Complex & Intelligent Systems (2022) 8:3603–3612
https://doi.org/10.1007/s40747-021-00393-y

ORIGINAL ARTICLE

Span identification and technique classification of propaganda in news articles

Wei Li[1] · Shiqian Li[1] · Chenhao Liu[1] · Longfei Lu[1] · Ziyu Shi[1] · Shiping Wen[2]

Received: 18 March 2021 / Accepted: 30 April 2021 / Published online: 8 May 2021

Abstract

Propaganda is a rhetorical technique designed to serve a specific topic, which is often used purposefully in news article to achieve our intended purpose because of its specific psychological effect. Therefore, it is significant to be clear where and what propaganda techniques are used in the news for people to understand its theme efficiently during our daily lives. Recently, some relevant researches are proposed for propaganda detection but unsatisfactorily. As a result, detection of propaganda techniques in news articles is badly in need of research. In this paper, we are going to introduce our systems for detection of propaganda techniques in news articles, which is split into two tasks, Span Identification and Technique Classification. For these two tasks, we design a system based on the popular pretrained BERT model, respectively. Furthermore, we adopt the over-sampling and EDA strategies, propose a sentence-level feature concatenating method in our systems. Experiments on the dataset of about 550 news articles offered by SEMEVAL show that our systems perform state-of-the-art.

Keywords Propaganda detection · Neural network · Span identification · Technique classification · BERT

Introduction

Recently, with the development of related models in the field of natural language processing, research on propaganda detection also goes ahead, which originates from document level [1], then develops to sentence level [6,21] and now to fragment level [13,26]. At present, identifying those specific fragments which contain at least one propaganda technique and identifying the applied propaganda technique in the fragment are main tasks of the fragment-level propaganda detection. As an extension of text classification task in the field of natural language processing, there are many relevant advanced algorithms [8,10,12,19,27] which can be used for reference.

SEMEVAL, the most influential and largest semantic evaluation competition all over the world, provides a news article corpus in which fragments containing one out of 14 propaganda techniques [14,18] have been annotated as shown in Fig. 1. Based on this dataset, numerous researchers have sprung up putting forward a large quantity of algorithms to search the usages of propaganda techniques. Among the algorithms, the great mass of them are based on the popular language models such as ELMO [16], GPT [17] and especially BERT [3]. As shown in Fig. 2, BERT Model raised by Google outperforms previous methods in 11 NLP tasks. Undoubtedly, it has achieved state-of-the-art performance on multiple NLP benchmarks [22]. In our systems, we choose BERT as our basic model as well.

In this work, we introduce our systems for span identification and technique classification of propaganda in news articles. As for the span identification task, we have set forth two of architectures working on it. The first is the BERT-based binary classifier, and the other one is the BERT-based three-token type classifier. The latter is our second-to-none

✉ Wei Li
weili@std.uestc.edu.cn ; liwei9719@126.com

Shiqian Li
2816425039@qq.com

Chenhao Liu
2448373440@qq.com

Longfei Lu
2803570984@qq.com

Ziyu Shi
1604875799@qq.com

Shiping Wen
shiiping.wen@uts.edu.au

[1] University of Electronic Science and Technology of China, Chengdu, Sichuan, China

