NTUAAILS at SemEval-2020 Task 11: Propaganda Detection and Classification with biLSTMs and ELMo

Anastasios Arsenos, Georgios Siolas
Artificial Intelligence and Learning Systems Laboratory,
National Technical University of Athens
Zografou, Athens, Greece
anarsenos@gmail.com, gsiolas@islab.ntua.gr

Abstract

This paper describes the NTUAAILS submission for SemEval 2020 Task 11 Detection of Propaganda Techniques in News Articles. This task comprises of two different sub-tasks, namely A: Span Identification (SI), B: Technique Classification (TC). The goal for the SI sub-task is to identify specific fragments, in a given plain text, containing at least one propaganda technique. The TC sub-task aims to identify the applied propaganda technique in a given text fragment. A different model was trained for each sub-task. Our best performing system for the SI task consists of pre-trained ELMo word embeddings followed by residual bidirectional LSTM network. For the TC sub-task pre-trained word embeddings from GloVe fed to a bidirectional LSTM neural network. The models achieved rank 28 among 36 teams with F1 score of 0.335 and rank 25 among 31 teams with 0.463 F1 score for SI and TC sub-tasks respectively. Our results indicate that the proposed deep learning models, although relatively simple in architecture and fast to train, achieve satisfactory results in the tasks on hand.

1 Introduction

Propaganda is the expression of an opinion or an action by individuals or groups deliberately designed to influence the opinions or the actions of other individuals or groups with reference to predetermined ends (Miller, C. R, 1937 1938). It is often combined with misinformation and fake news and all together have the potential to polarise public opinion, to promote violent extremism and hate speech.

To deal with this problem, automatic identification and categorisation of propaganda, fake news and hyperpartisan content has been heavily addressed in recent research, mainly in the article level (Da San Martino et al., 2019a; Rashkin et al., 2017; Kiesel et al., 2019). This task is a follow-up to the NLP4IF shared task on fine-grained propaganda detection(Da San Martino et al., 2019c) and aims to produce models capable of spotting propaganda techniques in text fragments and then categorise them in one or more of 14 propaganda techniques.

The first sub-task is called Span identification (SI) and its goal is to spot propaganda fragments given a plain text document. The evaluation metric for this task is a modified F1 measure appropriate for taking into account partial matching between the spans. Detailed information about this metric can be found in (Da San Martino et al., 2019b; Da San Martino et al., 2020). The second sub-task which is called Technique classification (TC) consists of text fragments as inputs, labelled as one or more of 14 propaganda classes such as Name Calling and Loaded Language. The goal is to classify the fragments to one or more of those techniques. Due to the fact that the distribution of the techniques in gold labels is rather imbalanced, results were evaluated with the micro-averaged F1 measure.

In the current paper, we propose two novel deep learning architectures to deal with the two subtasks of the competition. In addition, we compare our deep learning methods with classic machine learning methods such as Logistic Regression provided with word embeddings pre-trained on large external corpora. Transfer Learning architectures showed promising results in various natural language processing (NLP)

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1495–1501
Barcelona, Spain (Online), December 12, 2020.

tasks (Jiang et al., 2019; Baris et al., 2019; Oberstrass et al., 2019), so there was a strong intuition to try this kind of architectures in this demanding task as well.

The rest of the paper is organised as follows. Section 2 describes the previous related research work in the field of propaganda and fake news detection. Section 3 describes the data and the system overview. Section 4 analyses the results and the errors of the methods proposed. Finally, Section 5 draws conclusions and suggests future work and improvements in the existing one.

2 Related work

A very similar task was shared previously in 2019 called Fine-Grained Propaganda Detection. This was a part of natural language processing for internet freedom (NLP4IF'19) workshop and included to sub-tasks, Sentence Level Classification (SLC) and Fragment Level Classification (FLC). The dataset used is a subset for this task's dataset and FLC was basically a combination of SI and TC in terms of including both detection and classification of text fragments. This was the first time propaganda and fake news detection was addressed in a fragment level. Previous work in propaganda and hyperpartisan news have been tackled mostly in the article level (Kiesel et al., 2019). In Hyperpartisan news detection task at SemEval 2019 two datasets were used. One relatively small labelled manually and one large corpus suitable for deep learning methods that was labelled using distant supervision, assuming that all articles from a given news outlet share the label of that outlet. (Rashkin et al., 2017) created a corpus of news articles labelled as propaganda, trusted, hoax or satire with distant supervision as well. Apparently, this method introduces noise, and in the hyperpartisan news detection task a lot of participating teams claimed that using the large corpus resulted in worse performance than not using it at all.

