Unmasking Bias in News

Javier Sánchez-Junquera[1], Paolo Rosso[1],
Simone Paolo Ponzetto[2], and Manuel Montes-y-Gómez[3]

[1] PRHLT Research Center, Universitat Politècnica de València, Spain
jjsjunquera@gmail.com, prosso@dsic.upv.es
[2] Data and Web Science Group,University of Mannheim, Germany
simone@informatik.uni-mannheim.de
[3] Instituto Nacional de Astrofísica Óptica y Electrónica, Mexico
mmontesg@inaoep.mx

Abstract. We present experiments on detecting hyperpartisanship in news using a 'masking' methods that allows us to assess the role of style vs. content for the task at hand. Our results corroborate previous research on this task in that topical features yield better results that stylistic ones, while at the time improving the results by isolating topical components and reducing the data sparsity of a rather simple lexical classifier.

Keywords: Bias in information · hyperpartisan news · masking technique.

1 Introduction

Media such as radio, TV channels, and news media control which information spreads and how. The aim is often not only to inform readers but also to influence public opinion on specific topics from a hyperpartisan perspective.

Social media, in particular, have become the default channel for many people to access information and express ideas and opinions. The most relevant and positive effect is the democratization of information and knowledge but there are also undesired effects in social media. One of them is that social media foster information bubbles: every user may end up receiving only the information that matches her personal biases, beliefs, tastes and points of view. Because of this, social media are a breeding ground for the propagation of fake news: when a piece of news outrages us or matches our beliefs, we tend to share it without checking its veracity; and, on the other hand, content selection algorithms in social media give credit to this type of popularity because of the click-based economy on which their business are based. Another harmful effect is that the relative anonymity of social networks facilitates the propagation of toxic, hate and exclusion messages. Therefore, social networks contribute to the misinformation and polarization of society, as we have recently witnessed in the last presidential elections in USA or the Brexit referendum. Clearly, the polarization of society and its underlying discourses are not limited to social media, but rather reflected also in political

dynamics (e.g., like those found in the US Congress [1]): even in this domain, however, social media can provide a useful signal to estimate partisanship [4].

Closely related to the concept of controversy and the "filter bubble effect" is the concept of bias [2], which refers to the presentation of information according to the standpoints or interests of the journalists and the news agencies. Detecting bias is very important to help users to acquire balanced information. Moreover, how a piece of information is reported has the capacity to evoke different sentiments in the audience, which may have large social implications (especially in very controversial topics such as terror attacks and religion issues).

In this paper, we approach this very broad topic by focusing on the problem of detecting hyperpartisan news, namely news written with an extreme manipulation of the reality on the basis of an underlying, typically extreme, ideology. This problem has received little attention in the context of the automatic detection of fake news, although the potential correlation between them. The authors of [5] reported on a comparative style analysis of hyperpartisan news, evaluating features such as characters n-grams, stop words, part-of-speech, readability scores, and ratios of quoted words and external links. They found that the topic-based model outperforms the style-based model separating the left, right and mainstream orientations.

We build upon previous work and use the dataset from [5]: this way we can investigate hyperpartisan-biased news (i.e., extremely one-sided) that have been manually fact-checked by professional journalists from BuzzFeed. The articles originated from 9 well-known political publishers, 3 each from the mainstream, the hyperpartisan left-wing, and the hyperpartisan right-wing.

To detect hyperpartisanship, we apply a masking technique that transforms the original texts in a form where the textual structure is maintained, while letting the learning algorithm focus on the writing style or the topic-related information. This technique makes it possible for us to corroborate previous results that content matters more than style, while also improving the state-of-the art results for this task.

The rest of the paper is structured as follows. In Section 2 we describe our method. Section 3 presents the data, the experiments and the discussion. Finally, Section 4 concludes with some directions for future work.

2 Masking Technique

The masking technique that we propose here for hyperpartisan news detection task has been applied to text clustering [3], authorship attribution [7], and recently to deception detection [6] with encouraging results. The main idea of the proposed method is to transform the original texts to a form where the textual structure, related to a general style (or topic), is maintained while content-related (or style-related) words are masked. To this end, all the occurrences (in both training and test corpora) of non-desired terms are replaced by symbols.

Let W_k the set of the k most frequent words, we mask all the occurrences of a word $w \in W_k$ if we want to learn a *topic-related model*, or we mask $w \notin W_k$

Table 1: Examples of masking style-related information or topic-related information.