[2] University of Technology Sydney, Ultimo, Australia

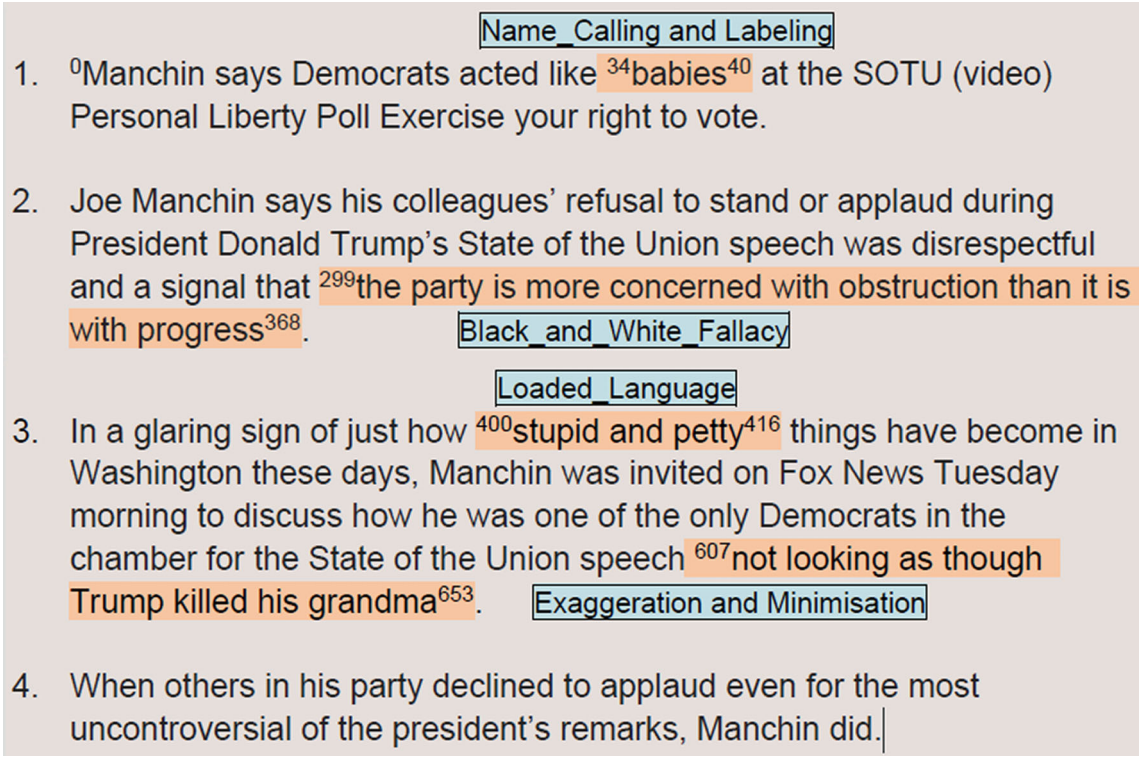

Fig. 1 The corpus of news articles which have been retrieved with the newspaper 3k library and sentence splitting has been performed automatically with NLTK sentence splitter

system. Besides joining the most popular BERT model, we have also optimized the sampling [2] process, combined EDA [23,24] to prevent the overfitting of our system and adopted the sentence-level feature concatenating (SLFC), in which case our model can learn characteristics better. As for the technique classification of propaganda task, we have designed a BERT-based architecture with a dimensionality reduction Full Connected (FC) layer and a linear classifier. Same as SI task, we have utilized EDA strategy in the data loading process. The final result in "Experiment and results" shows that it is very meaningful of our optimizing and improving of the pre-trained BERT model. At last, both of our systems for SI and TC have exceeded most of the existed models and made a breakthrough.

The contributions of our paper are as follows:

- We fine-tune the BERT with Linear layers and devise two accurate systems for the span identification and technique classification of propaganda in news articles.
- We change the binary sequence tagging task SI into a three-way classification task by adding 'invalid' token type and compare the binary tagging method with the three-token type method.
- We propose SLFC approach in SI system. To our best knowledge, it is the first work to integrate sentence-level classification features into each word.
- For our systems, we have obtained the optimal network parameters through experiments and comparative analysis.

Related work

The followings are the history and the correlative approaches about propaganda detection in news articles.

Propaganda detection

Propaganda techniques detection is born in the process of fake news detection. Some of the earlier workers judge a news article as authentic or not only according to its origin. As we can imagine that this approach is unscientific. Recently, with the rise of artificial intelligence and machine learning, propaganda detection has attracted researcher's eyes which promotes it to become a standalone research field.

In the early days of NLP neural networks, a bidirectional long-term short-term memory (BiLSTM) [5] layer was proposed to capture the semantic features of human language. Gradually, people began to utilize it to detect the using of propaganda in news articles. Initially, a corpus has been created for news articles automatically annotated with a novel multi-granularity neural network which is superior

to some powerful BERT-based baselines [14]. Simultaneously, Proppy [1], a system to unmask propaganda in online news, has appeared for document-level propaganda detection, which works by analysing various representations, from writing style and readability level to the presence of certain keywords. Later, to further improve the accuracy of detection, researchers began to pay attention to the detection of sentence level. Hou et al. proposed a model for sentence-level detecting which could understand semantic features of language better by constructing context-dependent input pairs (sentence-title pair and sentence-context pair) [6]. After the NLP4IF workshop, fragment-level classification (FLC) of propaganda occurred. For instance, different neural architectures (e.g., CNN, LSTM-CRF, and BERT) have been explored to further improve the effect of neural networks [15].

BERT-based model

In our experiment, we have been designing models specifically for SI and TC tasks based on BERT [3,4] model architecture, which incorporates the strength of the other language models. As shown in Fig. 2, BERT utilizes the transformer's attention mechanism [20] to decode the input word vector. Unlike the previous NLP models, BERT is able to run in parallel. More uniquely, the pre-training process of BERT includes two tasks, Masked Language Model (MLM) and Next Sentence Prediction (NSP), which make the BERT model more suitable for NLP tasks. After completing different pre-training and fine-tuning for different tasks, BERT has made great progress on many NLP tasks. Many researchers have discovered the huge potential of the two-stage new model (pre-training and then fine-tuning) on BERT. As a consequence, in recent years, based on BERT, many improved models occurred, such as MT-DNN [7], XLNET [25], ALBERT [11], etc.

In our system, using BERT is mainly for word feature extraction, thanks to that BERT adopts the popular feature extractor transformer, and also implements a bidirectional language model. It is the core concept of BERT to convert word into word vector input, which is added by Token Embedding, Segment Embedding, and Position Embedding to integrate the whole sentence semantics into each word in the same sentence. For our SI task, we process the obtained feature vector from BERT generator by incorporating sentence-level features into each word vector and then put it to a multi-class (prop, non-prop, invalid) classifier layer. To fit our TC task, we truncate the valid fragments and pad it for the latter FC layer and the classifier. Since there are two versions of BERT, taking our SI and TC tasks into account we use the 12-layer BERT pre-trained model as our basis.

