electronics

MDPI

Article

APSN: Adversarial Pseudo-Siamese Network for Fake News Stance Detection

Zhibo Zhou [1,†], Yang Yang [2,†] and Zhoujun Li [1,*]

[1] School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[2] SPD Bank, Shanghai 200002, China
* Correspondence: lizj@buaa.edu.cn
† These authors contributed equally to this work.

Abstract: Fake news is a longstanding issue that has existed on the social network, whose negative impact has been increasingly recognized since the US presidential election. During the election, numerous fake news about the candidates distributes vastly in the online social networks. Identifying inauthentic news quickly is an essential purpose for this research to enhance the trustworthiness of news in online social networks, which will be the task studied in this paper. The fake news stance detection can contribute to detect a startling amount of fake news, which aims at evaluating the relevance between the headline and text bodies. There exists a significant difference between news article headline and text body, since headlines with several key phrases are usually much shorter than the text bodies. Such an information imbalance challenge may cause serious problems for the stance detection task. Furthermore, news article data in online social networks is usually exposed to various types of noise and can be contaminated, which poses more challenges for the stance detection task. In this paper, we propose a novel fake news stance detection model, namely Adversarial Pseudo-Siamese Network model (APSN), to solve these challenges. With coupled input components with imbalanced parameters, APSN can learn and compute feature vectors and similarity score of news article headlines and text bodies simultaneously. In addition, by adopting adversarial setting, besides the regular training set, a set of noisy training instances will be generated and fed to APSN in the learning process, which can significantly enhance the robustness of the model. Extensive experiments have been conducted on a real-world fake news dataset, and the experimental results reveal that the presented model exceeds compared suspicious information detection models with significant advantages.

Keywords: fake news; stance detection; Pseudo-Siamese network; adversarial training

Citation: Zhou, Z.; Yang, Y.; Li, Z. APSN: Adversarial Pseudo-Siamese Network for Fake News Stance Detection. *Electronics* **2023**, *12*, 1043. https://doi.org/10.3390/electronics12041043

Academic Editors: Arkaitz Zubiaga and Dah-Jye Lee

Received: 18 January 2023
Revised: 13 February 2023
Accepted: 16 February 2023
Published: 20 February 2023

1. Introduction

Fake news usually employs ambiguous details to fool people in order to obtain benefits, including via traditional news media, e.g., print and broadcast, or deliberate misinformation or hoaxes spread on online social networks. Nowadays, with the fast expansion of the Internet and convenient development of mobile terminals, fake news is dispersing especially fast on social networks, especially in terms of politics.

During the 2016 and 2020 US president election, multifarious inauthentic news about the presidential candidate spread on social networks, which might have affected the outcome of the general election. The top 20 fake news about the 2016 U.S. presidential elections received more hits on Facebook than the top 20 election reports of 19 major media outlets, according to an analysis presented by BuzzFeed [1]. Only needing to add or modify a few words, social network users can easily change the content of the news, which affects the behavior of offline users. In some respects, those Internet providers even acquiesce in such behavior. How to improve the credibility of news on social networks has always been a difficult problem for practitioners. One method is to recognize fake news articles quickly, which will be the study researched in this paper.

The fake news detection problem is a difficult problem whose serious negative impact has been increasingly recognized since the 2016 election. Fake news has enormous differences from regular suspicious messages, such as spam email [2–6], studied by some researchers before, in different kinds of aspects: (1) impacts on society: spam emails are generally only transmitted to individuals or only to small forwarding groups. Their social impact is limited and the scope of spread is small. However, due to a massive social network users, the spread of fake news is usually widespread and influential. At the same time, reposting will also bring a new round of propagation [7–10]; (2) audiences' initiative: In the dissemination process of fake news, social network users will actively forward fake news and even look for fake news to spread. Most users who forward fake news merely want to obtain more reading with no sense about its correctness. However, most people will block spammers directly; and (3) spam or fake reviews are generally easier to identify. Nevertheless, it is more difficult to identify fake news. Identifying fake news requires finding enough evidence or requiring users to have relevant qualifications knowledge due to the shortage of related real news.

These features of the above-mentioned fake news lead to new challenges to the fake news detection task. Through a complete analysis of fake news dataset prior to preparing this paper, we found some common defects about fake news, which can be categorized as presentation defects. Literally, "presentation defect" denotes the instantly visible defect in news article presentations, which widely exists in each presentation modality, such as titles and textual contents [11–13]. Specifically, "presentation defect" covers information consistency defects among this news. Significantly different from the regular news articles written by professional journalists with well-polished words, live images, and videos, the information in fake news often suffers from inconsistency (such as text bodies and headlines being irrelevant), namely the presentation defects. These discovered fake news defects actually help give a direction for solving the "fake news detection" difficulty in our study.

Based on the above defects about fake news, the fake news stance detection can detect the fake news with information consistency defects. It aims to understand the relationship between the news headline and text body. This kind of detection can help to find some fake news articles whose headlines are irrelevant and even conflicting with their text body. In our paper, we propose a Pseudo-Siamese network model with coupled input components to do fake news stance detection, which can accept the news article input in various modalities and compute the multi-modality consistency based on the learned modality-specific signature representations, respectively.

Fake news stance detection is not easy and may suffer from several significant challenges. First, such stance detection is actually a multi-class classification problem, and a classical Siamese network [14] is proposed for binary classification. There are four kinds of relationships between news headline and their text body: unrelated, conflicting, neutral, and consistent. In addition, the conflicting, consistent, and neutral news belong to the related news because their headlines are related to their text bodies:

- Consistent News: The text body is consistent with the headline;
- Conflicting News: The text body contradicts with the headline;
- Neutral News: The text body discusses the same topic as the headline, but does not take a position;
- Unrelated News: The text body discusses a different topic rather than the headline.

If we design a binary classification model, the model can only distinguish unrelated news from the dataset. However, if we design a multi-class classification model, the model can distinguish both unrelated news and conflicting news and hence outperform the binary classification model. We need to define new loss functions for the Siamese network to train the multi-class classification model. Second, the model is likely to overfit since the size of fake news is small. Thus, we use data augmentation to extend our dataset. By adding some negative words in text body sentence, we obtain some conflicting news

from consistent news. With permutation of the headline and text body, we can acquire lots of unrelated news.

By extending the traditional Siamese network model to the fake news stance detection scenario, we propose an exponential Pseudo-Siamese model to address such stance detection. Furthermore, in our experiment, we find that the perturbation in news (e.g., stop-word, incomplete sentences) can have a bad impact on a model's performance so we use adversarial training to make a model more robust to perturbation. Our innovative contributions are summarized as follows:

- Size imbalance of headline and text body: We are the first to propose an exponential Pseudo-Siamese network for stance detection of fake news. The news headline is much shorter than its text body, which will lead to the imbalance of information. The exponential Pseudo-Siamese network we proposed can address such an imbalance:
- No human carefully selected features: Our model can learn the features automatically with pre-trained GloVe word vectors;
- Less training data with good performance: With only 60% of the training data, the proposed model can achieve a very good FNC score (89.7%), which is higher than the previous state-of-the-art method (89.0%) using all training data. With all of the data, our model can achieve the best FNC score (93.40%);
- Robustness to perturbation: Adversarial training method makes our model more robust against perturbation.

