton spam reviews. In addition, in existing research, the features of the comments text are usually not considered, while comment text includes different features that can extract and use to train the model.

In this research, a hybrid model is proposed that depends only on the text of the review and its label. The polarity of comments is also extracted from the comment text using a sentiment analysis model and added to the dataset. The model is implemented so that it can be trained for different languages with minor changes.

The proposed model is also trained for Persian. Existing methods for detecting spam reviews for the Persian language using traditional machine learning models have addressed this issue, so the results obtained in this study are significantly better than existing research for Persian.

2 Background and related works

The topic of review spam detection has been one of the most active topics for research in recent years. Much research has been done on this subject, and the number of these researches is increasing exponentially. In these researches, different learning methods and characteristics have been examined. Research to detect spam review can be divided into different categories by aspects such as the type of learning, type of features used,

identification techniques, etc. In terms of the type of learning, research is divided into supervised learning, Unsupervised Learning, and semi-supervised learning [4].

2.1 Types of learning

2.1.1 Supervised learning

Supervised learning is one of the most efficient methods of machine learning. This method uses labeled data. The problem with this type of learning is that there are not enough labeled data. For this reason, researchers are trying to use other methods as well. Numerous studies have used supervised learning methods to detect spam reviews. This type of learning performs better than other machine learning methods if there is a sufficiently labeled dataset. In this research, a supervised learning method and labeled dataset have been used. As mentioned, most existing methods try to use all available metadata. For example, Huang et al. [2] have used supervised learning. They collected data using crawlers from Epinions. They also gave their model information, such as how helpful the comment was and what rating it was given. Using extensive metadata does not necessarily improve model performance. Mukherjee et al. [3] showed that the low usefulness of a comment is not a reason for the comment to be spam because one of the methods used by spammers is to use group spamming in which several people write a comment and, in this situation, Spammers are more likely to rate each other's opinions higher and choose those opinions as applicable.

2.1.2 Unsupervised learning

One of the significant problems with machine learning models is the lack of labeled data. If there is enough labeled data, the best way is to use supervised learning, but real-world data is often unlabeled, so unsupervised learning does not require labeled data. Data labeling is a difficult task that is both time-consuming and costly. Lots of data related to user comments are also unlabeled. For this reason, researchers try to use the method of unsupervised learning. Although unsupervised methods have poorer performance than supervised methods, new research seeks to optimize these methods to perform better. In 2020, Saumya et al. [5] developed an unsupervised model using LSTM and Autoencoder networks that can be trained using comment text without labels. They used the Matthew correlation coefficient (MCC) metric to evaluate their model.

2.1.3 Semi-supervised learning

In recent years a method has been used called semi-supervised learning. This method uses labeled and unlabeled data. A small set of labeled data and a set of unlabeled data are given to the model. In this method, the unlabeled data is labeled using labeled data, and then the labeled data is used as training data [4]. In this way, more labeled data is given to the model for training. Research using this method has increased in

recent years. Different methods are used for semi-supervised learning, and in [6], the performance of each of these types of methods is compared.

2.2 Related works

In this research, the issue of detecting spam reviews for both Persian and English languages has been investigated, and the efficiency of the proposed model has been evaluated for both languages. So, in the related works section, related works for English and Persian languages are discussed in two separate sections.

2.2.1 Related works for English

The issue of spam detection was first formulated by Jindal et al. [7, 8, 9]. They divided spam into three categories: unrealistic, branded, and unrelated. They claimed that the second and third categories of comments do not pose a problem and are easily identifiable, but the first category are not easily recognizable, and a model must be created to identify them [10]. In research [9], measuring the similarity of opinions has been used to identify spam. The first public dataset to detect spam reviews was published by Ott et al. [11] in 2011. This dataset contains 800 truthful comments and 800 deceptive comments about 20 Chicago hotels. Deceptive reviews in this dataset are written by Amazon Mechanical Turk (AMT). A few years later, Li et al. [13], based on Ott et al. [11] dataset, introduced a dataset prepared in three domains: hotel, doctor, and restaurant. This dataset is one of the most widely used datasets in spam detection, and the amount of data in the Li et al. [13] dataset is more than the dataset of Ott et al. [11].

In the study of Wael et al. [12], the effect of different preprocessing stages on the data on the efficiency of the spam detection model was investigated. Several preprocessing methods on the text such as stop word removal, removing emphasis marks, stemming, etc. were examined in this study and the effect of each of these methods by teaching several different models of machine learning such as Naive Bayes Network, Support vector machine, random forest, etc. were measured. With the growth of deep neural networks, usage of these networks in spam review detection research has also increased. In general, deep neural networks have several advantages over traditional machine learning methods. First, neural network-based models have many nonlinear methods that can be modified and enhanced based on neural network depth. Second, neural networks can derive features from raw data. That is, the feature extraction step is done in the neural network itself, and the third thing that is most used in the field of working with text is that using deep learning if a good word embedding is used in model training, the model can easily understand the relationship between words and their proximity to each other and even sentence structure [14]. Lie et al. [15] Have used CNN networks to detect spam reviews. In their research, word vectors are given as input features to the network, and spam reviews are directly identified using CNN.

