Rouletabille at SemEval-2019 Task 4: Neural Network Baseline for Identification of Hyperpartisan Publishers

Jose G. Moreno and **Yoann Pitarch** and **Karen Pinel-Sauvagnat** and **Gilles Hubert**
IRIT / University of Toulouse
France
{jose.moreno, yoann.pitarch, karen.sauvagnat, gilles.hubert}@irit.fr

Abstract

This paper describes the Rouletabille participation to the Hyperpartisan News Detection task. We propose the use of different text classification methods for this task. Preliminary experiments using a similar collection used in Potthast et al. (2018) show that neural-based classification methods reach state-of-the art results. Our final submission is composed of a unique run that ranks among all runs at 3/49 position for the by-publisher test dataset and 43/96 for the by-article test dataset in terms of Accuracy.

1 Introduction

Printed press have been in the last decades the main way to access to news in written format. This tendency is changing with the appearance of online channels but usually the main factors of the journalistic content generation are still there: events, journalists, and editors. One of the problems of the generation of this content is the influence of each factor in the veracity of the generated content. Two main factors may influence the final view of an article: writer's preferences and affiliation of the editor house.

Identifying partisan preferences in news, based only on text content, has been shown to be a challenging task (Potthast et al., 2018). This problem requires to identify if a news article was written in such a way that it includes an overrated appreciation of one of the participants in the news (a political party, a person, a company, etc.). Despite the fact that sharply polarized documents are not necessarily fake, it is an early problem to solve for the identification of fake content. A recent paper (Potthast et al., 2018) claims that stylometric features are a key factor to tackle this task.

In this paper, we present the description of our participation to the Hyperpartisan classification task at SemEval-2019 (Kiesel et al., 2019). This task was composed of two subtasks, the first consist to identify hyperpartisan bias in documents classified by its individual content (bias of the writer or by-article category) and the second by the editorial house that published the article (bias of the editorial house or by-publisher category) as depicted in Figure 1[1]. To address this problem, we experimented with well-known models based on deep learning (Honnibal and Montani, 2017; Kim, 2014). They achieve state-of-the-art results on a publicly available collection (Potthast et al., 2018), showing that neural models can effectively address the task of hyperpartisan detection without including stylometric features. Our final submission ranked in the top-3 for the by-publisher category, and 43/96 for the by-article category (or 21/42 in the official ranking).

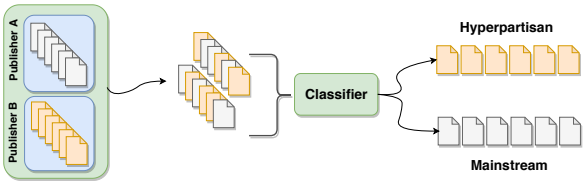

Figure 1: Publisher-based pipeline performed in training phase. During testing, different publishers were used and labels were unknown.

2 Classification Models

We have considered that the hyperpartisan classification task can be addressed as a binary classification task where only two classes ('hyperpartisan' and 'mainstream').

Three different models were considered for our participation. The first of them is based on a classical document-level representation and the

[1]More details of the dataset construction can be found in Kiesel et al. (2019)

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 981–984
Minneapolis, Minnesota, USA, June 6–7, 2019.

other two are based on word-level representations through the use of word embedding. All of them can be seen as baselines and no specific adaptation to the dataset[2] was performed.[3]

2.1 TF-IDF + Adaboost

For this model we represented our documents using the classical TF-IDF representation. Finally, the Adaboost classifier (Freund and Schapire, 1997) is used under the default configuration. Note that this is a very basic baseline, as it does not use recent representation techniques such as word embeddings.

2.2 SpaCy Model

In this case, we used the SpaCy (Honnibal and Montani, 2017) library[4]. We used the text categorisation algorithm implemented in SpaCy which is based on the hierarchical attention network proposed in Yang et al. (2016). The main improvement to the original model is the use of hash-based embeddings. We only defined two hyperparameters for the model: number of epochs and dropout rate. These parameters were set to 3 and 0.2, respectevelly.[5]

2.3 Convolutional Model

We also tested the neural classification model proposed by Kim (2014). This model uses convolutional neural networks that are finally reduced to a binary classification. This method is known as a highly competitive classification model for short documents. As SpaCy, this model is based on word embeddings representation. However, in this case we preferred to use the pre-calculated embeddings of GloVe (Pennington et al., 2014). Hyperparameters were defined using the training data.

3 Experiments and Results

3.1 Experimental Setup

Experiments were performed using two collections, the ACL2018 collection (Potthast et al., 2018) and the SemEval19 collection (Potthast et al., 2019). The first collection is composed of 1627 articles including 801 hyperpartisan and 826 mainstream manually annotated documents. As this collection is not originally split in training-test sets, results are presented using cross-validation. The second collection was split in train, validation, and test sets for the by-publisher category, and in train and test for the by-article category as presented in Table 1. Results in this second collection are exclusively calculated using the TIRA evaluation system (Potthast et al., 2019).

	training	validation	test
by-article	645	-	628
by-publisher	600000	150000	4000

Table 1: Number of documents used for training, validation, and test used in the SemEval19 collection.