2. Related Work

In recent years, some studies on fake news and stance detection have been launched. Some research is about the stance detection of tweets [15–17]. Mohammad et al. [18] designed an automatic Twitter stance detection system to detect whether the tweeter agrees, disagrees, or is irrelevant to the tweet. They had two tasks to verify the effectiveness of the system. For task A, the best classification F-score is 67.82, while the other task is 56.28. Augenstein et al. [15] experimented with conditional LSTM encoding, which built a representation of the tweet that was dependent on the target, and demonstrated that it outperformed, encoding the tweet and the target independently. Du et al. [19] brought a novel attention module to the neural network-based stance classification model, which combines target-specific information. Their model achieved the stoa performance on both the English and Chinese Stance Detection. Yang et al. [20] experimented with a two-step attention-based mechanism, which transforms tweet stance detection into two binary classification problems, and demonstrated that it outperformed some strong baselines.

However, the shortage of a corpus of deceptive news is the main challenge in this field for kinds of models to predict or detect. There are several ways to gather fake news: fake product reviews [21–23], fudged online resumes [24], opinion spamming [25–27], fake social network profiles [28–30], fake dating profiles [31], and forged scientific work. Some data are available but are restricted in content (e.g., to hotels and electronics reviews).

There are other studies on fake news detection. Rubin et al. [32] separates fake news into three classifications, namely Serious Fabrications, Large-Scale Hoaxes, and Humorous Fakes. According to their characteristics, Rubin et al. use them as a corpus for text analysis and prediction. Based on the theory of detection tool impact, Zahedi et al. [33] presented a method to study how the significant performance of detection tools and cost-related factors of the fake website affect users' thoughts of tools and threats, the efficiency of processing threats, and the dependence on such tools.

In addition, there is a contest named Fake News Challenge 1 (FNC-1) [34] (http://www.fakenewschallenge.org (accessed on 1 February 2023)) which concentrates on fake news and stance detection. Utilizing the dataset of this contest, Chopra et al. [34] leveraged an SVM trained on TF-IDF cosine similarity features to address stance detection and then employed various neural network architectures built on top of LSTM models and scored 86.58 according to the FNC-1's performance metric. Yuxi et al. (https://github.com/Cisco-Talos/fnc-1 (accessed on 1 February 2023)) designed a model founded on a weighted

average, which scored 82.02 of the FNC score. The model combines gradient-boosted decision trees and a deep convolutional neural network.

There are some previous studies about the Siamese network. In 1994, Bromley et al. [35] designed a rudimentary Siamese network to judge if two signatures came from one person. Their experiment showed that the Siamese network could recognize forgeries of signatures effectively. In recent years, the Siamese network has been applied in other questions [36–38]. Fu et al. [39] used the Siamese network on RGB-D object detection with joint learning and densely cooperative fusion. Ji et al. [36] put forward a Siamese-based cross-attention model for video salient object detection. Chen et al. [37] used a Siamese network with a spatial transformer layer for accurate pelvic fracture detection and achieved stoa performance. Huang et al. [38] employ correlational multimodal VAE through a triplet Siamese network for social image representation. Existing Text-based supervised fake news detection methods take the textual information of news as input to detect fake news. These methods often only focus on supervised fake news detection methods when there is enough labeled data.

Adversarial training is a meaningful way to enhance the robustness of neural networks. During adversarial training, samples are mixed with small perturbations, and the neural network is then adapted to the changes, making it robust to adversarial examples. Wang et al. [40] employed adversarial neural networks for multi-modal events which can generalize well for timely events. It consists of multimodal feature extractor, fake news detector, and event discriminator. Song et al. [41] proposed an adversarial multimodal framework for fake news detection which uses a knowledge augmented transformer to encode the information of news text. Wu et al. [42] used adversarial networks to reduce irrelevant features from the extracted features for information credibility evaluation.

3. Approach

We present an exponential Pseudo-Siamese network, a variation of the classic Siamese network in this paper. We innovatively exploit the specific contrastive loss for text information, which greatly improves the performance. In addition, adversarial training is embedded into the Siamese network to make the model more robust against perturbation.

3.1. Pseudo Siamese Network

The Siamese network usually contains more than two sub-networks, and the weights are shared between those sub-networks, including common parameters, configuration, and other information. It is a special neural network architecture. The parameter update is generally updated across subnet as displayed on the left-hand side of Figure 1. Finally, the Siamese network outputs a distance (e.g., Euclidean distance) to calculate the similarity of inputs. The more similar the two inputs are, the smaller this distance is. The figure shows the situation of two sub-networks, and some Siamese networks will have multiple sub-networks. The following Siamese network refers to the situation of two sub-networks.

The Siamese networks are well-known for the study of discovering similarities or associations between two comparable things. Bromley et al. [35] use Siamese for signature verification on American checks in order to determine whether the two signatures belong to the same person. The Siamese network is also used for scoring the repeater's performance in the paraphrase score judging system. In this case, the input is two sentences, and the output is the score. From these two cases, the Siamese network generally employs two sub-networks to process two inputs, and another module is used to integrate the output of the sub-networks to obtain the final result.

Siamese network architectures can achieve excellent results in these tasks because of the following reasons:

(1) First, sharing weights among sub-networks means that only a few parameters need to be trained, and less data are required. In addition, the tendency of overfitting can be reduced; (2) The sub-network is essentially a representation of the input. Therefore, it is reasonable to use similar models to process similar input types, such as similar sentences or signatures in each case of the previous cases.

In natural language processing, some recent studies have used Siamese architectures [43–46]. Das et al. [43] used the Siamese network to seek the semantic alikeness between the target and the generated questions. Shonibare et al. [46] employ Siamese and Triplet neural network architectures based on BERT(Bidirectional Encoder Representations from Transformers) to embed text into a vector.

The classical Siamese networks cannot directly solve the fake news stance detection in our task. The reason is that the headline is very short, and most of them contain less than 40 words. However, the text body is much longer, which contains much more information than the headline. In the classical Siamese networks, two subnetworks use shared parameters supposing that the two inputs of classical Siamese networks are similar in length and structure. The performance of classical Siamese networks on fake news is bad, and hence, we make the two branches of the Siamese network not share parameters with each other.

Different from the Siamese network, the left and right sides do not share weights, but two different neural networks. For the Pseudo-Siamese network, both sides can be different neural networks (such as one is LSTM, one is CNN) or the same type of neural network. As displayed on the right-hand side of Figure 1, the left branch deals with only the headline, while the right branch merely copes with the text bodies. Experiments show that the proposed Pseudo Siamese network outperforms the classical Siamese network in fake news.

