On Sentence Representations for Propaganda Detection: From Handcrafted Features to Word Embeddings

André Ferreira Cruz and **Gil Rocha** and **Henrique Lopes Cardoso**
Laboratório de Inteligência Artificial e Ciências dos Computadores (LIACC)
Departamento de Engenharia Informática,
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
{andre.ferreira.cruz, gil.rocha, hlc}@fe.up.pt

Abstract

Bias is ubiquitous in most online sources of natural language, from news media to social networks. Given the steady shift in news consumption behavior from traditional outlets to online sources, the automatic detection of propaganda, in which information is shaped to purposefully foster a predetermined agenda, is an increasingly crucial task. To this goal, we explore the task of *sentence-level propaganda detection*, and experiment with both handcrafted features and learned dense semantic representations. We also experiment with random undersampling of the majority class (non-propaganda) to curb the influence of class distribution on the system's performance, leading to marked improvements on the minority class (propaganda). Our best performing system uses pre-trained ELMo word embeddings, followed by a bidirectional LSTM and an attention layer. We have submitted a 5-model ensemble of our best performing system to the NLP4IF shared task on sentence-level propaganda detection (team *LIACC*), achieving rank 10 among 25 participants, with 59.5 F1-score.

1 Introduction

Propaganda shapes information in order to purposefully influence people's mindset and advance a predetermined agenda. The NLP4IF shared task on propaganda detection challenged participants to build systems capable of sentence-level (SLC) or fragment-level (FLC) detection of propagandistic texts (Da San Martino et al., 2019). We have participated on the SLC track, hence this will be the focus of this paper.

The rise of fake (Allcott and Gentzkow, 2017), hyperpartisan (Silverman et al., 2016), and propagandistic news on social media and online news outlets calls for improved automatic detection of bias in texts. However, any and all attempts at automated regulation of online content have freedom of speech implications, and risk unintended censorship (Akdeniz, 2010). Mindful of these considerations, we experiment with a set of handcrafted and interpretable stylometric features, together with a model based on Gradient Boosted Trees (Drucker and Cortes, 1996), thus facilitating inspection of what it is that the model has learned.

In addition, aiming for a better performance to the detriment of the model's interpretability, we experiment with deep neural networks, supplied with word embeddings learned on large external corpora, as this combination is the state-of-the-art for several natural language processing (NLP) tasks (Peters et al., 2018; Devlin et al., 2018; Akbik et al., 2019). Nonetheless, some degree of interpretability is maintained through the use of attention layers (Bahdanau et al., 2014), enabling inspection of which time-steps (words) the model is considering when making a prediction.

The provided train dataset consists of 350 articles, with a total of 16,965 sentences — 4,720 of which are labeled *propaganda*, and the remaining 12,245 labeled as *non-propaganda*. This class imbalance leads supervised learning models to favor predicting the majority class (*non-propaganda*), severely impacting performance on the minority class (Japkowicz and Stephen, 2002). In order to tackle this problem, we train all systems on a balanced version of the provided dataset, by means of random undersampling of the majority class, as this technique has been shown to have good results on several NLP tasks (Japkowicz and Stephen, 2002; Prusa et al., 2015).

The rest of the paper is organized as follows. Section 2 describes data pre-processing and feature selection, and details all tested models and their architectures. Section 3 analyzes our models' performance, analyzes attention-weight plots, and discusses results. Finally, Section 4 draws conclusions and sketches future work.

Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 107–112
Hong Kong, China, November 4, 2019

2 System Description

We propose an approach based on a selection of handcrafted features paired with a Gradient Boosted Trees (GBT) model, as well as an approach based on learned dense semantic representations (word embeddings) paired with different deep-learning models. This Section describes the data pre-processing and feature selection, the choice of word embeddings, and the tested models and their hyperparameters.

2.1 Data Pre-processing

We tokenize sentences into words using *Spacy* (Honnibal and Montani, 2017). We standardize quotation marks (left and right, single and double), as well as single grave and acute accents, as all these characters may be represented by different unicode characters while portraying the same meaning.

