PsuedoProp at SemEval-2020 Task 11:Propaganda Span Detection using BERT-CRF and Ensemble Sentence Level Classifier

Aniruddha Chauhan, Harshita Diddee
Department of Computer Science and Engineering
Bharati Vidyapeeth's College of Engineering
New Delhi, India
aniruddhac7@gmail.com, harshita.bvcoend@bvp.edu.in

Abstract

This paper presents our solution for Span Identification (SI) Task under "Task 11: Detection of Propaganda Techniques in News Articles" of SemEval 2020. This task aims to identify if a given sentence, taken from a corpus of news articles, contains a propaganda span and hence aims to identify the character level offsets of the identified propaganda element. Our solution proposes a sequential approach in which the span identification is preceded by an ensemble sentence level classifier (SLC). We only perform span identification on those samples which are flagged as propaganda samples by the SLC Model. We perform token level classification by fine-tuning BERT and use CRF to perform sequence tagging. Additionally, we present our analysis of different voting ensembles for the SLC model. Our system ranks 14th on the test set and 22nd on the development set and with an F1 score of 0.41 and 0.39 respectively.

1 Introduction

In contemporary times, fake news and propaganda have gained a lot of traction. A contributing factor to these problems is the easy dissemination of information on social media and various alternative news outlets on the Internet which house a vast repository of content which is tough to effectively moderate. Propaganda is often used to promulgate news articles or content that is misleading. In conjunction, Fake News not only contrives hysteria and spreads lies, in extreme cases it leads to physical violence (Kang and Goldman, 2016). Most of the workaround propaganda detection has been limited to document-level classification (Shu et al., 2017; Barrón-Cedeno et al., 2019; Rashkin et al., 2017). In the past, Shared tasks such as the NLP4IF 2019 have dealt with Sentence Level Classification (SLC) and Fragment Level Classification (FLC) of propaganda (Da San Martino et al., 2019). Fine-grained propaganda techniques provide a more suitable method of detecting propaganda because its classification provides the reasoning behind why an instance has been flagged as propaganda. The SemEval shared Task 11 makes progress in this aspect with its two tasks namely, Span Identification (SI) that has the objective of finding propaganda spans, and Technique Classification (TC) that labels the propaganda technique employed in a propaganda span (Da San Martino et al., 2020).

In this paper, we have focused on the Span Identification task that involves character level tagging of propaganda spans in text. To achieve this, we used ensemble transformer-based architectures to first perform SLC which is followed by token level tagging of spans of only propaganda sentences. These are fed to the BERT-CRF span identification model, predictions of which are later processed to obtain the character level tagging of propaganda fragments. In addition to this, we carry out various experiments to deal with the class imbalance in the provided data corpus and obtain a generic model that can be employed to detect propaganda fragments in any text.

The remaining paper is organised as follows. Section 2 gives a background on existing work on propaganda detection, the task we worked on and the novelty of our approach. Section 3 provides the rationale behind the system setup and the system's working. Following this, Section 4 describes the

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1779–1785
Barcelona, Spain (Online), December 12, 2020.

experimental setup to expatiate the basis on which we conducted our experiments. The next section, Section 5, provides the results of all our enlisted experiments and our final system along with the final model configurations that can be used to replicate our results. Section 6 finally concludes the paper and presents our error analysis.

2 Related Work

The Span Identification Task aims at providing a more fine-grained analysis of propaganda in text. Owing to the massive importance of regulating the quality of content being circulated among the populace, researchers have explored several techniques for achieving sequence labelling and sentence level classification, both of which provide an important background for the SI task.

The task organizers provide a corpus of 550 news articles that are labelled against character level offsets of the propaganda spans within the articles. It is a binary sequence tagging task. The annotation is done manually. The example below explains it clearly where character offsets from 19 to 40 contains a propaganda span "**nefarious connections**".