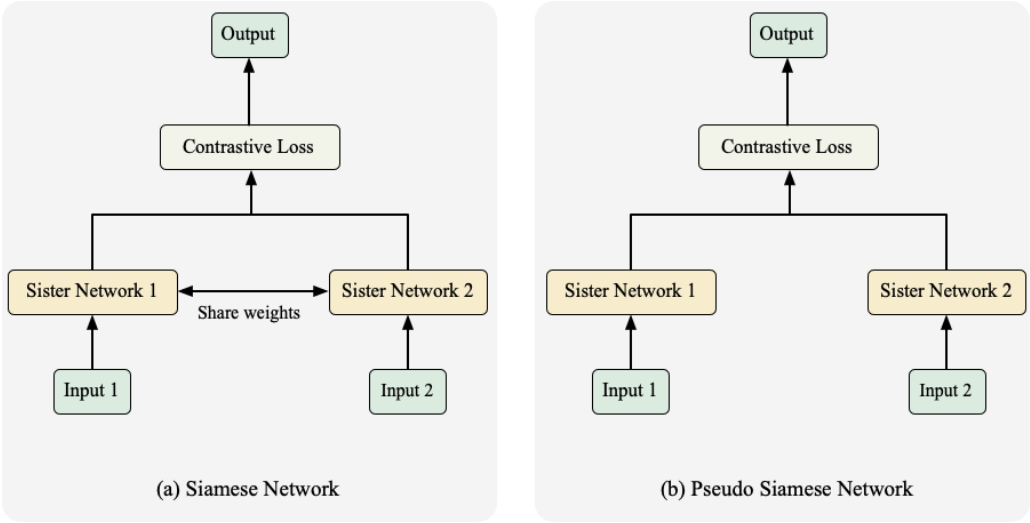

Figure 1. The architecture of the Siamese network.

3.2. Model Architecture

We present our exponential Pseudo-Siamese network architecture in detail in this section. We employ two parallel bidirectional LSTM to extract latent features from both news headline and text body at the same time. At last, we intend to combine news headline and text body representations to calculate the distance for fake news stance detection.

As Figure 2 shows, there are two major branches in our model, i.e., the headline and the text body branch. In each branch, news headline or text body word sequence as inputs, latent features are extracted by subnetwork for final predictions. We present our method by explaining the following four questions: (1) How can the latent features from the news text be obtained? (2) Why do we choose bidirectional LSTM as a subnetwork of the model? (3) How can the headline and text body features be combined?, and (4) Why do we add some noise into the output of the embedding layer? The symbols in this paper are shown in Table 1.

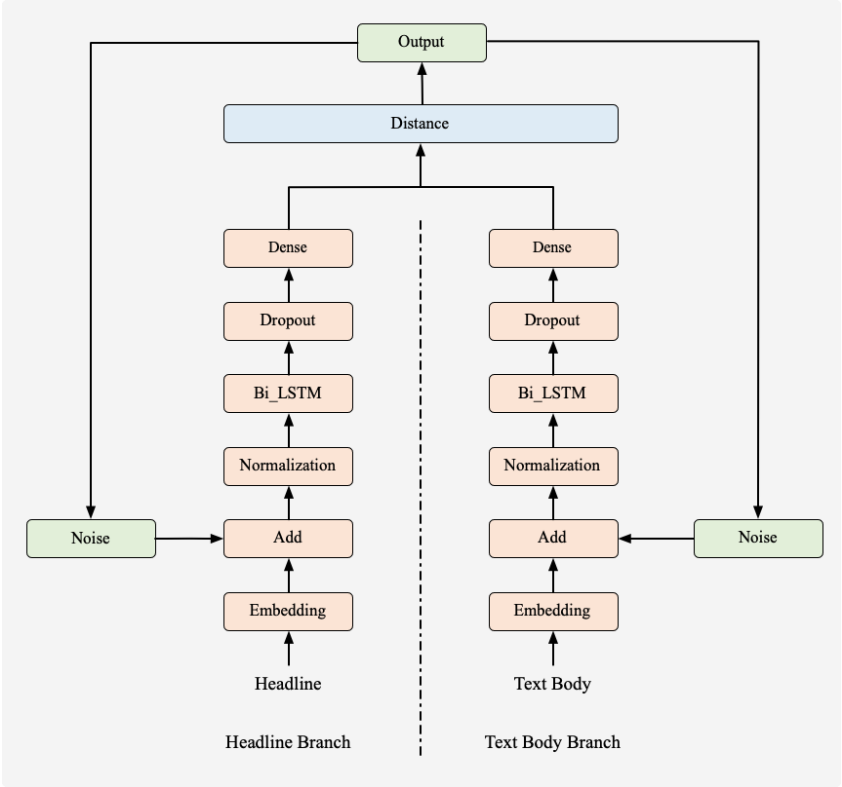

Figure 2. Model architecture.

Table 1. Symbols in this paper.

Symbol	Symbol Name
T_h	Word Sequence of News Headline
T_b	Word Sequence of News Text Body
V_h	Word Vector of Headline
V_b	Word Vector of Text Body
X_h	Headline Feature of Headline Branch
X_b	Text Body Feature of Text Body Branch
Y	Label of News Pair
w	Parameters of Model

3.2.1. Headline Branch

For the headline branch, the input is the word sequence of news headline T_h. In this sequence, every word is represented by its index number in the English Dictionary. Through the embedding layer, with the pre-trained GloVe word vectors as weights, such index numbers are converted into word vectors V_h, which represent latent features of headlines. After adding some noise on word vectors, these vectors are going to trained by the subnetwork. In our model, we employ bidirectional LSTM as the subnetworks. Long Short-Term Memory (LSTM) [47,48] is a special class of a recurrent neural network. Due to the special memory system, LSTM is fit for handling and forecasting things for extremely long periods. The recurrent neural network (RNN) is usually used to handle an input series of arbitrary size through the hidden state unit h_t. At each time step t, the inputs of the hidden unit are input vectors x_t (e.g., word vectors). The RNN obtains at time t and its last output vectors y_{t-1}. Then, this unit outputs vectors y_t based on the following equation, where W, U, and B are parameters of the hidden unit:

$$y_t = tanh(W * x_t + U * y_{t-1} + B) \quad (1)$$

With recursion of this process, RNN can pass information from one step to the next of the network and connects previous information to carry out the present calculation. However, if the sequence is too long and the needed previous information is too far, RNN might fail to find it, which leads to the gradient vector growing or decaying exponentially during training [49]. However, the vanishing or explosion of the gradient makes it hard for the RNN in a fake news stance detection task. The LSTM can address this problem through introducing a memory cell. Compared with RNN, LSTM can perform better in longer sequences. Here, we use Zaremba's version [50] to explain LSTM's process.

At every time step t, the LSTM unit is a set of vectors in $\mathbb{R}^d$ (d is the memory dimension of the LSTM). Unlike RNN, which only has one transmission state h_t, LSTM has two transmission states: one is memory cell state c_t, and the other is a hidden state h_t. The range of the gating vectors i_t, f_t, and o_t is [0, 1]. Specifically, the calculation formula for special time step t of LSTM is as follows:

$$\begin{aligned} i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + B^{(i)}) \\ f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + B^{(f)}) \\ o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + B^{(o)}) \\ u_t &= tanh(W^{(u)}x_t + U^{(u)}h_{t-1} + B^{(u)}) \\ c_t &= i_t \odot u_t + f_t \odot c_{t-1} \\ h_t &= o_t \odot tanh(c_t) \end{aligned} \tag{2}$$

where f_t, i_t, and o_t are forget gate, input gate, and output gate at time step t, respectively. σ is the logistic sigmoid function, and $\odot$ is elementwise multiplication. First, the forget gate combines the previous hidden state h_{t-1} with the current input x_t, and decides which old information to discard through the sigmoid function. The sigmoid value range is (0, 1), which means that the value closed to 0 is discarded and closed to 1 is kept. The input gate i_t and tanh function determine what new information is updated. Next, combine the forget gate and the updated information to obtain the cell state c_t at the current moment. Finally, the output gate multiple with the tanh value of cell state to determine which information is output.