The use of hybrid methods and the integration of several deep learning models have also been considered to detect spam reviews. Zhang and Ren [16] used document-level learning to detect spam. First, a document is given to the model, and using CNN combined with a network (Gated-RNN), the sentences and their structure are learned, and the document vectors are extracted by this method, then these vectors are used directly to detect spam reviews. Zhao et al. [17] have used a new method called using word order-preserving in the convolutional layers and merging CNN network, instead of using the usual concatenation layer in the convolutional network. This maintains the order of the words in the integration layer and improves the CNN network for spam detection.

The length of review texts is very different, so a maximum length is usually considered for the input text. This maximum length should be chosen so that the model has the most performance, but if this maximum length is small, a large part of the data will be lost, and if this maximum length is considerable, it will have a high computational cost. So, Kumar et al. [18] came up with using the full text of the reviews. They divided the text of each review into several smaller parts and assigned a label equal to the original review label for each of them. These scaled-down comments were given to a combined CNN and GRU networks model, and the final label was determined by max-voting. Barushka et al. [19] developed a deep neural networks (DNN) model. They have tried to use the content of the review to train their model, using both the bag of words (BOW) and the meaning of the words to teach the model. They also used N-gram and Skip-gram word embedding methods to obtain word vectors and train their model.

2.2.2 Related works for Persian

The proposed model is also trained and evaluated with Persian data in this research. Therefore, existing methods to identify spam opinion in Persian have been reviewed in this section.

Little research has been done to identify spam reviews in Persian. Existing research has also used traditional machine learning methods for this subject, so their results are not very good. Safarian et al. [20] have used feature ranking for review spam detection. They have tried to examine the various features used to train the model in the problem of spam review detection. They have used different models such as Naive Bayes, decision tree, support vector machine, etc. Each of these models is trained with different features such as overall product rating, the sentiment of comments, POS tags, etc. In their research, training data from users' opinions of the Digikala website (the most extensive retail site in Iran) has been used. Basiri et al. [32] Also tried to use various machine learning methods such as Naive Bayes, decision tree, support vector machine, and various features extracted from the comment text and other metadata available in the dataset. Their research has been done on balanced and unbalanced data, and according to the obtained results, the support vector machine for unbalanced

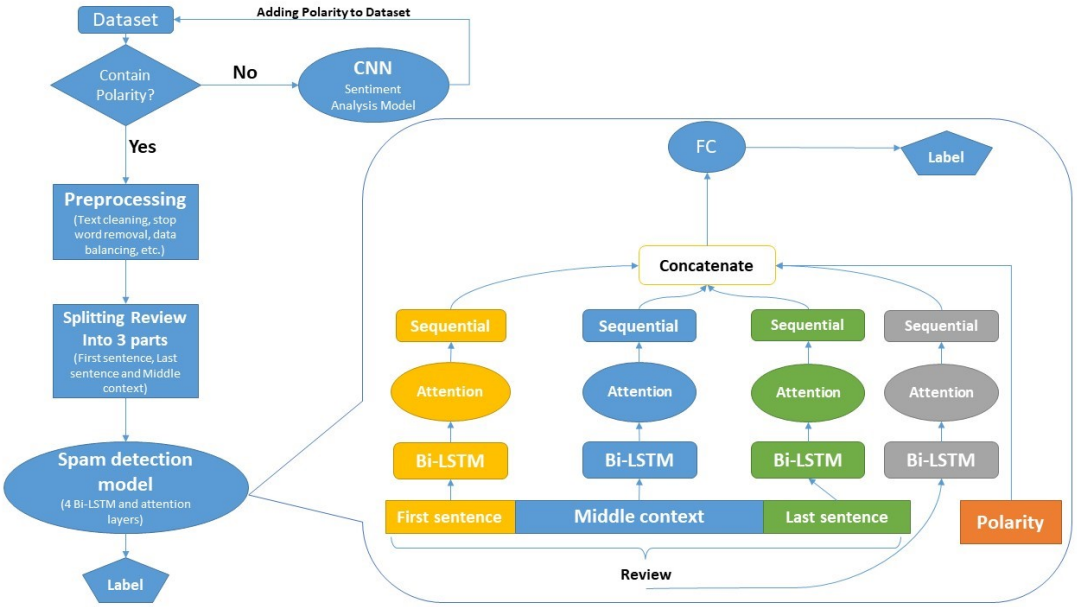

Figure 1: Proposed model architecture

data and the decision tree for balanced data have the best performance.

3 Methodology

This research aims to provide a hybrid model for detecting spam reviews. In the proposed model, only the text of the reviews and their labels is used to train the model. Due to the effect of opinion polarity on the problem of spam review detection [21], and given that the polarity of opinions may not be present in different datasets, in this study, the polarity of opinions is extracted by a sentiment analysis model and added to the dataset. In the text of a comment, the first sentence and the last sentence are more critical, and for this reason, in this research, training is done on the first and last sentence separately. As shown in Figure 1, in this research, the text of the comment is divided into three parts: first sentence, last sentence, and middle context, and each of these three parts is given to a bidirectional long short-term memory (BiLSTM), and the entire comment text is given to a BiLSTM. There is a total of 4 BiLSTMs in the proposed model. The output of each BiLSTM layer, after passing through a self-attention layer, eventually joins together to form a vector. The polarity of the review, which is calculated as binary (positive or negative), is also joined to this vector at this stage, and the resulting vector is given to a fully connected layer (classification layer) to produce the final output label.

