Computer Science and Information Systems 19(2):639–658 https://doi.org/10.2298/CSIS210501053K

Transfer Learning and GRU-CRF Augmentation for Covid-19 Fake News Detection

Andrea Stevens Karnyoto, Chengjie Sun, Bingquan Liu, and Xiaolong Wang

School of Computer Science and Technology,
Harbin Institute of Technology,
Harbin 150001, China.
andre@ukitoraja.ac.id
cjsun@insun.hit.edu.cn
liubq@hit.edu.cn
wangxl@insun.hit.edu.cn

Abstract. The spread of fake news on online media is very dangerous and can lead to casualties, effects on psychology, character assassination, elections for political parties, and state chaos. Fake news that concerning Covid-19 massively spread during the pandemic. Detecting misinformation on the Internet is an essential and challenging task since humans face difficulty detecting fake news. We applied BERT and GPT2 as pre-trained using the BiGRU-Att-CapsuleNet model and BiGRU-CRF features augmentation to solve Fake News detection in Constraint @ AAAI2021 - COVID19 Fake News Detection in English Dataset. This research proved that our hybrid model with augmentation got better accuracy compared to our baseline model. It also showed that BERT gave a better result than GPT2 in all models; the highest accuracy we achieved for BERT is 0.9196, and GPT2 is 0.8986.

Keywords: Covid-19 fake news, hybrid neural network, Transfer Learning, Augmentation.

1. Introduction

In recent years, many phenomena have emerged and spread on the Internet, especially regarding the proliferation of information dissemination in online media. Some of the negative phenomena are hoaxes, rumors, and misinformation. The spread of fake news on online media is very dangerous [1,2]. And the effects can lead to casualties [3], psychological effect [4,5], character assassination [5], elections for political parties [6], and state chaos [7]. Fake news that concerning Covid-19 spread massively resulted in misunderstandings of information to the national and global communities during the pandemic. Detecting this misinformation on the Internet is an important and challenging task since even humans face difficulty in detecting fake news. In other words, humans cannot accurately distinguish whether it is fake or true news, especially it needs a tedious activity such as collecting evidence and sifting through facts. Therefore, our research concerns detecting fake news that related to covid-19 by using the Constraint @ AAAI2021 - COVID19 Fake News Detection in English Dataset [8] with Natural Language Processing Approaches Based.

The dataset for training and testing is provided by the "Constraint shared task" organizer [9], which aims to fight fake news related to COVID-19 across social media platforms such as Twitter, Facebook, Instagram, and other popular press releases. The dataset consists of 10,700 social media posts categorized into two labels: real and fake, all those written in English. Several previous studies have contributed to this Constraint @ AAAI2021 - COVID19 Fake News Detection in English shared task. Azhan et al. [10] apply a Layer Differentiated training procedure for training a pre-trained ULMFiT, Kakwani et al. [11] compile the IndicGLUE benchmark for language, Baris et al. [12] propose a modeling framework for those features by using BERT language model and external sources. Considering the number of researches utilizing the dataset, we think it crucial to contribute to this shared-task by using another approach.

Natural Language Understanding (NLU) is a branch of artificial intelligence (AI) that uses computer software to understand input presented in the text or speech format [13]. NLU is applied in automated reasoning, machine translation, question answering, news-gathering, text categorization, voice-activation, archiving, and large-scale content analysis [14,15,16]. We used Natural Language Understanding to do text categorization because it is more intelligent and efficient, which significantly challenges the semantic understanding in the system's module. We apply and modify the deep learning model conducted by Pin Ni et al. (2020) [17]: Natural Language Understanding approaches based on Intent Detection and Slot Filling joint tasks. They used the model BERT-RCNN-(BiGRU-CRF) and BERT-BiGRU-Att-CapsuleNet-(BiGRU-CRF), and they got pretty significant results. Unlike Pin Ni et al. (2020), we not only applied BERT but also employed GPT2 to training and testing our model for input pre-trained model. Also, we used BiGRU-CRF not for filling joint tasks but for feature augmentation.

Our contributions are as follows:

1) We performed two model structures: BiGRU-Att-CapsuleNet-(BiGRU-CRF). Also, we tested the dataset on a simple LSTM, Bi-GRU, BiGRU-Attention, and BiGRU-Attention-Capsule as a baseline for comparison toward our approach. Our hybrid model structure proves high competitiveness.
2) We used BERT and OpenAI GPT2 as pre-trained to all models.
3) We are involved in Constraint @ AAAI2021 - COVID19 Fake News Detection dataset shared task.

Additionally, as a study concerning hybrid-based (BiGRU(RNN), Attention(CNN), Augmentations(RNN)) and focusing on features augmentation methods, the hybrid neural network-based task model can improve model accuracies. It is proven our proposed model accuracy better compare to the baseline accuracy. The rest of the paper is organized in the following manner. In Section 2, we formally define related work in fake news detection. Section 3 describes our proposed method (the dataset, the main model, and the explanation of each layer). Section 4 is Experiment and Task. In Section 5, we present the Result and Analysis. Section 6 concludes the paper.

