

On the Detection of Political and Social Bias

Javier Sánchez-Junquera

PRHLT Research Center, Universitat Politècnica de València, 46022 València, Spain

Abstract

Nowadays it is very easy to share, create and disseminate any kind of bias thanks to the increasing facilities of the technology. Political and social bias have a lamentable repercussion on the behaviours of people and our life quality. This research is focused on the detection of hyperpartisan news and immigrant stereotypes in political speeches. This work proposes two different explainable approaches: BERT-based models, known for their ability to capture semantic and syntactic patterns in the same representation but at the cost of great computational complexity and lack of transparency; and a masking-based model that has been recognized by its capabilities to deliver good and human-understandable results.

Keywords

political bias, social bias, immigrant stereotypes, hyperpartisan news, masking technique, BERT-based models

1. Proposed Research's Justification

Nowadays, people consume information with or without their intentions. The disseminated information has great repercussion on the perception of reality that we live, and therefore, in the decisions we make. Palpable examples of information that affects our day to day comes from news, rumors, political speeches, among others. In political contexts, the spread of information generated to make a political position or candidate seem more attractive, could have a lasting impact with proven effects on voter behaviour and consequent political outcomes. This biased information is present in electoral campaigns, fake news, partisan news, political debates, parliamentary speeches, among others.

By using specific linguistic means, politicians can fulfill their own goals, which are intended to shape people's thinking and persuade them to act as they want [1]. With hyperpartisan news, for example, politicians and journalists show an extreme manipulation of the reality based on an underlying and extreme ideology. It spreads much more successfully than mainstream news, and very often are inflammatory, emotional, and riddled with untruths [2]. Another information easily disseminated in social media and very often in the political context is regarding social phenomenons. Sometimes, the spread of social bias helps politicians to gain popularity over them, or to gain more visibility because of their controversial point of view. For example, it is not casual the coincident rise of far-right wing political parties with the rapid rate of European immigration [3]. These parties appeal to fears and anti-immigrant sentiments in the native


Doctoral Symposium on Natural Language Processing from the PLN.net network 2021 (RED2018-102418-T), 19-20 October 2021, Baeza (Jaén), Spain.

✉ juasanj3@doctor.upv.es (J. Sánchez-Junquera)

🆔 0000-0002-0845-9532 (J. Sánchez-Junquera)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

population, and support the spread of offenses, incitements to hate, and violent speech [4]. In addition, other findings suggest that increased social and political trust are associated with lower stereotyping and prejudice against immigrants [5].

In our research, we are interested in detecting both, political and social bias; in particular, we are focused on (i) hyperpartisanship detection in news, and (ii) the identification of immigrant stereotypes. Similar to applications of healthcare or security, in these tasks it is not enough to achieve high results, but it is also mandatory that results could be understood by human experts. Taking into account performance and explainability, the objective of this work is to compare two approaches diametrically opposite to each other in the text classification state of the art. On the one hand, BERT-based models, which have shown outstanding performance, but high complexity and poor explainability; and, on the other hand, a masking-based model, which requires fewer computational-resources and showed a good performance in related tasks like authorship attribution [6].

With the masking technique it is possible to transform 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 to know what are the most important words that the model preferred to highlight, and, in the case of hyperpartisan news detection, to corroborate previous results that content matters more than style. Moreover, we aim to find explainable predictions with the attention mechanism of the transformer-based models. With this purpose, we expect to derive the explanation by investigating the scores of different features used to output the final prediction. Based on this, we contrast the transparency of both approaches by comparing the relevant parts of the texts that they highlight.

2. Related Work

2.1. Hyperpartisanship Detection

The problem of hyperpartisanship detection has received little attention in the context of the automatic detection of fake news, despite the potential correlation between them. Seminal work from [2] presents 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. The results indicate that a topic-based model outperforms a style-based one to separate the left, right and mainstream orientations.

More recently, in [7] the authors summarize the features that participants used in SemEval-2019 task 4 on hyperpartisan news detection: n-grams, word embeddings, stylometry (e.g., punctuation and article structure), sentiment and emotion features, named entities, quotations, hyperlinks, and publication date. Using the same dataset from SemEval-2019, the authors of [8] found that dense document representations work better across domains and tasks than traditional sparse representations.

