applied sciences

MDPI

Article

Fake Sentence Detection Based on Transfer Learning: Applying to Korean COVID-19 Fake News

Jeong-Wook Lee and Jae-Hoon Kim *

Department of Computer Engineering and Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, Busan 49112, Korea; wjddnr96177@naver.com
* Correspondence: jhoon@kmou.ac.kr; Tel.: +82-051-410-4896

Abstract: With the increasing number of social media users in recent years, news in various fields, such as politics, economics, and so on, can be easily accessed by users. However, most news spread through social networks including Twitter, Facebook, and Instagram has unknown sources, thus having a significant impact on news consumers. Fake news on COVID-19, which is affecting the global population, is propagating quickly and causes social disorder. Thus, a lot of research is being conducted on the detection of fake news on COVID-19 but is facing the problem of a lack of datasets. In order to alleviate the problem, we built a dataset on COVID-19 fake news from fact-checking websites in Korea and propose deep learning for detecting fake news on COVID-19 using the datasets. The proposed model is pre-trained with large-scale data and then performs transfer learning through a BiLSTM model. Moreover, we propose a method for initializing the hidden and cell states of the BiLSTM model to a [CLS] token instead of a zero vector. Through experiments, the proposed model showed that the accuracy is 78.8%, which was improved by 8% compared with the linear model as a baseline model, and that transfer learning can be useful with a small amount of data as we know it. A [CLS] token containing sentence information as the initial state of the BiLSTM can contribute to a performance improvement in the model.

Keywords: COVID-19; fake news; fake news detection; transfer learning; KoCharELECTRA

Citation: Lee, J.-W.; Kim, J.-H. Fake Sentence Detection Based on Transfer Learning: Applying to Korean COVID-19 Fake News. *Appl. Sci.* **2022**, *12*, 6402. https://doi.org/10.3390/app12136402

Academic Editor: Antonio Fernández-Caballero

Received: 18 May 2022
Accepted: 20 June 2022
Published: 23 June 2022

1. Introduction

Users encounter news of various fields through a variety of platforms due to the development of social media. Unlike how the press and TV networks used to dominate news broadcasting, users of social media can directly participate in the news while the users quickly propagate the news due to the high accessibility and speed. Furthermore, fake news is propagated more quickly. The propagation of fake news inflicts significant damage to news consumers [1]. Individuals who produce fake news have been producing fake news for various purposes, such as political purposes, economic benefits, and so on. It is difficult to detect such fake news because they are generally written in exactly the same format as true news [2]. Accordingly, countries such as the United States and the United Kingdom have been paying attention to research on fake news detection [3], and fact-checking websites have emerged as a part of such research efforts [4].

The authenticity of the news is judged through fact-checking websites by verifying through the press. There are several fact-checking websites being operated overseas, but it is very insufficient compared to the amount of fake news being created. In spite of the insufficiency, research is ongoing on fake news detection using the data provided by fact-checking websites [5,6].

On the other hand, in Korea, about 10 fact-checking websites, including SNU FactCheck (https://factcheck.snu.ac.kr/ (accessed on 15 February 2022)) and FactChecknet (https://factcheck.snu.ac.kr/ (accessed on 15 February 2022)), are operated, and research on fake news detection has a serious problem of a lack of data because the number of verified

fact-checking websites is fewer and the amount of usable data is also inadequate. Hence, the frequently used fake news detection model is based on the term frequency-inverse document frequency (TF-IDF) [7]. The similarity of documents is checked using the frequency of words appearing in a document, which has the limitation of being incapable of reflecting the contextual information of sentences. Afterwards, various fake news detection models reflecting the contextual information of a document based on embedding techniques have been proposed [8,9].

The performance of fake news detection has drastically improved recently due to the development of language models such as BERT, which undergoes pre-training based on large-scale data [10]. Due to advancements in BERT, most research on fake news detection was conducted by embedding entire documents [11]. When embedding the entire document of fake news, however, discriminating fake news from real news becomes challenging because fake news maliciously includes truthful sentences [12]. In addition, embedding the entire document leads to performance degradation of a model since fake news is created so as to appear as real as possible. In particular, fake news detection in a specific category may not demonstrate a performance improvement in the BERT model due to insufficient data.

With the recent emergence of COVID-19, fake news on COVID-19 is spreading faster than real news on social media and in a few press outlets. Fake news starting with shocking headlines can lead to serious results since infections, in particular, are directly associated with human lives [13]. Research on detecting fake news on COVID-19 is actively conducted overseas since there is an ample amount of usable data amid the COVID-19 outbreak [14,15]. However, research is less actively conducted in Korea due to a lack of useful data. This paper, therefore, builds a dataset on COVID-19 fake news collected from multiple fact-checking websites and proposes a model for detecting fake sentences related to COVID-19. Embedding proceeds with KoCharELECTRA, which demonstrates excellent performance in embedding a sentence from the constructed dataset instead of embedding the entire document. Detecting fake news on COVID-19 is recognized as a binary classification problem, and the performance is compared using two models: fully connected neural network (FCNN) and bidirectional LSTM (BiLSTM).

Furthermore, a context vector, which is the contextual information of the encoder–decoder model, is judged to be similar to a [CLS] token, which is the contextual information of KoCharELECTRA; a [CLS] token is assigned as an input for the hidden initial state of BiLSTM for improving the model's performance. Consequently, setting a [CLS] token as the initial hidden and cell states of the BiLSTM model resulted in the highest accuracy of 78.8%.

The remainder of this paper is organized as follows: Related works on fake news detection are introduced in Section 2, while the method for constructing a Korean COVID-19 dataset from fact-checking websites is explained in Section 3. In Section 4, a deep learning model using the Korean COVID-19 dataset is described. In Section 5, the performance of the models is compared through experiments, and the models are analyzed by adjusting the hidden and cell states of the BiLSTM. Section 6 draws some conclusions and discusses future studies.

2. Related Works

In this chapter, we describe related works on the previous research on fake news detection and then briefly introduce Korean sentence embedding and transfer learning.