Original text	Masking topic-related words	Masking style-related words
Officers **went after** Christopher **Few after** watching **an** argument **between him and his** girlfriend outside **a** bar **just before the** 2015 shooting	* went after * Few after * an * between him and his * a * just before the # *	Officers * * Christopher * * watching * argument * * * * girlfriend outside * bar * * * 2015 shooting

Table 2: Statistics of the original dataset and its subset used in this paper.

	Left-wing	**Mainstream**	**Right-wing**	$\sum$
Original data [5]	256	826	545	1627
Cleaned data	252	787	516	1555

if we want to learn a *style-based model*. Whatever the case, the way in which we mask the terms in this work is called *Distorted View with Single Asterisks* and consists in replacing w with a single asterisk or a single # symbol if the term is a number. For further masking methods, refer to [7].

Table 1 shows a fragment of an original text and the result of masking style-related information or topic-related information. With the former we obtain distorted texts that allow for learning a *topic-based model*; on the other hand, with the latter, it is possible to learn a *style-based model*. One of the options to choose the terms to be masked or maintained without masking is to take the most frequent words of the target language [7]. In the original text from the table, we highlight some of the more frequent words in English.

3 Masking Content vs. Style in Hyperpartisan News

3.1 Hyperpartisan News Dataset

We use the BuzzedFeed-Webis Fake News Corpus 2016 collected by [5] whose articles were labeled with respect to three political orientations: mainstream, left-wing, and right-wing (see Table 2). Each article was taken from one of 9 publishers known as hyperpartisan left/right or mainstream in a period close to the US presidential elections 2016. Therefore, the content of all the articles is related to the same topic.

During initial data analysis and prototyping we identified a variety of issues with the original dataset: cleaning included excluding articles with empty (23 articles) or bogus ('*The document has moved here*') (14) text. Additionally, we removed duplicates (33), and files with the same text but inconsistent labels (2). As a result, we obtained a new dataset with 1555 articles out of 1627.[4] Following the settings of [5], we balance the training set using random duplicate oversampling.

[4] We will release the cleaned dataset by the publication date of this article.

Table 3: Results of the proposed masking technique ($k = 500$ and $n = 5$) applied to mask topic-related information or style-related information. The last two rows show the results reported in [5].

Masking Method	**Classif.**	**Macro F_1**	**Acc**	**Precision**			**Recall**			**F1**		
				left	right	main	left	right	main	left	right	main
Style-based model	NB	0.47	0.52	0.20	0.51	0.73	0.28	0.65	0.49	0.23	0.57	0.59
	RF	0.46	0.53	0.24	0.58	0.64	0.36	0.34	0.73	0.29	0.43	0.68
	SVM	0.57	0.66	0.33	0.66	0.75	0.26	0.61	0.84	0.29	0.62	0.79
Topic-based model	NB	0.54	0.60	0.26	0.63	0.74	0.36	0.62	0.65	0.29	0.62	0.69
	RF	0.53	0.55	0.27	0.64	0.71	**0.44**	0.60	0.58	0.33	0.61	0.64
	SVM	**0.66**	**0.74**	**0.48**	**0.73**	**0.81**	0.38	**0.78**	0.82	**0.42**	**0.75**	**0.82**
Results from [5]												
Style	RF	0.51	0.60	0.21	0.56	0.75	0.20	0.59	0.74	0.20	0.57	0.75
Topic	RF	0.52	0.64	0.24	0.62	0.72	0.15	0.54	**0.86**	0.19	0.58	0.79

3.2 Experiments and Discussion

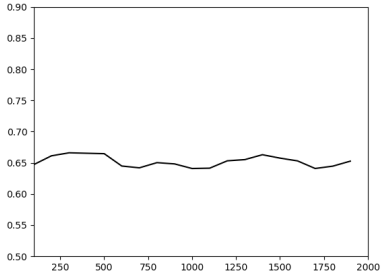
We report the results of the masking technique from two different perspectives. In one setting, we *mask topic-related information* in order to maintain the predominant writing style used in each orientation. We call this approach a *style-based model*. With that intention we select the k most frequent words from English[5], and then we transform the texts by masking the occurrences of the rest of the words. In another setting, we *mask style-related information* to allow the system to focus only on the topic-related differences between the orientations. We call this a *topic-based model*. For this, we mask the k most frequent words and maintain intact the rest.