2. Related Work

Fake news detection is classified into Text Classification or Text Categorization. Some Fake News Detection studies use Machine Learning [18,19,20], and others use Deep

Learning [21,22]. Those techniques can also be generally categorized as News Content-based learning and Social Context-based learning. News content-based approaches deal with the different writing styles of published news articles, focusing on extracting several fake news articles related to both information and the writing style. Whereas Social context-based approaches deal with the latent information between the user and news article. The social engagements on articles can be a significant feature for fake news detection (to find the semantic relationship between news articles and writers) [23]. In the Fake News Detection research field, many datasets can be used, such as PolitiFact [24,25], Fake News Kaggle [18,26], The Fake News Challenge (FNC-1) [27,28], and Constraint@AAAI2021 - COVID19 Fake News Detection [9,10,12,11]. Ahmad et al. [18] developed Fake News Detection Using Machine Learning Ensemble Methods consists of Logistic Regression, Support Vector Machine, Random Forest (RF), etc. Monti et al. [21] proposed learning fake news-specific propagation patterns by exploiting deep geometric learning, a novel class of deep learning methods designed to work on graph-structured data. Konkobo et al. [24] performed a model to extract users' opinions expressed in comments. They used CredRank Algorithm to evaluate users' credibility and built a small network of users involved in spreading a piece of given news. Xu et al. [27] presented a new system, FaNDS, that detects fake news efficiently. The system is based on several concepts used in some previous works but a different context. There are two main concepts: An Inconsistency Graph and Energy Flow. Azhan et al. [10] proposed a Layer Differentiated training procedure for training a pre-trained ULMFiT model. They also used unique tokens to annotate specific parts of the tweets to improve language understanding and gain insights into the model, making the tweets more interpretable.

2.1. Fake News with BERT-Based

Deep neural network architectures for Transfer Learning Approaches have achieved substantial advances in a range of natural language processing (NLP) tasks recently [29]. Google published BERT as a sophisticated pre-training transfer learning model, and it is the most significant update as one of the NLP algorithms in recent years. Pre-training and fine-tuning are the two steps in the BERT framework [30]. The prominent model architecture is based on a multi-layer bidirectional Transformer encoder, which comes from the original implementation described and delivered in the tensor2tensor library [31]. Several studies that have been done by using BERT for Fake News Detection: Kaliyar et al. [23] proposed a BERT-based (Bidirectional Encoder Representations from Transformers) deep learning approach (FakeBERT). They combined different parallel blocks of the single-layer deep Convolutional Neural Network (CNN) having different kernel sizes and filters with the BERT. Gundapu et al. [32] used an ensemble of three transformer models (BERT, ALBERT, and XLNET) to detecting fake news. This model was trained and evaluated in the context of the ConstraintAI 2021 shared task “COVID19 Fake News Detection in English”. Gupta et al. [33] presented a simple approach that uses BERT embeddings and a shallow neural network for classifying tweets using only text and discuss our findings and limitations of the method intext-based misinformation detection.

2.2. Fake News with OpenAI GPT2-Based

Recently, Pre-trained Generative Transformer-2(GPT-2) is a machine learning for text processing was developed by OpenAi. The capability of the GPT-2 is the ability to process up to 1024 tokens. Unlike other pre-trained models, GPT-2 technology can flow all tokens through the pre-training decoder generative block to provide good accuracy. Wang and Cho [34] declared that the GPT-2 generation gives good quality, powerful abilities, and minimal risk of error. On the other hand, the GPT-2 can generate text blocks such as short sentences that appear like written by humans, which means easy to generate fake text. Several studies solved fake news detection using OpenAI GPT-2: Harrag et al. [35] used GPT2-Small-Arabic generated fake Arabic Sentences. For evaluation, they compared different recurrent neural network (RNN) word embeddings-based baseline models, namely: LSTM, BI-LSTM, GRU, and BI-GRU, with a transformer-based model. Ranade et al. [44] generated fake CTI text descriptions using transformers automatically. They showed that given an initial prompt sentence, a public language model like GPT-2 with fine-tuning could generate plausible CTI text with the ability of corrupt cyber-defense systems.

2.3. Fake News with RNN-Based

A recurrent neural network (RNN) is an artificial neural network that uses sequential data or time-series data. It is powerful for modeling sequence data such as time series or natural language. The Advantages of RNN: Ability to process the input of any length, model size not increasing even it has different size of the input, all computation value is saved into account historical information, and the value of the weights are shared in all timeline. Singh et al. [36] proposed a framework that includes infrastructure to collect Twitter posts that spread false information. Their model implementation utilized the Transfer Learning scheme to transfer knowledge gained from a large and general fake news dataset to relatively more minor fake news events occurring during disasters as a means of overcoming the limited size of our training dataset. Ishiwatari et al. [37] proposed relational position encodings that provide Relational Graph Attention Networks (RGAT) with sequential information reflecting the relational graph structure. Accordingly, the RGAT model can capture both the speaker dependency and the sequential information.

2.4. Fake News with CNN-Based

CNN is a robust neural network with general-purpose functionalities in computer image and natural language processing; also, CNN can extract Euclidean structured data's spatial features. A small area sliding window is used in the CNN process to extract local features and then aggregates these features by pooling. The primary purpose of CNN is to reduce the number of parameters to a small extent but can effectively extract features over different matrix regions. Moreover, CNN plays a vital role in the information processing field. It employs 2D Convolution to process tokens of sentence embedding. It is pretty simple because every input only has a 2-dimensional matrix of tokens, and the output is also a 2-dimensional matrix having a smaller size than the input. In the fake news detection task, several studies experimented with CNN: Goldani et al. [45] used Convolutional Neural Networks (CNN) with margin loss and different embedding models proposed for detecting fake news. They compared static word embeddings with the non-static embeddings that provide the possibility of incrementally up-training and updating word embedding in the training phase. Lu et al. [38] developed a novel neural network-based model Graph-aware Co-Attention Networks (GCAN) to achieve the goal. Extensive experiments conducted on real tweet datasets exhibit that GCAN can significantly outperform state-of-the-art methods. Mandelli et al. [39] proposed a 2-channel-based CNN that learns how to compare camera fingerprint and image noise. The proposed solution turns out to be much faster than the conventional approach and ensures increased accuracy. The method makes the system particularly suitable in scenarios where large databases of images are analyzed, like over social networks.