2.2. Immigrant Stereotype Detection

There have been attempts to study stereotypes from a computational point of view, such as gender, racial, religion, and ethnic bias detection do [9, 10]. Those works predefine two opposite

categories (e.g., men vs. women) and use word embeddings to detect the words that tend to be more associated with one of the categories than with the other. In [11], the authors propose two different level tests for measuring bias. These tests are similar to the idea of [12] that consists in using natural language inference to measure entailment, contradiction, or neutral inferences to quantify the bias.

In the case of immigrant stereotypes, sentences like *¿Por qué ha muerto una persona joven?* (Why did a young person die?) do not contain an attribute of the immigrant group although from its context¹ it is possible to conclude that here immigrants are placed as victims of suffering. Also, it is not clear the representative word of the social group, since *persona joven* (young person) is neutral to immigrants and non-immigrants. From this, it is possible to conclude that immigrant stereotypes require other approaches to be faced with. Another work that confirm this conclusion is made from the participation of most participant teams from the HaSpeeDe shared task at EVALITA 2020 [13]: the participants adapted their hate speech models to the stereotype identification task, thus, representing (and reducing) stereotypes to characteristics of hate speech. The conclusions of [13] include that the immigration stereotype appeared as a more subtle phenomenon, which also needs to be approached as non-hurtful text.

3. Proposed Research’s Description

To carry out the two tasks introduced in the previous sections, the detection of (i) hyperpartisanship in news, (ii) and immigrant stereotypes, this research explores the trade-off between the performance of the models and the transparency of their results. Taking this into account, the idea is to apply two approaches diametrically opposite to each other in the text classification state of the art: BERT-based models, and a masking-based model.

The research questions aim to answer in this work are:

RQ1: Are the transformer more effective than the masking technique at identifying the hyperpartisan news and the immigrant stereotypes?

RQ2: Is it possible to obtain local explanations on the predictions of the models, to allow human interpretability about the hyperpartisan news and the immigrant stereotypes?

4. Methodology and Proposed Experiments

This research uses the texts from annotated datasets and use different classifiers to evaluate and compare their performance and results. The following list summarizes some important aspects of the methodology.

- **Datasets construction: The dataset used for the hyperpartisan news detection** is taken from [2] and cleaned by removing some useless articles. The news are written in English, and are labeled with respect to three political orientations: mainstream,

¹Fragment of a political speech from a Popular Parliamentary Group politician in 2006. The speaker is mentioning some of the conditions of immigrants in Spain in that period.

Table 1

Distribution of texts per label and the average length (with standard deviation) of their instances. The texts labeled as *Victims* or *Threat* are a subset of the texts labeled as *Stereotype*.

Label	Length	Texts	
<i>Stereotype</i>	45.62 ± 24.69	1673	3635
<i>Non-stereotype</i>	36.00 ± 21.17	1962	
<i>Victims</i>	48.93 ± 27.5	743	1479
<i>Threat</i>	45.84 ± 24.42	736	

left-wing, and right-wing. This dataset contains a total of 252 left-wing articles, 787 mainstream articles, and 516 right-wing articles (i.e., a total of 1555 articles). **For identifying stereotypes about immigrants**, this research proposes a new approach and taxonomy focused on the narrative contexts in which the immigrant group is repetitively situated in the public discourses of politicians, rather than the characteristics attributed to the immigrant group. Therefore, the annotation covers the whole spectrum of beliefs that make up the immigrant stereotype. The resultant dataset is called StereoImmigrants² and contains texts written in Spanish; its theoretical foundations and the annotation process are published in [14]. The StereoImmigrants dataset contains two annotations level: Stereotype vs. Non-stereotype; and Victims vs. Threat, attending to the attitudes that each sentence expresses. Table 1 shows the distribution per label of this dataset.

- **Employed models:** For each classification task two different approaches were applied: BERT-based models and a masking-based model. Bidirectional Encoder Representations from Transformers (BERT) is designed to pretrain deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers [15]. In this work three **BERT-based models** were used for the texts written in English: BERT; the multilingual BERT (M-BERT) [16]; and XLM-RoBERTa [17]. For the texts written in Spanish the BETO model was used [17]. **The masking-based model** consists of transforming the original texts to a distorted form where the textual structure is maintained while irrelevant words are masked by a neutral symbol e.g. “*”. After the masking process, traditional classifiers are employed receiving as input the transformed texts.
- **Discriminating words:** Since this research is also interested in supporting human comprehension of the results, both approaches were used to analyse the discriminating words that could help in the classification process. In the case of the masking-based model, the discriminative words were assumed from the list of non-masked words. In the case of the BERT-based models, the discriminating words were computed taking advantage of the attention mechanisms (see Section 4.3).