2.1. Previous Research on Fake News Detection

Many efforts are made to detect fake news in general since the damage caused by fake news is sharply rising [16]. Previous studies conducted overseas focused on detecting fake news by extracting different qualities of documents using traditional machine learning methods such as decision tree or SVM-based models [17,18]. In another study, a graph-based model was used for inferring the relationship between content and users who share

news [19]. Big tech companies such as Twitter and Facebook substantially improved the performance of fake news detection models by using deep learning algorithms [20]. Several studies are being conducted in the field of fake news detection in Korea as well. As efforts to prevent using fake news for political purposes continue, the National Election Commission in Korea has been operating a "TF team dedicated to slander/black propaganda", while the Ministry of Science and ICT in Korea has been researching fake news detection using artificial intelligence (AI) technology by holding an "AI R&D Challenge". Furthermore, the operation of non-governmental fact-checking websites has enabled the building of fake news detection data. Certain studies have refined data and applied them to machine learning and deep learning [21,22].

Fake news detection has been studied in diverse news categories. Due to the lack of appropriate data, the detection of fake news in specific categories has been performed based on news articles written in English. In particular, the detection of fake news related to COVID-19 is actively researched overseas. Shahi and Nandini (2020) constructed a COVID-19 dataset using a BERT-based model by collecting articles from fact-checking sites [23]. Al-Rakhami and Al-Amri (2020) extracted handcrafted features for fake news detection in order to detect fake news related to COVID-19 that is propagated through Twitter, which is a social media network [24]. Despite the prolonged COVID-19 crisis over the past several years, research on the detection of fake news related to COVID-19 has not been properly conducted in Korea due to insufficient data on news articles on COVID-19 written in Korean.

2.2. BERT-Based Korean Embedding

Fake news detection in Korea mostly involves the models that use TF-IDF, which identifies the nature of a document using the frequency of words [25]. TF-IDF has the disadvantage of not being able to grasp the context of sentences, so to solve this problem, a method of detecting fake news with FastText embedding methods based on self-collected datasets has been proposed [26]. Another study enhanced the performance of fake news through Word2Vec using a model that has been trained with the content of news articles [27]. In recent years, research is actively conducted on fake news detection using BERT-based models, which excel in identifying context by having been pre-trained with large-scale data [28].

BERT, which was released by Google in 2018, demonstrates excellent performance in various tasks as a pre-trained model. The pre-training of BERT involves randomly masking words using a masked language model, which then predicts the masked words, and applying the next sentence prediction to determine whether two sentences are connected. The models pre-trained with large-scale data of the problem being solved are fine-tuned and applied to various tasks. KoBERT [29] demonstrates excellent performance in the problems in Korea, as it was trained with an SKT large-scale Korean Wiki corpus.

Unlike KoBERT, KoELECTRA is a Korean pre-trained language model (PLM) based on ELECTRA [30]. ELECTRA consists of a generator and a discriminator capable of checking bidirectional context information. After converting the words of a specific ratio to a [MASK] token, a generator is induced to generate a word suitable for the [MASK]. In this process, a discriminator is trained by judging which token has been replaced based on the output of a generator. ELECTRA is more efficient, and the training speed is faster than BERT since training is applied for all tokens. KoELECTRA is a model that has been trained with around 34 GB of Korean sentences, including news, Korean Wiki, and NIKL Corpus (https://factcheck.snu.ac.kr/ (accessed on 15 February 2022)), in which it demonstrated an accuracy of 90.6% in the NSMC (https://factcheck.snu.ac.kr/ (accessed on 15 February 2022)) task, outperforming KoBERT, which demonstrated 89.5% accuracy [31]. In this paper, therefore, ELECTRA, demonstrating the best performance among the BERT-based models, was selected, and sentences were embedded using KoCharELECTRA for which a large-scale Korean corpus has been pre-trained in the syllable unit.

2.3. Transfer Learning

In recent research on fake news detection, deep learning-based techniques are gaining popularity to overcome the limitations of the extensive amount of time and cost required for verification by experts and that the subjectivity of the verifier may be introduced. The most significant element of a fake new detection model using deep learning is the amount of data. The research performance is outstanding when datasets of all categories with a large amount of data are used for deep learning models. However, the accuracy may be degraded due to insufficient data when the problem of a specific category, such as COVID-19 fake news detection, is being solved. There is lack of relevant research since the collection of COVID-19-related data is extremely limited in reality, and the authenticity of the data is not verified. Transfer learning can overcome the limitation of insufficient data [32]. Transfer learning refers to pre-training a model with related large-scale data and then reusing small-scale data for solving the main problem.

This paper proposes a model that has been trained through transfer learning based on small-scale data related to COVID-19. First, a fake news detection model is generated using a large-scale fake news dataset. Second, the validity of the fake news detection model is verified through transfer learning based on the COVID-19 data collected from fact-checking websites. Third, the performance is compared between the proposed model and the model for which the hidden state and cell state are modified using a small COVID-19 dataset.

3. Constructing Datasets

As mentioned above, fake news on COVID-19 is propagating quickly and causes social disorder in life. Thus, research is being conducted on the detection of fake news on COVID-19 but is facing the problem of a lack of datasets. In particular, there is no public dataset on COVID-19 fake news in Korea. Therefore, we build a process of building datasets on Korean COVID-19 fake news from fact-checking websites in Korea in this chapter.

Figure 1 shows the process of collecting documents from fact-checking websites and constructing datasets. For building a reliable dataset on COVID-19 in Korea, data are collected from the Korea Disease Control and Prevention Agency (KDCA), which is a government agency, and other fact-checking websites such as SNU Factcheck, vaccine fact-check, and FactChecknet. KDCA performs fact-checking by operating a separate platform for COVID-19-related news and issues and by clarifying false reports. Furthermore, details and links of fake news on COVID-19 are provided and relevant fact-checking materials of KDCA are suggested. Non-governmental fact-checking websites, including SNU Factcheck, vaccine fact-check, and FactChecknet, verify fake news of various news outlets and propose six types of verification results, including false, true, and deferred judgment. For constructing datasets correctly, only the news that has been verified as “true” and “false” are used as the data. Raw data that have been crawled from each site are saved as an xlsx file. The documents that are overlapped or unrelated to COVID-19 are removed from the saved raw data, and the formats of the data obtained from different platforms are adjusted to be uniform, as shown in Figure 1.

For data consistency, news data extracted from websites are configured with an index, article title, main content, source, and validation result, as shown in Table 1.

