

# Team QCRI-MIT at SemEval-2019 Task 4: Propaganda Analysis Meets Hyperpartisan News Detection

Abdelrhman Saleh<sup>1</sup>, Ramy Baly<sup>2</sup>, Alberto Barrón-Cedeño<sup>3</sup>,  
 Giovanni Da San Martino<sup>3</sup>, Mitra Mohtarami<sup>2</sup>, Preslav Nakov<sup>3</sup>, James Glass<sup>2</sup>

<sup>1</sup>Harvard University, MA, USA

<sup>2</sup>MIT Computer Science and Artificial Intelligence Laboratory, MA, USA

<sup>3</sup>Qatar Computing Research Institute, HBKU, Qatar

abdelrhman.saleh@college.harvard.edu,

{baly, mitram, glass}@mit.edu

{albarron, gmartino, pnakov}@hbku.edu.qa

## Abstract

In this paper, we describe our submission to SemEval-2019 Task 4 on Hyperpartisan News Detection. Our system relies on a variety of engineered features originally used to detect propaganda. This is based on the assumption that biased messages are propagandistic in the sense that they promote a particular political cause or viewpoint. We trained a logistic regression model with features ranging from simple bag-of-words to vocabulary richness and text readability features. Our system achieved 72.9% accuracy on the test data that is annotated manually and 60.8% on the test data that is annotated with distant supervision. Additional experiments showed that significant performance improvements can be achieved with better feature pre-processing.<sup>1</sup>

## 1 Introduction

The rise of social media has enabled people to easily share information with a large audience without regulations or quality control. This has allowed malicious users to spread disinformation and misinformation (a.k.a. “fake news”) at an unprecedented rate. Fake news is typically characterized as being hyperpartisan (one-sided), emotional and riddled with lies (Potthast et al., 2017a). The SemEval-2019 Task 4 on Hyperpartisan News Detection (Kiesel et al., 2019) focused on the challenge of automatically identifying whether a text is hyperpartisan or not. While hyperpartisanship is defined as “exhibiting one or more of blind, prejudiced, or unreasoning allegiance to one party, faction, cause, or person”, we model this task as a binary document classification problem.

Scholars have argued that all biased messages can be considered propagandistic, regardless of whether the bias was intentional or not (Ellul,

1965, p. XV). As a result, we approached the task departing from an existing model for propaganda identification (Barrón-Cedeño et al., 2019). Our hypothesis is that as propaganda is inherent in hyperpartisanship – the two problems are two sides of the same coin, and solving one of them would help solve the other. Our system consists of a logistic regression model that is trained with a variety of engineered features that range from word and character TFIDF  $n$ -grams and lexicon-based features to more sophisticated features that represent different aspects of the article’s text such as the richness of its vocabulary and the complexity of its language.

Our official submission achieved an accuracy of 72.9% (while the winning system achieved 82.2%). This was achieved using word and character  $n$ -grams. Additional, post-submission experiments show that further performance improvements can be achieved by careful pre-processing of the engineered features.

## 2 Related Work

The analysis of bias and disinformation has attracted significant attention, especially after the 2016 US presidential election (Brill, 2001; Finberg et al., 2002; Castillo et al., 2011; Baly et al., 2018a; Kulkarni et al., 2018; Mihaylov et al., 2018). Most of the proposed approaches have focused on predicting credibility, bias or stance. Popat et al. (2017) assessed the credibility of claims based on the occurrence of assertive and factive verbs, hedges, implicative words, report verbs and discourse markers, which were extracted using manually crafted gazetteers (referred to as stylistic features).

Stance detection was considered as an intermediate step for detecting fake claims, where the veracity of a claim is checked by aggregating the

---

<sup>1</sup>Our system is available at <https://github.com/AbdulSaleh/QCRI-MIT-SemEval2019-Task4>

stances of retrieved relevant articles (Baly et al., 2018b). Several stance detection models have been proposed as part of the Fake News Challenge (FNC)<sup>2</sup> including deep convolutional neural networks (Baird et al., 2017), multi-layer perceptrons (Hanselowski et al., 2018), and end-to-end memory networks (Mohtarami et al., 2018)

The stylometric analysis model of Koppel et al. (2007) was used by Potthast et al. (2017b) when looking for hyperpartisanship. They used articles from nine news sources whose factuality has been manually verified by professional journalists. Writing style and complexity was also considered by Horne and Adal (2017) to differentiate real news from fake news and satire. They used features such as the number of occurrences of different part-of-speech tags, swearing and slang words, stop words, punctuation, and negation as stylistic markers. They also used a number of readability measures. Rashkin et al. (2017) focused on a multi-class setting: real news, satire, hoax, or propaganda. Their supervised model relied on word  $n$ -grams.