4.1. Results in the Hyperpartisan News Detection

Table 2 shows the results of the proposed method and the system from [2]³ in our cleaned dataset. In one setting, we masked *topic-related information* in order to maintain the predominant writing

²<https://github.com/jjsjunquera/StereoImmigrants>.

³<https://github.com/webis-de/ACL-18>

style used in each orientation. We call this approach a *style-based model*. With that intention we selected the k most frequent words from the target language, and then we transformed the texts by masking the occurrences of the rest of the words. In another setting, we masked *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 masked the k most frequent words and maintained intact the rest. For these experiments there were extracted 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. Follow the steps of [18], k is set to 500.

Table 2

Results of the proposed masking technique applied to mask topic-related information or style-related information. NB: Naive Bayes; RF: Random Forest; SVM: Support Vector Machine. The last two rows show the results obtained by applying the system from [2] to our cleaned dataset.

Masking Method	Classifier	Macro F_1	Accuracy	Precision			Recall			F_1		
				left	right	main	left	right	main	left	right	main
Baseline model	NB	0.52	0.56	0.28	0.57	0.81	0.49	0.58	0.56	0.35	0.57	0.66
	RF	0.56	0.62	0.28	0.61	0.80	0.36	0.72	0.63	0.32	0.66	0.70
	SVM	0.70	0.77	0.55	0.75	0.84	0.42	0.79	0.87	0.47	0.77	0.85
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
System from [2] (applied to our cleaned dataset)												
Style	RF	0.61	0.63	0.29	0.62	0.71	0.16	0.62	0.80	0.20	0.61	0.74
Topic	RF	0.63	0.65	0.27	0.65	0.72	0.15	0.62	0.84	0.19	0.63	0.77
Transformer-based models												
	M-BERT	0.76	0.83	0.65	0.75	0.93	0.49	0.86	0.92	0.56	0.93	0.80
	XLM-RoBERTa	0.80	0.86	0.80	0.76	0.95	0.50	0.91	0.94	0.61	0.83	0.95
	BERT	0.86	0.89	0.77	0.87	0.94	0.75	0.86	0.96	0.76	0.87	0.95

Similar to [2], the topic-based model achieves better results than the style-related model. The highest scores of the masking technique were consistently achieved using the SVM classifier and masking the style-related information (i.e., applying the topic-related model). This could be explained with 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 [2], the left-wing orientation is harder to predict, possibly because this class is represented with fewer examples in the dataset.

The last three rows of the Table 2 show the results of the BERT-based models. As we can see, these models achieved the highest results (**RQ1**), in particular the BERT model, with a Macro $F_1 = 0.86$. These models are known for their ability to capture complex syntactic and semantic patterns, therefore, these results are somehow justified to be the highest compared to the masking approach. However, what is interesting at this point is the effectiveness of the BERT-based models at predicting the correct orientation using just the beginning of the news ($max_length = 200$). This is aligned to the work of [19] that focused on analyzing the initial part of false news articles. The authors assumption is that false news tend to present a unique emotional pattern for each false information type in order to trigger specific emotions to the readers; in hyperpartisan news this probably happens to gain readers’ attention and sympathy. More details about this work can be found in [20] as a result of our experiments.

4.2. Results in the Immigrant Stereotypes Detection

In preliminary experiments, the highest results were achieved by masking the words out of the following lists: (i) the words with higher relative frequency (*RelFreq*), i.e., the k words with a frequency in one class remarkably higher than its frequency in the opposite class; and (ii) the k words with the highest absolute frequency (*AbsFreq*) in all the collection, excluding stopwords (i.e., stopwords were masked). Therefore, both lists were used in all the dataset obtaining the best results with $k = 1000$. The results of the models are shown in Table 3. It is possible to see high results of LR with the original texts. However, we observe that masking the terms out of the list *RelFreq* is slightly better than using the original text. These results suggest that the masking technique improves the quality of the stereotype detection and its dimensions.

Table 3

F-measure in both classification tasks: Stereotype vs. Non-stereotype (S/N), and Victims vs. Threat (V/T).