The COVID-19 news dataset collected in the Korean language consists of a total of 2500 fact-checking news entries and 3500 main sentences. For data balance, the ratio of fake data to true data was set to 1:1 in which 1500 sentences are fake data and another 1500 sentences are true data.

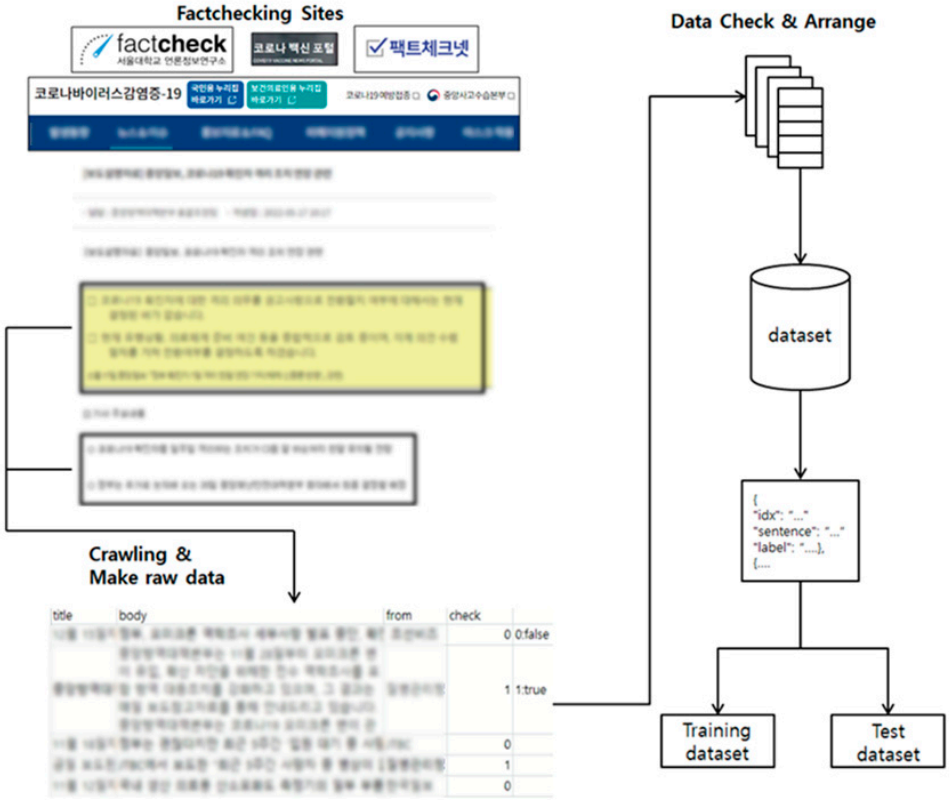

Figure 1. Dataset construction process from fact-checking websites.

Table 1. An example of datasets extracted from fact-checking websites.

Category	Description	Content
Index	Sequence No.	1
Title	The title of a document	3월 22일 SBS, 연합뉴스, 뉴시스 "4차 백신...수십만 명분 폐기 불가피" (Translated in English: March 22, SBS, Yonhap News, Newsis "Fourth vaccine dose... Disposal of vaccines for hundreds of thousands inevitable")
Body	The main content of a document	전국 요양병원과 시설에 공급된 4차 백신이 남아돌고 있음. 접종 대상자들이 잇따라 확진 판정을 받거나, 접종을 꺼리기 때문인데, 이번 주 안에 수십만 명분의 백신이 폐기될 것으로 보임 (Translated in English: Excessive fourth vaccine doses have been provided to nursing hospitals and other institutions. Individuals subject to vaccination have recently tested positive for COVID-19 or are reluctant to get vaccinated, which will result in the disposal of vaccines for hundreds of thousands of people this week.)
Speaker	The source of a document	SBS, Yonhap News, Newsis
Veracity	The validation of a document	Fake

In order to solve the problem of insufficient data, the data of FactChecknet, which was supported by SNU Factcheck Center are used. A total of 80,000 sentences of AI fact-check data provided by NEWSTOF (https://factcheck.snu.ac.kr/ (accessed on 15 February 2022)) for fake news detection research were used as training data, and transfer learning proceeds using the collected COVID-19 dataset. For data consistency, the data structure is configured in a JSON format, as shown in Figure 2. For embedding sentences instead of embedding the entire document in the constructed dataset, the dataset is configured to include sentence order idx, sentence, and sentence label of either fake or true in a dictionary format.

```
[
 {
  "idx": "1",
  "sentence": "Excessive fourth vaccine doses have been provided to nursing hospitals and other institutions.",
  "label": "Fake"
 },
 {
  "idx": "2",
   "sentence": "Individuals subject to vaccination have recently been tested positive for COVID-19 or are reluctant
    to get vaccinated, which will result in disposal of vaccines for hundreds of thousands of people this week.",
  "label": "Fake"
 }
]
```


Figure 2. A part of extracted datasets in a JSON format.

Training data and test data are configured as shown in Table 2 for training the model. For training the model, a total of 51,290 sentences with a 1:1 ratio of true to false data are used among 70,000 sentences of the training data of AI FactChecknet. The collected COVID-19 dataset is configured with a 1:1 ratio of true to false data for transfer learning from which a total of 2500 sentences are used. For evaluating the model, a total of 3332 sentences with a 1:1 ratio of true to false data among 10,000 sentences from AI FactChecknet are used in addition to 500 sentences from the collected COVID-19 dataset.

Table 2. Dataset configurations.

Source	Datasets	No. of Sentences
AI FactChecknet	Training	51,290
COVID		2500
AI FactChecknet	Test	3332
COVID		500

4. Fake Sentence Detection Based on Deep Learning

Previously, research on fake news detection mostly took the approach of a binary classification problem using linear models. Then, a performance improvement was achieved by using RNN and CNN models [33]. This paper uses BiLSTM, which resolved the issue of the vanishing gradient of RNN and long-term dependency, which is the problem for which the desired output depends on inputs presented at times far in the past. The overall structure of the model is shown in Figure 3. The model consists of an input layer, embedding layer, contextual layer, and classification layer. Sentences are tokenized in the input layer by syllable unit to be used as an input of the embedding layer.