Similarly to Potthast et al. (2017b), we believe that there is an inherent style in propaganda, regardless of the source publishing it. Many stylistic features were proposed for authorship identification, i.e., the task of predicting whether a piece of text has been written by a particular author. One of the most successful representations for such a task are character-level  $n$ -grams (Stamatatos, 2009), and they turn out to represent some of our most important stylistic features.

More details about research on fact-checking and the spread of fake news online can be found in (Lazer et al., 2018; Vosoughi et al., 2018; Thorne and Vlachos, 2018).

### 3 System Description

We developed our system for detecting hyperpartisanship in news articles by training a logistic regression classifier using a set of engineered features that included the following: character and word  $n$ -grams, lexicon-based indicators, and readability and vocabulary richness measures. Below, we describe these features in detail.

**Character 3-grams.** Stamatatos (2009) argued that, for tasks where the topic is irrelevant, character-level representations are more sensitive

than token-level ones. We hypothesize that this applies to hyperpartisan news detection, since articles on both sides of the political spectrum may be discussing the same topics. Stamatatos (2009) found that “the most frequent character  $n$ -grams are the most important features for stylistic purposes”. These features capture different style markers, such as prefixes, suffixes and punctuation marks. Following the analysis in Barrón-Cedeño et al. (2019), we include TFIDF-weighted character 3-grams in our feature set.

**Word  $n$ -grams** Bag-of-words (BoW) features are widely used for text classification. We extracted the  $k$  most frequent [1, 2]-grams, and we represented them using their TFIDF scores. We ignored  $n$ -grams that appeared in more than 90% of the documents, most of which contained stop-words and were irrelevant with respect to hyperpartisanship. Furthermore, we incorporated Naive Bayes by weighing the  $n$ -grams based on their importance for classification, as proposed by Wang and Manning (2012). We define  $\mathbf{x}_i \in \mathbb{R}^{|V|}$  as a row vector in the TFIDF feature matrix, representing the  $i^{th}$  training sample with a target label  $y_i \in \{0, 1\}$ , where  $V$  is the vocabulary size. We also define vectors  $\mathbf{p} = \alpha + \sum_{i:y_i=1} \mathbf{x}_i$  and  $\mathbf{q} = \alpha + \sum_{i:y_i=0} \mathbf{x}_i$ , and we set the smoothing parameter  $\alpha$  to 1. Finally, we calculate the vector:

$$\mathbf{r} = \log \left( \frac{\mathbf{p}/\|\mathbf{p}\|}{\mathbf{q}/\|\mathbf{q}\|} \right) \quad (1)$$

which is used to scale the TFIDF features to create the NB-TFIDF features as follows:

$$\mathbf{x}'_i = \mathbf{r} \circ \mathbf{x}_i, \quad \forall i \quad (2)$$

**Bias Analysis** We analyze the bias in the language used in the documents by (i) creating bias lexicons that contain *left* and *right* bias cues, and (ii) using these lexicons to compute two scores for each document, indicating the intensity of bias towards each ideology. To generate the list of cues that signal biased language, we use Semantic Orientation (SO) (Turney, 2002) to identify the words that are strongly associated with each of the left and right documents in the training dataset. Those SO values can be either positive or negative, indicating association with right or left biases, respectively. Then, we select words whose absolute SO value is  $\geq 0.4$  to create two bias lexicons:  $BL_{left}$  and  $BL_{right}$ . Finally, we use these lexicons to compute two bias scores per document ac-

---

<sup>2</sup><http://www.fakenewschallenge.org>

cording to Equation (3), where for each document  $D_j$ , the frequency of cues in the lexicon  $BL_i$  that are present in  $D_j$  is normalized by the total number of words in  $D_j$ :

$$bias_i(D_j) = \frac{\sum_{cue \in BL_i} count(cue, D_j)}{\sum_{w_k \in D_j} count(w_k, D_j)} \quad (3)$$

**Lexicon-based Features.** Rashkin et al. (2017) studied the occurrence of specific types of words in different kinds of articles, and showed that words from certain lexicons (e.g., negation and swear words) appear more frequently in propaganda, satire, and hoax articles than in trustworthy articles. We capture this by extracting features that reflect the frequency of words from particular lexicons. We use 18 lexicons from the Wiktionary, Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001), Wilson’s subjectives (Wilson et al., 2005), Hyland’s hedges (Hyland, 2015), and Hooper’s assertives (Hooper, 1975). For each lexicon, we count the total number of words in the article that appear in the lexicon. This resulted in 18 features, one for each lexicon.