	S/N	V/T
Original text	0.82	0.79
Masking Technique with <i>AbsFreq</i>	0.79	0.75
Masking Technique with <i>RelFreq</i>	0.84	0.81
BETO	0.86	0.83

In comparison with *AbsFreq*, maintaining unmasked the *RelFreq* words helps to ignore more words that are less discriminative for classification tasks. This could be explained because *AbsFreq* includes words similarly frequent in both classes, which could not help at predicting immigrant stereotypes: *países* (countries), *gobierno* (government), *señor* (mister), *partido* (party); or at identifying the immigrant-stereotype dimension: *fronteras* (frontiers), *política* (politic), *seguridad* (security), *grupo* (group).

BETO achieves the highest results in both classification tasks (**RQ1**). This is not surprising because the transformer-based models are known for their properties at capturing semantic and syntactic information, and richer patterns in which the context of the words are taken into account. However, we do not observe a significant difference between the results of such a resource-hungry model, and the combination of the masking technique with the traditional LR classifier. Considering the computational capabilities that BETO demands, and the less complexity of the masking technique, the latter shows a better trade-off between effectiveness and efficiency than the latter. More details about this work can be found in [21] as a result of our experiments.

4.3. Discriminating Words

Motivated by the similar results of BETO and the masking technique, in the experiments related to the immigrant stereotypes, it was developed a step related to observe and compare what portions of the texts they could be focusing on. For this purpose, it was observed the last layer of BETO and computed the average of the attention heads. Therefore, for each text, it is obtained a matrix from which it is possible to compute the attention that the model gave to each word in that texts. Figure 1 shows examples of texts where the two models agreed on the right label. From the figure, it is possible to see what words were relevant for both approaches (**RQ2**).

Stereotype:

BETO:

pues bien , el fenómeno de la inmigración es hoy sin ninguna duda , no solamente por las encuestas del cis sino por distintos diagnósticos de la opinión pública , un asunto que preocupa más a los ciudadanos que los problemas del terrorismo y del paro .

Masking:

**** bien ** fenómeno ** ** inmigración ** hoy *** ninguna **** **
***** ** ** encuestas *** **
** opinión pública ** asunto *** preocupa *** ** ciudadanos *** ** prob-
lemas *** **

Non-stereotype:

BETO:

cuando se habla de inmigración , de qué está hablando el grupo parlamentario de izquierda ?

Masking:

***** ** habla ** inmigración **
** izquierda

Victims:

BETO: hay una situación de desamparo en muchas personas a la que necesitamos dar una solución .

Masking:

*** **
***** dar *** solución

Threat:

BETO: el tiempo nos ha dado la razón , se ha convertido en un problema muy serio , en un problema muy importante tanto para europa como para españa .

Masking:

** tiempo *** **
problema *** **
españa

Figure 1: Examples of attention visualization and masking transformation over the same texts. These examples were correctly classified by both models. The more intense the color, the greater is the weight of attention given by the model.

It is possible to observe that some content-related words can be helpful for expert’s analysis. For instance, the text labeled as Stereotype has as relevant words *fenómeno* (phenomenon), *inmigración* (immigration), *problema* (problem), *terrorismo* (terrorism), *paro* (unemployment), among others. The text labeled as Victims contains *desamparo* (abandonment), *personas* (people), *necesitamos dar una solución* (we need to give a solution), reflecting how immigrants were seen as people more than their illegal status, and the target of problems that need solutions. Moreover, in the example of Threat, some of the words and phrases receiving more importance (such as *problema muy serio*, *problema muy importante*) reflect how immigrants were seen as a problem to the continent and the country, but not the country where immigrants come from.

In a similar way, the attention mechanisms were used to observe the relevant part of hyper-partisan and mainstream news. The left-wing orientation remarks the names of the opposite politicians, and it was possible to see which of the parties is the favourite of the journalist. In particular, the leader of the right-wing (i.e., Trump) is referred in a negative way (he does not know his own words) while Hillary Clinton, the representative of the left-wing is favored by the news. Similar to this, other news do the same but in the opposite direction; i.e., Hillary Clinton is put as a very negative “character” who *loves taxes* and is *the most despicable liar*

ever. In mainstream news, Trump’s campaign is mentioned without describing the stance of the author whether Trump did well or not in his topic selection. This suggests that the style used to speak about the leaders can differ from the more biased (hyperpartisan) news to the less biased (mainstream).