*^{0}I detailed Obama's ^{19}**nefarious connections**^{40} and resulting worldview in my book, The Post-American Presidency: The Obama Administration's War on America, which was ignored by the mainstream media.*

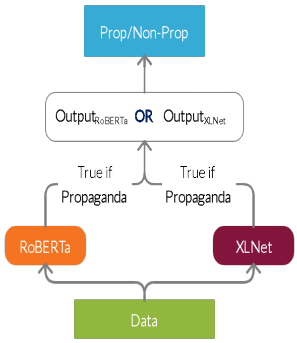
For Sentence Level Classification, Rashkin (2017) used an LSTM model and presented a comparison of its performance with Naive Bayes and Maximum Entropy models. Da San Martino (2019) used multi-granularity BERT for fine-grained propaganda detection. Work by Graves (2005) demonstrated the use of LSTMs in sequence tagging. To further this work and leverage the learning from both the future and past inputs in a sequence, Graves (2013) discussed the use of Bi-LSTMs. Recently, Huang (2015) proposed BiLSTM-CRFs which promise bidirectional comprehension whilst making use of the sentence level tag information mapped by the CRF layer. CRF's efficacy was further demonstrated by Lample (2016) who reported higher F1 scores in NER with four different languages without leveraging any knowledge specific to those languages.

3 System Overview

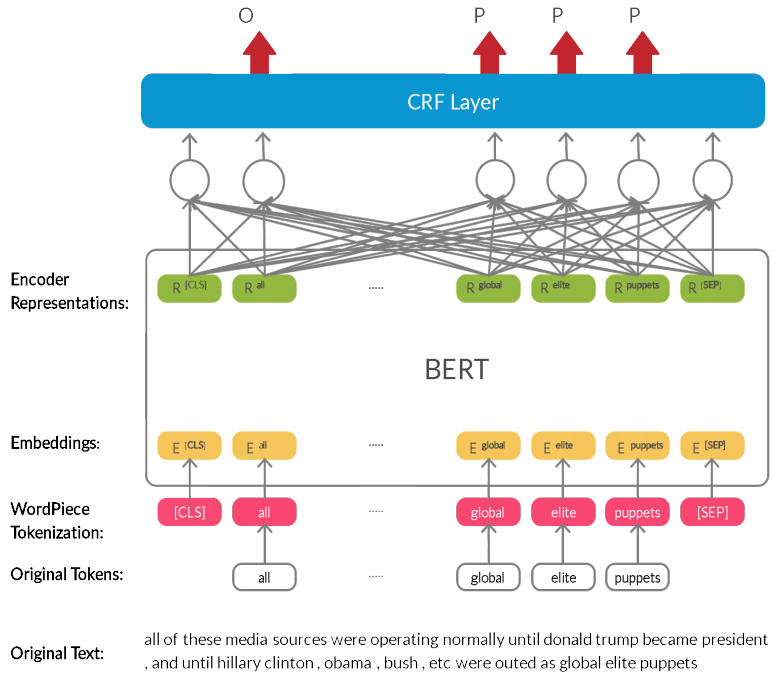
We adopt a two-step method of detecting propaganda spans by first performing SLC and then detecting spans in sentences which have been flagged as propaganda sentences. Comparing the results of span identification: with and without SLC - we observe an F1 Score improvement of nearly 0.13 using the former method.

3.1 Classification of Sentences

One of the most recent strides in NLP has been that of transfer learning. Transformers like BERT, RoBERTa, XLNet and AlBERT are trained on a large corpus of data and these language models can be fine-tuned on different downstream tasks (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019; Lan et al., 2019) to achieve advanced results in the field. One advantage of these large language models is that they tend to generalize well on smaller datasets like ours, this was one of the deciding factors for us to choose these language models. Upon experimentation, we came to the conclusion that an ensemble model having XLNet and RoBERTa as the base models - performed better for the classification of sentences when compared to AlBERT, BERT, and several other ensemble permutations of these models. We primarily credit RoBERTa's advanced performance to the fact that Liu (2019) pretrained RoBERTa on a larger corpus of data that includes the CC-News dataset which has public news articles which is similar to the data provided in our task. Secondly, RoBERTa is trained with larger mini-batches and learning rates. As far as XLNet is concerned, Yang (2019) uses a novel permutation language modelling objective that helps the autoregressive (AR) model to capture bidirectional context which may be otherwise lost in AR models. Besides, XLNet and RoBERTa's superior performance to BERT on several downstream tasks was an indication that it may perform better on the task given to us.



(a) Model classifying sentences as Propaganda or Non-Propaganda based on the output of fine-tuned RoBERTa and XLNet models.



(b) BERT-CRF architecture for Span Identification

Figure 1: (a) Provides an overview of the model used for the classification of sentences. (b) Depicts the model architecture of BERT-CRF.

3.2 Ensemble of Transformers

Different base models were used to analyze which ensemble configuration worked best. While we explored conventional ensemble criteria for the SLC model results of which can be reviewed in the results section, we also considered two other criterion including an OR Based Ensemble method in which if either of the base models predicted an instance as being propagandistic in nature the sentence would be flagged as a propaganda element. On similar lines, AND Based Ensemble was considered in which only if both of the base models predicted the sample to be a propaganda sentence would the final sentence be deemed as a propaganda instance.