Table 1: Statistics of HDR dataset

	Turker	*Expert*	*Customer*
Hotel (P/N)	400/400	140/140	400/400
Restaurant (P/N)	200/0	120/0	200/200
Doctor (P/N)	200/0	32/0	200/0

3.1 Dataset

In this research, two datasets OpSpam [22] and (Hotel, Doctor, Restaurant (HDR)) [13], have been used. The reason that several datasets are used in this research is to the proposed model be comparable with different models and different researches, and also the performance of the model is measured in different domains.

OpSpam [22] is a balanced database that contains 1,600 reviews of Chicago hotels. This dataset contains 800 spam comments and 800 real comments. There are 400 negative comments and 400 positive comments in each of these categories. In this dataset, real comments are collected from the Yelp website, and deceptive comments are generated by Amazon Mechanical Turk (AMT).

Data sets (Hotel, Doctor, Restaurant (HDR)) [13] have also been used in this research, which is one of the most widely used datasets in research in the field of spam review detection. This dataset is collected in three domains of comments related to hotels, restaurants, and doctors. real comments in this dataset are collected from customers of each domain and deceptive comments are written by AMT or employees of each domain (expert). The statistics of this dataset are given in Table 1.

3.2 Data preparation

As shown in Figure 2, the dataset is first examined to see if it includes the polarity of the comments. If the dataset does not have the polarity of comments, the sentiment analysis model automatically extracts the polarity of the comments in binary (positive or negative) from the comments text and adds it to the dataset. The dataset is then divided into two parts: training data and evaluation data. In this study, 20% of the data is considered evaluation data, and the rest is considered training data.

After this step, the data balance is checked, and if the dataset is unbalanced, the data are balanced using the OverSampling method. This method is one of the standard methods of data balancing.

After balancing the data, the comment text is divided into the first sentence, middle context, and final sentence. Each of these sections is tokenized. The entire text of the comment is also tokenized at this stage. In this research, the SpaCy library has been used for preprocessing in English, and also Hazm and Parsivar libraries have been used for Persian. After data tokenization, the stop words are removed, and word vectorization is performed. Finally, these vectors are given as input data to the spam review detection model.

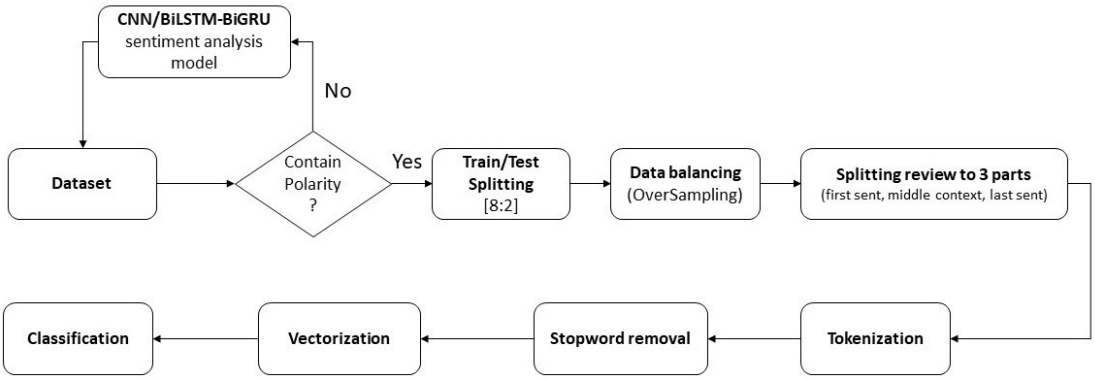

Figure 2: Data preparation flowchart

3.3 Comment polarity extraction

Due to the effect of opinion polarity in detecting spam reviews [21] in this study, the polarity of comments has been used to detect spam opinions. Because the polarity of opinions may not be present in many datasets, first, the dataset is examined. If the polarity of opinions is not available, using a sentiment analysis model, the polarity of opinions is extracted from the text of the opinion and added to the dataset. Finally, the model is trained to detect spam using this new dataset.

For English, an open-source sentiment analysis model has been used, implemented using CNN, and has an accuracy of about 85%. For Persian, an ensemble sentiment analysis model has been used, implemented using BiLSTM and BiGRU networks, and has an accuracy of about 92%. In this research, sentiment is considered binary, and one opinion can be positive or negative.