2.5. Fake News with Hybrid-based

Hybrid models are constructed of different neural networks (linear neural network and multi-layer neural network). The proposed hybrid system can be applied to many applications such as function approximation, time series prediction, and pattern classification, and text classification [40]. The hybrid model's output is formed of two or more different network yields. In one hybrid model, at least two types of neural network sets were considered together [41]. Nasir et al. [42] proposed a novel hybrid deep learning model that combines convolutional and recurrent neural networks for fake news classification. The model was successfully validated on two fake news datasets (ISO and FA-KES), achieving detection results that are significantly better than other non-hybrid baseline methods. Song et al. [43] proposed a multimodal fake news detection framework based on Crossmodal Attention Residual and Multichannel Convolutional Neural Networks (CARMN). The Crossmodal Attention Residual Network (CARN) can selectively extract the relevant information related to a target modality from another source modality while maintaining its unique information. In this research, we used the same model in Ni et al. [17]. Ni et al. use the hybrid model to solve the problem "Natural language understanding approaches based on the joint task of intent detection and slot filling for IoT voice interaction."

3. Proposed Method

3.1. Dataset Statistics

The Constraint @ AAAI2021 - COVID19 Fake News Detection in English Dataset [8] provided the shared task, which contains 10,700 humans annotated from media articles and posts acquired from multiple platforms. It is divided into data training (6,420 rows), validation (2,140 rows), and test (2,140 rows). The unique words in the training dataset are 30,046, most length tokens are 1,481, and balance data distribution for Real and Fake labels. The dataset contains the post ID, tweet, and their corresponding label fields.

Table 1. Data Distribution for Constraint @ AAAI2021 - COVID19 Fake News Detection

Data	Real	Fake	Total	Unique Word
Train	3,360	3,060	6,420	30,046
Validation	1,120	1,120	2,140	13,697
Testing	1,120	1,120	2,140	14,121

Table 2. Some Post Fake and Real

Label	Post
Real	This #FourthOfJuly weekend if you choose to spend time outdoors at an event or gathering stay 6 ft apart &amp; wear a cloth face cover to slow the spread of #COVID19. Learn more at https://t.co/c4F0aouMLd. https://t.co/u5tTl3m572
Real	We launched the #COVID19 Solidarity Response Fund which has so far mobilized $225+M from more than 563000 individuals companies &amp; philanthropies. In addition we mobilized $1+ billion from Member States &amp; other generous to support countries-@DrTedros https://t.co/xgPkPdvn0r
Fake	@realDonaldTrump has shifted his focus at different moments in the #CoronavirusOutbreak. We updated our running timeline of his response to the virus. https://t.co/pgXjssaRCB Reply to us with any recent Trump moments you think belong on this running list. https://t.co/g4WYcppDSO
Fake	RT @EllenCutch: Coronavirus misinformation is moving offline. A reddit user posted this flyer to the site and told us it had been delive…

Table 2 shows post tweets containing URL, Mention, Retweet, Hashtag, HTML special entities, and Number.

3.2. Data Preprocessing

First, we executed our tweet preprocessing and text preprocessing for transformer-based models by removing useless punctuation marks for text classification. We kept symbols ’@’ and’#’ because those have specific function in tweets. Second, we changed the text into lowercase and replaced URLs, mentions, and emojis into particular tokens. Third, we utilized the Python emoji library to exchange the emoji with a short textual description: redheart:, :thumbsup:, etc. Furthermore, we transformed hashtags into words (”#DESEASE”→”DESEASE”).

3.3. Main Model

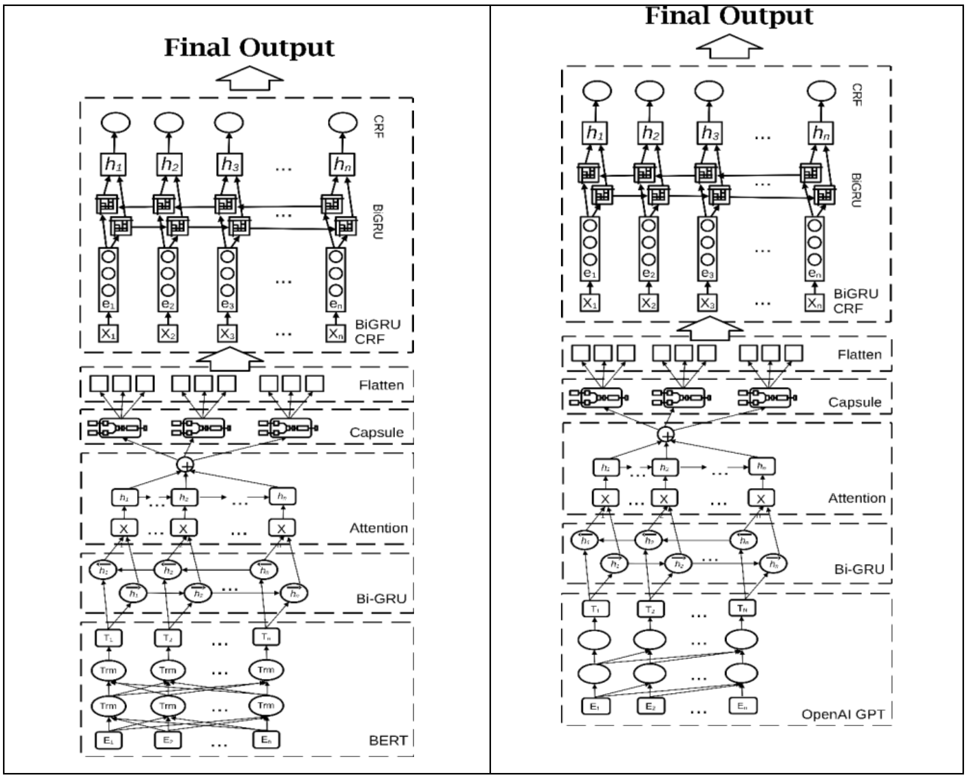

Fig 1. Main model

Fig 1. shows our two models that we were performed. The difference between the two models is the pre-training part before the first Bi-GRU layer was processed. The first model used BERT pre-trained, and the other used OpenAI GPT2 pre-trained.

