NSIT@NLP4IF-2019: Propaganda Detection from News Articles using Transfer Learning

Kartik Aggarwal[1] and **Anubhav Sadana**[2]

[1]Netaji Subhas Institute of Technology, Delhi, India
[1]*kartik.mp.16@nsit.net.in*
[2]SAP Labs
[2]*anubhav.sadana@sap.com*

Abstract

In this paper, we describe our approach and system description for NLP4IF 2019 Workshop: Shared Task on Fine-Grained Propaganda Detection. Given a sentence from a news article, the task is to detect whether the sentence contains a propagandistic agenda or not. The main contribution of our work is to evaluate the effectiveness of various transfer learning approaches like ELMo, BERT, and RoBERTa for propaganda detection. We show the use of Document Embeddings on the top of Stacked Embeddings combined with LSTM for identification of propagandistic context in the sentence. We further provide analysis of these models to show the effect of oversampling on the provided dataset. In the final test-set evaluation, our system ranked 21st with F_1-score of 0.43 in the SLC Task.

1 Introduction and Background

Propaganda is the deliberate spreading of ideas, facts or allegations with the aim of influencing the opinions or the actions of an individual or a group. Propaganda uses rhetorical and psychological techniques that are intended to go unnoticed to achieve maximum effect. Social media has contributed immensely in spreading these propagandistic articles reaching million users instantaneously. These articles may also lead to fake news circulation, election bias or misinformation thereby having adverse societal and political impact (Lewandowsky et al., 2017). Hence, there is an urgent need to detect these propagandistic articles and stop them from proliferating.

Propaganda Detection is the technique to automatically detect the use of propaganda in news articles. This will help to identify news outlets or articles that are biased and are trying to influence people's mindset and spread awareness limiting the impact of propaganda and help in fighting disinformation. Generally, propagandistic news articles use techniques like whataboutism, loaded-language, name-calling or bandwagon, etc (Da San Martino et al., 2019b). Detecting these techniques can help to easily identify propagandistic articles. This work aims to provide an approach that can accurately classify articles as Propagandistic or Non-Propagandistic.

Recently, there has been a lot of interest in studying bias and disinformation in news articles and social media (Baly et al., 2018; Gupta and Kumaraguru, 2018). Terms such as Propaganda detection, Fact-Checking, Fake News identification, etc. have started to gain huge attention in the domain of NLP (Rashkin et al., 2017; Volkova et al., 2017). Our work is an enhancement in this domain with the employment of recent state-of-the-art deep learning methods and architectures like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018).

Fine-Grained Analysis of propaganda in news articles (Da San Martino et al., 2019a) focuses on identifying the instances of use of specific propaganda techniques in the news article through a multi-granularity network. In this direction, Proppy - a system to unmask propaganda in online news (Barrón-Cedeño et al., 2019) was developed which monitors a number of news sources, deduplicates and clusters them into events on the basis of propagandistic content likelihood using various NLP techniques. With this motivation, two shared tasks for Fine-Grained Propaganda Detection were conducted as a part of *"Second Workshop on NLP for Internet Freedom (NLP4IF): Censorship, Dis-*

Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 143–147
Hong Kong, China, November 4, 2019

information, and Propaganda", EMNLP-IJCNLP 2019 (Da San Martino et al., 2019a). Our team participated in the Sentence Level Classification (SLC) Task of the workshop. The details for the task is as follows:

***Problem Definition SLC Task**: Given a labelled training dataset D with a set of sentences, the objective of the task is to learn a binary classification/prediction function that predicts a label l, l ∈ {propaganda, non-propaganda} for a given sentence S, where propaganda: denotes the sentence containing propagandistic fragment and non-propaganda: denotes the sentence not containing any propagandistic fragment*

Towards this objective we make the following contributions in this work:

1. We train transformer-based models like ELMo, BERT and RoBERTa with the provided dataset and show the effectiveness of transfer learning on downstream tasks in the domain of propaganda detection.

2. We show the use of document embeddings on a combination of multiple models for identifying whether the sentence contains propagandistic fragments or not.