2.2 Feature Selection

We use a small set of linguistically-inspired style and complexity features, already proven to have good performance on a similar bias-detection task – hyperpartisan news detection (Cruz et al., 2019). Some of the features portray the article in which each sentence is incorporated, while others portray the sentence itself. Our features are as follow:

- *num_sentences*: total number of sentences in the article;
- *avg_sent_char_len*: average character-length of article's sentences;
- *var_sent_char_len*: variance of character-length of the article's sentences;
- *actual_sent_char_len*: character-length of current sentence;
- *avg_word_len*: average of character-length of this sentence's words;
- *var_word_len*: variance of character-length of this sentence's words;
- *punct_freq*: this sentence's punctuation frequency;
- *capital_freq*: this sentence's capital-case frequency;
- *type-token-ratio* over lemmatized words — a measure of vocabulary diversity and richness (Johnson, 1944).
- TF-IDF (Robertson, 2004) vector for the 50 most frequent *unigrams* and *bi-grams*, whose document frequency does not exceed 95%.

2.3 Contextualized Word Representations

Deep-learning models proposed in this paper are supplied with dense word representations, generated from the pre-trained ELMo model (Peters et al., 2018). We use the *Flair* library (Akbik et al., 2019) to generate contextualized 3072-dimensional representations for each input word (concatenation of outputs from three 1024-dimensional layers). These embeddings are a function not only of the word itself but also of its context, enabling word disambiguation into different semantic representations.

We crop sentences to a maximum of 50 words, as a compromise between the representation's expressiveness and its computational cost (affecting only 3.7% of longer samples, see Figure 1). Shorter sentences are padded out with zeros.

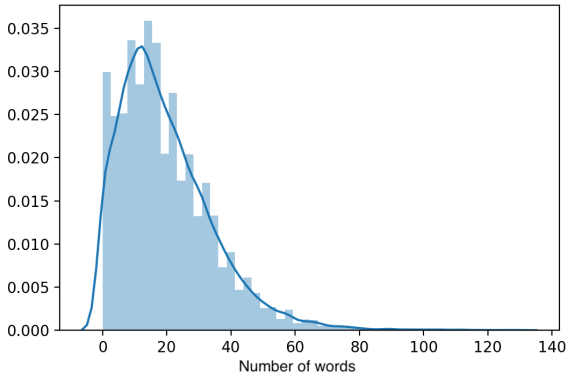

Figure 1: Distribution of sentence length.

2.4 Models & Architectures

As mentioned, we pair the data from handcrafted features with a Gradient Boosted Trees (GBT) model (Drucker and Cortes, 1996). Table 1 shows all hyperparameter values set for the GBT model. These values are the result of extensive grid searching, optimizing for F1-score (the task's official metric), and selecting the best performing model on 5-fold cross-validated results.

Additionally, we devise two deep-learning models to pair with word embedding representations.

estimators	100
learning-rate	0.1
loss	*exponential*
max. tree depth	10
min. samples at leaf	10
min. samples to split	2

Table 1: Hyperparameter values for GBT.

The *BiLSTM* model consists of a bidirectional long short-term memory layer (Gers et al., 2000). The last hidden time-step, concatenated from both directions, is then passed through a final fully connected layer followed by a sigmoid activation function. The *ABL* (**A**ttention-based **B**idirectional **L**STM) model is similar to the *BiLSTM* model, with an added attention layer (Bahdanau et al., 2014) operating over the hidden LSTM representations. Figure 2 shows this model's architecture. We use 40% dropout (Srivastava et al., 2014) on the initial embeddings, and 20% dropout on all remaining hidden-layers. All LSTM layers use 50 as the number of features of the hidden state.

For training, we use the Adam optimizer (Kingma and Ba, 2014) with default parameters, and Binary Cross-Entropy as the loss function. The batch size was set to 16, and training was stopped after 25 epochs, with early stopping upon 5 consecutive non-improving epochs on validation loss.

Deep-learning models were implemented using *PyTorch* (Paszke et al., 2017), and GBT using *scikit-learn* (Pedregosa et al., 2011).

3 Results and Discussion

Table 2 shows the results of all models over 5-fold cross-validation on the provided SLC training data. The top rows correspond to systems trained on a *balanced* version of the provided dataset, by means of random undersampling of the majority class (Japkowicz and Stephen, 2002), as an attempt to tackle the class imbalance on the original dataset (only 27.8% of which corresponds to *propaganda* sentences).