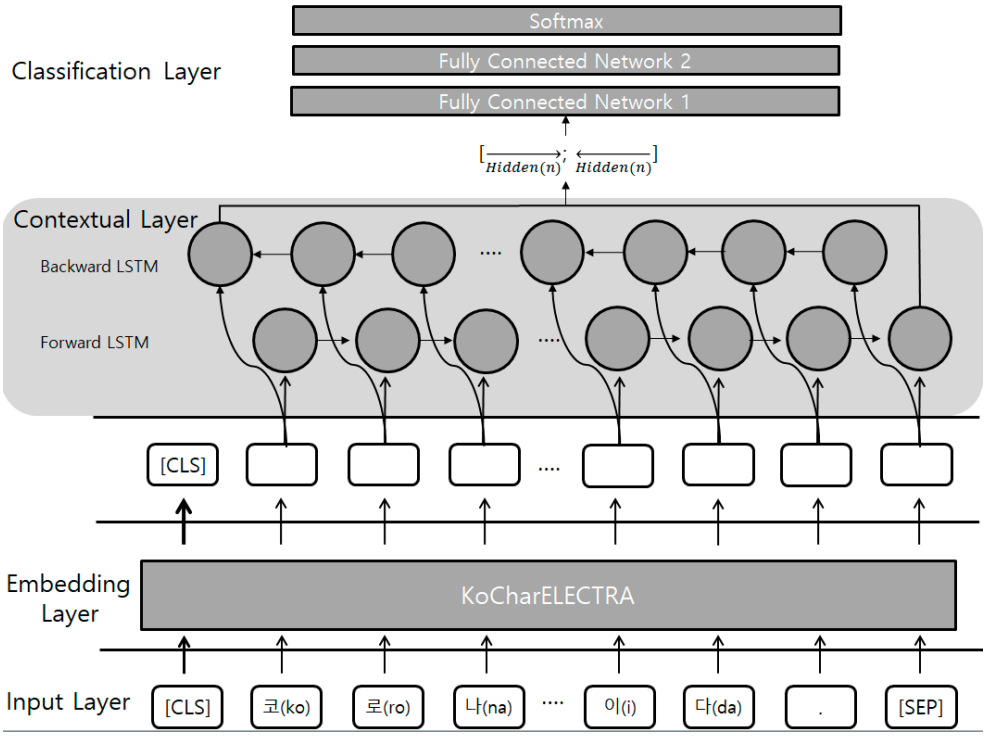

Figure 3. The proposed model in detail.

4.1. Embedding Layer

The embedding layer uses KoCharELECTRA, as described in Section 2.2. KoCharELECTRA is a Korean ELECTRA model that has been trained with a character-level tokenizer instead of a WordPiece-level tokenizer. A KoCharELECTRA-based model has a maximum length of 512, and the vocab size is 11,568, in which Chinese characters are excluded not only in vacab but also during pre-processing. The dimension of an output vector is expressed as 768 dimensions in the base model. Sentences are tokenized in the input layer by syllable unit using the KoCharELECTRA tokenizer to be used as an input of the embedding layer. For example, a sentence "코로나는 독감이다. (Coronavirus is an influenza)" is tokenized into "[CLS] 코(ko) 로(ro) 나(na) 는(neun) 독(dok) 감(gam) 이(i) 다(da). [SEP]" at the syllable level. Tokens that are tokenized by syllable are converted into integers of KoCharELECTRA Vocab for training the model. Subsequently, integers that have been tokenized by syllable in the embedding layer are output as pre-trained vector values.

4.2. Contextual Layer

A vector in syllables generated from the embedding layer is bi-directionally trained by being sequentially input in BiLSTM. A [CLS] containing the contextual information in the sentence is not used as an input of BiLSTM in the contextual layer. The last hidden state generated from the forward LSTM and the last hidden state generated from the backward LSTM are concatenated to be delivered as an input for the classification layer.

BiLSTM

For comparing the performance of the proposed model, a comparative model was designed by adjusting the initial hidden and cell states of BiLSTM. The internal structure of BiLSTM is shown in Figure 4. BiLSTM renews the cell state using the input gate for inputting the current cell and the forget gate for removing the information of the past cell. The initial hidden state and input vector are concatenated to determine whether to leave the memory using the sigmoid function and determine new information to remember using the sigmoid function and tanh function.

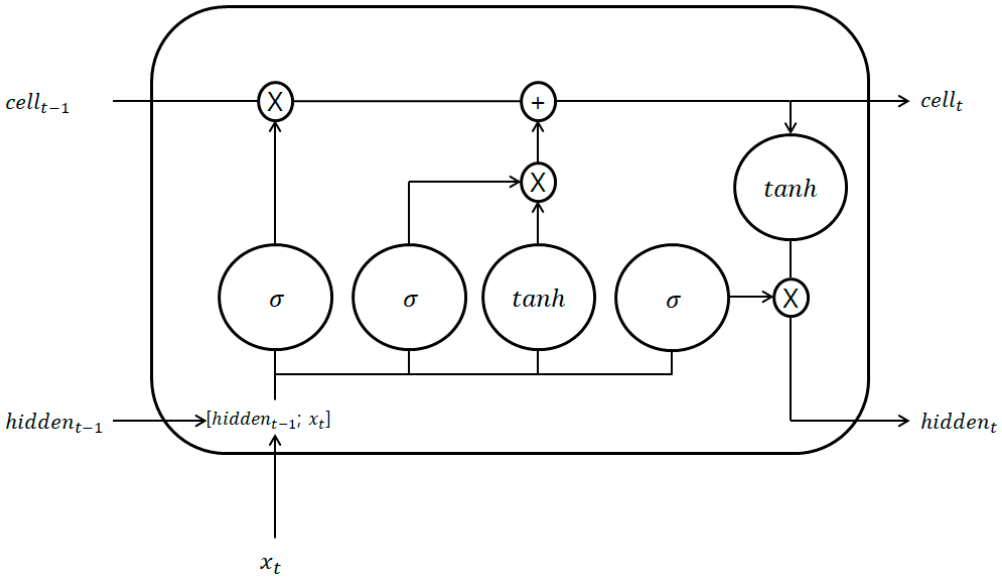

Figure 4. The structure of BiLSTM.