3.4 Bidirectional Long Short-Term Memory (BiLSTM) layer

Long Short-term memory networks (LSTMs) are commonly used for sequence models, and since a text is also a sequence of words and letters, LSTM networks perform well for text classification issues. These networks are a particular type of recurrent neural network (RNN) that has solved the problem of gradient vanishing by introducing memory cells and gate mechanisms. In this type of network, the information generated at the output is stored in a memory cell. This storage operation is controlled by three gates (g_i, g_f, g_o) and determines the amount of forgetting or storage of information defined in Equations 1 to 3. In these equations, x_j is the input at position j of the sequence given to the model. h_{j-1} is also the state of the previous cell.

$$g_i = \sigma(x_j W^{x_i} + h_{j-1} W^{h_i}) \tag{1}$$

$$g_f = \sigma(x_j W^{x_f} + h_{j-1} W^{h_f}) \tag{2}$$

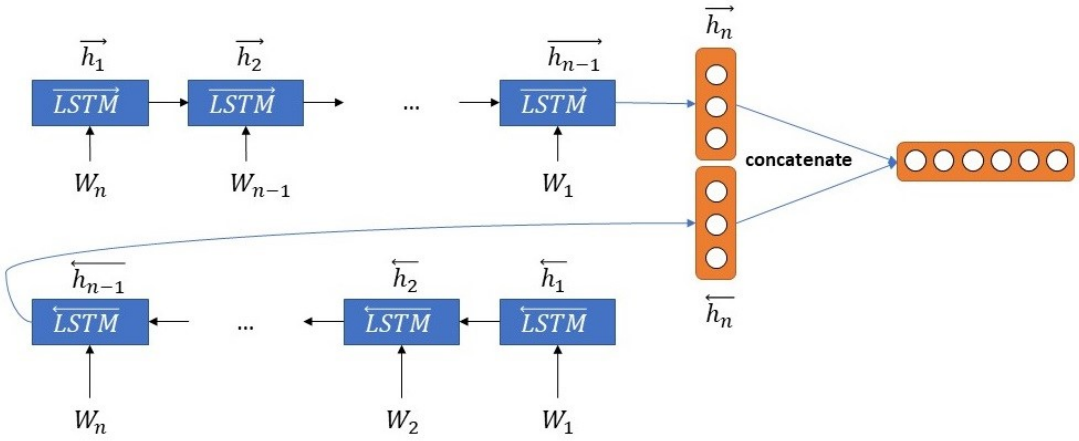

Figure 3: Bidirectional Long Short-term memory (BiLSTM) architecture

$$g_o = \sigma(x_j W^{x_o} + h_{j-1} W^{h_o}) \tag{3}$$

The new model is also a linear equation of x_j and h_{j-1} given to a tanh activation function (Equation 4).

$$z = \tanh(x_j W^{x_z} + h_{j-1} W^{h_z}) \tag{4}$$

The value of z in a linear combination with the previous amount of memory creates a new amount of memory. Equation 5 shows this linear combination in which c_j is the new value, and c_{j-1} is the previous value of the memory cell. g_f controls the amount of forgetting the previous amount of memory, and g_i specifies the new amount to be stored in the memory cell.

$$c_j = g_f c_{j-1} + g_i z \tag{5}$$

The final output, as mentioned, is controlled using the g_o gate, and the c_j value is generated using the tanh activation function, which is shown in Equation 6. In this equation, h_j represents the LSTM output at position j.

$$h_j = g_o(\tanh(c_j)) \tag{6}$$

The BiLSTM model has been used in this research. BiLSTM traverses the sequence in two directions. Two LSTMs are used in this model, one of which follows the sequence from beginning to end and the other from end to end. Moreover, the training process is done this way. The information of these two LSTMs is concatenated in each step. The architecture of the BiLSTM model is shown in Figure 3.

BiLSTM output in each position is the concatenation of the output of forwarding LSTM ($\overrightarrow{LSTM}$) and the output of backward LSTM ($\overleftarrow{LSTM}$) (Equations 7 to 9).

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(e_t, \overrightarrow{h_{t-1}}) \tag{7}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(e_t, \overleftarrow{h_{t-1}}) \tag{8}$$

$$H_t = (\overrightarrow{h_t} : \overleftarrow{h_t}) \tag{9}$$

In this research, the text of each comment is divided into three parts: the first sentence, the middle context, and the final sentence. Each of these parts is given to a BiLSTM. The entire comment text is also given to a BiLSTM. That is, in total, the proposed model includes 4 BiLSTM models.

3.5 Self-attention mechanism

Self-attention is a particular attention mechanism that can efficiently detect dependence in different parts of a sequence such as convolutional neural networks or recurrent neural networks, with the difference that in comparison with recurrent neural networks or convolutional neural networks have fewer parameters and less complexity. The output of the self-attention layer is a weighted average of different positions of the sequence.