Method

In this section, we will introduce the details of our solutions and show the model architectures designed for the span identification (SI) and technique classification (TC) tasks.

Data process

S On account of that only a small portion of the texts use propaganda in SI and some of the techniques rarely appear in the given fragments in TC which lead to the imbalance of dataset, we have proposed two methods aimed at these two problems.

Over-sampling In the task of SI, we utilize the over-sampling (OS) [2,26] method to get more balanced and suitable dataset for training. Since sentences with propaganda techniques are relatively few, we sample them with a higher probability, and the number of non-propaganda ones is correspondingly reduced considering the whole training process. Nevertheless, during our experiment, we find that if over-sampling is overused, the labeled part will be too much in the sample which will cause overfitting, and the F1-score will decline to an undesirable level as a result. Therefore, when training our model, we take the strategy that the first half of the epoch uses the over-sampling and the latter part uses the sequential sampling. While TC is merely a classification task and each fragment in the given dataset corresponds to a specific propaganda technique, over-sampling is superfluous.

Data augmentation Since the pre-trained BERT model is easy to overfit, we have adopted a data augmentation scheme to improve the generalization ability and robustness [24] of the model. In the task of SI, we apply EDA Synonym Replacement (SR) [9] and Random Swap (RS) [23] to our model. Namely, each word has equal probability of being swapped or replaced by its synonyms without changing the label. Compared to short sentences, long sentences absorb more noise, which can better balance the dataset. After processing, sentences different from before are added to the training dataset. While in TC task, the data augmentation strategy is the same as that in SI which is a random process, initially. However, some of the techniques still cannot be detected such as Appeal_to_Authority, Bandwagon, Reductio_ad_hitlerum and Black-and-White_Fallacy. Aiming at this problem, on the basis of random data augmentation, we compulsively add them into the set to be data augmented. In this way, the purpose of increasing the valid noise of the training dataset is achieved. Meanwhile, the training time gets shortened as well.

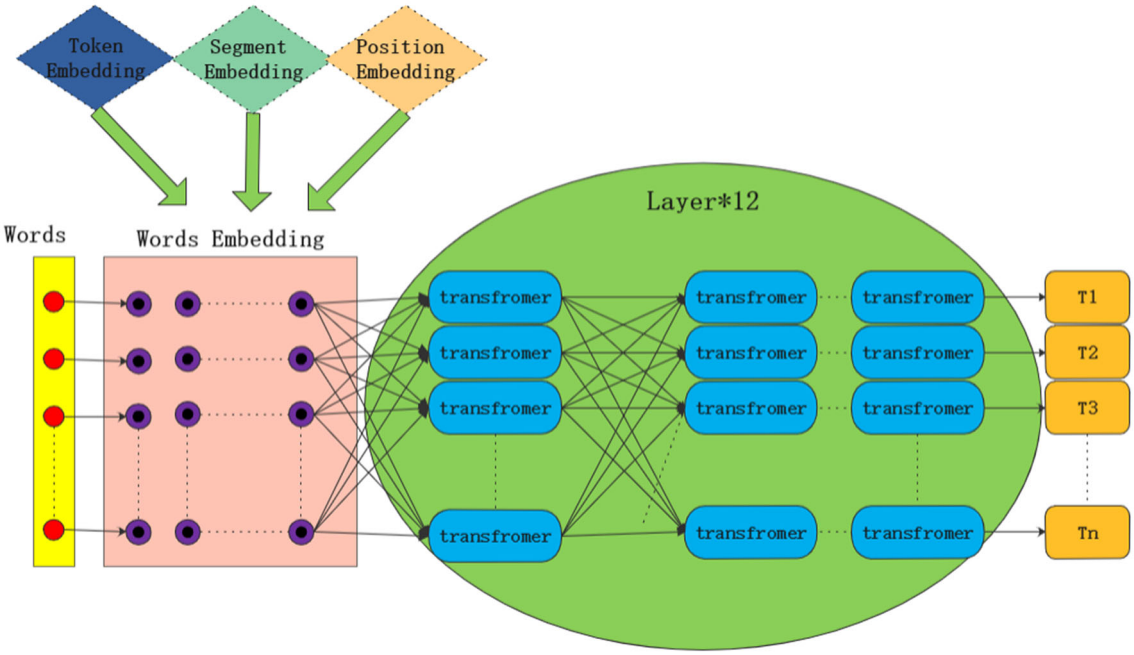

Fig. 2 The architecture of the pre-trained BERT model with the Word Embedding layer for the gain of word vectors and 12-layer Encoders making up of parallel transformers for the fusion of semantic

Approach of span identification (SI)

To deal with the SI task, we first defined it as a binary classification task [13], but after experiment we found Precision and F1-score of this solution were unexpected. After analyzing the cause and effect of this issue, we propose a three-classification model to classify each word in the news articles into three token types. The concrete architecture of our model is shown in Fig. 3. Two of them are 'prop' and 'non-prop'; the other one is 'invalid' which means the label of some invalid words like 'CLS', 'SEP' [3] and those used to ensure the input sentence with the same length. Classifying these invalid words into a 'invalid' token type reduces the noise and improves Precision and F1-score. Furthermore, we have utilized sampling skill and EDA to optimize the dataset.

