Tintin at SemEval-2019 Task 4: Detecting Hyperpartisan News Article with only Simple Tokens

Yves Bestgen
Centre for English Corpus Linguistics
Université catholique de Louvain
Place Cardinal Mercier, 10 1348 Louvain-la-Neuve, Belgium
yves.bestgen@uclouvain.be

Abstract

Tintin, the system proposed by the CECL for the Hyperpartisan News Detection task of SemEval 2019, is exclusively based on the tokens that make up the documents and a standard supervised learning procedure. It obtained very contrasting results: poor on the main task, but much more effective at distinguishing documents published by hyperpartisan media outlets from unbiased ones, as it ranked first. An analysis of the most important features highlighted the positive aspects, but also some potential limitations of the approach.

1 Introduction

This report presents the participation of Tintin (Centre for English Corpus Linguistics) in Task 4 of SemEval 2019 entitled Hyperpartisan News Detection. This task is defined as follows by the organizers[1]: "Given a news article text, decide whether it follows a hyperpartisan argumentation, i.e., whether it exhibits blind, prejudiced, or unreasoning allegiance to one party, faction, cause, or person."

This question is related to the detection of fake news, a hot topic in our internet and social media world (Pérez-Rosas et al., 2017). There are, however, essential differences between these two tasks. An article can be hyperpartisan without mentioning any fake content. Another difference is that it is a news article (or even a claim) that is fake whereas a news article but also a media outlet (or publisher) can be considered as hyperpartisan. The challenge organizers took these two possibilities (i.e. an article or a publisher can be hyperpartisan) into account by offering two test sets. The main test set, the labels-by-article one, contained documents that had been assessed as hyperpartisan or not by human judges, while the documents in the secondary test set, the labels-by-publisher one, had been categorized according to whether their publishers were considered to be hyperpartisan or not by organizations that disseminate this type of evaluation. In both these test sets, participants had to decide whether a document expresses a hyperpartisan point-of-view or not.

If the main task is particularly interesting, the secondary task is also relevant because it is about achieving through an automatic procedure what a series of organizations manually perform in a way that is sometimes called into question as to its impartiality and quality (Wilner, 2018). However, in this context, the task would preferably be evaluated, not at the document level, but at the publisher level by providing several documents from a publisher and asking whether the publisher is biased or not. Nevertheless, it can be assumed that many systems developed for categorizing publishers will start by evaluating each document separately and thus getting good performance in the current secondary task is at least a first step.

To take up these tasks, the question is how to determine automatically whether a document is hyperpartisan or not. This question has not attracted much attention in the literature, but, very recently, Potthast et al. (2018) proposed to use stylometric features such as characters, stop words and POS-tag n-grams, and readability measures. They compared the effectiveness of this approach to several baselines including a classical bag-of-words feature approach[2] (Burfoot and Baldwin, 2009). Their stylistic approach obtained an accuracy of 0.75 in 3-fold cross-validation in which publishers present in the validation fold were unseen during the learning phase. The bag-of-words feature

[1] https://pan.webis.de/semeval19/semeval19-web

[2] More specifically, Potthast et al. (2018) used the frequency, normalized by the document length, of the tokens of at least two characters that occurred in at least 2.5% of the documents in the collection.

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 1062–1066
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

approach obtained an accuracy of 0.71, which is not much lower. These results were obtained on a small size corpus (due to the cost of the manual fact-checking needed for the fake-news part of the study) containing only nine different publishers. It is therefore not evident that this corpus was large enough to evaluate the degree of generalizability of the bag-of-words approach, especially since Potthast et al. (2018, p. 233) emphasizes that using bag-of-words features potentially related to the topic of the documents renders the resulting classifier not generalizable. In contrast, the datasets prepared for the present challenge are significantly larger since the latest versions available contain more than 750,000 documents and more than 240 different media outlets.

Therefore, it seemed interesting to evaluate the effectiveness of a bag-of-words approach for the labels-by-publisher task, the one used by Potthast et al. (2018). This is the purpose of this study. Another reason why I chose to focus on the labels-by-publisher task is that I was unclear about what could be learned on the basis of the labels-by-publisher sets for the labels-by-article test set. If one can think that some publishers almost always distribute hyperpartisan articles, it seems doubtful that this is the case for all of them.

The next sections of this paper describe the datasets, the developed system, and the obtained results as well as an analysis of the most important features.

2 Data

As explained in Kiesel et al. (2019), several datasets of very different sizes were available for this challenge. The learning labels-by-publisher set contained 600,000 documents form 158 media outlets in its final version. The corresponding validation set contained 150,000 documents from 83 media outlets, and the test set consisted of 4,000 documents. The first labels-by-article set provided to the participants contained 645 documents and was intended for fine-tuning systems developed on the labels-by-publisher sets. The test set contained 628 documents.