BiLSTM [34] learns not only the relationship between sequentially input values and previous values but also the relationship with subsequent values by combining forward LSTM and backward LSTM. The greatest difference between BiLSTM and previous RNN models is the cell state. In general, the initial states of cell and hidden are set to a zero vector or a random vector when using a BiLSTM model. In this paper, a [CLS] token is used as the initial hidden and cell states, considering the correlation between the context vector containing contextual information generated from the encoder in the encoder–decoder model and the [CLS] token reflecting the contextual information. Figure 5 shows the structure of the Seq2Seq model [35]. The last hidden state in the encoder consisting of LSTM becomes a context vector. The generated context vector is used as the hidden initial state of the LSTM in the decoder.

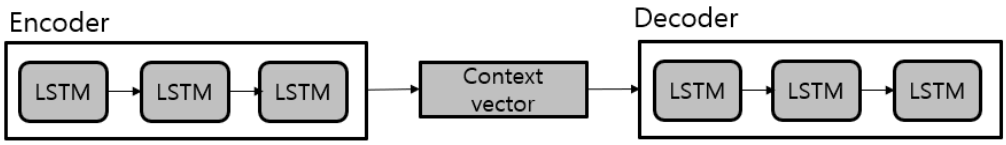

Figure 5. The structure of Seq2Seq.

Figure 6 shows the model in which the initial states of cell and hidden are used as the [CLS] token of KoCharELECTRA.

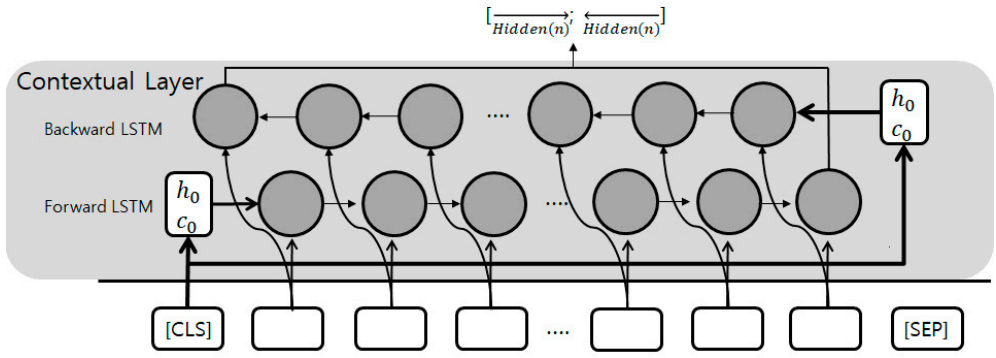

Figure 6. The contextual layer with hidden and cell states initialized to [CLS].

4.3. Classification Layer

The classification layer consists of a total of three layers. The concatenation of the forward and backward hidden states, $\left[\overrightarrow{Hidden(n)}; \overleftarrow{Hidden(n)}\right]$, is used as an input of FCNN and is configured with two layers. A softmax function is used for the binary classification of the hidden state in the last layer. Equation (1) represents the softmax function. The range of output values that have passed through the softmax function is between 0 and 1, while the sum of all probabilities is equal to 1. Through one-hot encoding, the label with the highest probability among the softmax probability values is set to 1, while the rest are set to 0.

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}} \quad (1)$$

Cross-entropy is used as a loss function. Equation (2) represents the cross-entropy, where S is the predicted value of the model and L is the actual value of the label. The model is trained by minimizing the loss based on the comparison of the distance of the probability distributions between the predicted value and actual value.

$$\mathbf{D}(\mathbf{S}, \mathbf{L}) = -\sum_{\mathbf{i}} \mathbf{L_i} \cdot \mathbf{log}(\mathbf{S_i}) \quad (2)$$

5. Experiments and Evaluation

In this chapter, we evaluate the proposed model and interpret its improvements both quantitatively and qualitatively. We set a linear model with the input of a [CLS] token as a base model and compare with the model that initialized the hidden and cell states of the BiLSTM model to a [CLS] token instead of a zero vector.

5.1. Dataset Configuration

Table 3 shows dataset configuration for transfer learning. Experiments were conducted for examining the validity of the proposed model when large-scale fake news data were learned, and transfer learning was performed with the collected COVID-19 dataset. A total of four experiments were conducted in which AI FactChecknet training data and AI FactChecknet test data were used in test 1-1. In test 1-2, AI FactChecknet training data were used for training, while a small amount of COVID-19 test data were used for testing. In test 2, AI test data were additionally used for training to learn a greater amount of data, and COVID-19 test data were used for testing. For the validity of transfer learning in test 3, the collected COVID-19 data were applied to the test 2 model for transfer learning, and COVID-19 test data were used for testing.

Table 3. Dataset configurations for transfer learning.

Datasets	Test 1-1	Test 1-2	Test 2	Test 3
Training	AI FactChecknet (training data)		+AI FactChecknet (test data)	+COVID (training data)
Test	AI FactChecknet (test data)	COVID (test data)	COVID (test data)	COVID (test data)

5.2. Performance Evaluation

Using the datasets described above, Table 4 shows the performance comparison of a linear model of a baseline model that only used a [CLS] token and the BiLSTM model that used syllable vectors. The linear model, which is most frequently used in binary classification problems, was configured with two layers of FCNN and softmax layers and only used a [CLS] token of KoCharELECTRA as an input for a linear layer. In the BiLSTM model, which is proposed in this paper, only syllable vectors were used as inputs without using a [CLS] token.

Table 4. Performance comparison of linear and BiLSTM models in accuracy.

Model	Test 1-1	Test 1-2	Test 2	Test 3
Linear	59.34	49.61	50.78	70.70
BiLSTM	60.54	50.78	52.21	74.51

Accuracy, which is most widely used in classification problems, is used as an evaluation metric. The accuracy represents the number of correctly predicted samples among all samples.

In test 1-1 where the model was trained with large-scale fake news data and the performance was evaluated using the AI FactChecknet test data, the accuracies of the linear model and BiLSTM were 59.34% and 60.54%, respectively, in which BiLSTM slightly outperformed.

Similarly, in test 1-2, the accuracy of the proposed model was higher when the performance was evaluated using the small amount of the collected COVID-19 dataset. In test 2, a higher accuracy was measured for both the linear and proposed models compared to test 1-2. This result indicates that performance improves when trained with a large amount of data. In addition, the performance was significantly improved when a greater amount of data was used in the BiLSTM model compared to the linear model.