In order to determine the best configuration to our participation using the SemEval collection, we decided to perform experiments and fix hyperparameters using the ACL2018 collection.

3.2 Results in the ACL2018 Collection

Table 2 reported results of the 3 classification models presented in section 2 (lines labelled TF-IDF+Adaboost, SpaCy and CNN-Kim), as well as results of the approach presented in Potthast et al. (2018) (line labelled ACL18), on the ACL2018 collection.

We only used the first fold produced by the authors' code[6]. As our results are not directly comparable with the values reported in Potthast et al. (2018), we re-evaluated their approach on this single fold.

Values of the three F-measures were calculated with sklearn[7]. Note that in binary classification, micro F-measure values are equivalent to accuracy values.

Two state-of-the-art models (SpaCy and Kim (2014)) outperform the approach presented in Potthast et al. (2018), showing that stylometric features are probably not necessary for the task.

3.3 Results in the Semeval2019 Collection

Experiments on the official collection were performed through the use of TIRA (Potthast et al., 2019)[8]. As our previous experiments have not shown clear improvement with the convolutional model, we submitted our official runs using

[2]Different to the classical training of the involved classifiers.

[3]Further experiments were performed using network-based models but as results did not show improvement in an existing collection, we decided to not include these results.

[4]https://spacy.io/

[5]We based on SpaCy's guidelines.

[6]https://github.com/webis-de/ACL-18

[7]https://scikit-learn.org/

[8]https://www.tira.io/task/hyperpartisan-news-detection/

	F-measure		
	macro	accuracy /micro	weighted
ACL18	0.7605	0.7509	0.7480
TF-IDF + Adaboost	0.7069	0.7130	0.7039
SpaCy (dp =0.2, epochs=3)	0.8087	0.8091	0.8081
CNN-Kim	**0.8273**	**0.8306**	**0.8290**

Table 2: Macro, micro and weighted F-measure for the ACL2018 collection.

	accuracy/micro	f1
top1	**0.7060**	0.6825
top2	0.6998	0.6587
our (rank 3/49)	0.6805	**0.7213**
top4	0.6640	0.7061

Table 3: Official results for the by-publisher test dataset.

SpaCy: it can be seen as an 'easy-to-implement' but strong baseline. The same model was trained on the by-publisher training set for both submissions (on the by-publisher and by-article dataset).

Tables 3 and 4 respectively present official results on the by-publisher[9] and the by-article datasets.

One can see that relative results (i.e. regarding the official ranking) are strongly better on the by-publisher dataset than on the by-article one. This can be easily explained by the fact that collections were differently annotated.

If we now compare accuracy scores of the SpaCy model between the ACL2018 collection and the SemEval2019 one, we can notice a decrease in performance (0.6640 vs 0.8091 on the by-publisher dataset for example), leading us to think that there exist some differences between the two collections. Both collections seem to be complementary for the evaluation of hyperpartisan detection.

Another important observation is that the SpaCy model performs remarkably well on the by-publisher set, although not specifically tuned for the hyperpartisan detection task. Indeed, we are ranked first on the F1 metric, and 3rd on the Accuracy one. Some other experiments are needed to get a fine-tuned model for the task, but this version can already be considered as a strong baseline for the by-publiser subtask.

[9]https://www.tira.io/task/hyperpartisan-news-detection/dataset/pan19-hyperpartisan-news-detection-by-publisher-test-dataset-2018-12-12/ last visit 19/02/2019.

	accuracy /micro	f1
top1	**0.8217**	0.8089
top2	0.8201	**0.8215**
top3	0.8089	0.8046
our (rank 43/96)	0.7245	0.6905

Table 4: Official results for the by-article test dataset.

4 Conclusion

Our experiments and participation to the Hyperpartisan task led us to conclude that:

- stylometric features seem not to be necessary to achieve state-of-the-art results for hyperpartisan detection in the ACL2018 collection. This deserves a set of extra experiments to better understand the real contribution of stylometric features when combined with strong representations/classifiers to validate the work of Potthast et al. (2018).
- a state-of-the-art classification model in its default configuration (SpaCy) can be considered as a strong baseline for next experiments. Indeed, SpaCy is top-ranked according to the F1 metric on the by-publisher dataset. One question is thus now if other top-ranked approaches are also from the text classification literature or dedicated ones.

References

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139.

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

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 *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval 2019)*. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. 2019. TIRA Integrated Research Architecture. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*. Springer.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 231–240. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.