Some of these datasets could be downloaded while those used to perform the final test were hidden on a TIRA server (Potthast et al., 2019). An important feature of these data is that no publisher in a dataset is present in any other dataset. This has the effect of penalizing (usefully) any system that learns to categorize on the basis of the publishers since generalization to unseen media outlets should be problematic.

3 System

3.1 The Bag-of-Words Feature Approach

The developed system, which implements the bag-of-words approach, is very classical. It includes steps for preprocessing the data, reading the documents, creating a dictionary of tokens (only unigram tokens as bigrams did not appear to improve performance), and producing the file for the supervised learning procedure. It was written in C, with an initial data cleaning step in Perl, and was thus very easy to install on a TIRA server. In this section, only a few implementation details are mentioned.

During preprocessing, a series of character sequences like *;amp;amp;amp;*, *&amp;#160;* and *&amp;amp;lt;* were regularized. When reading a document (both the title and the text), strings were split by separating the following characters when they were at the beginning or end of the strings and they were outputted separately: ’ * ” ? . ; : / ! ,) (} { [ ] -. Alphabetic characters were lowercased. A binary feature weighting scheme was used.

3.2 Supervised Learning Procedure

During the development and test phases of the challenge, the models were build using two solvers available in the LIBLINEAR package (Fan et al., 2008), the L2-regularized L2-loss support vector classification (-s 1) and the L2-regularized logistic regression (-s 7), which resulted in equivalent performance. The regularization parameter C was optimized on the labels-by-publisher validation set using a grid search.

4 Analyses and Results

4.1 Official Results

On the main task of the challenge, the Tintin system obtained an accuracy of 0.656, ranking 27th out of 42 teams, very far from the best teams who scored 0.82.

Twenty-nine teams submitted a system for the labels-by-publisher task. Tintin ranked first, with an accuracy of 0.706. This level of performance is identical to that obtained by Potthast et al. (2018) bag-of-words model in their experiments on a significantly smaller dataset.

In general, the performances of the different teams on the second task were much lower than on the main task. Tintin, on the other hand, achieved a better score on the second task. It is not the only system in this case since, of the 28 teams that participated in the two tasks, three others also scored better in the second task and one team only participated in this task. Reading the papers describing these systems will make it possible to know if these teams have also chosen to favor the secondary task. It is also noteworthy that the difference between the two best teams is much greater in the secondary task (0.706 vs. 0.681) than in the main task (0.822 vs. 0.820).

4.2 Analysis of the Most Important Features

In order to get an idea of the kind of features underlying the system's efficiency in the secondary task, the 200 features (and thus tokens) that received the highest weights (in absolute value) in the logistic regression model[3] were examined.

Table 1 shows the ten features that received the highest weights as well as a series of features selected because of their interest to understand how the system works. Positive weights indicate that the feature predicts the hyperpartisan category, while negative weights are attributed to features that are typical of the non-biased category. The table gives in addition to the token and the weight, the number of publishers (#Pub) and the number of documents (#Doc) in which that token appears for each of the two categories to be predicted. The maximum percentage of documents a publisher represents in each category is also provided (Max%). The percentage for the category that this feature predicts is boldfaced.

As expected, some of the most important features are typical of a single publisher like *globalpost*, which is present in 750 times more non-biased than hyperpartisan documents, but 99.76% of the non-biased documents come from the same publisher (*pri.org*). Other tokens are not so strongly associated with a single publisher. In the 8th position, the token *h/t*, a way of acknowledging a source, is present in 53 hyperpartisan media outlets and 63% of the documents of this category in which it occurs are not found in the publisher that contains the most (*dailywire.com*). *Jan* is an even more obvious example of features that are not tied to a single publisher.

There are also in these particularly important features some tokens that might not be seen as unexpected such as *leftists(s)*, *shit*, *beast*, *right-wing*, *hell...* Other features, such as *fla*, *beacon*, *alternet* or *via*, are not related to a single publisher, but their usefulness for categorizing unseen media outlets is no less debatable. For instance, *via* can be used in many different contexts such as *via twitter*, *transmitted to humans via fleas*, *linking Damascus to Latakia and Aleppo via Homs*. It is therefore widespread. However, its usefulness in categorizing *unseen* media outlets is not necessarily obvious since some part of its weight results from its occurrence in all of the 976 documents from *thenewcivilrightsmov* as each of these documents offers to *subscribe to the New Civil Rights Movement via email.*

These observations lead to wonder whether the system does not show a strong variability of efficiency according to the unseen publishers to predict, working well for some, but badly for others. It was not possible to evaluate this conjecture by analyzing the system accuracy for the different publishers in the test set since it is not publicly available. However, an indirect argument in its favor is provided by the meta-learning analyses done by the task's organizers that suggest that some publishers are much easier to predict than others. For these analyses, each set was randomly split into two samples (2668 vs. 1332 for the labels-by-publisher test set) and submitted to a majority voting procedure. As this procedure is unsupervised, the expected value of the difference in accuracy between the two samples is 0. This was not the case for the labels-by-publisher task since it was larger than 0.23, an extremely significant difference (Chi-square test). The most obvious explanation is to consider that the need to put each publisher in only one sample leads to a nonrandom distribution in which the publishers of one sample are much easier to predict.