Test 3 reflected the performance when transfer learning was performed with the COVID-19 dataset for the model in test 2. The accuracies of the linear model and the BiLSTM models were 70.70% and 74.51%, respectively, thus demonstrating a significant improvement in both models.

When the accuracies of test 1-2 and test 3 were compared, pre-training with large-scale data and then performing transfer learning with a small amount of data can contribute significantly to performance improvement. Furthermore, performance was more outstanding in the BiLSTM model, which only used syllable vectors, compared to the linear model, which used only a [CLS] token of KoCharELECTRA.

5.3. Performance Improvements

Table 5 presents the input of the classification layer and the initial values of the hidden and cell states in the proposed BiLSTM model. In model 1, initial hidden and cell states of BiLSTM were set to 0. For the input of the classification layer, the value obtained by concatenating the last hidden state of BiLSTM was used as in $\left[\underset{Hidden(n)}{\longrightarrow}; \underset{Hidden(n)}{\longleftarrow}\right]$. Similar to model 1, the initial hidden and cell states were set to 0 in model 2. Since a significant performance improvement was observed in the model that only used a [CLS] token, the last hidden state of BiLSTM was concatenated with a [CLS] token as the input of the classification layer as in $\left[\text{CLS}; \underset{Hidden(n)}{\longrightarrow}; \underset{Hidden(n)}{\longleftarrow}\right]$. Model 3 was configured to evaluate the performance of the case using a [CLS] token containing sentence information instead of 0 for the initial hidden and cell states. In model 3, the initial hidden state was set to a [CLS] token, while the initial cell state was set to 0. In model 4, the initial hidden and cell states were set to a [CLS] token to evaluate the performance.

Table 5. Inputs of the classification layer and initial states of BiLSTM.

Model	Input of the Classification Layer	Initial State of BiLSTM	
		Hidden	Cell
Model 1	$\left[\underset{Hidden(n)}{\longrightarrow}; \underset{Hidden(n)}{\longleftarrow}\right]$	0	0
Model 2	$\left[\text{CLS}; \underset{Hidden(n)}{\longrightarrow}; \underset{Hidden(n)}{\longleftarrow}\right]$	0	0
Model 3	$\left[\underset{Hidden(n)}{\longrightarrow}; \underset{Hidden(n)}{\longleftarrow}\right]$	[CLS]	0
Model 4	$\left[\underset{Hidden(n)}{\longrightarrow}; \underset{Hidden(n)}{\longleftarrow}\right]$	[CLS]	[CLS]

Table 6 shows the results of the experiments of the models described in Table 5. Accuracy was used as the performance evaluation metric as in the previous experiments.

Table 6. Performance of the proposed and comparative models.

Model	Test 1-1	Test 1-2	Test 2	Test 3
Model 1	60.54	50.78	52.21	74.51
Model 2	61.72	49.22	52.0	75.78
Model 3	61.7	53.52	53.12	75.61
Model 4	**64.2**	**55.86**	**55.86**	**78.80**

Compared to model 1, which is proposed in this paper, model 2 demonstrated approximately a 1.2% performance improvement in test 1-1 and test 3. Both models resulted in similar levels of accuracy in test 1-2 and test 2, for which transfer learning was not performed. When model 1 and model 2 are compared, concatenating a [CLS] token with the last hidden state of the BiLSTM can improve the performance. However, the speed was lower than the proposed model during training since the input size of the classification layer was enlarged by 1.5 times.

Similar to model 2, model 3 also demonstrated approximately a 1.2% performance improvement in test 1-1 and test 3. A similar level of accuracy to model 2 was observed in test 1-2 and test 2, for which transfer learning was not performed. However, the training speed was faster than in model 2 since the input size of the classification layer was identical to that of model 1. Model 4 shows the accuracy when both the hidden and cell states are set to a [CLS], which is the essential idea of the BiLSTM. The accuracy was improved by 4% to 78.52% in test 3, which was the highest when compared to the proposed model, and the performance was improved by 8% compared to the linear model. A similar level of accuracy improvement was observed in test 1-1 as well. The accuracy was 2.8% higher than with model 3, which implies that using a [CLS] token as the initial cell state of BiLSTM can drastically improve the performance of the model. Conducting the above experiments demonstrates that using a [CLS] containing sentence information is highly beneficial for improving the performance since the cell of BiLSTM plays the role of a memory for recalling previous values.

Additionally, we conducted an experiment to order to demonstrate the increasing performance in proportion to the size of the training datasets in Figure 7. The x-axis and the y-axis show the number of COVID-19 training datasets required for further learning after transfer learning and the accuracy of Model 4, respectively. In fact, this is very common in the field of machine learning, but in transfer learning with a small number of datasets, the increase takes effect without exception.

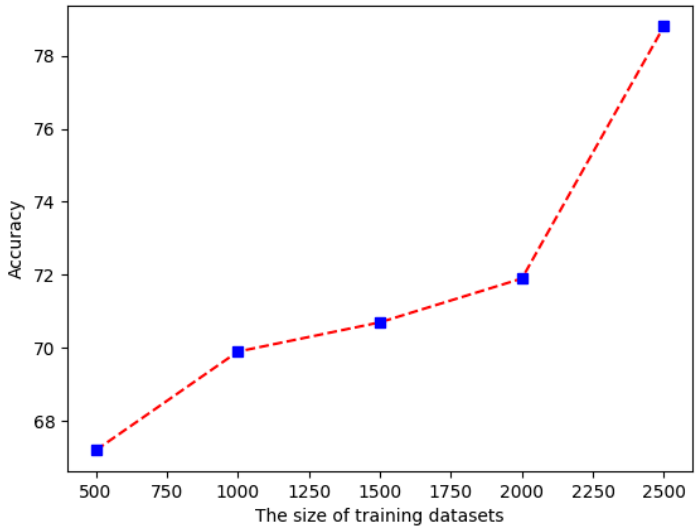

Figure 7. Increasing performance in proportion to the size of training datasets.