3.3 Span Identification

Span identification is a binary sequence labelling task, which we tackle using BERT-CRF. In the context of our task, Transfer Learning proves to be a powerful and efficient approach because of the lack of training data. Fine-tuning is used as a transfer learning method in this task. BERT is used as the encoder to do fine-tuning and a CRF layer is used to decode and get sequence predictions. The architecture can be observed in Figure 2, where the BERT language model is connected to a fully connected layer that is finally connected to the CRF layer.

We use linear-chain CRF as a decoder. Every character of the input sequence x is converted into a vector w. The posterior probability of y given x is:

$$P(\mathbf{y}|\mathbf{x}; A) = \frac{e^{h^1(y_1;\mathbf{X})+\sum_{k=1}^{n-1} h^{k+1}(y_{k+1};\mathbf{X})+A_{y_k,y_{k+1}}}}{Z(\mathbf{x})} \quad (1)$$

$Z(\mathbf{x})$ is the normalization factor for $\mathbf{x}$, $h^k(y_k;\mathbf{x})$ is the output of the previous layer of Softmax and gives the probability of y_k at k position and n is the sequence length. The transition score matrix A can be learned by the model or set manually, we let the model learn the parameter itself. The probability from tag y_k to y_{k+1} is given by $A_{y_k,y_{k+1}}$. The most probable tag sequence of x is represented by $\hat{y}$ (Sutton et al., 2012).

$$\hat{\mathbf{y}} = argmax P(\mathbf{y}|\mathbf{x}) \quad (2)$$

4 Experimental Setup

This section describes our train-test setup which includes our analysis of the dataset, the data processing steps and the various models that we use to achieve our results.

4.1 Data

Preliminary data analysis on the provided corpus concluded a class imbalance between the propaganda and non-propaganda samples with only 3211 sentences containing propaganda spans out of the 15275 training samples that we had. To generate balanced classes - we explored two techniques :

a Minority Class oversampling: The sentences that contained propaganda spans were a clear minority in the provided dataset. Hence, We oversampled this class and concluded that the resulting oversampled train corpus produced a higher F1 score. Results from our experiments with the use of both, oversampled and non-oversampled can be found in Table 3.1 in the results sections.

b Paraphrasing: Wei et al. (2019) propose data augmentation techniques such as synonym replacement, random insertion, random swap, and random deletion which were explored to compensate for the lack of propaganda samples [1].

4.2 Models

For sentence classification we experiment with BERT, RoBERTa, XLNet and AlBERT transformer architectures and our experiments conclude that RoBERTa and XLNet be chosen for the base models for the final ensemble. All sentences in the training corpus are fed to RoBERTa and XLNet for training the SLC model in the proposed pipeline. While the ensemble model is trained on the entire training corpus - The SI model is only trained on the set of sentences flagged as containing a propaganda span. In the prediction cycle, SLC is carried out on the entire test set - Following which, we feed only the sentences which have been flagged as having a propaganda element by the SLC model to the BERT-CRF model.

5 Results

In this section, we present the results of all the experiments discussed in Section 4 which justify our choice of the proposed model and its configurations. The metrics used in the results are the same metrics used to evaluate the task results for the leaderboard[2]. In Table 1 we discuss results which inspired our choice of using a sequential approach with an SLC model. The training configurations for these results include a learning rate of 1.00e-05, batch size of 16, number of epochs as 10 and RoBERTa was used as the SLC Model. All the results presented here are evaluated on the dev set that was provided by the organizers.

SI Model	F1 Score (SI)	Precision (SI)	Recall (SI)	SLC	Oversampling
BERT-CRF	0.1028	0.0981	0.1079	No	No
BERT-CRF	0.2267	0.1505	0.4593	No	Yes
BERT-CRF	0.3772	0.3237	0.4519	Yes	Yes

Table 1: Oversampling and SLC conjunction Experiment results with BERT-CRF

As observed, The SI model's performance improved by nearly 0.15 when used with the SLC Model. A notable difference was also noticed with the use of oversampling, after which we sought to explore some data augmentation techniques to create a more balanced training corpus. The most relevant results are summarised in Table 2.

As observed in Table 2 - The model's performance was much better with oversampling in comparison to its performance with paraphrasing. Since Wei (2019) had already discussed that EDA may not be as effective for use with pre-trained models - further experiments were not conducted for the same.