5 Conclusion

The Tintin system, developed for the Hyperpartisan News Detection task, is extremely simple since it is exclusively based on the document tokens. If its performance on the main task was poor, it ranked first when it was used to discriminate documents published by hyperpartisan media outlets from unbiased ones. An analysis of the

[3] As the features are binary coded, weight is the sole factor that affect the classification function for an instance.

			Unbiased			Hyperpartisan		
Rank	**Token**	**Score**	**#Pub**	**#Doc**	**Max%**	**#Pub**	**#Doc**	**Max%**
1	globalpost	-2.40	4	10506	**99.7**	10	14	21.4
2	n.m	-1.94	17	17951	**96.1**	18	107	21.4
3	upi	-1.64	15	7541	**94.6**	24	186	48.3
4	>	-1.36	13	2022	**87.2**	27	242	51.2
5	_	-1.17	8	7516	**99.4**	9	170	34.1
6	©	1.16	15	1821	76.7	25	8840	**53.1**
7	h/t	1.06	11	71	33.8	53	3016	**36.8**
8	fe	-1.03	20	13797	**97.4**	28	294	40.4
9	et	0.95	29	2326	28.3	76	13306	**84.4**
10	jan	-0.90	38	19428	**27.2**	75	4937	45.7
22	trump's	0.73	28	2966	31.6	62	9164	**39.2**
35	via	0.61	37	10738	31.3	104	24279	**18.1**
62	fla	-0.51	26	3481	**36.2**	47	797	29.2
66	leftists	0.50	16	147	27.8	74	2984	**30.6**
67	leftist	0.49	19	895	26.0	79	4904	**35.0**
76	shit	0.47	18	136	27.2	67	2167	**39.9**
82	beast	0.46	26	887	25.7	79	3151	**30.5**
95	right-wing	0.44	26	1542	23.0	88	8608	**32.7**
97	beacon	0.43	26	569	27.5	72	2048	**25.1**
143	yesterday	0.37	32	3849	18.9	102	10018	**21.4**
171	hell	0.34	31	2518	28.9	97	8382	**35.0**
192	alternet	0.33	9	15	26.6	28	795	**33.0**

Table 1: Some of the 200 most useful features for predicting hyperpartisanship.

Figure 1: Main entrance of the Musée Hergé[4].

most important features for predicting hyperpartisanship emphasizes the presence of tokens specific to certain publishers, but also of tokens that could have some degree of generalizability.

In future work, it might be interesting to use other weighting functions than the binary one such as the bi-normal separation feature scaling (Forman, 2008) that has been shown to be particularly effective for satire detection (Burfoot and Baldwin, 2009) or BM25 which has proved useful in the VarDial challenge (Bestgen, 2017). Such development, however, would only be justified if the system is stable, that is to say, if it achieves good performance for many publishers not seen during learning. Designing a weighting function that would favor the hyperpartisan distinction while simultaneously reducing the impact of the media outlets could perhaps improve this stability.

6 Namesake: Tintin

I chose this fictitious reporter as my namesake for this task because the Musée Hergé, an unusual looking building in front of which a huge fresco represents this cartoon character, is located a few tens of meters from my office in Louvain-la-Neuve. *Tintin* is also a French interjection that means *nothing* or *No way!*

Acknowledgments

The author is a Research Associate of the Fonds de la Recherche Scientifique - FNRS.

[4] ©V. Pypaert, CC BY-SA 4.0, from Wikimedia.

References

Yves Bestgen. 2017. Improving the character ngram model for the DSL task with BM25 weighting and less frequently used feature sets. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 115–123, Valencia, Spain. Association for Computational Linguistics.

Clint Burfoot and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

George Forman. 2008. BNS feature scaling: an improved representation over tf-idf for svm text classification. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 263–270, New York, NY, USA. ACM.

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.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401. Association for Computational Linguistics.

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.

Tamar Wilner. 2018. We can probably measure media bias. but do we want to? *Columbia Journalism Review*, Retrieved 2019-02-01.