Due to that the labels of the plain-text document offered by SEMEVAL are at char level, converting them into word level for word embedding in pre-trained BERT model is the first step. Before inputting them into the classifier, we combine the word vectors in each sentence with the feature vector of the sentence where they are. Then the word vectors with semantic integration of the sentence are normalized for the last classifier layer. As shown in Figure 3 on the right, the concatenating process generates the new concatenated vectors by placing the sentence vector in front of the word vectors. The following formula (1) shows the concatenating process mathematically:

$$\begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \text{concat} \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,200} \\ w_{2,1} & w_{2,2} & \dots & w_{2,200} \\ \vdots & \vdots & \vdots & \vdots \\ w_{768,1} & w_{768,2} & \dots & w_{768,200} \end{bmatrix} = \begin{bmatrix} s_1 & s_1 & \dots & s_1 \\ s_2 & s_2 & \dots & s_2 \\ w_{1,1} & w_{1,2} & \dots & w_{1,200} \\ w_{2,1} & w_{2,2} & \dots & w_{2,200} \\ \vdots & \vdots & \vdots & \vdots \\ w_{768,1} & w_{768,2} & \dots & w_{768,200}, \end{bmatrix} \quad (1)$$

where s_1, s_2 represent the elements in a sentence vector which contains the classifying result of sentence-level prediction. And the right matrix (768×200) contains 200 word vectors (768 dim) on behalf of one sentence. By concatenating, the input matrix (770×200) of the final classifier is made as the below one. This concatenating step is reasonable considering that sentence-level prediction is more undemanding and accurate than word-level prediction. The result also shows

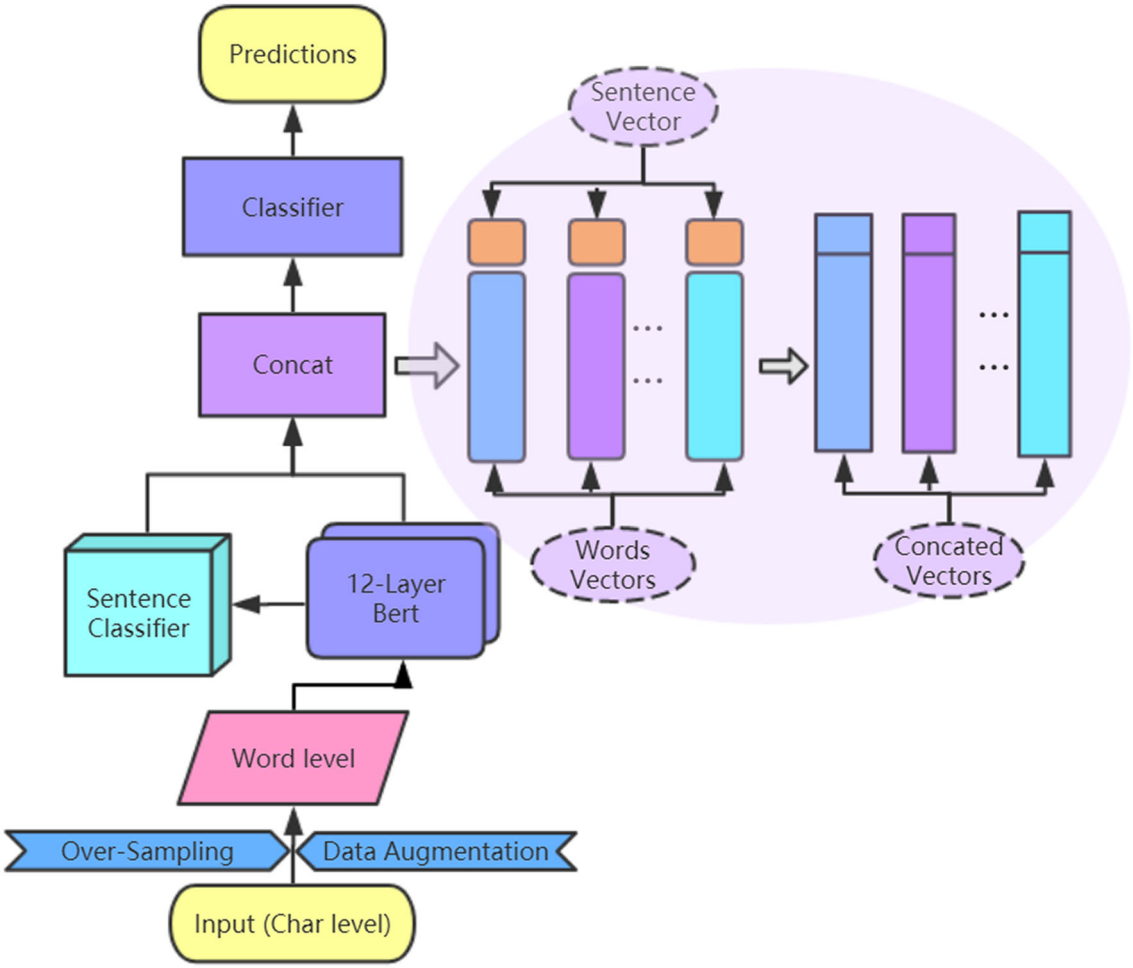

Fig. 3 The architecture of Span Identification task adopting over-sampling, data augmentation and sentence-level feature concatenating. The Concat means adding the classification feature of the sentence to its every word vector

that concatenating step plays a key role in the promotion of word-level prediction accuracy. Finally, by merging the successive words with identical propaganda technique, those specific fragments which include at least one propaganda technique are identified.