In this research, multilayer perceptron has been used as the primary attention function, and softmax function has been used for normalization. The output vectors of the first sentence, the middle context, and the last sentence generated by BiLSTM are represented by s_1, s_2, and s_3. The input of the attention layer itself is a combination of three vectors, s_1, s_2, and s_3, which are displayed as $S = [s_1 : s_2 : s_3]$. Equations 10 to 12 show these steps.

$$Adp = \tanh(W \cdot S^T + b) \tag{10}$$

$$Attention = \text{softmax}(Adp(S)) \tag{11}$$

$$Attention_i = \frac{\exp(Adp(S_i))}{\sum_{i=1}^{3} \exp(Adp(S_i))} \tag{12}$$

The output of the self-attention mechanism is the weighted average S, while the weight matrix is Attention. In practice, the output of the self-attention mechanism is still a sequence, and each element can be seen as a representation of the document. The final output of the self-attention mechanism is displayed with Z.

The output of the BiLSTM corresponding to the entire comment text is displayed with s_c. s_c and Z are both representations of the comment text that concatenate together. At this point, the polarity of the review represented by p is also added to this sequence. Equation 13 specifies the final output generation steps.

$$O = [Z : s_c : p] \tag{13}$$

This output (O) is given to a fully connected (FC) layer to generate the output label. The final label is generated in binary and specifies whether the comment is spam or genuine.

Table 2: Optimal hyper-parameter values for each dataset

Dataset	Embedding size	Learning rate	Hidden size	Epochs
OpSpam	100	0.0007	64	35
HDR → Hotel	100	0.0005	64	40
HDR → Doctor	100	0.0005	64	40
HDR → Restaurant	100	0.001	32	40
Digikala (Persian lang)	50	0.005	40	15

4 Results

As mentioned earlier, the proposed model for English is evaluated on the OpSpam and HDR datasets, and for Persian on the Digikala dataset. Because the proposed model for Persian and English has been evaluated, the comparison of results for each language is given in separate sections. Compared to the base model [10] and other models, the results showed that the proposed model's performance for the OpSpam dataset and the Hotel domain of the HDR dataset is better than other models. Also, for Persian, the obtained results showed that the performance of the proposed model is much better than the existing methods.

4.1 Evaluation metrics and Hyperparameters

In this research, several evaluation metrics have been used to make the results more reliable and to be able to compare these results with other research. Accuracy, $F1$, *Recall*, and *Precision* metrics are used (equation 14 to 17). The use of multiple evaluation metrics is crucial in research that uses unbalanced data sets because the use of one metric cannot show reliable results.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{17}$$

The proposed model hyperparameters are adjusted to obtain the best result for each dataset. Table 2 shows metadata per dataset. This table lists essential parameters such as learning rate, embedding layer size, hidden layer size, and the number of configured epochs per dataset.

Table 3: The proposed model result for English datasets

Dataset	F1	Accuracy	Precision	Recall
OpSpam	88.6	89.4	86.6	90.6
HDR → Hotel	87.7	87.7	87.9	87.5
HDR → Doctor	88.54	89.51	96.66	81.69
HDR → Restaurant	86.42	86.25	85.36	87.5

Table 4: Results comparison for English

Dataset	Model	Year	F1	Accuracy	Precision	Recall
HDR (Hotel)	EnsDOS [25]	2019	85.7	85.7	85.5	86.1
	SOMCNN [29]	2021	85	86	85	86
	CNN_BiLSTM [30]	2021	86.1	83	86.8	85.3
	Proposed Model	-	**87.7**	**87.7**	**87.9**	**87.5**
HDR (Doctor)	EnsDOS [25]	2019	85	84.7	83.6	86.5
	SOMCNN [29]	2021	**93.0**	**94.0**	93	**93.0**
	CNN_BiLSTM [30]	2021	92.8	91	**97.0**	88.9
	Proposed Model	-	88.54	89.51	96.66	81.69
HDR (Restaurant)	EnsDOS [25]	2019	85.8	85.5	84.1	**88.5**
	SOMCNN [29]	2021	**88.0**	**88.0**	89	87
	CNN_BiLSTM [30]	2021	80.9	77.5	**90.5**	73.1
	Proposed Model	-	86.42	86.25	85.36	87.5
OpSpam	SingleCNN [23]	2017	81.1	81.2	78.2	84.3
	IMP [24]	2018	-	83.5	-	-
	MFCNN [26]	2020	86.5	83.5	84.6	88.4
	DOSDL [27]	2020	87.1	87.2	**87.3**	87.5
	DOSLSTM [28]	2020	-	83.3	78	81
	Proposed Model	-	**88.6**	**89.4**	86.6	**90.6**

4.2 Result comparison for English

Table 3 shows the results obtained for the proposed model for different English datasets. The training and evaluation of this model for English have been done on two widely used datasets, OpSpam and HDR. The HDR dataset includes three domains: Hotel, Doctor, and Restaurant. Due to the differences in the domains of this dataset, training/evaluation of each domain has been done separately. Table 3 shows the results for each dataset in separate rows.

In the following, a comparison is made between the results of the proposed model for English and the existing methods (Table 4). However, before explaining the comparison table, it should be noted that the results of research in spam review detection are highly dependent on the dataset. For this reason, in Table 4, the results of other research are presented based on their datasets, and the best results obtained for each research are shown in this table.