Our model consists of four main parts:

1) Bidirectional GRU in the first layer, this layer receives the input from BERT.
2) The attention Layer is used to extract important features from the content.
3) The next layer is the capsule network layer, which is used to present each output of neurons with different intensity connections.
4) And the last layer is BiGRU-CRF; this layer is features segmentation that processes the capsule network features output.

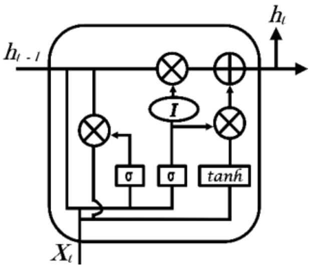

Fig 2. Gate Recurrent Unit

The first layer of our model is the Bidirectional Gate Recurrent Unit (BiGRU) layer. GRU is one of the LSTM variants with more advantages such as more superficial structural characteristics than standard LSTM, GRU has fewer parameters, and GRU has better convergence (Fig 2 shows GRU neuron structure). The two main parts of the GRU are the update gate and the reset gate. GRU uses the *z* Update Gate to manage the degree of impact at the previous time (*t - 1*) at the current hidden layer. The reset gate *r* is used as a control mechanism to ignore the output of the hidden layer information or not. The larger the update gate's value, the more the hidden layer's output is influenced by the previous layer. Moreover, smaller result value of the update gate means a lot of information was ignored at the last hidden layer. For more details, see the following formula:

$$
\begin{aligned}
r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\
z_t &= \sigma(W_i \cdot [h_{t-1}, x_t]) \\
\tilde{h}_t &= \tanh(W_c \cdot [r_t \cdot h_{t-1}, x_t]) \\
h_t &= (1 - z_t) \cdot c_{t-1} + z_t \cdot \tilde{h}_t
\end{aligned}
\tag{1}
$$

We created a forward or backward GRU network model for the context in order to accomplish the bidirectional process of the text (from the beginning to end and vice versa). Both unidirectional GRUs together determines the output performance in the right and opposite directions, respectively. This layer is not only providing input information at each moment but also the bi-unidirectional GRUs jointly to determining the output for next moment (*t+1*). Layer that consists two unidirectional GRUs can considerably the bidirectional GRU, the weighted summation of the hidden layer state's output in the forward direction h_{t-1} and the hidden layer state in the backward direction (h_{t-1}): obtains the hidden state of the BiGRU at time *t*. See the following formula:

$$
\begin{aligned}
\overrightarrow{h}_t &= GRU(x_i, \overrightarrow{h}_{t-1}) \\
\overleftarrow{h}_t &= GRU(x_i, \overleftarrow{h}_{t-1}) \\
h_i &= w_t \overrightarrow{h}_t + v_t \overleftarrow{h}_t + b_t
\end{aligned}
\tag{2}
$$

The input word vector from BERT/OpenAI GPT2 is transformed into nonlinear by using GRU. The length of input and output vector of GRU is different, the output size is adjusted to the next layer input size. The hidden layer forward state processing w_t and v_t in current time generates h_t. Meanwhile, h_t is also used for GRU hidden layer backward state, which processing at time t. The hidden layer at present t also has a bias b_t. Next, to produce vector h_i for each word, the forward and backward GRU outputs are combined. Finally, each recurrent unit can process dependencies at different times.

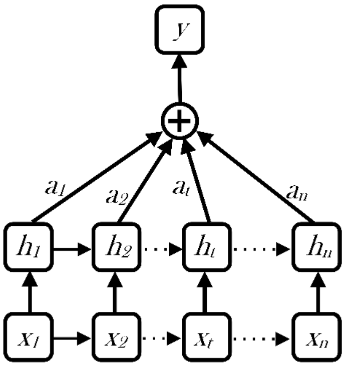

Fig 3. Attention Layer

The next layer is Attention as show in Fig 3. This mechanism was first proposed by the Google Mind Team. The main task of attention is to get important features from the bidirectional GRU layer output, and the motivation behind the incorporation of the attention make a network capable of learning object attentive features. In other words, attention mechanism focusses important information by simulate the attention characteristics of human brain. The output of each previous BiGRU layer is imported into the attention mechanism, and the result of this layer are specific array which will process in next layer. For instance, the terms “wonderful algorithm but the code was difficult”, “wonderful” and “difficult” are all sentimentally inclined. It is sentiment polarity is more likely to be positive for the target “algorithm” because “wonderful” is closer to “algorithm”.

$$s = \sum_{i=1}^{l} \alpha_i h_i$$

$$\alpha i = \frac{\exp(e_i)}{\sum_{j=1}^{n} (e_j)}$$

$$e_i = v_i \tanh(w_i h_i + b_i) \tag{3}$$

The summation of the multiplication the coefficient α_1 and hidden layer state h_1 result from initial hidden layer state to updated hidden layer state in the initial input generates vector S. Weight coefficient matrix in *i-th* time is denoted by v_i and w_i. Corresponding offset at the *i-th* time denoted by b_i and e_i represents the value determined by the hidden layer state vector h_i at the *i-th* time.