Approach of technique classification (TC)

For the multi-class classification task TC, we have utilized a Full Connection layer and a linear classifier based on BERT model, as shown in Fig. 4. Since the dimension of the valid fragment vector is large, we utilize the former Full Connection layer for dimensionality reduction, and the second for classifying them into 14 classes. And we handle those propaganda techniques that rarely appear in the dataset by utilizing EDA so as to solve the imbalance of dataset. Comparing our model without and with EDA, the latter gets an improvement of around 4 points in F1 score as shown in Table 3.

For details of TC task, we take the given text fragment identified as propaganda and its document context as the input of the pre-trained BERT generator. Different from SI task which is a word-level classification task, the TC task is fragment-level. Hence, incorporating sentence-level features into each word vector is ineffectual for TC task. As for the fragment which belongs to several sentences, we divide it into different sentences in the training process, while evaluating we treat it as a whole. Then for the sake of obtaining the valid fragment with propaganda techniques, we make segmentation of the output of BERT and pad it with invalid zero vector to a settled length (120). With the dimensionality reduction of our Full Connection layer, a linear classifier is used for 14 token types classification.

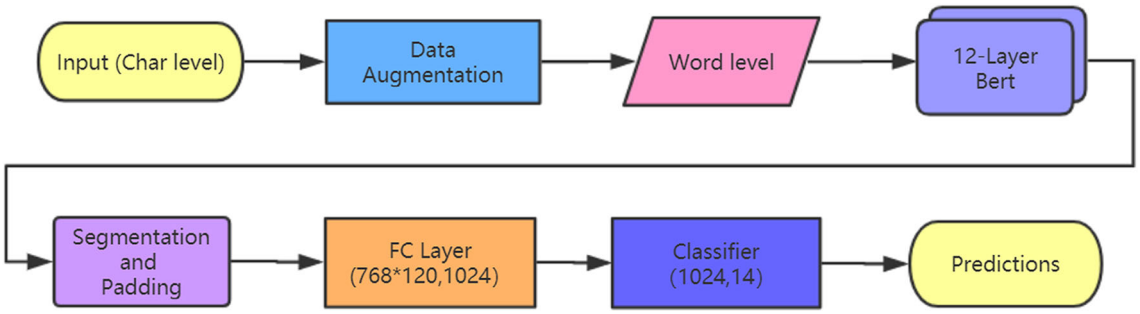

Fig. 4 The architecture of Technique Classification task with segmentation and padding operations, an FC layer and a linear classifier layer

Table 1 Corpus statistics including instances per technique and its proportion

Technique	Instances	Proportion (%)
Appeal_to_Authority	144	2.35
Appeal_to_fear-prejudice	294	4.80
Bandwagon, Reductio_ad_hitlerum	72	1.17
Black-and-White_Fallacy	107	1.75
Causal_Oversimplification	209	3.41
Doubt	493	8.04
Exaggeration, minimisation	466	7.60
Flag-Waving	229	3.74%
Loaded_Language	2123	34.64
Name_Calling, Labeling	1058	17.26
Repetition	621	10.13
Slogans	129	2.10
Thought-terminating_Cliches	76	1.24
Whataboutism, Straw_Men, Red_Herring	108	1.76
Total	6129	

Experiment and results

In this section, we will show the experiment details and the achieved experiment results by comparing our surpassing systems respectively for SI and TC to several other attempts.

Experiment details

In our experiment, we have trained our models parallelly with 4 Nvidia GTX 1080Ti GPUs to reduce the time required. Based on the PyTorch Framework and CrossEntropy Optimizer [28] (after trying the focal loss), we have fine-tuned the pre-trained BERT model for our SI and TC tasks.

Dataset The datasets for both of the SI and TC tasks are news articles in plain text format, including train-articles, dev-articles and the label files. To begin with, we have split each article into individual sentences to reduce parameters of our model. And before the experiment, we divided the annotated corpus of about 550 articles into 80% train set for model training and 20% test set for model evaluation, respectively. By calculating the instances of each technique, we find that the dataset for TC is imbalanced as shown in Table 1. Some of the techniques such as "Loaded_Language" has a high proportion of 34.64%, while some of the techniques such as "Black-and-White_Fallacy", "Slogans" and "Whataboutism, Straw_Men, Red_Herring" show up less often. What is worse, neither "Bandwagon, Reductio_ad_hitlerum" nor "Thought-terminating_Cliches" has no more than 80 instances which may badly influence the training process. During training, in order to enhance the generalization capability of our model, we utilized EDA to make train set extended and more well-proportioned. Besides, particularly for SI task, we adopted the over-sampling strategy for tagged sentences.