According to the characteristic of LSTM, it is well-suited to learn some advanced features from time series, which is suitable to process word vectors. However, LSTM's output is only based on the previous and current status, which does not take future status into account. To make up for that shortage, we use bidirectional LSTM as a subnetwork. Bidirectional LSTM is based on the idea that the output does not simply depend on the previous output in the sequence, but is also related to future elements. A bidirectional LSTM [51] comprises two LSTMs running simultaneously. One LSTM is designed to process the regular input sequence, while the other LSTM is designed to process the opposite direction of the input sequence. The hidden layer needs to save two values at time step t. One participates in the forward calculation, and the other participates in the reverse calculation. Thus, the bidirectional LSTM can understand the sentences better and output headline feature X_h.

3.2.2. Text Body Branch

The architecture of the text body branch is similar to the headline branch and has its own parameters. In addition, its input is a word sequence of news text body T_b, and the output is text body feature X_b. The subnetwork of this branch also uses the bidirectional LSTM network.

3.2.3. Exponential Distance

In a classical Siamese network, two branches output two vectors, and the network outputs the Euclidean distance of two vectors at last. However, in our multi-class classification, the Euclidean margin between two classes is too small, which harms the model's

performance and makes tuning the model's parameters harder. Thus, we calculate the exponential distance of two branches' output X_h and X_b as model output. It can increase two classes' margins effectively.

3.2.4. Adversarial Training

As is shown in Figure 2, before entering the subnetwork of each branch, the word vectors V_h and V_b add some noise, which is called adversarial training and can make the proposed model more robust to perturbation. Goodfellow et al. [52] proposed the adversarial network architectures in 2014. It has been used in many ways. Huang et al. [53,54] incorporate the attention mechanism and the adversarial networks for multimodal representation. We utilize $X = (X_h, X_b)$ as the input pairs, y as the label, and W as the parameter of the neural networks. The adversarial training loss is as follows:

$$\begin{aligned} &-log\ \ f(y|x_h + r^h_{adv}; W)\ \ , where \\ &r^h_{adv} = \underset{r,||r||\leq\epsilon}{argmin}\ \ log(p(y|x_h + r; \hat{W})) \end{aligned} \tag{3}$$

$$\begin{aligned} &-log\ \ f(y|x_b + r^b_{adv}; W)\ \ , where \\ &r^b_{adv} = \underset{r,||r||\leq\epsilon}{argmin}\ \ log(p(y|x_b + r; \hat{W})) \end{aligned} \tag{4}$$

r^h_{adv} and r^b_{adv} are the perturbation on headline and on text body, respectively. The $\hat{W}$ is a constant value of parameter W in the back-propagation process. The purpose of the perturbation is to add perturbation on the input and challenge the model to be robust to learn something in the most difficult situations. The r_{adv} is the worst case perturbations on the model. Miyato et al. [55] proposed an approximation algorithm to estimate the r_{adv} by linearizing $log\ p(y|x; \hat{W})$. A L_2 norm is used to normalize the perturbation. In addition, ϵ is the intensity of perturbation:

$$r_{adv} = -\epsilon g/||g||_2, \text{where } g\ =\ \nabla_x L \tag{5}$$

Then, we can derive the loss for headline and text body branch, respectively:

$$L^h_{adv}(w) = -\frac{1}{N}\sum_{n=1}^{N} logf(y_n|x_h, n + r_{adv,n}; w) \tag{6}$$

$$L^b_{adv}(w) = -\frac{1}{N}\sum_{n=1}^{N} logf(y_n|x_b, n + r_{adv,n}; w) \tag{7}$$

3.3. Contrastive Loss

The traditional machine learning loss function is to sum over all the differences of samples between the predicted value and true value. The loss function of the Siamese network is designed based on the distance between pairs of samples. Suppose that T_h is the word sequence of the news headline, and T_b is the word sequence of news text body. T_h and T_b are inputs of the model. X_h and X_b are vectors output by the two subnetworks of the Siamese network. The distance function output by the Siamese network is usually defined as the Euclidean distance between the X_h and X_b. Y is the label of each [T_h, T_b] pair. For a traditional Siamese network, if the pair is dissimilar, $Y = 0$. Otherwise, $Y = 1$:

$$D_w(X_h, X_b) = e^{-\alpha||\hat{X}_h - \hat{X}_b||_2^2} \tag{8}$$

where $\hat{X}_h$ and $\hat{X}_b$ are the the partial derivative of X_h and X_b, respectively:

$$(\hat{X}_h, \hat{X}_b) = -logf(y|x_h + r^h_{adv}, x_b + r^b_{adv}, W) \tag{9}$$

The loss function is as follows. m is the number of samples, and w the parameters of model. $(Y, X_h, X_b)^i$ is the i-th sample, which is composed of a [HEADLINE, TEXT BODY] pair and a label:

$$L(w) = \sum_{i=1}^{m} L(w, (Y, X_h, X_b)^i) \tag{10}$$

$$L(w, (Y, X_h, X_b))^i = (1 - Y) * L_D(D_W^i) + Y * L_S(D_W^i) \tag{11}$$

$$L_D(D_W^i) = max(0, margin - D_W^i)^2 \tag{12}$$

$$L_S(D_W^i) = (D_W^i)^2 \tag{13}$$

$L_S(D_W^i)$ is the partial loss function for a similar pair, while $L_D(D_W^i)$ is the partial loss function for a dissimilar pair. When Y equals 1, the inputs are similar and the distance between them should be as small as possible. Thus, $L(w, (Y, X_h, X_b))^i$ equals $(D_W^i)^2$, which means that the loss of this sample is directly proportional to the square of distance. When Y equals 0, the inputs are dissimilar, and the distance between them should be as large as possible. Hence, we set a positive number *margin* and, unless the distance of two dissimilar inputs is more significant than this *margin*, the loss will not reach the minimum value.