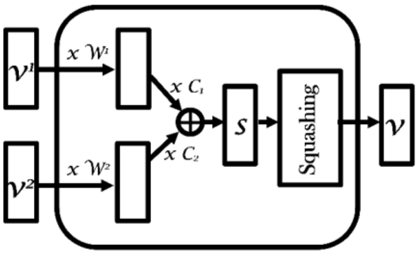

Fig 4. Capsule Network

Fig 4 shows the Capsule Network is a CNN model alternative with a slightly different operation than regular CNN, and it has a hierarchical relationship. In this study, dynamic routing aims to train neural networks to analyze the relationship between words in a vector and get characteristics in a text. The input of this layer is a vector generated from the previous attention layer. Simultaneously, the output is a vector consist of a probability of observation and a position for that observation. The following equation provides an overview of the process of the capsule network. The input capsule network is written with the V_i^t symbol.

$$\begin{aligned} \hat{u}_{j|i}^t &= W_{ij}^t v_i^t \\ s_j^t &= \sum_i c_{ij} \hat{u}_{j|i}^t \\ c_{ij} &= \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \end{aligned} \tag{4}$$

To updated through the dynamic routing process, we are using coupling coefficient is denoted by C_{ij}. W_{ij}^t is a weight matrix that transforms input features from beginning to end. S_j^t is a global feature based on all input features. The coupling coefficients between global features T_j and all the input features sum into 1 and are determined by a "routing softmax" with B_{ij} initialized to 0. Then, to scale the globally represented modulus length between 0 and 1 uses the squash function:

$$g_j^t = squash(s_j^t) \tag{5}$$

Algorithm 1. shows how the dynamic routing process is executed.

```
procedure ROUTING(Uji, r)
 for all input feature i in input layer:
 for all global feature j in output layer: bij=0
   for r iterations do
     cij = softmax(bij)
     stj = Sigma(Cij x Utji)
     gtj = squash(stj)
     bij = bij + (Utji x Gtj)
```

Variable **r** is the number of iterations, and **Gtj** is one of the global features based on the input features. Therefore, for **j** = 1, …n_c, and all **U**tj in $\mathbf{U}^c$ or $\mathbf{U}^l$, we can generate two global representations, respectively.

BIGRU-CRF

Our approach utilizes a BiGRU architecture and CRF (Conditional Random Fields) for features segmentation. More precisely, this last layer involves three sub-layers: 1) an input layer containing flatten of capsule vectors, 2) a hidden layer where the Bi-GRU maps vectors to hidden sequences, and 3) an output layer that calculate the probability of label base of previous hidden sequences. The CRF model is a discriminant undirected graphical probability. It has been successfully applied in various natural language processing. Linear chain CRFs are most popular in NLP tasks, which implement sequential dependencies in the predictions and consist of undirected graph learning, based on the maximum entropy and hidden Markov, but simpler compare standard hidden Markov models.

$$p(\hat{y} \mid \hat{x}; w) = \frac{\exp(\sum_i \sum_j w_j f_j(y_i - 1, y_i, \hat{x}, i))}{\sum_{y_z \in y} \exp(\sum_i \sum_j w_j f_j(y_i - 1, y_i, \hat{x}, i))} \tag{6}$$

Formula 6 shows a form of CRF model. Let $x = \{x_1, x_2, \cdots, x_n\}$ denote the observation sequence and $y = \{y_1, y_2, \cdots, y_n\}$ be the set of finite states. W is the weight vector for weighing the output feature vector f from the BiGRU. The Viterbi algorithm will perform training and decoding.

4. Experiment and Task

4.1. Experiment Setup

The experiments run on Intel Core (TM) i7 8700, 6 core 3.20GHz Processor, 16 GB RAM, Nvidia GeForce GTX 1050 Ti GPU 4 GB. Base program and training use Python 3.7.8 and TensorFlow Version: 2.1.0. Pre-training and tokenizer use BERT transformer 3.4.0 with PyTorch 1.6.0+cu101. Several important hyper-parameters determine this architecture: Training model learning-rate= 0.001, epoch 10, batch-size 8. The dimension of word embedding for model inputs is different depend on datasets.

4.2. Description of Task

We divided tasks into two steps. 1) Prepare a pre-trained process using BERT and GPT2, 2) Train the Model using datasets. First, we took a maximum of 300 features (words) using the NLTK tool in the preprocessing before entering the pre-training process. However, BERT and GPT2 tokenizers generate a different number of tokens in the pre-training. BERT generates 375 tokens while GPT2 generates 413 tokens from those 300 features, so the system automatically adjusted the maximum padding and input layer number referred to BERT and GPT2 output. Secondly, we conducted model training. We train for BERT on the BERT-LSTM, BERT-BiGRU, BERT-BiGRU-Attention, BERT-BiGRU, BERT-BiGRU-Attention-Capsule, BERT-BiGRU-Attention-Capsule-BiGRU-CRF models. Also, we were performing for OpenAI-GPT2 on the GPT2-LSTM, GPT2-BiGRU, GPT2-BiGRU-Attention, GPT2-BiGRU, GPT2-BiGRU-Attention-Capsule, GPT2-BiGRU-Attention-Capsule-BiGRU-CRF models. After completing all training processes towards models, we calculate and compare the testing results, such as accuracy, loss, F1, and the recall score.

5. Result and Analysis

To further prove that our proposed model can better accuracy by capturing more features details and enhancing the dependency between layers. We evaluated our proposed model (BiGRU-Attention-CapsNet-(BiGRU-CRF)) and baseline systems (LSTM, BiGRU, BiGRU-Attention, and BiGRU-Attention-CapsNet) toward The Constraint @ AAAI2021 - COVID19 Fake News Detection in English Dataset. Experiments applied the same hyperparameters such as fine-tuning, learning-rate, and batch-size settings. We present experimental results in Table 3 to prove that the techniques discussed in our proposed method contribute to increasing Neural-Network-based binary classification performance and then compare all models on the datasets mentioned above to get an overall impression of their performance. Recall, F1-Score and accuracy have been determined from the confusion matrix, and we used those results to decide classification results. The BERT-BASED section in Table 3 shows although the highest training accuracy is BERT-LSTM (baseline), our proposed method got the best accuracy for testing. It indicates no rigid relationship between training accuracy and testing accuracy. When the training accuracy is the highest, it does not mean it will get the highest accuracy result for testing. In contrast to GPT2-BASED, our approach got the highest accuracy for training, validation, and testing. Meanwhile, the BERT-BiGRU-Attention-CapsNET model got unsatisfied accuracy for training both for BERT and GPT2 Pre-training, although the accuracy for testing is still higher than LSTM. We assume that the shuffling process for features is still in capsules form and has not entered the augmentation process. Moreover, we concluded that BERT Pre-trained achieves better accuracy than GPT2 Pre-trained, although both use the same models (baseline and our proposed model). Comparing the two Pre-trained testing accuracies for our proposed model is 0.0208, and we concluded that BERT has significant enough, even on the F1 and Recall results.