Evaluation metric So as to make a fair comparison, we use different evaluation criteria in different comparison experiment. For both of SI and TC tasks, we adopt the F1-score (F1) as the main metric. In addition, the general Precision (P) and Recall (R) are the secondary metric for SI task. The

Table 2 Contrasting with other models on test dataset, our three-token type classification system for SI task with Over-sampling, EDA and SLFC gains the top performance

Model	F1	Precision	Recall
Baseline	0.007862	0.099663	0.004092
BERT + binary classifier	0.370578	0.385497	0.356771
BERT + three classifier	0.408815	0.401099	0.416834
Our SI system	0.441732	0.432075	0.451831

F1-score is denoted by the following formula:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{2}$$

Results: span identification (SI)

For the purpose of achieving the SI task, we have presented two diverse architectures and optimized one of them with over-sampling, EDA, and sentence-level feature concatenating (SLFC). As is shown in Table 2, our top perform system is three-token type classification system with Over-sampling, EDA and SLFC. We have contrasted our SI system with BERT-based Binary classifier model, and BERT-based Three-token type classifier model.

As we have seen, the BERT-based three-token type classifier reaching 40.8815% F1-score, 40.1099% Precision and 41.6834% Recall behaves better than baseline which is merely BERT-based with no fine tuning and the Binary classifier model. We owe this success to the 'invalid' token type which impairs the noise of the invalid words by classifying the irrelevant words individually. Besides, after using EDA, it only took two epochs or so to reach the peak of the Recall, without which it took about six epochs to reach the peak and the results were not as expected. Ultimately, our SI system, based on our three-token type classifier and utilizing our strategies of over-sampling, EDA and SLFC, prevails over others, which scores 44.1732% of F1-score on the test set.

Next, we will give a deep analysis of the usage of SLFC and how it benefits our system on Recall and F1-score shown in Fig. 5. Generally speaking, the word-level prediction requires more accurate detection and there is a bigger margin of error than sentence-level prediction, which is why we give each word more information about whether the sentence it is in has propaganda with the aid of SLFC. Namely, the sentence classification prediction provides reference for the word prediction. If a sentence is propagandistic, it is of high probability containing propaganda fragments. On the contrary, if a sentence is non-propagandistic, the words in it are not of propagandistic as well. Based on this knowledge, we successively apply SLFC to our model, which does increase the F1-score by around three points and the Recall by around four points, respectively. Meanwhile, the precision does not decrease significantly. All in all, compared with no SLFC,

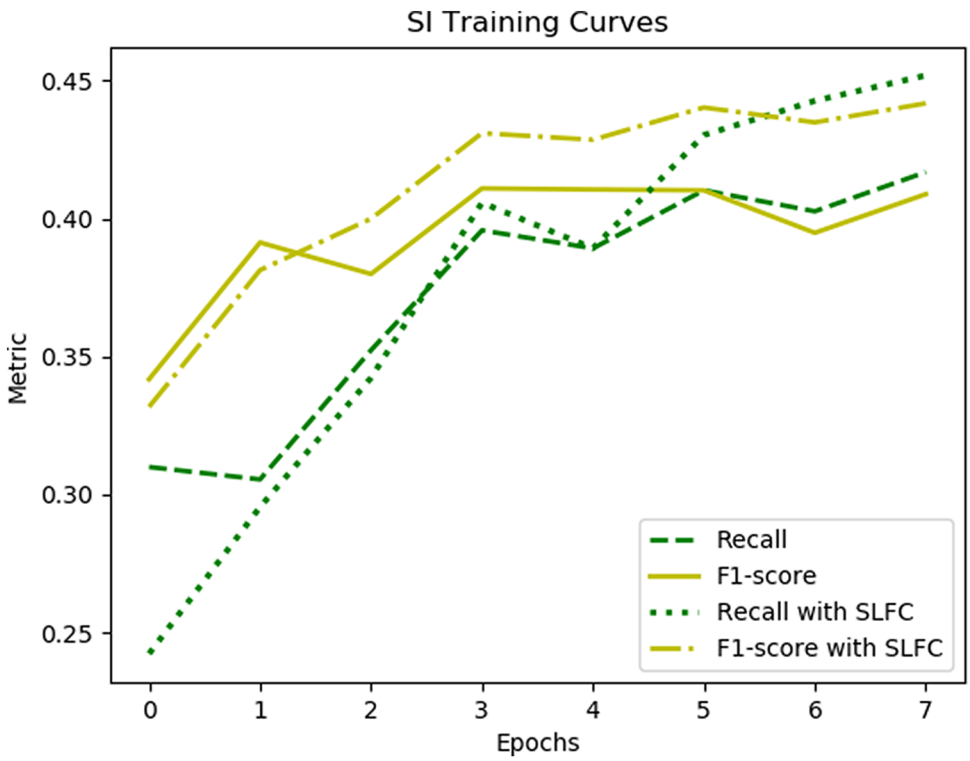

Fig. 5 The comparison of SI training process between our systems with and without SLFC

Table 3 Compared with other models on dev dataset, our system with EDA makes many breakthroughs

Model	F1
Baseline	0.262326
BERT without EDA	0.539535
Our TC system	**0.575729**

The bold values shows that our better model or method

our system identifies the propaganda spans more accurate which consequently promotes the F1-score and Recall.