After the transformation of the texts, we use the *CountVectorizer* from *sklearn* for the text representation using a *tf* weighting scheme. We extract the character n-grams having a document frequency strictly lower than 50; the other parameters were set by default. In addition to the results of the Random Forest (RF) classifier (used in [5] for the same dataset), we also report the results of Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers.

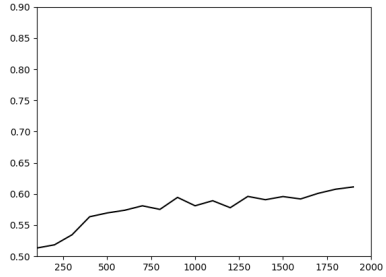
Table 3 shows the results of the proposed method. We compare with [5] against their topic and style-based methods. In order to compare our results with those reported in [5], we report the same measures the authors used. We also include the macro F_1 score because of the unbalance test set. For these experiments we extract the character 5-grams from the transformed texts, taking into account that as more narrow is the domain more sense has the use of longer n-grams. We follow the steps of [8] and set k = 500 for this comparison results.

Similar to [5], the topic-based model achieves better results than the style-related model. However, the differences between the results of the two evaluated approaches are much higher (0.66 vs. 0.57 according to Macro F_1) than those

[5] We extract the most frequent words of the BNC corpus (https://www.kilgarriff.co.uk/bnc-readme.html)



Varying k values and masking the most frequent words: topic-based model.



Varying k values and maintaining without masking the most frequent words: style-based-model.

Fig. 2: Macro F_1 results of the proposed masking technique. We set n=5 for comparing results of different values of k.

shown in [5]. The highest scores were consistently achieved using the SVM classifier and masking the style-related information (i.e., the topic-related approach). This could be due to the fact that all the articles are about the same political event in a very limited period of time. In line with what was already pointed out in [5], the left-wing orientation is harder to predict, possibly because this class is represented with fewer examples in the dataset.

Figures 2 shows how sensitive is the masking technique in the hyperpartisan news detection when different values of k and n are chosen. Varying values of $k \in \{100; 200; 300; ...; 2000\}$ with $n = 5$, we conclude that masking style-related information is more robust to the value of k. Varying the size of n-grams we confirm that character 5-grams are more useful for this dataset and combining all the n-grams with $n \in \{2, 3, 4, 5\}$ the method does not improve the predictions.

4 Conclusions

In this paper we presented initial experiments on the task of hyperpartisan news detection: for this, we explored the use of a masking techniques to boost the performance of a lexicalised classifier. Our results corroborate previous research on the importance of content features to detect extreme content: masking, in addition, shows the benefits of reducing data sparsity for this task. Future work will explore more complex learning architectures (e.g., representation learning of masked texts), as well as the application and adaptation of unsupervised methods for detecting idelogical positioning from political texts 'in the wild' for the online news domain.

References

1. Andris, C., Lee, D., Hamilton, M.J., Martino, M., Gunning, C.E., Selden, J.A.: The rise of partisanship and Super-Cooperators in the U.S. house of representatives. PLoS ONE **10**(4), e0123507 (2015)
2. Baeza-Yates, R.: Bias on the web. Communications of the ACM **61**(6), 54–61 (2018). https://doi.org/10.1145/3209581
3. Granados, A., Cebrian, M., Camacho, D., de Borja Rodriguez, F.: Reducing the loss of information through annealing text distortion. IEEE Transactions on Knowledge and Data Engineering **23**(7), 1090–1102 (2011)
4. Hemphill, L., Culotta, A., Heston, M.: #polarscores: Measuring partisanship using social media content. Journal of Information Technology & Politics **13**(4), 365–377 (2016)
5. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B.: A stylometric inquiry into hyperpartisan and fake news. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 231–240 (2018)
6. Sánchez-Junquera, J.: Adaptación de dominio para la detección automática de textos engañosos. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica (2018), http://inaoe.repositorioinstitucional.mx/jspui/handle/1009/1470 (in Spanish)
7. Stamatatos, E.: Authorship attribution using text distortion. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. vol. 1, pp. 1138–1149 (2017)
8. Stamatatos, E.: Authorship attribution using text distortion. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 1138–1149. Association for Computational Linguistics (2017)