As shown in Table 4, the performance of the proposed model for the OpSpam and

Table 5: The proposed model results for Persian (Digikala dataset)

Dataset	F1	Accuracy	Precision	Recall
Digikala	88.6	89.4	86.6	90.6

Table 6: Results comparison for Persian

Model	Year	Digikala			
		F1	Accuracy	Precision	Recall
FRRSD [32]	2019	82.4	83.3	-	82.4
SURSD [33]	2019	78.0	-	-	-
Proposed	-	**87.4**	**87.7**	**88.6**	**86.2**

Hotel domain of HDR datasets is better than the other methods, but for the Restaurant and Doctor domains of (HDR) dataset, the proposed model performs worse than the other methods. One reason for this difference is the size of datasets. The OpSpam and Hotel (HDR) datasets are larger than the Restaurant (HDR) and Doctor (HDR) datasets. Therefore, it can be said that according to the obtained results, the performance of the proposed model is better on larger datasets and does not perform well on a small dataset. According to the results in Table 4 for the Restaurant (HDR) and Doctor (HDR) datasets, none of the models is better than the other in all metrics.

4.3 Result comparison for Persian

As mentioned in this study, the proposed model on a Persian dataset was also trained and evaluated. Since not much research has been done for Persian on this subject, there are not many datasets to detect spam reviews. The only dataset used by researchers in this field is the Digikala dataset (Digikala.com, the largest retail website in Iran). This research has used this dataset to train and evaluate the model. Table 5 shows the results obtained by the proposed model for the Digikala dataset.

The following compares the results of the proposed model and existing methods in the field of spam review detection for Persian (Table 6). Not much research has been done to detect spam reviews for Persian, and existing methods have used traditional machine learning methods. In all methods presented in this table, the Digikala dataset has been used.

There is a big difference between the performance of the proposed model and other methods of detecting spam for Persian, and the proposed model has a better performance than other methods. The main reason for this difference in performance is that other methods (FRRSD, SURSD) use traditional machine learning methods to detect spam reviews. Although metadata is also used in these methods, their performance is significantly lower than the proposed method. This indicates that the use of metadata does not increase efficiency and, in some cases, may reduce model performance due to challenges such as singleton spam reviews or group spamming.

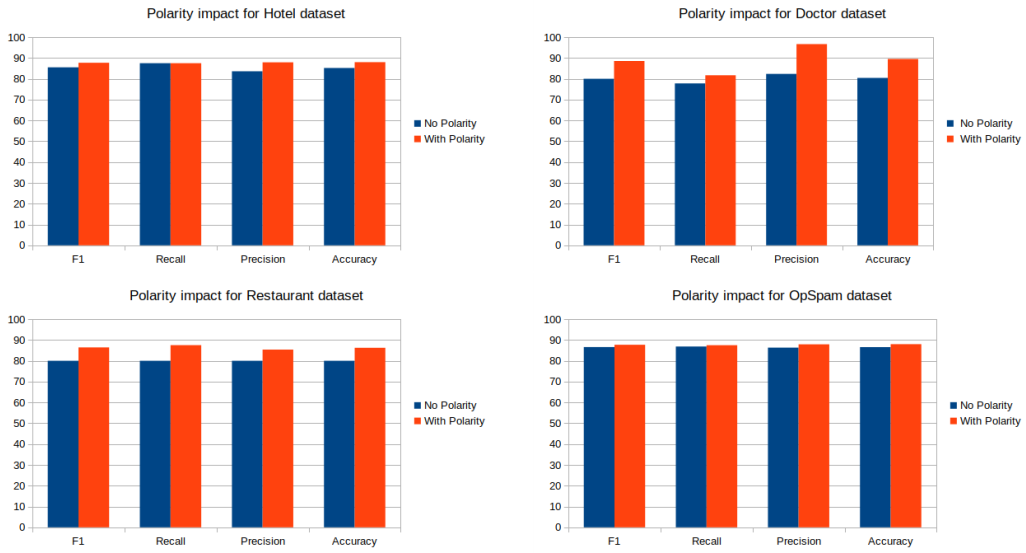

Figure 4: Impact of using opinion polarity in proposed model for different dataset

4.4 Impact of each Technique

This section examines the impact of "data balancing" and "review polarity" used in the proposed model.

4.4.1 Using review polarity

To determine the effect of using the polarity of review in the proposed model to detect spam reviews, the model is trained once without using polarity and once using polarity, and the results can be seen in Figure 4. As shown in this figure, the impact of using the polarity of reviews is considerable.

4.4.2 Data balancing

Data balancing can increase the model's efficiency because if the data set is balanced, the model will be trained equally on each class. In this section, the impact of using data balancing is examined. The OpSpam dataset is balanced, so there is no need to use balancing, but the Doctor (HDR) and Restaurant (HDR) datasets are not balanced and need to be balanced. As explained, this study used OverSampling to balance the data. This section shows the impact of using data balancing (Figure 5). As shown in Figure 5, the use of data balancing in unbalanced domains of the HDR dataset (Doctor and Restaurant domains) has significantly impacted model performance.