**Vocabulary Richness** Potthast et al. (2017b) showed that hyperpartisan outlets tend to use a writing style that is different from mainstream outlets. Different topic-independent features have been proposed to characterize the vocabulary richness, style and complexity of a text. For this task, we used the following vocabulary richness features: (i) type–token ratio (*TTR*): the ratio of types to tokens in a text, (ii) *Hapax Legomena*: number of types appearing once in a text, (iii) *Hapax Dislegomena*: number of types appearing twice in a text, (iv) *Honore’s R*: A combination of types, tokens and hapax legomena (Honore, 1979):

$$\text{Honore's R} = \frac{100 \times \log(|\text{tokens}|)}{1 - |\text{Legomena}|/|\text{types}|} \quad (4)$$

and (v) *Yule’s characteristic K*: The chance of a word occurring in a text following a Poisson distribution (Yule, 1944):

$$\text{Yule's K} = 10^4 \cdot \frac{\sum_i i^2 |\text{types}_i| - |\text{tokens}|}{|\text{tokens}|^2}, \quad (5)$$

where tokens refer to all words in a text (including repetitions), types refer to distinct words,  $i$  are the tokens’ frequency ranks (1 being the least frequent), and  $\text{types}_i$  are the number of tokens with the  $i^{th}$  frequency.

**Readability** We also used the following readability features that were originally designed to estimate the level of text complexity: 1) *Flesch–Kincaid grade level*: represents the US grade level necessary to understand a text (Kincaid et al., 1975), 2) *Flesch reading ease*: is a score for measuring how difficult a text is to read (Kincaid et al., 1975), and 3) *Gunning fog index*: estimates the years of formal education necessary to understand a text (Gunning, 1968).

## 4 Experiments and Results

### 4.1 Dataset

We trained our models on the Hyperpartisan News Dataset from SemEval-2019, Task 4 (Kiesel et al., 2019), which is split by the task organizers into: 1) *Labeled by-Publisher*: contains 750K articles labeled via distant supervision, i.e. using labels of their publisher<sup>3</sup>. Labels are evenly distributed across the “hyperpartisan” and “not-hyperpartisan” classes. This set is further split into 600K for training and 150K for validation. 2) *Labeled by-Article*: This set contains 645 articles labeled through crowd-sourcing (37% are hyperpartisan and 63% are not). Only articles with a consensus among annotators were included.

### 4.2 Experimental Setting

We train a logistic regression (LR) model with a Stochastic Average Gradient solver (Schmidt et al., 2017) due to the large size of the dataset. In order to reduce overfitting we use  $\mathbb{L}_2$  regularization (with  $C = 1$  as the regularization parameter). Feature normalization was needed since the different features represent different aspects of text, hence have very different scales. We tried to normalize each feature set by subtracting the mean and scaling it to unit variance. However, we found that multiplying the features by constant scaling factors resulted in better performance. The scaling factor for each family of features was a hyperparameter that was tuned during the fine-tuning experiments.

---

<sup>3</sup>Publishers labels are identified by BuzzFeed journalists or by the Media Bias/Fact Check project

Features	Labeled by-Article				Labeled by-Publisher			
	Accuracy	Prec.	Rec.	F1	Accuracy	Prec.	Rec.	F1
1 BoW (TFIDF)	67.8	53.8	89.1	67.1	56.7	55.1	72.5	62.6
2 BoW (NB-TFIDF)	69.6	56.1	80.7	66.2	57.1	56.4	61.9	59.0
3 ↴+ Char trigrams	74.0	62.5	73.5	67.6	54.8	54.3	60.8	57.4
4 ↴+ Bias	75.2	67.7	62.6	65.1	54.5	55.0	50.4	52.6
5 ↴+ Lexical	75.2	67.0	64.7	65.8	52.3	52.3	51.5	51.9
6 ↴+ Vocab. Richness	75.8	67.1	67.6	67.4	50.9	50.8	52.5	51.7
7 ↴+ Readability	76.0	66.4	70.6	68.4	51.6	51.5	53.9	52.7

Table 1: An incremental analysis showing the performance of different feature combinations, evaluated on the validation sets labeled by *article* and by *publisher*.

We trained the classifier using the 600K training examples annotated *by-Publisher*, then used the remaining 150K examples for evaluation. We fine-tuned the hyperparameters on the 645 *by-Article* examples. The hyper-parameters include  $k \in [50, 200, 700] \times 10^3$  as the most frequent word  $n$ -grams and the scaling parameters of the different features except for the  $n$ -grams. Best fine-tuning results suggested using the 200K most-frequent word [1, 2]-grams. We assessed the different feature sets, described in Section 3, by incrementally adding each set, one at a time, to the mix of all features.

### 4.3 Results

Table 1 illustrates the results obtained on both the *by-Article* set (which we used to fine-tune the model’s hyperparameters) and the *by-Publisher* set (which we used for evaluation). Our results suggest that scaling the TFIDF values through Naive Bayes is better than using raw TFIDF scores. Hence, these features were used for all subsequent experiments. It can also be observed that adding each group of features introduces a consistent improvement in accuracy on the *by-Article* data. However, we observed an opposite behaviour on the *by-Publisher* data. We believe this is due to the significant amount of noisy labels introduced by the distant supervision labeling strategy. Therefore, we based our decisions on the results obtained on the *by-Article* data since its labels are more accurate.