5.4. Cohen's Kappa

Cohen's kappa coefficient is used to measure the reliability of the model. Cohen's kappa coefficient is a statistical method for finding the reliability between evaluators by measuring the agreement of measurement category values between two observers [36]. Equation (3) shows the Cohen's kappa coefficient where $\mathbf{P_{observed}}$ is the probability of a match between evaluators and $\mathbf{P_{chance}}$ is the probability of a match by chance.

$$\kappa = \frac{\mathbf{P_{observed}} - \mathbf{P_{chance}}}{1 - \mathbf{P_{chance}}} \tag{3}$$

Cohen's kappa coefficient grades are shown in Table 7, which follow the interpretation of Landis and Koch [37].

Table 7. Kappa grades.

κ	Strength of Agreement
>0.000	Poor
0.000–0.200	Slight
0.201–0.400	Fair
0.401–0.600	Moderate
0.601–0.800	Substantial
0.801–1.000	Almost Perfect

κ, representing Cohen's kappa coefficient, has a value between 0 and 1. As a κ value becomes closer 1, the reliability of a model increases. A Cohen's kappa coefficient, κ, between 0.000 and 0.200 indicates slight agreement, while a value greater than or equal to 0.8 indicates almost perfect agreement.

A previous study also used Cohen's kappa coefficient to measure the reliability of a binary classification model [38]. The reliability of a model is determined by measuring the agreement between the label of the test data and the label predicted by the deep learning model. Figure 8 illustrates the result of generating a confusion matrix using the COVID-19 test data consisting of 500 sentences for measuring the reliability of the proposed model.

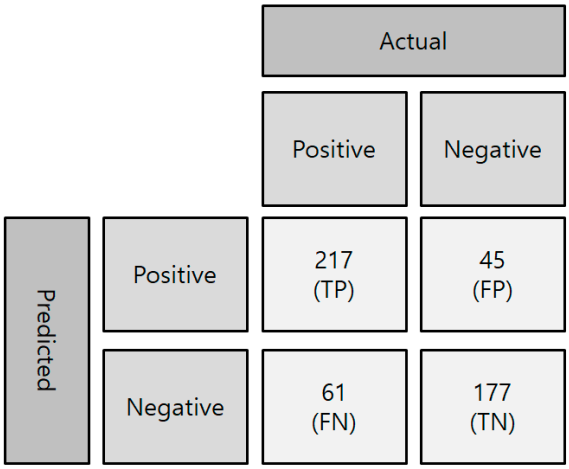

Figure 8. Confusion matrix for binary classification.

In the confusion matrix, four kinds of results can be taken as follows:

- True Positive (TP): a result that correctly indicates the positive;
- True Negative (TN): a result that correctly indicates the negative;
- False Positive (FP): a result that wrongly indicates the positive when in fact the result belongs to the negative;
- False Negative (FN): a result that wrongly indicates the negative when in fact the result belongs to the positive.

Through the confusion matrix, Cohen's kappa coefficient κ can be expressed as shown in Equation (4).

$$\kappa = \frac{2 \times (\mathbf{TP} \times \mathbf{TN} - \mathbf{FN} \times \mathbf{FP})}{(\mathbf{TP} + \mathbf{FP}) \times (\mathbf{FP} + \mathbf{TN}) + (\mathbf{TP} + \mathbf{FN}) \times (\mathbf{FN} + \mathbf{TN})} \quad (4)$$

In this paper, the experiment result produced κ = 0.5737 when calculated using Equation (4), which shows moderate agreement according to the interpretation of Landis and Koch.

5.5. Discussion

Unlike previously existing research, this paper proposed a fake sentence detection model for COVID-19. As mentioned in introduction, most previous studies were based documents and used document-level embeddings. Fake and real sentences, however, are mixed in fake news documents and would be represented in the embeddings of them. This seriously affects the performance of the model. We also solved the problem for a specific category called COVID-19, unlike the existing research. It is difficult to generalize news in all categories because there are limitations in distinguishing fake and real news, and the performance of the model varies from category to category (e.g., sports, politics, economics, and society). Therefore, in this paper, we constructed datasets and proposed the model for discriminating fake sentences about COVID-19. As a result, it showed a high performance for COVID-19 fake sentence detection.

6. Conclusions

This paper proposed a model for detecting Korean fake news about COVID-19. As mentioned in Section 3, the datasets for Korean fake news about COVID-19 were constructed by the researchers because of their non-availability. The proposed model classified sentences into fake or true instead of documents because all sentences in a document are not fake, only one or two sentences. Moreover, unlike a conventional linear model that only used a [CLS] token, we used the BiLSTM and initialized the hidden and cell states of a BiLSTM model to a [CLS] token instead of a zero vector in order to reflect syllables in a

sentence as well as a [CLS] token. Transfer learning was performed in order to effectively apply the COVID-19 data to the model due to the lack of datasets. Through experiments, the proposed model was shown to have an accuracy of 78.8%, which was improved by 8% compared with the linear model as a baseline model. However, the experiments showed that a greater amount of data results in a better performance for the BiLSTM model; thus, the model performance can be further improved by constructing a dataset with a large amount of labeled COVID-19 fake news data.