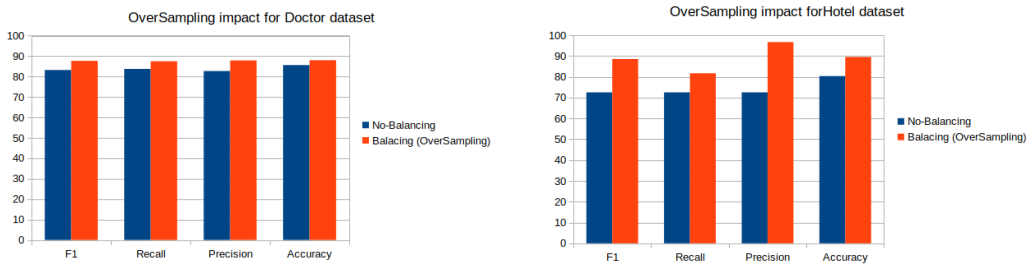

Figure 5: Impact of oversampling for imbalances datasets

5 Conclusion and future works

This research aims to provide a model using deep learning to detect spam opinions, in which only the text of the comments and their labels are used for training. Since in the comment text, the first and last sentences are more important than the middle context, in the proposed model, the comment text is divided into three parts, the first sentence, the middle context, and the last sentence, and each of these sections is given to a BiLSTM model. The entire comment text is also given to a BiLSTM. Using a sentiment analysis model, the polarity of opinions is also extracted in the absence and given to the spam review detection model. Finally, the output of the four BiLSTMs and the polarity of the review are concatenated together to form a vector. This vector is then given to a fully connected (FC) network, generating the final label. Various techniques such as balancing, using the self-attention mechanism, etc., have been used to increase the efficiency of the proposed model.

The proposed model for Persian and English languages has been trained and evaluated in this research. For English, two datasets, OpSpam and HDR, were used. Comparing the proposed model with similar methods shows that the proposed model performs better than similar works for the OpSpam dataset and Hotel domain of the HDR dataset. Although performance enhancement is minor in the proposed model, it is worth noting that only the text of the comments was used for learning in this study, and no metadata was used. For Persian, considering that the research done so far has all used traditional machine learning methods, the performance of the proposed model was much better compared to them. Although model performance is currently acceptable, some points can improve model performance and be referred to as future work. One of these tasks is to use algorithms to find the optimal value of hyper-parameters or use meta-learning. It is also possible to augment the data by translating the comment into different languages and then returning it to the original language and using it to balance the data.

References

[1] Alessandro Bondielli, Francesco Marcelloni, A survey on fake news and rumour detection techniques, Information Sciences, Volume 497, 2019, Pages 38-55, ISSN 0020-0255,

https://doi.org/10.1016/j.ins.2019.05.035.

[2] Li, Fangtao & Huang, Minlie & Yang, Yi & Zhu, Xiaoyan. (2011). Learning to Identify Review Spam. IJCAI Proceedings-International Joint Conference on Artificial Intelligence. 2488-2493. 10.5591/978-1-57735-516-8/IJCAI11-414.

[3] Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Spotting opinion spammers using behavioral footprints. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13). Association for Computing Machinery, New York, NY, USA, 632–640. DOI:https://doi.org/10.1145/2487575.2487580

[4] Raga S. H. Istanto, Wayan Firdaus Mahmudy, and Fitra A. Bachtiar. 2020. Detection of online review spam: a literature review. In Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology (SIET '20). Association for Computing Machinery, New York, NY, USA, 57–63. DOI:https://doi.org/10.1145/3427423.3427434

[5] Saumya, Sunil & Singh, Jyoti. (2022). Spam review detection using LSTM autoencoder: an unsupervised approach. Electronic Commerce Research. 22. 10.1007/s10660-020-09413-4.

[6] Alexander Ligthart, Cagatay Catal, Bedir Tekinerdogan,Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification, Applied Soft Computing, Volume 101, 2021, 107023, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2020.107023.

[7] N. Jindal and B. Liu, "Analyzing and Detecting Review Spam", Seventh IEEE International Conference on Data Mining (ICDM 2007), 2007, pp. 547-552, doi: 10.1109/ICDM.2007.68.

[8] Nitin Jindal and Bing Liu. 2007. Review spam detection. In Proceedings of the 16th international conference on World Wide Web (WWW '07). Association for Computing Machinery, New York, NY, USA, 1189–1190. DOI:https://doi.org/10.1145/1242572.1242759

[9] Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08). Association for Computing Machinery, New York, NY, USA, 219–230. DOI:https://doi.org/10.1145/1341531.1341560

[10] Zeng, Zhi-Yuan & Lin, Jyun-Jie & Chen, Mu-Sheng & Chen, Zorro & Lan, Yan-Qi & Liu, Jun-Lin. (2019). A Review Structure Based Ensemble Model for Deceptive Review Spam. Information. 10. 243. 10.3390/info10070243.

[11] M. Ott, Y. Choi, C. Cardie, and J.T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

[12] Wael Etaiwi, Ghazi Naymat, The Impact of applying Different Preprocessing Steps on Review Spam Detection, Procedia Computer Science, Volume 113, 2017, Pages 273-279, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2017.08.368. (https://www.sciencedirect.com/science/article/pii/S1877050917317787)

[13] Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. 2014. Towards a General Rule for Identifying Deceptive Opinion Spam. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1566–1576, Baltimore, Maryland. Association for Computational Linguistics.