The BERT-Based accuracy and GPT2-Based accuracy for all Baseline and Our Proposed model curves after training versus the number of epochs for the classification task based are shown in Fig 5. We can observe the curves that the model has learned well and does not have any significant result swing of accuracy at the end of the epoch.

Table 3. Train, Validation, and Test Result

	Train		**Validation**		**Testing**		
	Acc	**Loss**	**Acc**	**Loss**	**Acc**	**F1**	**Recall**
BERT BASED							
BERT-LSTM	**0.9937**	**0.0193**	0.9061	0.5393	0.9046	0.9045	0.9050
BERT-BiGRU	0.9827	0.0895	0.9192	0.2975	0.9117	**0.9116**	0.9122
BERT-BiGRU-Attention	0.9748	0.0640	0.9168	0.6256	0.9079	0.9076	0.9074
BERT-BiGRU-Attention-CapsNET	0.9221	0.1489	0.9084	0.2350	0.9061	0.9057	0.9054
BERT-BiGRU-Attention-CapsNET-(BiGRU-CRF)	0.9899	0.1011	**0.9238**	**0.2013**	**0.9196**	0.9113	**0.9193**
GPT2 BASED							
GPT2-LSTM	0.9123	0.2523	0.8860	0.2594	0.8841	0.8841	0.8800
GPT2-BiGRU	0.9165	0.2131	0.9192	**0.2411**	0.8766	0.8741	0.8718
GPT2-BiGRU-Attention	0.9226	0.2110	0.9168	0.2692	0.8855	0.8855	0.8878
GPT2-BiGRU-Attention-CapsNET	0.8876	**0.2089**	0.9084	0.2725	0.8846	0.8843	0.8841
GPT2-BiGRU-Attention-CapsNET-(BiGRU-CRF)	**0.9409**	0.2098	**0.9238**	0.2424	**0.8986**	**0.8924**	**0.8897**

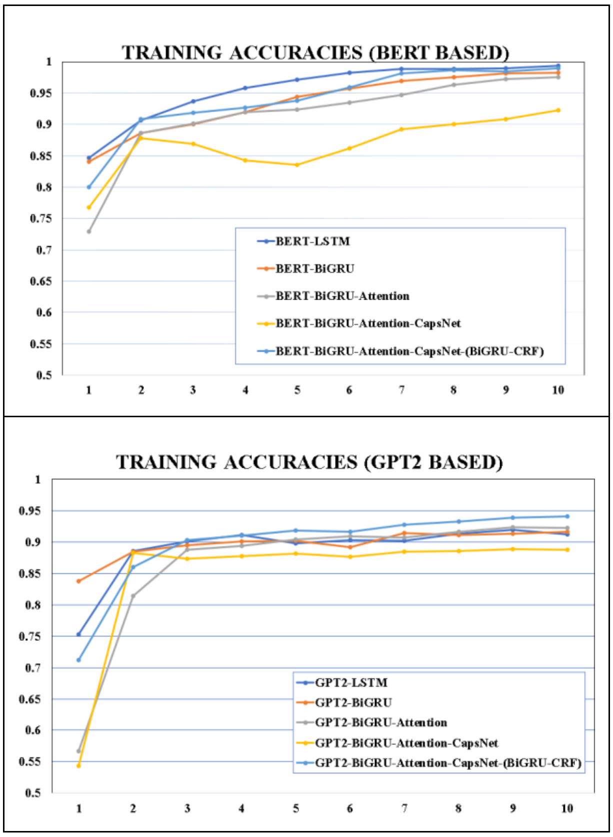

Fig 5. Training Accuracies

As shown in Fig 5, all models yield accuracies with only slight differences from one to another. It shows LSTM gets the best training results. The training system on LSTM (simple model) can only train non-hidden and distinctive features, leading to ease of learning for neural networks. BiGRU-Attention-CapsNet got the worst outcomes. We assume the model is not complete, and this neural network is still just starting to analyze the relationship between words and try to get the characteristics of the text. Although our approach (BiGRU-Attention-CapsNet-(BiGRU-CRF)) ranks number two for accuracy during training both for BERT and GPT2, this model obtained good accuracy during testing. However, the overall model provides increased accuracy in subsequent epochs.

5.1. Effect of Sentence Length

In our testing dataset, we found 1210 rows (56.54%) contain 15 words or less, 878 rows (41.02%) have 16-30 words, 47 rows (2.19%) contain 31-45 words and five rows (0.23%) contain more than 45.

Table 4. Effect of Sentences length toward classification accuracies

	Sentence Length			
	<=15	**16-30**	**31-45**	**>=46**
BERT-BASED				
BERT-LSTM	0.8726	**0.9568**	0.9352	0.6000
BERT-BiGRU	0.8769	**0.9636**	0.8723	0.6000
BERT-BiGRU-Attention	0.8727	**0.9567**	0.9362	0.6000
BERT-BiGRU-Attention-CapsNET	**0.9116**	0.8998	0.8936	0.8000
BERT-BiGRU-Attention-CapsNET-(BiGRU-CRF)	0.8810	**0.9761**	0.8723	0.8000
GPT2-BASED				
GPT2-LSTM	0.8791	**0.8942**	0.8733	0.6000
GPT2-BiGRU	0.8686	0.8884	**0.8936**	0.6000
GPT2-BiGRU-Attention	0.8760	**0.8998**	0.8723	0.8000
GPT2-BiGRU-Attention-CapsNET	0.8793	**0.8941**	0.8723	0.6000
GPT2-BiGRU-Attention-CapsNET-(BiGRU-CRF)	**0.9083**	0.8884	0.8723	0.6000

Table 4 shows the effect of sentence length from each model experiment. The sentences that contain 16-30 words got better accuracy for most models. Sentences that contain over 45 words have the lowest accuracy due to the uneven distribution of data. It indicates models we tested gain the best learning rates and predictions when the text contains 16-30 words in length. Our proposed model got the highest result in the BERT section, which is 0.9761. And the lowest is obtained by BERT-LSTM, BERT-BiGRU, BERT-BiGRU-Attention, which is 0.6. In the GPT2 section, our proposed model also gets the best result, which is 0.9083. In contrast to BERT, our proposed model gains the best results in the 1–15-word group for GPT2. But even so, the difference between one result to another is only slight. Because sentences longer than 45 are few in the dataset, the learning process in the network only has a few samples.