3 Methodology and Data

3.1 Data

The input data for the SI consists of news articles in plain-text format. Specifically, the data for this sub-task includes 371, 75 and 90 articles for training, development and test partitions respectively. More details about the data collection, the annotation and statistics about the corpus can be found in (Da San Martino et al., 2019d). TC input data consists of text fragments identified as propaganda within their document context. Table 1 illustrates the total number of instances per technique in the train set, the percentage with respect to the total number and the evaluation results achieved by our best performing model in development and test set.

3.2 Preprocessing

Given that the dataset has been partially preprocessed (title and sentences splitting has been performed automatically with NLTK sentence splitter by the organisers) we only did minimum additional pre-processing by removing most punctuation marks which do not include any useful information for text classification. Our solution for the SI task performs token-level classification while the data labels are at the character level. Due to the conversion from character level labels to word level labels (model training), as well as the reverse process (for prediction), a small information loss incurs, that affects the performance of the models used in this sub-task. In addition, our approach for the TC task does not consider overlapping labels, that occur when the token belongs to multiple propaganda techniques simultaneously. These suggest that there is much space for future improvement.

3.3 Models

3.3.1 Span Identification

During the competition three model architectures have been explored for this sequence labelling sub-task. The first is a very simple approach based on classic machine learning algorithms such as Logistic Regression and pre-trained word vectors from Word2Vec (Mikolov et al., 2013). This model was used as a baseline in order to compare it with more state-of-the art deep learning methods. The biggest drawback of this approach is that a lot of words of the corpus are lost because only pre-trained vectors are used and about 10,000 words do not match with any of the pre-trained embeddings. The main reason for this loss

Technique	#Train	(%)	F1(Dev)	F1(Test)
Bandwagon, Reductio ad Hitlerum	72	1.17	0.44	0.09
Whataboutism, Straw Men, Red Herring	108	1.76	0.08	0.04
Causal Oversimplification	209	3.41	0.17	0.00
Doubt	493	8.04	0.51	0.48
Appeal to Authority	144	2.34	0.06	0.36
Black-and-white Fallacy, Dictatorship	107	1.74	0.07	0.00
Name calling or labeling	1058	17.26	0.59	0.54
Loaded Language	2123	34.63	0.69	0.65
Exaggeration or Minimisation	466	7.60	0.43	0.29
Flag-waving	229	3.73	0.73	0.46
Appeal to fear/prejudice	294	4.79	0.21	0.28
Slogans	129	2.10	0.27	0.13
Thought-terminating cliché	76	1.24	0.00	0.11
Repetition	621	10.13	0.27	0.18
overall	6129		0.53	0.46
overall (best team:ApplicaAI)			0.70	0.62

Table 1: Statistics about the propaganda techniques and the evaluation results on the development and test set.

of words is that no fine tuning on these vectors is done on our data. In addition, the sequential (Lipton et al., 2015) nature of the textual data suggests the utilisation of Recurrent Neural Networks. Taken that into consideration, the second model utilises a bidirectional LSTM network architecture. Pre-trained word vectors from GloVe (Pennington et al., 2014) were used for encoding words to a vector space. Results were slightly better when using 300-dimensional GloVe vectors (in comparison with experiments done using different dimensions of GloVe model). These vectors were fine tuned on our corpus through the embedding layer. LSTM model outperforms the baseline significantly even without any fine tuning of its hyperparameters. The final model proposed in this paper consists of an embedding layer, two bidirectional layers, a residual connection to the first BiLSTM layer and a final fully connected layer followed by a softmax activation function. Figure 1 illustrates the model's architecture. This model is supplied with contextualized word representations generated from the pre-trained ELMo model (Peters et al., 2018). These embeddings are a function not only of the word itself but also of its context, enabling word disambiguation into different semantic representations. The input sequences of BiLSTM layers are the sentences of the corpus. We set the size of the sentences to a maximum of 80 words, as a compromise between the representation's expressiveness and its computational cost (losing only few longer samples). Shorter sentences are padded.

Architecture. As Figure 1 illustrates, our model is based in Bidirectional LSTM architecture. In particular, two BILSTMs layers are used in order to make predictions that take past and future information into account, since the context covers past and future labels in a sequence. A bidirectional LSTM is a combination of two LSTMs, ones runs forward and one backwards. In addition, the residual connection allows the network to skip training of the layers that are not useful and do not add value in overall accuracy. The ELMo (Embedding from Language Models) is the key element of our approach. One of biggest benefits is that there is no need of feature engineering but only the sequences and the token-level labels are needed. There are three word representations in the ELMo model, contextual: each word depends on the entire context, deep: the word representations combine all layers of a deep pre-trained neural network, character-based: ELMo are purely character based allowing the network to use morphological clues to form representations for tokens not seen in training.