On the *balanced* dataset, the *ABL* model is the best-performing on both F1-score (official task metric) and accuracy, while *BiLSTM* achieved the best F1-score on the original data. *GBT* has a surprisingly inferior F1-score on the original data (32.6 points vs 53.0 points on the F1-metric for *BiLSTM*), but suffers the largest boost when compared with its training on the *balanced* data (67.7 F1-score). Nonetheless, models based on word embeddings (*BiLSTM* & *ABL*) perform far better than those based on a handcrafted selection of features (*GBT*). This is expected, as *n-grams* fail to encode the text as a sequence, and fail to carry the meaning and relations between each word, which are known to be encoded in word embeddings (Peters et al., 2018).

Model	***F1***	*P*	*R*	*A*
ABL *Balanced*	75.0	71.9	78.5	73.9
BiLSTM *Balanced*	74.7	69.5	80.7	72.6
GBT *Balanced*	67.7	65.8	69.6	66.7
BiLSTM	53.0	60.7	48.3	76.5
ABL	52.1	62.6	46.0	77.0
GBT	32.6	38.0	28.7	67.1

Table 2: Propaganda detection performance over 5-fold cross-validation. Models are ordered by decreasing F1-score (the task's official metric).

Model	***F1***	*P*	*R*
Best (team *ltuorp*)	63.2	60.3	66.5
Ours (ABL-*Balanced-Ens*)	59.5	50.9	71.6

Table 3: Official results for propaganda detection task (on withheld test data).

Regarding the effectiveness of training on a balanced dataset, all systems saw dramatically increased performance on metrics relative to the positive class (labeled *propaganda*), accompanied by small decreases of overall accuracy. This is expected, as we are effectively depriving the model of useful samples from the majority class (labeled *non-propaganda*), but remarkably beneficial as can be seen by the improved F1-scores.

Our submission to the task was a 5-member *ABL* ensemble (ABL-*Balanced-Ens*), from 5 cross-validation iterations, trained on the *balanced* data. This system's predictions were the average of each model's independent prediction. This follows numerous works demonstrating consistent performance improvements when using ensembles of deep-learning classifiers (Peters et al., 2018).

Table 3 presents ours results on the official test data. Our system achieved 59.5 F1-score, ranking 10^{th} among 25 participants, but lagging only 3.7 F1 points behind the best-performing system.

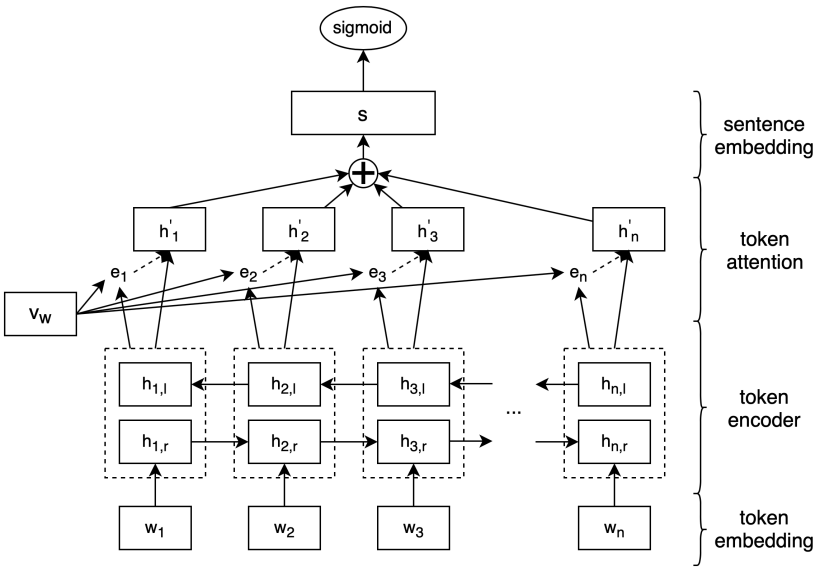

Figure 2: Visualization of ABL (bidirectional LSTM with attention).

3.1 Analyzing Attention Weights

Although the predictions of deep learning models are notoriously opaque, the attention layer present in *ABL* renders some degree of interpretability possible. By analyzing the attention energy associated with each word, we can intuitively extract conclusions regarding which parts of a sentence a model is taking into consideration.