Based on the loss of the traditional Siamese network, we design a new loss of multi-class classification for our model. The labels of the dataset are encoded with a number as $y = \{0, 1, 2, 3\}$, which represent consistent, conflicting, neutral, and unrelated, respectively. The indicator function is as follows:

$$\begin{aligned} f_0(y) &= (1 - y)(2 - y)(3 - y)/6 \\ f_1(y) &= (y - 0)(2 - y)(3 - y)/2 \\ f_2(y) &= (0 - y)(1 - y)(3 - y)/2 \\ f_3(y) &= (y - 0)(y - 1)(y - 2)/6 \end{aligned} \tag{14}$$

The loss function is as follows:

$$\begin{aligned} L = &\alpha * f_0(y) * max(0, 0 - g(D_w))^2 + \\ &\beta * f_1(y) * max(0, (g(D_w) - l_2) * (g(D_w) - l_3))^2 + \\ &\gamma * f_2(y) * max(0, l_1 - g(D_w))^2 + \\ &\delta * f_3(y) * max(0, (g(D_w) - l_4) * (g(D_w) - l_5))^2 \end{aligned} \tag{15}$$

α, β, γ, δ are the weights of each class. (l_2, l_3), (l_4, l_5) and $(l_1, +\infty)$ are intervals of distance for each class. $g(D_w)$ is a transformation of D_w^2. Because of different classes corresponding to different partial loss functions, $f_i(y)$ is used to choose the right partial loss function. Similarly, a partial loss function will reach a minimum value only when the distance of each sample is in a corresponding interval.

4. Experiments

4.1. Case Study

Our dataset is from a Fake News Challenge (TNC) contest (http://www.fakenewschallenge.org/?imm_mid=0ed405&cmp=em-data-na-na-newsltr_ai_20170213 (accessed on 1 February 2023)). The data set consists of news headline and text body, and one sample is a [HEADLINE, TEXT BODY] pair, as is shown in Table 2. In this table, column "type" describes a relationship between headline and text body. For example, in the last news, what its headline and text body talk about are different things, so the value of "type" is "unrelated". For the news with type of "unrelated" or "conflicting", we judge them as fake news.

Table 2. Case study.

Headline	Text Body	Type
"Robert Plant Ripped up $800M Led Zeppelin Reunion, Contract"	"... Led Zeppelin's Robert Plant turned down 500 MILLION to reform supergroup. ..."	Consistent
	"... No, Robert Plant did not rip up an $800 million deal to get Led Zeppelin back together. ..."	Conflicting
	"... Robert Plant reportedly tore up an $800 million Led Zeppelin reunion deal. ..."	Neutral
	"... Richard Branson's Virgin Galactic is set to launch SpaceShipTwo today. ..."	Unrelated

In our experiment, we use data augmentation to prevent overfitting. The most types of news we collect are "neutral" and "consistent". By adding some negative words in a text body sentence, we obtain some "conflicting" news from "consistent" news. With permutation of the headline and text body, we can acquire lots of "unrelated" news. After our counting, about 70% of news is "unrelated" news. The percentage of each type of news is shown in Table 3. In addition, we totally obtain 49,979 pairs of news.

Table 3. Distribution of data set.

Type of News	Percentage
Consistent	7.41%
Conflicting	2.04%
Neutral	17.74%
Unrelated	72.81%

4.2. Experimental Setup

We employ 70% of the dataset for training, 10% of the dataset for validation, and the rest for testing. All experiments used Keras on GPUs of two NVIDIA GTX 1080Ti. We use Adam optimizer [56] with an initial learning rate of 0.00001 to train. Batch Normalization [57] is also employed to reduce the occurrence of covariate shift within the neural networks. To evaluate our method on the testing set, we choose the snapshot of the trained model which performs best on the validation set. All of the experiments are conducted at least 10 times individually in our experiment. As is shown in Figure 2, our model uses bidirectional LSTM as the subnetwork. The model outputs the distance which can measure the relevance between the headline and text body. Before being imported into the model, every word in the headline and text body is going to be transformed into its index number in the dictionary. In addition, each headline sequence is padded to 40 words, and each text body sequence is padded to 400 words. We elaborate the set process of the parameters of every layer in two aspects. As Table 4 shows, our compared method is a Gradient Boosting (GB) Classifier, Gradient Boosting Decision Tree (GBDT), and CNN. The most significant advantage of GBDT is that it prevents overfitting, has a solid expressive ability, and does not require complex feature engineering and feature transformation. It has strong interpretability and can automatically sort feature importance. However, Boost is a serial process, which is not easy to parallelize, and has high computational complexity. At the same time, it is not suitable for high-dimensional sparse features. If too many features exist, each regression tree will consume a lot of time.

(1) Headline branch: For this branch, the dimension of GloVe embedding is set to 100. We introduce detailed information on selecting parameters in the sensitivity analysis section. In adversarial training, after a batch of data are trained, the perturbation will be calculated and added into the next batch of embedding output. In the bidirectional LSTM layer, the number of units is set to 128. Then, a dropout whose drop rate is 0.1 can avoid overfitting. In the end of this branch, we add a dense layer with 128 neurons. The outputs of the headline branch and text body branch are combined by calculating the exponential

distance. In our experiment, the distance is $e^{2.5-\|O_1-O_2\|}$, where O_1 and O_2 are outputs of headline branch and text body branch, respectively.

(2) Text Body branch: For this branch, the dimension of GloVe embedding is set to 100. The method of generating perturbation is the same as headline branch. The bidirectional LSTM layer has 128 units and the dropout, and dense layers are the same as the headline branch.

Table 4. Results of experiments.

Model	FNC Score	Data Size (News Pair)	Hand-Coded Features	Table Note
GB Classifier	79.53	49,979	Word(ngram) Overlap Features and Indicator Features for Polarity and Refutation	Models Specification. GBDT: Gradient Boosting Decision Tree. CS: Cosine Siamese network. ES: Exponential Siamese network. AT: Adversarial Training.
CNN + GBDT	82.02	49,979	Count, TF-IDF, Sentiment	
CS	89	49979	Weighted Bag of Word	
ES + LSTM	90.12	49,979	None	
ES + LSTM + AT	89.12	33,000	None	
ES + LSTM + AT	93.40	49,979	None	

4.3. Evaluation

Because our data are imbalanced, a novel scoring system is designed (http://www.fakenewschallenge.org (accessed on 1 February 2023)). We divide the final evaluation score into three classes based on whether the [HEADLINE, TEXT BODY] pair in the test dataset has a related target label or not. If there is an unrelated label, the final evaluation score is 0.25. If there is a related label, the final evaluation score is 1.00. Otherwise, the final evaluation score will be 0.00. The mean of every pair's final evaluation score is the final score to evaluate the model's performance.

4.4. Experimental Results

We make a comparison about our results and several competitive methods in Table 4. The baseline model uses hand-coded features and a Gradient Boosting classifier. In the FNC contest, the score of best model is 82.02 (https://github.com/Cisco-Talos/fnc-1 (accessed on 1 February 2023)). A Stanford team uses a Cosine Siamese network and weighted bag of words feature to obtain a score of 89 (https://web.stanford.edu/class/cs224n/reports/2759862.pdf (accessed on 1 February 2023)), which is their highest score. In the beginning, we just combine an Exponential Siamese network with bidirectional LSTM and obtain a score of 90.12. After adding the adversarial training into our model, with only 47% of data set, we can obtain a score of 89.12; however, with an entire data set, the score of our model is 93.40. Compared to other methods, the Siamese network is useful among tasks that involve finding similarity or a relationship between two similar things. In the fake news stance detection task, the headline is very short, and most of them contain less than 40 words, while the text body is much longer, which contains much more information than the headline. Therefore, Gradient Boosting Decision Tree and Cosine Siamese network are not handled well in these tasks.