Implementation. 10% dropout is used on all hidden layers. In addition, the BiLSTM layers use 512 units and 10% recurrent_dropout. The ELMo contextualized embeddings are used with default number of features (1024) for every token. Adam optimizer is selected with default parameters and

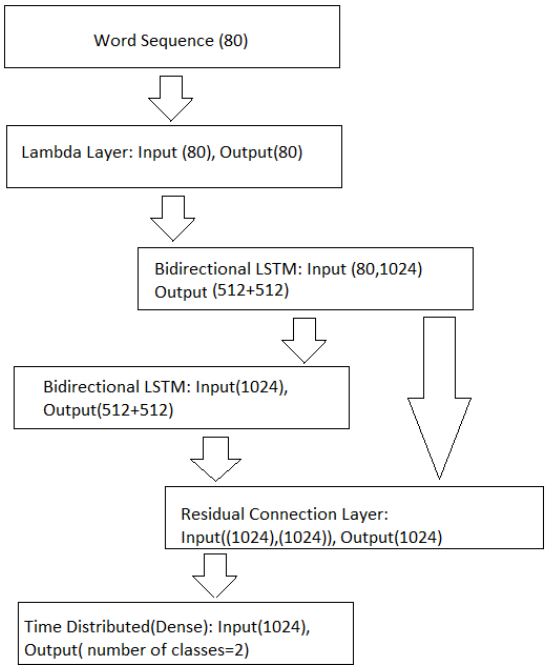

Figure 1: Visualisation of Bidirectional LSTM with ELMo model

sparse_categorical_crossentropy as the loss function. The batch size is set to 32 and training lasts 8 epochs. Deep learning models are implemented using Keras with Tensorfolw backend and the Logistic Regression using scikit-learn. The metrics used for training are custom precision and recall so that we can experiment different thresholds to classify each sample of propaganda and no-propaganda classes.

Model	F1	Precision	Recall
Logistic Regression + Word2Vec	0.16	0.09	0.47
BiLSTM + GloVe	0.30	0.25	0.37
BiLSTM + ELMo	0.32	0.27	0.38

Table 2: Evaluation results for Span Identification task on development set

Model	micro-averaged F1 measure
Logistic Regression + Word2Vec	0.41
MLP + Word2Vec	0.44
BiLSTM + GloVE	0.53

Table 3: Evaluation results for Technique Classification task on development set

3.3.2 Technique Classification

The algorithm for this multiclass classification task is based on components used in SI task as well. Our first approach consists of a Logistic Regression classifier and Word2Vec pre-trained embeddings. This model surpassed the given baseline by 15 per cent. In addition, we implemented one of the most widely used artificial neural networks, the Multi-layer Perceptron classifier from scikit-learn which is suitable for a multiclass classification problem. The features of the MLP classifier model are the pre-trained vectors of Wrd2Vec model. For the word representation, the best model used embeddings from GloVe project of 300 dimensions which are fine tuned on the input data through the embedding layer. Each sentence is padded to the maximum length sentence since the data size is very small. Each token (word) was replaced with its vector. The resulting sequence is then fed to the bidirectional LSTM layer.

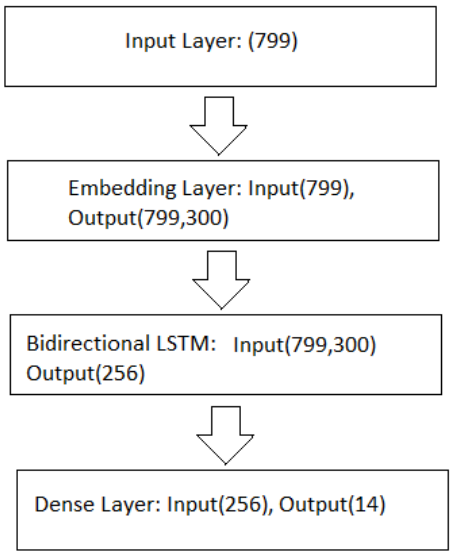

Figure 2: Visualisation of Bidirectional LSTM network with GloVe embeddings for TC task

Architecture. Figure 2 illustrates the architecture of our best model for the Technique classification task. The input sequences are fed into the embedding layer which is initialised with the weights of the pre-trained GloVe model and then trained with respect to our corpus. Due to the small size of the data there is only one bidirectional layer used followed by a Dense layer where the classification is done through the softmax function.

Implementation. Our implementation is based on Keras framework with Tensorflow back-end. To accelerate training, model is trained in accuracy metric function and using the following hyper-parameters: batch size 32, units 128, recurrent dropout 0.1, dropout 0.1, sequence length 799, and early stopping . The best model is trained for 4 epochs. Adam optimiser is used with a learning rate 0.003. During the competition participants had access only to the training set labels for both of the sub-tasks.