Figure 3 shows a plot of attention energies over a sample article. The model seems to track writing style mostly through verb conjugations (*e.g.* 'needs', 'given', 'unprecedented'), as well as words with strong connotation which often portray the writer's opinion (*e.g.* 'wretched deals', 'machination', 'horrify').

From the sentences shown in Figure 3, the model incorrectly classifies the 4^{th} and 5^{th} sentences as *non-propaganda* (marked •), although with markedly low confidence (8% and 18% respectively). All remaining sentences are correctly classified. Through inspection of several attention-plots, intuitively, the model seems to pay close attention to a single opinion-inducing word when classifying a sentence as *propaganda*, while featuring a broader spread of attention weights when classifying a sentence as *non-propaganda*. The latter happens for both the 4^{th} and 5^{th} sentences.

•	He needs medical attention, the kind of treatment you get only in a hospital.
•	But it has been made clear to him that if he attempts to go to a hospital he will not be given free passage and he will be arrested.
+	His treatment amounts to the most unprecedented persecution.
•	Julian could leave the embassy if his own government, the government of his homeland, Australia, applied legitimate diplomatic pressure on behalf of its citizen.
•	We must ask ourselves why this hasn' t happened.
+	But that might be one of the so-called "wretched deals" that are being offered Assange.
•	Some very strange things are being said by senior members of these two governments.
+	The new foreign secretary of the United Kingdom, Jeremy Hunt, said sarcastically that the British police would offer Julian "a warm welcome" when he came out, when he would face serious charges.
•	Was Hunt referring to a deal which has already been done with the United States on extradition?
+	But this is the milieu of machination around someone who has the right of natural justice concerning his freedom.
+	Putting aside freedom of speech, the persecution of this man has been something that should horrify all free-thinking people.

Figure 3: Plots of attention-weights. Sentences are marked with + if **predicted** to be propaganda, and • otherwise. Symbol is colored red if prediction is wrong.

4 Conclusions and Future Work

We experimented with several models for sentence-level propaganda detection, exploring both handcrafted features and word embeddings. As expected, deep learning models improve performance to the detriment of feature inter-

pretability. The best performing model is based on a bidirectional LSTM followed by an attention layer. We have submitted a 5-member ensemble of this model to the NLP4IF shared task, achieving 59.5 F1-score on the official test data, and ranking 10^{th} among 25 participants.

Additionally, we have experimented with random undersampling to tackle the class imbalance on the provided training data. This lead to dramatic performance improvements on all models for metrics related to the minority class, accompanied by a small decrease in accuracy.

For future work, we intend to explore the integration of handcrafted features with word embeddings, to improve both model performance and transparency. We also intend to experiment with ensembles of independent classifiers, from independent feature-sets, in order to capture different facets of this complex problem.

Acknowledgments

André Ferreira Cruz is supported by the Calouste Gulbenkian Foundation, under grant number 226338. Gil Rocha is supported by a PhD studentship (with reference SFRH/BD/140125/2018) from Fundação para a Ciência e a Tecnologia (FCT). This research is partially supported by project DARGMINTS (POCI/01/0145/FEDER/031460), funded by FCT.

References

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 724–728.

Yaman Akdeniz. 2010. To block or not to block: European approaches to content regulation, and implications for freedom of expression. *Computer Law & Security Review*, 26(3):260–272.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

André Ferreira Cruz, Gil Rocha, Rui Sousa-Silva, and Henrique Lopes Cardoso. 2019. Team fernandopessa at SemEval-2019 task 4: Back to basics in hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 999–1003, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, EMNLP-IJCNLP 2019, 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*.

Harris Drucker and Corinna Cortes. 1996. Boosting decision trees. In *Advances in neural information processing systems*, pages 479–485.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with lstm. *Neural Computation*, 12(10):2451–2471.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Wendell Johnson. 1944. Studies in language behavior: A program of research. *Psychological Monographs*, 56(2):1–15.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Joseph Prusa, Taghi M Khoshgoftaar, David J Dittman, and Amri Napolitano. 2015. Using random undersampling to alleviate class imbalance on tweet sentiment data. In *2015 IEEE international conference*

on information reuse and integration, pages 197–202. IEEE.

Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520.

Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, and Jeremy Singer-Vine. 2016. Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. *Buzzfeed News*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.