[14] Y. Ren and D. Ji, "Learning to Detect Deceptive Opinion Spam: A Survey", in IEEE Access, vol. 7, pp. 42934-42945, 2019, doi: 10.1109/ACCESS.2019.2908495.

[15] Li, L., Ren, W., Qin, B., & Liu, T. (2015). Learning Document Representation for Deceptive Opinion Spam Detection. CCL.

[16] Yafeng Ren and Yue Zhang. 2016. Deceptive Opinion Spam Detection Using Neural Network. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 140–150, Osaka, Japan. The COLING 2016 Organizing Committee.

[17] Zhao, Siyuan & Xu, Zhiwei & Liu, Limin & Guo, Mengjie. (2018). Towards Accurate Deceptive Opinion Spam Detection based on Word Order-preserving CNN. Mathematical Problems in Engineering. 2018. 10.1155/2018/2410206.

[18] Jain N., Kumar A., Singh S., Singh C., Tripathi S. (2019) Deceptive Reviews Detection Using Deep Learning Techniques. In: Métais E., Meziane F., Vadera S., Sugumaran V., Saraee M. (eds) Natural Language Processing and Information Systems. NLDB 2019. Lecture Notes in Computer Science, vol 11608. Springer, Cham. https://doi.org/10.1007/978-3-030-23281-8_7

[19] Barushka A., Hajek P. (2019) Review Spam Detection Using Word Embeddings and Deep Neural Networks. In: MacIntyre J., Maglogiannis I., Iliadis L., Pimenidis E. (eds) Artificial Intelligence Applications and Innovations. AIAI 2019. IFIP Advances in Information and Communication Technology, vol 559. Springer, Cham. https://doi.org/10.1007/978-3-030-19823-7_28

[20] Safarian Neshat, Basiri Mohammad Ehsan, KHOSRAVI HADI. Feature ranking for Persian Spam Review detection. JOURNAL OF SOFT COMPUTING AND INFORMATION TECHNOLOGY (JSCIT). 2019 [cited 2022March12];8(2):1-16. Available from: https://www.sid.ir/en/journal/ViewPaper.aspx?id=745032

[21] Hernández-Castañeda, Ángel & Calvo, Hiram & Gambino, Omar. (2018). Impact of polarity in deception detection. Journal of Intelligent & Fuzzy Systems. 35. 1-10. 10.3233/JIFS-169610.

[22] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

[23] Bhargava, Rupal, Baoni, Anushka, and Sharma, Yashvardhan. Composite sequential modeling for identifying fake reviews. Journal of Intelligent Systems, 28, 04 2018.

[24] Hernández-Castañeda, Ángel, Calvo, Hiram, and Gambino, Omar. Impact of polarity in deception detection. Journal of Intelligent & Fuzzy Systems, 35:1–10, 07 2018.

[25] Zeng, Zhi-Yuan, Lin, Jyun-Jie, Chen, Mu-Sheng, Chen, Zorro, Lan, Yan-Qi, and Liu, Jun-Lin. A review structure based ensemble model for deceptive review spam. Information, 10:243, 07 2019.

[26] Ye Wang, Bixin Liu, Hongjia Wu Shan Zhao Zhiping Cai, Donghui Li Cheang Chak Fong. An opinion spam detection method ased on multi-filters convolutional neural network. Computers, Materials & Continua, 65(1):355–367, 2020.

[27] Anass, Fahfouh, Jamal, Riffi, Mahraz, Mohamed Adnane, Ali, Yahyaouy, and Tairi, Hamid. Deceptive opinion spam based on deep learning. In 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS), pages 1–5, 2020.

[28] P Mahalakshmi, Varri Sampreeth, Challa Venkataramana. Detection of opinion spam using lstm networks. 29:67 – 75, Apr. 2020.

[29] Neisari, Ashraf, Rueda, Luis, and Saad, Sherif. Spam review detection using self-organizing maps and convolutional neural networks. Computers & Security, 106:102274, 2021.

[30] Liu, Yuxin, Wang, Li, Shi, Tengfei, and Li, Jinyan. Detection of spam reviews through a hierarchical attention architecture with n-gram cnn and bi-lstm. Information Systems, page 101865, 2021.

[31] Safarian Neshat, Basiri Mohammad Ehsan, KHOSRAVI HADI. Feature ranking for Persian Spam Review detection. JOURNAL OF SOFT COMPUTING AND INFORMATION TECHNOLOGY (JSCIT). 2019 [cited 2022March12];8(2):1-16. Available from: https://www.sid.ir/en/journal/ViewPaper.aspx?id=745032

[32] M. E. Basiri, N. Safarian and H. K. Farsani, "A Supervised Framework for Review Spam Detection in the Persian Language", 2019 5th International Conference on Web Research (ICWR), 2019, pp. 203-207, doi: 10.1109/ICWR.2019.8765275.