Team	F1	Precision	Recall
Hitachi (best)	0.51	0.56	0.47
NTUAAILS	0.33	0.46	0.26

Table 4: Official results for Span Identification task on test set

4 Discussion and Results

4.1 SI Task

Multiple experiments have been made in order to maximize the performance of our models for this task. Table 2 illustrates the results of the three main algorithms over the development set. As already mentioned, deep learning models have shown dramatic increase in precision metric and a significant improvement on the F1 score. The first algorithm which consists of a logistic regression classifier and Word2Vec achieved a Precision of only 0.09 and a relatively high recall of 0.47. The lack of a high precision mainly caused by the fact that the model classifies one word at a time and no context information is taken into account in the classification. The LSTM algorithm in the other hand solves this problem even without any fine tuning of its hyperparameters. This increase in precision of at least +0.10, is followed by a simultaneous decrease of the recall metric from 0.47 to 0.37. Despite this inverse proportional relation the F1 average metric is

significantly increased using the LSTM algorithm. After fine tuning the LSTM algorithm, experimented with different word vectors and used different techniques to avoid overfitting (such as early stopping and a small number of LSTM layers (2)), the LSTM achieves performance of 0.30, 0.25, 0.37 for F1, precision and recall respectively. Finally, the last model which includes ELMo contextualised embeddings and LSTM has the best F1 score of 0.31213 on the development set, so we used this model for our final submission on the test set. There is a fluctuation on the results of almost all the participating teams' submissions between development (best) and test set in terms of precision and recall. Our team achieved 0.33596, 0.46052 and 0.26444 in F1, precision and recall respectively, and placed 28th in SI task. Table 4 illustrates the results in comparison with the first team.

4.2 TC Task

In this task the goal is to specify the propaganda type of a given fragment. Initially, 18 propaganda techniques were given but due to the fact that some classes had very few samples, they were reduced to 14. Even though, the data still remained very imbalanced between the classes and the dataset very small. That was the main reason that only one layer of BiLSTM used in our final model. In addition, it was observed that almost all the teams had decreased results in the test set submission results in comparison with the ones in development set. That lack of robustness is very likely to arise from the small size and imbalance of the data. In Table 1 we demonstrate the performance of our final algorithm for every class and the number of samples of each of these classes. The results of our models in the development set are illustrated in Table 3. Our team achieved 0.46 micro-averaged F1 score in the final test set submission and placed 25th among 31 teams. The conclusion is that the classes with more samples are easier to predict.

5 Conclusion and Future Work

In this paper, we presented methods which combined transfer learning and recurrent neural networks capable of detecting and classifying propaganda fragments in news articles. In the competition various techniques and architectures were explored and the results show that deep learning methods improve performance as was initially expected. For future work, we plan to explore different methods such as BERT based sequence tagger and to experiment with techniques that will help to tackle the class imbalance problem, mainly in the TC task.

References

Ipek Baris, Lukas Schmelzeisen, and Steffen Staab. 2019. Clearumor at semeval-2019 task 7:convolving elmo against rumors. In *Proceedings of SemEval 2019 Conference*, Institute for Web Science and Technologies (WeST), University of Koblenz-Landau, Germany2Web and Internet Science Group (WAIS), University of Southampton, United Kingdom.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Israa Jaradat, and Preslav Nakov. 2019a. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Giovanni Da San Martino, Albertono Barrón-Cedeno, and Preslav Nakov. 2019b. Evaluation of propaganda detection tasks. *Shared Task at SemEval 2020 Task 11:"Detection Of Propaganda Techniques In News Articles"*, (5).

Giovanni Da San Martino, Albertono Barrón-Cedeno, and Preslav Nakov. 2019c. Findings of the nlp4if-2019 shared taskon fine-grained propaganda detection. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 201*, (5).

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019d. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 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, December.

Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. 2019. Team bertha von suttner at semeval-2019 task 4: Hyperpartisan news detection using elmo sentence representation convolutional network. In *Proceedings of SemEval 2019 Conference*, University of Sheffield Sheffield , UK.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. *In 12th International Workshop on Semantic Evaluation*, (5).

Zachary C. Lipton, John Berkowitz, and Charles Elkan. 2015. A critical review of recurrent neural networks for sequence learning.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Miller, C. R. 1937-1938. *THE PUBLICATIONS OF THE INSTITUTE FOR PROPAGANDA ANALYSIS, INC.*, volume I. INSTITUTE FOR PROPAGANDA ANALYSIS, INC.

Alexander Oberstrass, Julia Romberg, Anke Stoll, and Sefan Conrad. 2019. Hhu at semeval-2019 task 6: Context does matter - tackling offensivelanguage identification and categorization with elmo. In *Proceedings of SemEval 2019 Conference*, Institute of Computer Science, Heinrich Heine University D usseldorf, Germany.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. cite arxiv:1802.05365Comment: NAACL 2018. Originally posted to openreview 27 Oct 2017. v2 updated for NAACL camera ready.

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. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 2931–2937.