[1]Code for paraphrasing available at: `https://github.com/jasonwei20/eda_nlp`

[2]Description of evaluation metrics: `https://propaganda.qcri.org/semeval2020-task11/data/propaganda_tasks_evaluation.pdf`

Category	Data Split (number of sentences)	F1 Score (SI)	Precision (SI)	Recall (SI)
Paraphrased Large	Total - 371584 Propaganda - 185791	0.3377	0.3248	0.3516
Paraphrased Small	Total - 84080 Propaganda - 42039	0.3589	0.3298	0.3936
Oversampled	Total - 24128 Propaganda - 12064	0.3720	0.3107	0.4634

Table 2: Results of experiments with data augmentation techniques

Additionally, Since our SI Model was now being analysed in conjunction with the SLC Model, we explored various ensembles along with individual base models to improve the SLC model's performance. Those results are encapsulated in Table 3.

Models	F1 Score (SI)	Precision (SI)	Recall (SI)
XLNet	0.3798	0.3169	0.4739
BERT	0.3421	0.2984	0.4008
AlBERT	0.3582	0.2833	0.4872
RoBERTa	0.3783	0.3115	0.4816
BERT + XLNet + AlBERT + RoBERTA	0.3868	0.2941	0.5647
XLNet + RoBERTa (OR Based Ensemble)	0.3932	0.3427	0.4611
XLNet + RoBERTa (AND Based Ensemble)	0.3651	0.3186	0.4274

Table 3: Results of Ensemble experiments

As observed, the XLNet-RoBERTa ensemble produced the best F1 score for the SI Task and hence it was employed in our final model pipeline. We also attempted to use the BERT - Large for the task but achieved an F1 of 0.35. Additionally, We analyzed several hyperparameters including sets of higher learning rates such as that of 1e-4 (produced F1 - 0.38), more number of training epochs such as 20 epochs producing an F1 score of 0.372 and smaller batch sizes such as 8 which resulted in an F1 score of 0.377. Having analysed all these results - our proposed model's configurations were decided as shown in Table 4.

Model Used	Task	Training Configuration
BERT-CRF (BERT-Base Uncased-uncased$_L - 12_H - 768_A - 12$)	SI	Epochs - 10, BS - 64, LR - 1e-05, Max Seq Len - 128, Dropout - 0.5
RoBERTa (roberta-base L-12$_H$ − 768$_A$ − 12)	SLC	Epochs : 2, BS - 8, LR - 4e-5, Max Seq Len - 128
XLNet (xlnet-base-cased L-12$_H$ − 768$_A$ − 12)	SLC	Epochs - 2, BS - 8, LR - 4e-5, Max Seq Len - 128

Table 4: A summary of the final models used and their training configurations

6 Conclusion and Future Work

This paper explains our teams' submission to the Shared Task of Fine-Grained Propaganda Detection in which we propose a sequential BERT-CRF based Span Identification model where the fine-grained detection is carried out only on articles that are flagged as containing propaganda by an ensemble SLC model. We propose this setup bearing in mind the practicality of this approach in identifying propaganda spans in the exponentially increasing content base where the fine-tuned analysis of the entire data repository may not be the optimal choice due to its massive computational resource requirement.

In the future, we intend to explore more advanced and efficient transformer models including T5 and Reformer respectively.

6.1 Error Analysis

We identify two possible scopes of errors which arise from assumptions that we make during data processing. Firstly, we suspect programming fallacies in the data post-processing steps where the BERT token-based predictions are mapped to their original token form (where each token is a word from the sentence and against which we have the character offsets) to get the original character offsets against the predicted tokens. Errant punctuation processing may have produced errors in the computed character offsets. Further, some assumptions are made while doing this post-processing such as assuming an 'X' token to be a 'p' token if it is succeeded and preceded by a 'p' token respectively, which may not necessarily be the case.

7 Acknowledgements

This work wouldn't have been possible without the guidance and support of Dr. Monika Aggarwal of CARE, Indian Institute of Technology, Delhi.

References

Alberto Barrón-Cedeno, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9847–9848.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, EMNLP-IJCNLP 2019, Hong Kong, China, November.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval 2020, Barcelona, Spain, September.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Cecilia Kang and Adam Goldman. 2016. In washington pizzeria attack, fake news brought real guns. *New York Times*, 5.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Charles Sutton, Andrew McCallum, et al. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Jason W Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.