5.2. Most Common Terms in Fake News our Model can Detect

We also explored the most frequent words in our fake news dataset that our model can detect. However, we found the terms frequently used in fake news are similar to those that appeared most often in the entire dataset.

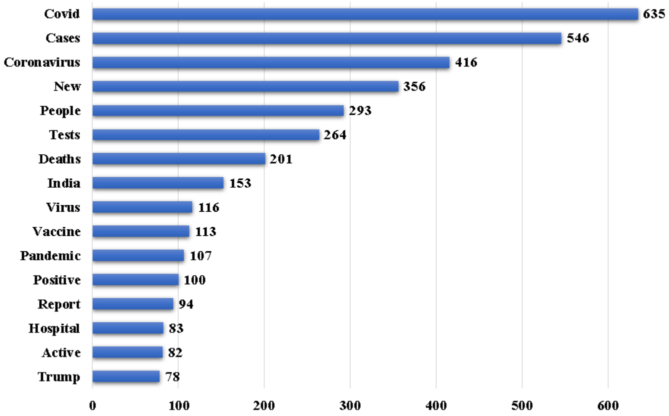

Fig 6. Most common term was detected by our models in The Constraint @ AAAI2021 - COVID19 Fake News Detection Dataset

Fig 6 shows the most frequent words in the testing dataset which our model can detect. The single text contains terms, which means every text in the dataset may consist of more than one most common word. Because the dataset for this research is related to covid-19, we only show the whole words related to covid-19. We noticed that "Covid," "Cases," and "Coronavirus" words are the most commonly appeared in this fake news dataset; however, those words are also common to appear in the news media and attract readers the most both for real or fake news. After the "India" term, the next frequently used words were not significantly different because some were described sufficiently in the top 3 words.

In the experiment, we train and test the performance of our proposed model and various baseline models. It showed the results of our proposed models are better than the baseline model. It also confirmed that our model has good observation and can catch complex augmentation and robust detection to improve the quality of the text classification.

6. Conclusion

With the growing popularity of online media such as online news, Facebook, Twitter, and other social media, more and more people get information from online media instead of newspapers and television. However, unresponsible people also used online media to spread fake news, and the effects make a negative impact on individual users and broader society. In this paper, we first exposed the interest and descriptions of automatic fake news detection. Then we compared and discuss our proposed (BiGRU-Att-CapsuleNet-(BiGRU-CRF)) model and our baseline (BiGRU, BiGRU-Attention, BiGRU-Attention-CapsuleNet). We also used BERT, GPT as pre-trained, and Constraint @ AAAI2021 - COVID19 Fake News Detection in English as a dataset to test our model and baseline. Based on our observations, our proposed method got better accuracies compared to baseline.

References

1. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data, 8(3), 171-188.
2. Riedel, B., Augenstein, I., Spithourakis, G. P., & Riedel, S. (2017). A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. arXiv preprint arXiv:1707.03264.
3. Apuke, O. D., & Omar, B. (2020). Fake news proliferation in Nigeria: Consequences, motivations, and prevention through awareness strategies. Humanities and Social Sciences Reviews, 8(2), 318-327.
4. Ozbay, F. A., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: Statistical Mechanics and its Applications, 540, 123174.
5. Pulido, C. M., Ruiz-Eugenio, L., Redondo-Sama, G., & Villarejo-Carballido, B. (2020). A new application of social impact in social media for overcoming fake news in health. International journal of environmental research and public health, 17(7), 2430.
6. Maldonado, M. A. (2019). Understanding fake news: Technology, affects, and the politics of the untruth. Historia y Comunicación Social, 24(2), 533.
7. Waisbord, S. (2018). Truth is what happens to news: On journalism, fake news, and post-truth. Journalism studies, 19(13), 1866-1878.
8. Constraint-shared-task-2021, Available: https://constraint-shared-task-2021.github.io/ (current April 2021)
9. Akhtar, M. S., & Chakraborty, T. (2021). Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers (p. 42). Springer Nature.
10. Azhan, M., & Ahmad, M. (2021). LaDiff ULMFiT: A Layer Differentiated training approach for ULMFiT. arXiv preprint arXiv:2101.04965.
11. Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020, November). iNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (pp. 4948-4961).
12. Baris, I., & Boukhers, Z. (2021). ECOL: Early Detection of COVID Lies Using Content, Prior Knowledge and Source Information. arXiv preprint arXiv:2101.05499.
13. Ovchinnikova, E. (2012). Integration of world knowledge for natural language understanding (Vol. 3). Springer Science & Business Media.
14. Van Harmelen, F., Lifschitz, V., & Porter, B. (Eds.). (2008). Handbook of knowledge representation. Elsevier.
15. Petrović, Đ., & Stanković, M. (2018). Use of linguistic forms mining in the link analysis of legal documents. Computer Science and Information Systems, 15(2), 369-392.
16. Zhao, H., Cao, J., Xu, M., & Lu, J. (2020). Variational neural decoder for abstractive text summarization. Computer Science and Information Systems, 17(2), 537-552.
17. Ni, P., Li, Y., Li, G., & Chang, V. (2020). Natural language understanding approaches based on joint task of intent detection and slot filling for IoT voice interaction. Neural Computing and Applications, 1-18.
18. Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake News Detection Using Machine Learning Ensemble Methods. Complexity, 2020.