Author Contributions: Conceptualization, J.-W.L. and J.-H.K.; methodology, J.-W.L.; software, J.-W.L.; validation, J.-W.L. and J.-H.K.; formal analysis, J.-W.L.; investigation, J.-W.L.; resources, J.-W.L.; data curation, J.-W.L.; writing—original draft preparation, J.-W.L.; writing—review and editing, J.-H.K.; visualization, J.-W.L.; supervision, J.-H.K.; project administration, J.-H.K.; funding acquisition, J.-H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Akram, W.; Kumar, R. A study on positive and negative effects of social media on society. *Int. J. Comput. Sci. Eng.* **2017**, *5*, 351–354. [CrossRef]
2. Jwa, H.; Oh, D.; Lim, H. Research analysis in automatic fake news detection. *J. Korea Converg. Soc.* **2008**, *10*, 15–21.
3. Chen, Y.; Conroy, N.J.; Rubin, V.L. Misleading online content: Recognizing clickbait as "false news". In Proceedings of the ICMI '15: International Conference on Multimodal Interaction, Seattle, WA, USA, 13 November 2015; pp. 15–19.
4. Choi, S.; Youn, S. The implications of collaborative fact-check service: Case of <SNU FactCheck>. *J. Cybercommun. Acad. Soc.* **2017**, *34*, 173–205.
5. Islam, N.; Shaikh, A.; Qaiser, A.; Asiri, Y.; Almakdi, S.; Sulaiman, A.; Moazzam, V.; Babar, S.A. Ternion: An autonomous model for fake news detection. *Appl. Sci.* **2021**, *11*, 9292. [CrossRef]
6. Ahmed, B.; Ali, G.; Hussain, A.; Baseer, A.; Ahmed, J. Analysis of text feature extractors using deep learning on fake news. *Eng. Technol. Appl. Sci. Res.* **2021**, *11*, 7001–7005. [CrossRef]
7. Jung, H. Fake News Detection Using Content-Based Feature Extraction Method. Master's Thesis, Ewha Womans University, Seoul, Korea, 2019.
8. Goldberg, Y.; Levy, O. Word2Vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv* **2014**, arXiv:1402.3722.
9. Lau, J.H.; Baldwin, T. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv* **2016**, arXiv:1607.05368.
10. Truică, C.-O.; Apostol, E.-S. MisRoBÆRTa: Transformers versus misinformation. *Mathematics* **2022**, *10*, 569. [CrossRef]
11. Kula, S.; Choraś, M.; Kozik, R. Application of the BERT-based architecture in fake news detection. In Proceedings of the Computational Intelligence in Security for Information Systems Conference, Seville, Spain, 13–15 May 2019; pp. 239–249.
12. Shu, K.; Wang, S.; Liu, H. Beyond news contents: The role of social context for fake news detection. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 11–15 February 2019; pp. 312–320.
13. Kim, Y. Third-person effect on fake news in social media: Focusing on false information related to infectious diseases. *Korean J. Broadcast. Telecommun. Stud.* **2021**, *35*, 5–32.
14. Bang, Y.; Ishii, E.; Cahyawijaya, S.; Ji, Z.; Fung, P. Model generalization on COVID-19 fake news detection. In Proceedings of the International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation, Virtual Event, 8 February 2021; pp. 128–140.
15. Al-Ahmad, B.; Al-Zoubi, A.M.; Abu Khurma, R.; Aljarah, I. An evolutionary fake news detection method for COVID-19 pandemic information. *Symmetry* **2021**, *13*, 1091. [CrossRef]
16. Rubin, V.L.; Conroy, N.J.; Chen, Y.; Cornwell, S. Fake news or truth? Using satirical cues to detect potentially misleading news. In Proceedings of the Second Workshop on Computational Approaches to Deception Detection, San Diego, CA, USA, 12–17 June 2016; pp. 7–17.
17. Tacchini, E.; Ballarin, G.; della Vedova, M.L.; Moret, S.; de Alfaro, L. Some like it hoax: Automated fake news detection in social networks. *arXiv* **2017**, arXiv:1704.07506.

18. Vo, N.; Lee, K. The rise of guardians: Fact-checking URL recommendation to combat fake news. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 275–284.
19. Kang, M.; Seo, J.; Lim, H. Korean fake news detection with user graph. *Hum. Lang. Technol.* **2021**, 97–102.
20. Nguyen, V.H.; Sugiyama, K.; Nakov, P.; Kan, M.Y. Fang: Leveraging social context for fake news detection using graph representation. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual Event, 19–23 October 2020; pp. 1165–1174.
21. Kumar, S.; Asthana, R.; Upadhyay, S.; Upreti, N.; Akbar, M. Fake news detection using deep learning models: A novel approach. *Trans. Emerg. Telecommun. Technol.* **2020**, *31*, e3767. [CrossRef]
22. Rodríguez, Á.I.; Iglesias, L.L. Fake news detection using deep learning. *arXiv* **2019**, arXiv:1910.03496.
23. Shahi, G.K.; Nandini, D. FakeCovid—A multilingual cross-domain fact check news dataset for COVID-19. *arXiv* **2020**, arXiv:2006.11343.
24. Al-Rakhami, M.S.; Al-Amri, A.M. Lies kill, facts save: Detecting COVID-19 misinformation in twitter. *IEEE Access* **2020**, *8*, 155961–155970. [CrossRef]
25. Shim, J.; Lee, J.; Jeong, I.; Ahn, H. A study on Korean fake news detection model using word embedding. *Korean Soc. Comput. Inf.* **2020**, *28*, 199–202.
26. Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; Mikolov, T. FastText.zip: Compressing text classification models. *arXiv* **2016**, arXiv:1612.03651.
27. Lim, D.; Kim, G.; Choi, K. Development of a fake news detection model using text mining and deep learning algorithms. *Inf. Syst. Rev.* **2021**, *23*, 127–146.
28. Park, C.; Kang, J.; Lee, D.; Lee, M.; Han, J. COVID-19 Korean fake news detection using named entity and user reproliferation information. *Hum. Lang. Technol.* **2021**, 85–90.
29. Hur, Y.; Son, S.; Shim, M.; Lim, J.; Lim, H. K-EPIC: Entity-perceived context representation in Korean relation extraction. *Appl. Sci.* **2021**, *11*, 11472. [CrossRef]
30. Clark, K.; Luong, M.-T.; Le, Q.V.; Manning, C.D. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv* **2020**, arXiv:2003.10555.
31. Park, J.; Kim, M.; Oh, Y.; Lee, S.; Min, J.; Oh, Y. An empirical study of topic classification for Korean newspaper headlines. *Hum. Lang. Technol.* **2021**, 287–292.
32. Weiss, K.; Khoshgoftaar, T.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 9. [CrossRef]
33. Endo, P.; Santos, G.L.; Xavier, M.E.D.L.; Campos, G.R.N.; de Lima, L.C.; Silva, I.; Egli, A.; Lynn, T. Illusion of Truth: Analysing and classifying COVID-19 fake news in Brazilian Portuguese language. *Big Data Cogn. Comput.* **2022**, *6*, 36. [CrossRef]
34. Graves, A.; Jaitly, N.; Mohamed, A.R. Hybrid speech recognition with deep bidirectional LSTM. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 8–12 December 2013; pp. 273–278.
35. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; p. 27.
36. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
37. Landis, J.R.; Koch, G.G. The Measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
38. Chicco, D.; Warrens, M.J.; Jurman, G. The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. *IEEE Access* **2021**, *9*, 78368–78381. [CrossRef]