Results: technique classification (TC)

To better complete the TC task, we have presented two architectures, one without EDA and another with EDA. Comparing them with the baseline which is merely BERT-based with no fine tuning, both of our systems with EDA or not have reached a new high state, improving the F1-score by two times approximately as shown in Table 3. During our experiment, we have made an experimental comparison and analysis for our strategy of utilizing EDA in the data loading process of our TC system. The final result has indicated that our TC system with EDA improved F1-score by around 3% compared to the absent EDA system. It stands to reason that our TC system reaches the state-of-the-art in the end, which scores 57.5729% of F1-score on the test set.

The respective promotions with EDA strategy in F1-score for each of propaganda technique are shown in Table 4. Compared with our no EDA model, in spite of the fact that three of techniques ('Doubt', 'Flag-Waving', and 'Whataboutism, Straw_Men, Red_Herring') have slightly decreased to some extent, more than half of the techniques have made progress in F1-score. For details, most of them have gotten about eight-point improvement on average, such as 'Appeal_to_fear-prejudice', 'Exaggeration, Minimisation', 'Repetition' and so on. What is worth mentioning is that the techniques named 'Causal_Oversimplification' and 'Thought-terminating_Cliches' have gotten about 14-point improvement. Thus, our TC system makes many breakthroughs on the whole, giving the credit to EDA which can enhance the data set, prompt the model to converge faster and improve the generalization ability and robustness of the model.

Parameter analysis

After a series of experiments, we have given a set of optimal parameters [epoch, learning rate (lr), sentence length (sentlen)] for the models of the two tasks. The optimal parameter combinations are shown in bold in Table 5.

For the sentence length, which is the length of the single input into the BERT and is usually set to 256, we have set it to 200 and 210 for our SI and TC tasks, respectively. In SI task, it is attributed to that the whole sentences in dataset do not exceed 200 in length, and too much padding will lead to greater classification error. As for TC task, the maximum length of valid fragments in the dataset is 210, so we choose it as the limit for padding. In terms of learning rate, both of our choices are 3×10^{-5} because our valid dataset is small. Through the analysis of the SLFC method for SI task in Sect. 4.2, we have found that the model began to converge around the epoch 7, so we set the training epoch to 8 to prevent overfitting in our SI system. Besides, in the experiment process of TC task, we have found the epoch parameter greater than 15 caused F1-score decreased, so we set it as the best choice for our TC system. Based on the above optimal parameters, our SI and TC systems finally obtained the F1-score of 0.441732 and 0.575729, and both of the training processes have taken around 2.5 h using 4 Nvidia GTX 1080Ti graphics cards (i.e. around 10 GPU hours).

Conclusion and future work

Based on the BERT model, we have set forth two specific systems for Span Identification and Technique Classification of Propaganda in news articles. In the data loading process, we have tried two strategies, over-sampling in SI task and EDA in both of SI and TC tasks, in order to deal with the imbalance between data with and without tags and enlarge our training dataset. For SI task, we have afresh defined it as a three-token type sequence tagging task with our SI system, and adopted sentence-level feature concatenating method. For TC task, we have devised a system based on BERT with a dimensionality reduction FC layer and a linear classifier. Ultimately, we have achieved two efficient and accurate systems for propaganda detection in news articles. And the final result also confirmed that our research further perfects the BERT model.

In the future, we are going to improve the Precision, Recall and F1-score further by drawing lessons from the SpanBERT model, which may perform better. Namely, in the process of masking, we would like to cover consecutive words randomly instead of scattered words. And we are thinking about searching for a more suitable architecture of BERT adopting the popular Neural Architecture Search (NAS). Besides, we hope our model can be compressed to some extent. For instance, we can prune the classifier layer, quantify or share the parameters of our model. In these cases our model can be applied widely and conveniently in our daily lives.

Table 4 The respective F1-score of the 14 propagandistic techniques with and without EDA strategy

Technique	F1 with EDA	F1 without EDA
Appeal_to_Authority	0	0
Appeal_to_fear-prejudice	**0.3870967741935484**	0.31460674157303375
Bandwagon, Reductio_ad_hitlerum	0	0
Black-and-White_Fallacy	0	0
Causal_Oversimplification	**0.26666666666666666**	0.12000000000000002
Doubt	0.4647887323943662	0.4678362573099415
Exaggeration, Minimisation	**0.39655172413793105**	0.30894308943089427
Flag-Waving	0.632258064516129	0.6369426751592357
Loaded_Language	**0.7414772727272728**	0.7005208333333333
Name_Calling, Labeling	**0.6504065040650406**	0.6129032258064515
Repetition	**0.48366013071895425**	0.4129554655870445
Slogans	**0.588235294117647**	0.5245901639344264
Thought-terminating_Cliches	**0.3636363636363636**	0.21052631578947367
Whataboutism, Straw_Men, Red_Herring	0.09523809523809525	0.1111111111111111

The bold values shows that our better model or method

Table 5 For the two models we have proposed, the parameters are analyzed experimentally