4.5. Sensitivity Analysis

In this section, we study the effectiveness of several parameters in the proposed model: length of news headline and text body, exponential distance parameter, perturbation parameter ϵ, and the dropout probability.

(1) Length of News Headline and Text Body: Before embedding layers, each headline sequence has the same number of words and so does text body sequences. We use grid search to choose the length of headline and text body. As is shown in Figure 3, the model performs best when each headline has 37–49 words, and each text body has 295–404 words. From the figure, we can see that the model will not perform best if the sequence is too long or too short. Specifically, we test our model with a 10-word headline and 100-word text body. Then, the FNC score is just 92.74, which means that a short sequence without enough information can not lead to the best FNC score. On the other hand, an overlong sequence often has redundant information, and the FNC score of the model does not increase significantly. Thus, we choose a 40-word headline and 400-word text body, which can result in a high FNC score and shorten the training time.

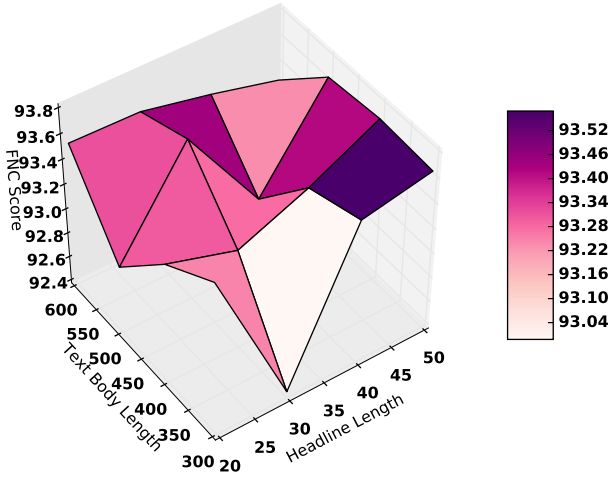

Figure 3. Length of news headline and text body and performance of a model.

(2) Exponential distance parameter: In our model, the exponential distance is set to $e^{2.5-\|O_1-O_2\|}$, and "2.5" is the exponential distance parameter. The bigger this parameter is, the faster the curve of exponential distance declines. At the beginning, we use the Euclidean distance and find that every category is close to each other and can not be classified effectively. After we use the exponential distance instead, it can expand the distance of two categories and contribute to classification. As is shown in Figure 4, we test a set of values of this parameter and find that, when it is in [1.5, 2.5], the model performs well. In this paper, we set the exponential distance parameter as 2.5.

(3) Perturbation parameter ϵ: In adversarial training, we use a parameter ϵ to control the intensity of perturbation. With the same data set but without the adversarial training, the FNC score of the model is 91.7583. Then, we test a set of ϵ values to see how the model's performance changes. As is shown in Figure 5, adversarial training can improve model's performance when ϵ is in $[1 \times 10^{-5}, 1 \times 10^{-1}]$. However, if the perturbation is too intense (e.g., ϵ value is 1×10^{2}), the adversarial training will worsen the performance of the model.

(4) Dropout probability: We analyze the dropout probabilities, and the dropout layer is shown in Figure 2 (model architecture). In Figure 6, D_α is the probability of the dropout layer connected to the text body Bi_LSTM layer, while D_β is the probability of dropout layer connected to the headline Bi_LSTM layer. During the experiment, we employ the grid search to select the appropriate dropout probability. Our experiments show that, when D_α is in the range of [0.1, 0.3] and D_β is also in the range of [0.1, 0.3], the model performs well. Therefore, we set the two dropout probabilities to (0.1, 0.1), which can improve generalization ability of the model and speed up the training process in this paper.

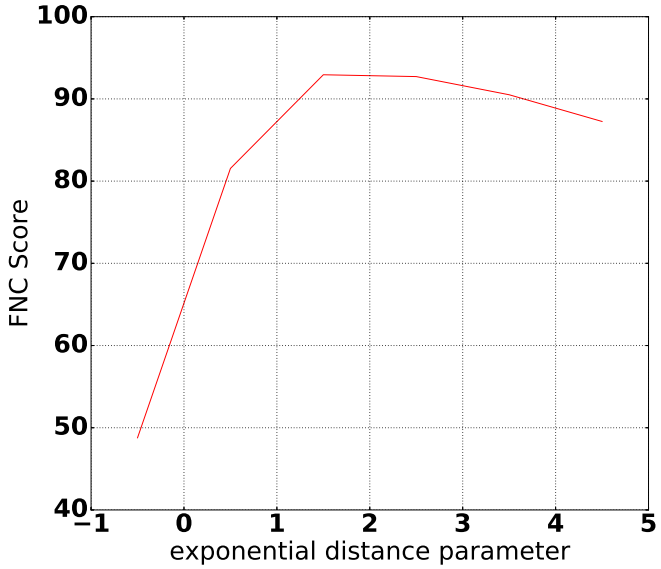

Figure 4. Exponential distance parameter and performance of the model.

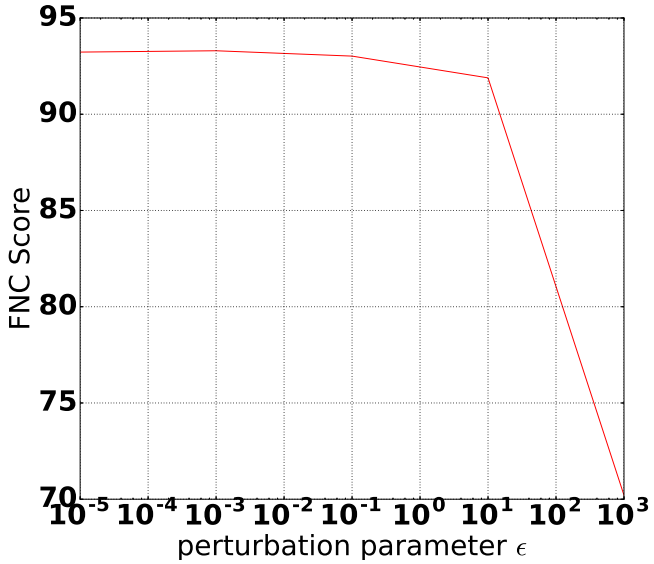

Figure 5. Perturbation parameter and performance of the model.

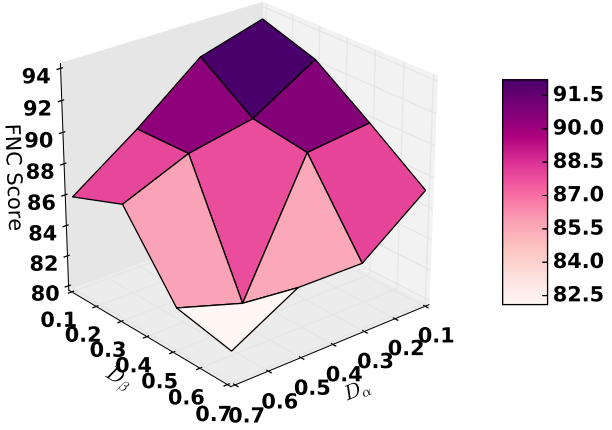

Figure 6. Dropout probabilities (D_α, D_β) and the performance of the model.