3. We also show that these models do not perform very well on highly imbalanced datasets and thus require re-sampling techniques such as class oversampling to give better results on classification tasks.

4. We also present the comparison of these pretrained transformer-based architectures with classical algorithms such as Naive Bayes, Logistic Regression and SVM.

Further, we have organised the paper as follows: In Section-2 we discuss the experimental setup adopted for this task. Section-3 details about the results for the experimented models followed by error analysis of the best model. Finally, Section-5 highlights the concluding remarks and the future work of the performed study.

2 Experimental Setup

This section provides an overview of the dataset used for training and evaluation along with the details of the various models used in this work.

Label	Train
Propaganda	4720
Non-Propaganda	12245

Table 1: Data Distribution

2.1 Dataset

The dataset for the SLC Task used in all of our experiments is provided by the organisers of NLP4IF. This data comes in the form of news articles given in TXT format. Each article starts with the title followed by an empty line and news-article body with the Labels for each article provided in a separate file.

The dataset is divided into training and development set where the labels are distributed as {*propaganda, non-propaganda*}. The training set consists of 16,965 examples of which 4,720 contain one or more propagandistic fragments and the remaining (12,245) do not. Figure 1 (Blue) exhibits the distribution of the data in the training set. The unlabelled development and test set were used for evaluation in our experiments. The standard evaluation measure for this task was F1-score even though precision and recall are reported.

As it is clearly evident from Fig.1 (Blue), there is a high imbalance between distribution of sentences that are propaganda and non-propaganda, which also happens in case of a real world dataset. We deal with this high data disproportion by the technique of class oversampling. For this, we just randomly select and duplicate the propaganda sentences so that the ratio changes from 3:1 to 3:2 approximately. Fig.1 (Red) shows the distribution between both the classes after oversampling.

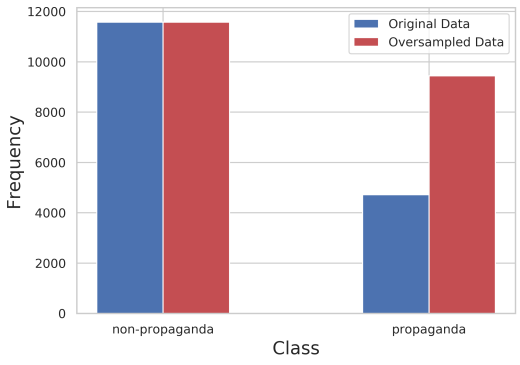

Figure 1: Distribution of Classes in Training Set

Table 2: Model Architectures used for training and their optimal hyperparameters

Model	Hyperparameters
BERT-1	BERT-Base-Uncased *batch-size=32, learning-rate=2e-5,epochs=3*
BERT-2	DocumentEmbeddings {Stacked Embeddings BERT + GRU + Dropout (p=0.5)} *batch-size=32, learning-rate=0.01,epochs=2, anneal-factor=0.5, patience=5*
ELMo-1	DocumentEmbeddings {Stacked Embeddings ELMo + GRU + Dropout (p=0.5)} *batch-size=64, learning-rate=1e-1,epochs=2, anneal-factor=0.5, patience=5*
ELMo-2	DocumentEmbeddings {Stacked Embeddings ELMo + FLAIR Embeddings (forward + backward) + GRU + Dropout (p=0.5)} *batch-size=64, learning-rate=0.001,epochs=3, anneal-factor=0.5, patience=5*
RoBERTa	DocumentEmbeddings {Stacked Embeddings RoBERTa + GRU + Dropout(p=0.5)} *batch-size=64, learning-rate=0.001,epochs=2, anneal-factor=0.5, patience=5*

2.2 Training Models

Transfer Learning has recently been one of the most effective methods in NLP. The key idea is to use a language model pretrained on a large corpus to transfer the information onto a downstream task. Fine-tuning these large pre-trained models produce very good results especially when there are small datasets available for training. Hence, for this task, we mainly use transformer-based models such as RoBERTa (Liu et al., 2019), BERT and ELMo models as they have shown great success in handling language based tasks across various domains. Training was largely done using Flair framework[1] (Akbik et al., 2019) along with AllenNLP library[2] (Gardner et al., 2018). Pretrained Stacked Embeddings are used to combine embeddings from multiple models. Document representation is then generated by applying LSTM over the stacked word embeddings in the document. Now we describe each of the models in brief:

Embeddings from Language Model (ELMo): We use the FLAIR implementation of ELMo by fine-tuning the pretrained stacked weights on Document Embeddings (ELMo-1). ELMo goes beyond the traditional word embeddings approach by producing context-sensitive features in a bi-directional manner. Left-to-right and right-to-left representations are concatenated to form an immediate word vector which are then fed to subsequent layers. Thus, ELMo can be effective for detecting words with propagandist context in the sentence even though the word by itself does not contain any propagandistic sentiment. We find the optimal parameters and train the model over original and oversampled dataset. Apart from this, we also experiment with a combination of Pretrained ELMo embeddings with FLAIR word-embeddings (ELMo-2).

Bidirectional Encoder Representations from Transformers (BERT) outperformed most of the existing systems on various NLP tasks by using a masked language model (MLM) pre-training method. Moreover, instead of reading the sentence in a sequential manner (left-to-right or right-to-left), BERT reads the entire sequence at once in a unidirectional manner. In addition, BERT goes deeper by expanding the base model to 12 layers while ELMo is a shallower model with only 2 LSTM layers. We use the Tensorflow [3] implementation of the BERT-base-uncased model by fine-tuning it with best parameters (BERT-1). DocumentRNN implementation of the Stacked pre-trained BERT along with LSTM is done using FLAIR (BERT-2).

RoBERTa moves one step ahead of BERT by pre-training the model over larger data and with bigger batches. This approach improved previous state-of-the-art on certain tasks by choosing better training strategies and design choices. We trained a RoBERTa classifier by finding the best parameters over both original and oversampled dataset using the FLAIR framework.

[1] https://github.com/zalandoresearch/flair

[2] https://github.com/allenai/allennlp

[3] https://github.com/google-research/bert

We also experiment with classical algorithms such as MultinomialNB, Logistic Regression and Support Vector Classifier for comparison.

3 Results

In this section, we briefly summarize the evaluation and results of the models used for the task. The metric used for evaluation is standard F_1 score. In addition, precision (Pr) and recall (Rc) are also reported.

Table 3 represents the performance of all the models trained on the training dataset and evaluated on the development data for the SLC Task. We see that the RoBERTa model gives the best performance on the oversampled dataset for the detection of propaganda in news articles with an F_1 score of 0.60 and a recall of 0.79. The highest precision of 0.66 was recorded by SVM and BERT-1 model. The results obtained from Table 3 show that models such as Naive Bayes, Logistic Regression and SVM perform decent with respect to deep learning-based models for the classification of propaganda in sentences.

Table 3: Performance of different models on development data for SLC Task

Model	F_1	Pr	Rc
Naive Bayes (count vectorizer)	0.44	0.57	0.36
Logistic Regression (count vectorizer)	0.41	0.58	0.31
SVM (Linear Kernel) (tf-idf vectorizer)	0.40	**0.66**	0.28
BERT-1	0.57	**0.66**	0.51
BERT-2	0.55	0.45	0.73
ELMo-1	0.51	0.46	0.56
ELMo-2	0.49	0.61	0.40
RoBERTa	**0.60**	0.49	**0.79**

Further, the performance of the transformer models were also evaluated on the original training dataset to observe the effect of oversampling. Fig. 2 helps us to compare the F_1 scores of these models. We observe that oversampling the examples of the minority class i.e. propaganda in this dataset, provides a significant improvement in the classification performance.