5. Specific Research Items Proposed for Discussion

I find interesting the following aspects to be considered for discussion and future work:

1. BERT-based models could be used to investigate with more detail bias by using the attention mechanisms. In this sense, a study of the attention scores obtained in different layers (in this study only the last one was used) could give some information related to other semantic or syntactic patterns.
2. To explore more deeply the advantages of the attention mechanisms to increase the performance; and to use discriminative words to find debiasing strategies to mitigate the immigrant stereotypes in social media and political speeches.
3. To evaluate how necessary is to use all the news (and not only the beginning), e.g. with the Transformer-XL model.
4. To extend the study of stereotypes detection to other social groups (e.g., LGBTQ+) which are currently being victim⁴ of discrimination, violence and crimes derived in part from the heteronormativity (e.g., transphobia, homophobia, and serophobia).

References

- [1] F. H. Al-Hindawi, N. M. Al-Aadili, The pragmatics of deception in american presidential electoral speeches, *International Journal of English Linguistics* 7 (2017) 207.
- [2] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, B. Stein, 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)*, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 231–240.
- [3] L. Davis, S. S. Deole, Immigration and the rise of far-right parties in europe, *ifo DICE Report* 15 (2017) 10–15.
- [4] C. A. Calderón, G. de la Vega, D. B. Herrero, Topic modeling and characterization of hate speech against immigrants on twitter around the emergence of a far-right party in spain, *Social Sciences* 9 (2020).
- [5] S. Ahmed, V. C. Hsueh-Hua, A. I. Chib, Xenophobia in the time of a pandemic: Social media use, stereotypes, and prejudice against immigrants during the covid-19 crisis, *International Journal of Public Opinion Research* (2021).
- [6] E. Stamatatos, Authorship attribution using text distortion, 15th Conference of the European Chapter of the Association for Computational Linguistics, *EACL 2017 - Proceedings of Conference 1* (2017) 1138–1149.

⁴https://www.report-it.org.uk/files/online-crime-2020_0.pdf

- [7] J. Kiesel, M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, M. Potthast, Semeval-2019 task 4: Hyperpartisan news detection, 2019, pp. 829–839.
- [8] T. Anthonio, Robust Document Representations for Hyperpartisan and Fake News Detection, Master’s thesis, University of the Basque Country UPV/EHU, 2019.
- [9] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, *Proceedings of the National Academy of Sciences* 115 (2018) E3635–E3644.
- [10] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, *Advances in neural information processing systems* 29 (2016) 4349–4357.
- [11] M. Nadeem, A. Bethke, S. Reddy, Stereoset: Measuring stereotypical bias in pretrained language models, *arXiv preprint arXiv:2004.09456* (2020).
- [12] S. Dev, T. Li, J. M. Phillips, V. Srikumar, On measuring and mitigating biased inferences of word embeddings., in: *AAAI*, 2020, pp. 7659–7666.
- [13] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. A. Stranisci, C. Bosco, C. Tommaso, V. Patti, R. Irene, et al., Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task, in: *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, CEUR, 2020, pp. 1–9.
- [14] J. Sánchez-Junquera, B. Chulvi, P. Rosso, S. P. Ponzetto, How do you speak about immigrants? taxonomy and stereoisimmigrants dataset for identifying stereotypes about immigrants, *Applied Sciences* 11 (2021).
- [15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, Association for Computational Linguistics, 2019, pp. 4171–4186.
- [16] K. K. Z. Wang, S. Mayhew, D. Roth, Cross-lingual ability of multilingual bert: An empirical study, in: *International Conference on Learning Representations*, 2020.
- [17] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: *PML4DC at ICLR 2020*, 2020.
- [18] E. Stamatatos, 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*, Association for Computational Linguistics, 2017, pp. 1138–1149.
- [19] B. Ghanem, S. P. Ponzetto, P. Rosso, F. Rangel, FakeFlow: Fake news detection by modeling the flow of affective information, in: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Association for Computational Linguistics, Online, 2021, pp. 679–689.
- [20] J. Sánchez-Junquera, P. Rosso, M. M. y Gómez, S. P. Ponzetto, Unmasking bias in news (2019). *arXiv:1906.04836*, proceedings pending to be published.
- [21] J. Sánchez-Junquera, P. Rosso, M. Montes-y-Gómez, B. Chulvi, Masking and bert-based models for stereotype identification., in: *In: Procesamiento del Lenguaje Natural (SEPLN)*, vol. 67, 2021.