5. Conclusions

With the rapid development of social networks, fake news spread all over the world in a very short time. It is very important to identify this fake news in time, i.e., fake news detection. In this paper, we focus on the fake news stance detection, which detects fake news by evaluating the relevance between news headline and text bodies. We novelly propose the Pseudo-Siamese network to project the features of headline and text bodies into the same space. Then, an exponential projection function is applied to project the points in high-dimensional space into the two-dimensional space. We conduct experiments on a fake news challenge dataset. The experimental results outperform many competitive baselines. The highest score of the proposed model is 93.40.

The Siamese network is also a promising solution to fuse multi-view data in order to evaluate the relevance between image and text. Furthermore, for fake news detection tasks, it is difficult to collect a large amount of data. Generative Adversarial Nets (GAN) is a possible way to generate real and fake headlines from text body, which may greatly improve the performance.

In the future work, we will continue to investigate more specious fake news. First, some fake news consists of images which are irrelevant to its text bodies or headlines. Data fusion is helpful to identify such fake news. Second, even though some news' headlines and text bodies are consistent, some details in text bodies are changed. Such news are also fake news. For instance, the news is "... Robert Plant reportedly tore up an $800 billion Led Zeppelin reunion deal. ..." but the truth is "... Robert Plant reportedly tore up an $800 million Led Zeppelin reunion deal. ...". Because a figure has changed, this news is fake. We hope that the sentiment analysis method could detect this kind of fake news, for such fake news usually contains some grandiloquent sentiment sentences.

Author Contributions: Z.Z. and Y.Y. conceived and designed the study, collected and organized data, wrote the paper, and conducted data analysis. Z.L. reviewed and edited the manuscript. All authors read and approved the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments: An earlier version of this paper has been presented as a preprint online [58].

Conflicts of Interest: The authors declare that there are no conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Chang, J.; Lefferman, J.; Pedersen, C.; Martz, G. When fake news stories make real news headlines. *Nightline. ABC News* **2016**. Available online: https://abcnews.go.com/Technology/fake-news-stories-make-real-news-headlines/story?id=43845383 (accessed on 29 November 2016).
2. Rao, S.; Verma, A.K.; Bhatia, T. A review on social spam detection: Challenges, open issues, and future directions. *Expert Syst. Appl.* **2021**, *186*, 115742. [CrossRef]
3. Bindu, P.V.; Mishra, R.; Thilagam, P.S. Discovering spammer communities in twitter. *J. Intell. Inf. Syst.* **2018**, *51*, 503–527. [CrossRef]
4. Gangavarapu, T.; Jaidhar, C.D.; Chanduka, B. Applicability of machine learning in spam and phishing email filtering: Review and approaches. *Artif. Intell. Rev.* **2020**, *53*, 5019–5081. [CrossRef]
5. Ren, Y.; Ji, D. Neural networks for deceptive opinion spam detection: An empirical study. *Inf. Sci.* **2017**, *385*, 213–224. [CrossRef]
6. Kaur, R.; Singh, S.; Kumar, H. Rise of spam and compromised accounts in online social networks: A state-of-the-art review of different combating approaches. *J. Netw. Comput. Appl.* **2018**, *112*, 53–88. [CrossRef]
7. Rathore, S.; Loia, V.; Park, J.H. SpamSpotter: An efficient spammer detection framework based on intelligent decision support system on Facebook. *Appl. Soft Comput.* **2018**, *67*, 920–932. [CrossRef]
8. Ferrara, E. The history of digital spam. *Commun. ACM* **2019**, *62*, 82–91. [CrossRef]
9. Harada, J.; Darmon, D.; Girvan, M.; Rand, W. Prediction of Elevated Activity in Online Social Media Using Aggregated and Individualized Models. In *Trends in Social Network Analysis*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 169–187.

10. Fu, M.; Feng, J.; Lande, D.; Dmytrenko, O.; Manko, D.; Prakapovich, R. Dynamic model with super spreaders and lurker users for preferential information propagation analysis. *Phys. A Stat. Mech. Its Appl.* **2021**, *561*, 125266. [CrossRef]
11. Tian, S.; Yin, X.C.; Su, Y.; Hao, H.W. A unified framework for tracking based text detection and recognition from Web videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 542–554. [CrossRef]
12. Zhang, Z.; Zhang, C.; Shen, W.; Yao, C.; Liu, W.; Bai, X. Multi-oriented text detection with fully convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4159–4167.
13. Qin, X.; Zhou, Y.; Guo, Y.; Wu, D.; Wang, W. FC 2 RN: A Fully Convolutional Corner Refinement Network for Accurate Multi-Oriented Scene Text Detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, 6–11 June 2021; pp. 4350–4354. [CrossRef]
14. Chopra, S.; Hadsell, R.; Lecun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 539–546.
15. Augenstein, I.; Rocktäschel, T.; Vlachos, A.; Bontcheva, K. Stance Detection with Bidirectional Conditional Encoding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, TX, USA, 1–4 November 2016; Su, J., Carreras, X., Duh, K., Eds.; The Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 876–885. [CrossRef]
16. Zotova, E.; Agerri, R.; Rigau, G. Semi-automatic generation of multilingual datasets for stance detection in Twitter. *Expert Syst. Appl.* **2021**, *170*, 114547. [CrossRef]
17. Al-Ghadir, A.I.; Azmi, A.M.; Hussain, A. A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments. *Inf. Fusion* **2021**, *67*, 29–40. [CrossRef]
18. Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.D.; Cherry, C. SemEval-2016 Task 6: Detecting Stance in Tweets. In Proceedings of the SemEval@ NAACL-HLT, San Diego, CA, USA, 16–17 June 2016; pp. 31–41.
19. Du, J.; Xu, R.; He, Y.; Gui, L. Stance Classification with Target-Specific Neural Attention Networks. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017.
20. Yang, Y.; Wu, B.; Zhao, K.; Guo, W. Tweet Stance Detection: A Two-stage DC-BILSTM Model Based on Semantic Attention. In Proceedings of the 5th IEEE International Conference on Data Science in Cyberspace, DSC 2020, Hong Kong, China, 27–30 July 2020; pp. 22–29. [CrossRef]
21. Wu, Y.; Ngai, E.W.T.; Wu, P.; Wu, C. Fake online reviews: Literature review, synthesis, and directions for future research. *Decis. Support Syst.* **2020**, *132*, 113280. [CrossRef]
22. Mohawesh, R.; Tran, S.N.; Ollington, R.; Xu, S. Analysis of concept drift in fake reviews detection. *Expert Syst. Appl.* **2021**, *169*, 114318. [CrossRef]
23. Abri, F.; Gutiérrez, L.F.; Namin, A.S.; Jones, K.S.; Sears, D.R.W. Linguistic Features for Detecting Fake Reviews. In Proceedings of the 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020, Miami, FL, USA, 14–17 December 2020; Wani, M.A., Luo, F., Li, X.A., Dou, D., Bonchi, F., Eds.; pp. 352–359. [CrossRef]
24. Guillory, J.; Hancock, J.T. The effect of Linkedin on deception in resumes. *Cyberpsychol. Behav. Soc. Netw.* **2012**, *15*, 135–140. [CrossRef]
25. Noekhah, S.; Salim, N.B.; Zakaria, N.H. Opinion spam detection: Using multi-iterative graph-based model. *Inf. Process. Manag.* **2020**, *57*, 102140. [CrossRef]
26. Xu, G.; Hu, M.; Ma, C. Secure and smart autonomous multi-robot systems for opinion spammer detection. *Inf. Sci.* **2021**, *576*, 681–693. [CrossRef]
27. Byun, H.; Jeong, S.; Kim, C. SC-Com: Spotting Collusive Community in Opinion Spam Detection. *Inf. Process. Manag.* **2021**, *58*, 102593. [CrossRef]
28. Ojo, A.K. Improved model for detecting fake profiles in online social network: A case study of twitter. *J. Adv. Math. Comput. Sci.* **2019**, *33*, 1–17. [CrossRef]
29. Awan, M.J.; Khan, M.A.; Ansari, Z.K.; Yasin, A.; Shehzad, H.M.F. Fake profile recognition using big data analytics in social media platforms. *Int. J. Comput. Appl. Technol.* **2021**, *68*, 215–222. [CrossRef]
30. Joshi, S.; Nagariya, H.G.; Dhanotiya, N.; Jain, S. Identifying Fake Profile in Online Social Network: An Overview and Survey. In Proceedings of the International Conference on Machine Learning, Image Processing, Network Security and Data Sciences, Silchar, India, 30–31 July 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 17–28.
31. Toma, C.L.; Hancock, J.T. What lies beneath: The linguistic traces of deception in online dating profiles. *J. Commun.* **2012**, *62*, 78–97. [CrossRef]
32. Rubin, V.L.; Chen, Y.; Conroy, N.J. Deception detection for news: Three types of fakes. *Proc. Assoc. Inf. Sci. Technol.* **2015**, *52*, 1–4. [CrossRef]
33. Zahedi, F.M.; Abbasi, A.; Chen, Y. Fake-Website Detection Tools: Identifying Elements that Promote Individuals' Use and Enhance Their Performance. *J. Arab. Islam. Stud.* **2015**, *16*, 2. [CrossRef]
34. Chopra, S.; Jain, S.; Sholar, J.M. *Towards Automatic Identification of Fake News: Headline-Article Stance Detection with LSTM Attention Models*; Tech. Rep.; Stanford Univ.: Stanford, CA, USA, 2017.

35. Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; Shah, R. Signature verification using a "Siamese" time delay neural network. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 28 November–1 December 1994; pp. 737–744.
36. Ji, Y.; Zhang, H.; Jie, Z.; Ma, L.; Wu, Q.M.J. CASNet: A Cross-Attention Siamese Network for Video Salient Object Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2676–2690. [CrossRef] [PubMed]
37. Chen, H.; Wang, Y.; Zheng, K.; Li, W.; Chang, C.; Harrison, A.P.; Xiao, J.; Hager, G.D.; Lu, L.; Liao, C.; et al. Anatomy-Aware Siamese Network: Exploiting Semantic Asymmetry for Accurate Pelvic Fracture Detection in X-ray Images. In *Lecture Notes in Computer Science, Proceedings of the Computer Vision—ECCV 2020—16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XXIII; Vedaldi, A., Bischof, H., Brox, T., Frahm, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12368, pp. 239–255. ._15. [CrossRef]
38. Huang, F.; Zhang, X.; Xu, J.; Zhao, Z.; Li, Z. Multimodal Learning of Social Image Representation by Exploiting Social Relations. *IEEE Trans. Cybern.* **2021**, *51*, 1506–1518. [CrossRef]
39. Fu, K.; Fan, D.P.; Ji, G.P.; Zhao, Q.; Shen, J.; Zhu, C. Siamese network for RGB-D salient object detection and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *4*, 5541–5559. [CrossRef]
40. Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; Gao, J. Eann: Event adversarial neural networks for multi-modal fake news detection. In Proceedings of the 24th ACM Sigkdd International Conference On Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 849–857.
41. Song, C.; Ning, N.; Zhang, Y.; Wu, B. Knowledge augmented transformer for adversarial multidomain multiclassification multimodal fake news detection. *Neurocomputing* **2021**, *462*, 88–100. [CrossRef]
42. Wu, L.; Rao, Y.; Nazir, A.; Jin, H. Discovering differential features: Adversarial learning for information credibility evaluation. *Inf. Sci.* **2020**, *516*, 453–473. [CrossRef]
43. Das, A.; Yenala, H.; Chinnakotla, M.K.; Shrivastava, M. Together we stand: Siamese Networks for Similar Question Retrieval. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 378–387.
44. Lu, Z.; Li, H. A deep architecture for matching short texts. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 1367–1375.
45. Dadashov, E.; Sakshuwong, S.; Yu, K. Quora Question Duplication. 2017. Available online: https://sukolsak.com/files/quora_question_duplication.pdf (accessed on 1 February 2023).
46. Shonibare, O. ASBERT: Siamese and Triplet network embedding for open question answering. *arXiv* **2021**, arXiv:2104.08558.
47. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
48. Wikipedia. Long Short-Term Memory—Wikipedia, The Free Encyclopedia. 2017. Available online: https://en.wikipedia.org/wiki/Long_short-term_memory (accessed on 10 October 2017).
49. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef]
50. Zaremba, W.; Sutskever, I. Learning to execute. *arXiv* **2014**, arXiv:1410.4615.
51. 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 (ASRU), Olomouc, Czech Republic, 8–12 December 2013; pp. 273–278.
52. Goodfellow, I.J.; Pougetabadie, J.; Mirza, M.; Xu, B.; Wardefarley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative Adversarial Nets. *Adv. Neural* **2014**, 2672–2680.
53. Huang, F.; Zhang, X.; Li, Z. Learning Joint Multimodal Representation with Adversarial Attention Networks. In Proceedings of the 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, 22–26 October 2018; Boll, S., Lee, K.M., Luo, J., Zhu, W., Byun, H., Chen, C.W., Lienhart, R., Mei, T., Eds.; ACM: New York, NY, USA, 2018; pp. 1874–1882. [CrossRef]
54. Huang, F.; Jolfaei, A.; Bashir, A.K. Robust Multimodal Representation Learning With Evolutionary Adversarial Attention Networks. *IEEE Trans. Evol. Comput.* **2021**, *25*, 856–868. [CrossRef]
55. Miyato, T.; Dai, A.M.; Goodfellow, I.J. Adversarial Training Methods for Semi-Supervised Text Classification. *arXiv* **2017**, arXiv:1605.07725.
56. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
57. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015; JMLR Workshop and Conference Proceedings; Bach, F.R., Blei, D.M., Eds.; Volume 37, pp. 448–456.
58. Zhou, Z.; Yang, Y.; Huang, F.; Li, Z.J. APSN: Adversarial Pseudo-Siamese Network for Fake News Stance Detection. *Res. Sq.* **2022**, 1–9. [CrossRef]