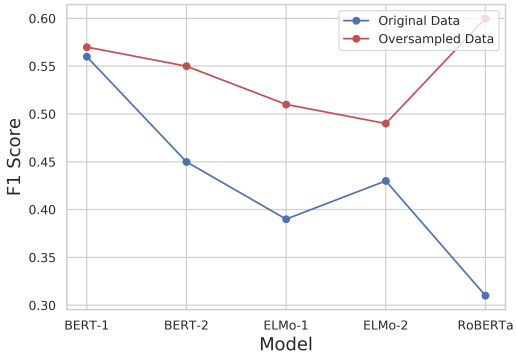

Figure 2: Effect of oversampling on the training data for different models

4 Error Analysis

In this section, we briefly highlight the error analysis of our best performing model "RoBERTa" with oversampled data. Since the labels for the development and the test set were not provided, the analysis is done on the test set synthetically created from the training dataset. 20 percent of the sentences were randomly chosen as the test set for prediction. Fig.3 shows the confusion matrix for the test data. In general, the most incorrect predictions were made for the non-propaganda classes while the model performed pretty good on detecting the propagandistic sentences.

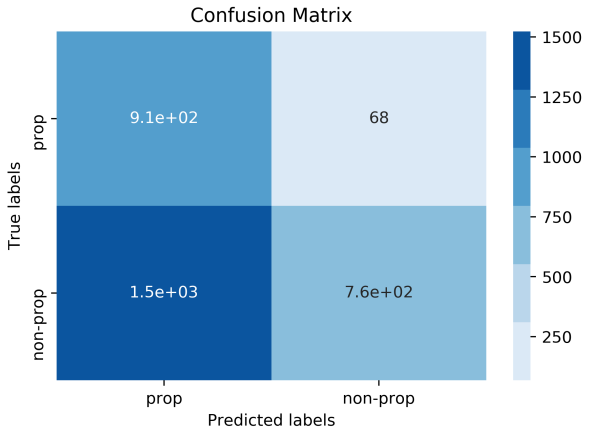

Figure 3: Confusion Matrix on 8:2 Training to Testing split

5 Conclusion and Future Work

In this work, we report our models and their respective performance in SLC task of "Second Workshop on NLP for Internet Freedom

(NLP4IF): Censorship, Disinformation, and Propaganda", EMNLP-IJCNLP 2019. We showed how transfer learning of transformer-based pretrained models perform well with the provided dataset. Our final submission on test set was made from BERT-1 weights and the team ranked 21st with an F_1 score of 0.43 in the SLC Task in the final evaluation of the test set. Hence, there is a significant room for improvement.

In the future, we would like to investigate the effectiveness of these models on the FLC Task of the workshop where the aim is to detect fine-grained propaganda techniques from 18 different classes. In particular, we intend to conduct a comprehensive analysis of the task by cleaning the annotated data and drawing out patterns specific to the given problem of propaganda detection. We would also like to experiment with other machine learning architectures like OpenAIGPT2, XLNet, etc for better performances specific to the dataset.

6 Code and Reproducibility

We provide the code for FLAIR based models on the Github Repository located at https://github.com/Kartikaggarwal98/Propaganda_Detection-NLP4IF. The results can be reproduced using the weights for the models provided in the github repository. The Tensorflow implementation of the BERT-1 model can be reproduced using https://github.com/google-research/bert. The datasets for the tasks are not provided according to the workshop guidelines.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. *arXiv preprint arXiv:1810.01765*.

Alberto Barrón-Cedeño, 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, Alberto Barron-Cedeno, and Preslav Nakov. 2019a. Findings of the nlp4if-2019 shared task on fine-grained propaganda detection. In *Proceedings of the 2nd Workshop on NLP for Internet Freedom (NLP4IF): Censorship, Disinformation, and Propaganda*, NLP4IFEMNLP '19, Hong Kong, China.

Giovanni Da San Martino, Seunghak Yu, Alberto Barron-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, EMNLP '19, Hong Kong, China.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Aditi Gupta and Ponnurangam Kumaraguru. 2018. Misinformation in social networks: Analyzing twitter during crisis events. *Encyclopedia of Social Network Analysis and Mining*, pages 1329–1338.

Stephan Lewandowsky, Ullrich KH Ecker, and John Cook. 2017. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4):353–369.

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*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

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.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653.