Model	OS	EDA	SLFC	epoch	lr	sent-len	F1
Baseline (SI)							0.007862
BERT + binary classifier	√	√	√	10	3×10^{-5}	256	0.370578
BERT + three classifier				10	3×10^{-5}	256	0.408815
BERT + three classifier	√	√	√	10	3×10^{-5}	256	0.427860
BERT + three classifier	√	√	√	8	3×10^{-5}	256	0.429325
Our SI system	√	√	√	8	3×10^{-5}	200	0.441732
Baseline (TC)							0.262326
BERT				20	3×10^{-3}	256	0.425729
BERT		√		20	3×10^{-5}	256	0.540473
BERT		√		20	3×10^{-5}	210	0.560931
Our TC system		√		15	3×10^{-5}	210	0.575729

Declarations

Conflict of interest We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. Barron-Cedeno A, Da San Martino G, Jaradat I, Nakov P (2019) Proppy: a system to unmask propaganda in online news. In: Proceedings of the AAAI conference on artificial intelligence, pp 9847–9848
2. Corney D, Albakour D, Martinez-Alvarez M, Moussa S (2016) What do a million news articles look like? In: NewsIR@ ECIR, pp 42–47
3. Devlin J, Chang M.W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. North American chapter of the association for computational linguistics
4. Fadel A, Tuffaha I, Al-Ayyoub M (2019) Pretrained ensemble learning for fine-grained propaganda detection. In: Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda, pp 139–142
5. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput:1735–1780
6. Hou W, Chen Y (2019) Caunlp at nlp4if 2019 shared task: context-dependent bert for sentence-level propaganda detection. In: Proceedings of the second workshop on natural language

processing for internet freedom: censorship, disinformation, and propaganda, pp 83–86
7. Huang Y, Wang W, Wang L, Tan T (2013) Multi-task deep neural network for multi-label learning. In: 2013 IEEE International conference on image processing, pp 2897–2900
8. Khalid A, Khan FA, Imran M, Alharbi M, Khan M, Ahmad A, Jeon G (2019) Reference terms identification of cited articles as topics from citation contexts. Comput Electr Eng 74:569–580
9. Kobayashi, S (2018) Contextual augmentation: data augmentation by words with paradigmatic relations. North American chapter of the association for computational linguistics, pp 452–457
10. Kurian D, Sattari F, Lefsrud L, Ma Y (2020) Using machine learning and keyword analysis to analyze incidents and reduce risk in oil sands operations. Saf Sci
11. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2020) Albert: a lite bert for self-supervised learning of language representations. ICLR
12. Liu S, Lee K, Lee I (2020) Document-level multi-topic sentiment classification of email data with bilstm and data augmentation. Knowl Based Syst:105918
13. Mapes N, White A, Medury R, Dua S (2019) Divisive language and propaganda detection using multi-head attention transformers with deep learning bert-based language models for binary classification. In: Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda, pp 103–106
14. Martino DSG, Yu S, Barron-Cedeno A, Petrov R, Nakov P (2019) Fine-grained analysis of propaganda in news articles. EMNLP/IJCNLP 1:5635–5645
15. Pankaj G, Khushbu S, Usama Y, Thomas R, Hinrich S (2019) Neural architectures for fine-grained propaganda detection in news. In: Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda
16. Peters E.M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer S.L (2018) Deep contextualized word representations. North American chapter of the association for computational linguistics
17. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pretraining
18. San G.D.M, Alberto B.C, Preslav N (2019) Findings of the nlp4if-2019 shared task on fine-grained propaganda detection. In: Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda
19. Tchiehe N.D, Gauthier F (2017) Classification of risk acceptability and risk tolerability factors in occupational health and safety. Saf Sci:138–147
20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez N.A, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems 30 (NIPS 2017), pp 5998–6008
21. Vlad G.A, Tanase M.A, Onose C, Cercel D.C (2019) Sentence-level propaganda detection in news articles with transfer learning and bert-bilstm-capsule model. In: Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda, pp 148–154
22. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman R.S (2018) Glue: A multi-task benchmark and analysis platform for natural language understanding. In: International conference on learning representations
23. Wei WJ, Zou K (2019) Eda: easy data augmentation techniques for boosting performance on text classification tasks. EMNLP/IJCNLP 1:6381–6387
24. Xie Z, Wang I.S, Li J, Lévy D, Nie A, Jurafsky D, Ng Y.A (2017) Data noising as smoothing in neural network language models. ICLR
25. Yang Z, Dai Z, Yang Y, Carbonell GJ, Salakhutdinov R, Le VQ (2019) Xlnet: generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems 32 (NIPS 2019), pp 5754–5764
26. Yoosuf S, Yang Y (2019) Fine-grained propaganda detection with fine-tuned bert. In: Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda, pp 87–91
27. Zhan Z, Hou Z, Yang Q, Zhao J, Zhang Y, Hu C (2020) Knowledge attention sandwich neural network for text classification. Neurocomputing:1–11
28. Zhang Z, Sabuncu M (2018) Generalized cross entropy loss for training deep neural networks with noisy labels. In: Advances in neural information processing systems, pp 8778–8788