The normalization strategy, i.e., scaling the features using calibrated scaling parameters, introduced significant performance improvements. Unfortunately, we were not able to perform these calibration experiments by the competition’s deadline,

hence we submitted the system that was available at that time, which is based on the BoW (NB-TFIDF) and character 3-gram features, as shown in row 3 in Table 1. Our system achieved a 72.9% accuracy on the test *by-Article* data, ranking 20<sup>th</sup>/42. It also achieved 60.8% accuracy on the test *by-Publisher* data, ranking 15<sup>th</sup>/42. All subsequent, and superior, results (rows 4–7) were obtained after the deadline.

## 5 Conclusion

In this paper, we present our submission to SemEval-2019 Task 4 on Hyperpartisan News Detection. We trained a logistic regression model with a feature set that included word and character  $n$ -grams, represented with TFIDF. This system achieved a 72.9% and 60.8% accuracy on the test data that is labeled *by-Article* and *by-Publisher*, respectively.

We also evaluated additional features that represent different aspects of the article’s text such as its vocabulary richness, the kind of language it uses according to different lexicons, and its level of complexity. Initial experiments showed that these features hurt the model. However, with proper pre-processing and scaling we were able to achieve significant performance improvements of up to 2% in absolute accuracy. These results were obtained after the competition’s deadline, hence were not considered as part of our submission.

## 6 Acknowledgment

This research was carried out in collaboration between the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) and the HBKU Qatar Computing Research Institute (QCRI).

## References

- Sean Baird, Doug Sibley, and Yuxi Pan. 2017. Talos targets disinformation with fake news challenge victory. <https:////blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html>.
- Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018a. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 3528–3539.
- Ramy Baly, Mitra Mohtarami, James Glass, Lluís Márquez, Alessandro Moschitti, and Preslav Nakov. 2018b. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 21–27, New Orleans, LA, USA.
- Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: Organizing news coverage on the basis of their propagandistic content. *Information Processing and Management*.
- 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 Thirty-Third AAAI Conference on Artificial Intelligence (AAAI'19)*, AAAI'19, Honolulu, HI, USA.
- Ann M Brill. 2001. Online journalists embrace new marketing function. *Newspaper Research Journal*, 22(2):28.
- Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW'11, pages 675–684, Hyderabad, India.
- Jacques Ellul. 1965. *Propaganda: The Formation of Men's Attitudes*. Vintage Books, United States.
- Howard Finberg, Martha L Stone, and Diane Lynch. 2002. Digital journalism credibility study. *Online News Association*. Retrieved November, 3:2003.
- Robert Gunning. 1968. *The Technique of Clear Writing*. McGraw-Hill.
- Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance detection task. *arXiv preprint arXiv:1806.05180*.
- Anthony Honore. 1979. Some Simple Measures of Richness of Vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2):172–177.
- Joan B. Hooper. 1975. On assertive predicates. In J. Kimball, editor, *Syntax and Semantics*, volume 4, page 91124. Academic Press, New York.
- Benjamin D Horne and Sibel Adal. 2017. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. In *Proceedings of the International Workshop on News and Public Opinion at ICWSM*, Montreal, Canada.
- Ken Hyland. 2015. *The International Encyclopedia of Language and Social Interaction*, chapter Metadiscourse. American Cancer Society.
- Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, David Corney, Payam Adineh, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of The 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Memphis TN Naval Air Station, Research B*.
- Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors elisheva bonchek-dokow. *Journal of Machine Learning Research*, 8:1261–1276.
- Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, Brussels, Belgium.
- David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.
- Todor Miaylov, Tsvetomila Miaylova, Preslav Nakov, Lluís Márquez, Georgi Georgiev, and Ivan Koychev. 2018. The dark side of news community forums: Opinion manipulation trolls. *Internet Research*, 28(5):1292–1312.
- Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Márquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 767–776, New Orleans, LA, USA.

- James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *LIWC Operators Manual 2001*.
- Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the Web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1003–1012, Perth, Australia.
- Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017a. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.
- Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017b. [A stylometric inquiry into hyperpartisan and fake news](#). *CoRR*, abs/1702.05638.
- 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*, EMNLP '17, pages 2931–2937, Copenhagen, Denmark.
- Mark Schmidt, Nicolas Le Roux, and Francis Bach. 2017. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1–2):83–112.
- Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.
- James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING '18, pages 3346–3359, Santa Fe, NM, USA.
- Peter D. Turney. 2002. [Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews](#). In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.
- Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.
- George Udny Yule. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge University Press.