19. Gilda, S. (2017, December). Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection. In 2017 IEEE 15th student conference on research and development (SCOReD) (pp. 110-115). IEEE.
20. Aphiwongsophon, S., & Chongstitvatana, P. (2018, July). Detecting fake news with machine learning method. In 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON) (pp. 528-531). IEEE.
21. Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. arXiv preprint arXiv:1902.06673.
22. Sahoo, S. R., & Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. Applied Soft Computing, 100, 106983.
23. Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 1-24.
24. Konkobo, P. M., Zhang, R., Huang, S., Minoungou, T. T., Ouedraogo, J. A., & Li, L. (2020, November). A Deep Learning Model for Early Detection of Fake News on Social Media. In 2020 7th International Conference on Behavioural and Social Computing (BESC) (pp. 1-6). IEEE.
25. Oriola, O. Exploring N-gram, Word Embedding and Topic Models for Content-based Fake News Detection in FakeNewsNet Evaluation. International Journal of Computer Applications, 975, 8887.
26. Shakeel, D., & Jain, N. Fake news detection and fact verification using knowledge graphs and machine learning.
27. Xu, J., Zadorozhny, V., Zhang, D., & Grant, J. (2020). FaNDS: Fake News Detection System Using Energy Flow. arXiv preprint arXiv:2010.02097.
28. Hassan, F. M., & Lee, M. (2020, September). Multi-stage News-Stance Classification Based on Lexical and Neural Features. In Conference on Complex, Intelligent, and Software Intensive Systems (pp. 218-228). Springer, Cham.
29. Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., ... & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:1912.07076.
30. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.
32. Gundapu, S., & Mamid, R. (2021). Transformer based Automatic COVID-19 Fake News Detection System. arXiv preprint arXiv:2101.00180.
33. Gupta, A., Sukumaran, R., John, K., & Teki, S. (2021). Hostility Detection and Covid-19 Fake News Detection in Social Media. arXiv preprint arXiv:2101.05953.
34. Wang, A., & Cho, K. (2019). Bert has a mouth, and it must speak: Bert as a markov random field language model. arXiv preprint arXiv:1902.04094.
35. Harrag, F., Debbah, M., Darwish, K., & Abdelali, A. (2021). Bert transformer model for detecting Arabic GPT2 auto-generated tweets. arXiv preprint arXiv:2101.09345.
36. Singh, D., Shams, S., Kim, J., Park, S. J., & Yang, S. Fighting for Information Credibility: An End-to-End Framework to Identify Fake News during Natural Disasters.
37. Ishiwatari, T., Yasuda, Y., Miyazaki, T., & Goto, J. (2020, November). Relation-aware Graph Attention Networks with Relational Position Encodings for Emotion Recognition in Conversations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 7360-7370).

38. Lu, Y. J., & Li, C. T. (2020). GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. arXiv preprint arXiv:2004.11648.
39. Mandelli, S., Cozzolino, D., Bestagini, P., Verdoliva, L., & Tubaro, S. (2020). CNN-based fast source device identification. IEEE Signal Processing Letters, 27, 1285-1289.
40. Chen, Y., Kak, S., & Wang, L. (2008). Hybrid neural network architecture for on-line learning. arXiv preprint arXiv:0809.5087.
41. Rojek, I. (2010, June). Hybrid neural networks as prediction models. In International Conference on Artificial Intelligence and Soft Computing (pp. 88-95). Springer, Berlin, Heidelberg.
42. Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. International Journal of Information Management Data Insights, 1(1), 100007.
43. Song, C., Ning, N., Zhang, Y., & Wu, B. (2021). A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. Information Processing & Management, 58(1), 102437.
44. Ranade, P., Piplai, A., Mittal, S., Joshi, A., & Finin, T. (2021). Generating Fake Cyber Threat Intelligence Using Transformer-Based Models. arXiv preprint arXiv:2102.04351.
45. Goldani, M. H., Safabakhsh, R., & Momtazi, S. (2021). Convolutional neural network with margin loss for fake news detection. Information Processing & Management, 58(1), 102418.

Andrea Stevens Karnyoto received the Master Engineering Degree in computer science from Universitas Hasanuddin, Makassar, Indonesia, in 2010. He is currently working toward the Ph.D. degree at the School of Computer Science, Harbin Institute of Technology, Harbin, China. His current research interests include Fake News Prevention and Detection, Text Mining, Natural Language Processing, and Artificial Intelligence.

Chengjie Sun received the Ph.D. degree from Harbin Institute of Technology, Harbin, China, where he iscurrently working on discriminative learning models for text mining. Since 2009, he has been an Associate Professor with Harbin Institute of Technology. His research interests include developing machine learning techniques for natural language processing and understanding.

Bingquan Liu received the Ph.D. degree in computer application technology from Harbin Institute of Technology, Harbin, China, in 2003. He is currently an Associate Professor with the School of Computer Science and Technology, and the Deputy Dean of Intelligent Technology and Natural Language Processing Research Group, Harbin Institute of Technology. His research interests include Question and Answering, Natural Language Processing, and Artificial Intelligence.

Xiaolong Wang received the Ph.D. degree in computer application technology from Harbin Institute of Technology, Harbin, China, in 1989. He is currentlya Professor with the School of Computer Science and Technology, Harbin Institute of Technology, China. He was honored as the outstanding contribution doctor in 1991, and the special allowance expert of the State Council in 1993. He was the inventor of the Chinese sentence level input method that embeddedin Microsoft Windows since 1996. His research interests include intelligent input method, online finance information platform, question answering, and artificial intelligence.

Received: May 01, 2021; Accepted: September 30, 2021.