

# Masking and Transformer-based Models for Hyperpartisanship Detection in News

**Javier Sánchez-Junquera**

Universitat Politècnica de València  
jjsjunquera@gmail.com

**Paolo Rosso**

Universitat Politècnica de València  
proso@dsic.upv.es

**Manuel Montes-y-Gómez**

Instituto Nacional de Astrofísica, Óptica y Electrónica  
mmontesg@inaoep.mx

**Simone P. Ponzetto**

University of Mannheim  
simone@informatik.uni-mannheim.de

## Abstract

Hyperpartisan news show an extreme manipulation of reality based on an underlying and extreme ideological orientation. Because of its harmful effects at reinforcing one's bias and the posterior behavior of people, hyperpartisan news detection has become an important task for computational linguists. In this paper, we evaluate two different approaches to detect hyperpartisan news. First, a text masking technique that allows us to compare style vs. topic-related features in a different perspective from previous work. Second, the transformer-based models BERT, XLM-RoBERTa, and M-BERT, known for their ability to capture semantic and syntactic patterns in the same representation. Our results corroborate previous research on this task in that topic-related features yield better results than style-based ones, although they also highlight the relevance of using higher-length n-grams. Furthermore, they show that transformer-based models are more effective than traditional methods, but this at the cost of greater computational complexity and lack of transparency. Based on our experiments, we conclude that the beginning of the news show relevant information for the transformers at distinguishing effectively between left-wing, mainstream, and right-wing orientations.

## 1 Introduction

Media such as radio, TV channels, and newspapers control which information spreads and how it does it. 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. One of them is that social media foster information bubbles: every user may end up receiving only the information that matches his/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 media 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 (Andris et al., 2015)): even in this domain, however, social media can provide a useful signal to estimate partisanship (Hemphill et al., 2016).

Closely related to the concept of controversy and the “filter bubble effect” is the concept of bias (Baeza-Yates, 2018), 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, despite the potential correlation between them. Seminal work from (Potthast et al., 2018) 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 (Kiesel et al., 2019), the features that participants used in SemEval-2019 task 4 on hyperpartisan news detection have been summarized: 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, (Anthonio, 2019) evaluated features like bag-of-words, bag-of-clusters, word embeddings and contextual character-based embeddings, POS n-grams, stylistic features and the sentiment; the authors found that dense document representations work better across domains and tasks than traditional sparse representations. Finally, (Hosseinia, 2020) found effective to use personality information in hyperpartisan news detection after topic-based sub-sampling of the news training data. The datasets proposed in (Kiesel et al., 2019) were manually labeled and the largest one was labeled in a semi-automated manner via distant supervision.

Instead of employing the datasets from (Kiesel et al., 2019), we build upon previous work and use the dataset from (Potthast et al., 2018): this way we can investigate hyperpartisan-biased news (i.e., extremely one-sided) that have been manually fact-checked by journalists from BuzzFeed, and contrast our results with what they achieved. The articles originated from 9 well-known political publishers, three each from the mainstream, the hyperpartisan left-wing, and the hyperpartisan

right-wing. To detect hyperpartisanship, we aim to explore the trade-off between the performance of the models and the transparency of their results. Taking this into account, we apply two approaches diametrically opposite to each other in the text classification state of the art. On the one hand, we use three transformer-based models, which have shown outstanding performance, but high complexity and lack of transparency. On the other hand, we use a masking-based model that requires fewer computational-resources and showed a good performance in related tasks such as authorship attribution (Stamatatos, 2017a).

The masking technique 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. Moreover, we aim to find explainable predictions of hyperpartisanship 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.

The rest of the paper is structured as follows. In Section 2 we describe our method to hyperpartisan news detection based on masking. Section 3 presents details on the dataset and the experimental setup. In Section 4 we show the obtained results and discuss about them. Finally, Section 5 concludes with some directions for future work.

## 2 Masking and Transformer-based Models

### 2.1 Investigating Masking for Hyperpartisanship Detection

The masking technique that we propose here for the hyperpartisan news detection task has been applied to text clustering (Granados et al., 2011), authorship attribution (Stamatatos, 2017a), and deception detection (Sánchez-Junquera, 2018) 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 of non-desired terms are re-

placed by symbols. Let  $W_k$  be 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 all  $w \notin W_k$  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 word or a number, respectively. For further masking methods, refer to (Stamatatos, 2017a).

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 (Stamatatos, 2017a). In the original text from the table, we highlight some of the most frequent words in English.

## 2.2 Transformer-based Models

Transformer-based models have been trained with huge general language datasets. Such is the case of the Bidirectional Encoder Representations from Transformers (BERT). BERT is designed to pre-train deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers (Devlin et al., 2018). This text representation allows the model to capture complex patterns going beyond merely the use of words and capturing semantic and syntactic patterns in the same representation.

The framework of BERT consists of two steps: pre-training and fine-tuning. For the pre-training, the collected data included BooksCorpus (800M words) and English Wikipedia (2,500M words). The BERT<sub>BASE</sub> model has 12 layers with 12 self-attention heads, and uses 768 as hidden size, with a total of 110M parameters; and the BERT<sub>LARGE</sub> model has 24 layers with 16 self-attention heads, and uses 1024 as hidden size, with a total of 340M parameters. The vocabulary contains 30K tokens. For fine-tuning, the model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream task, which in our case are 1555 news annotated with the political orientation. The first token of every sequence is always a special classifica-

tion token ([CLS]), which is used as the aggregate sequence representation for classification tasks. In our work, we add to the [CLS] representation two dense layers and a Softmax function to obtain the binary classification.

In this paper we evaluate three transformer-based models: BERT; the multilingual BERT (M-BERT) (Devlin et al., 2018), which was pretrained on the concatenation of monolingual Wikipedia datasets from 104 languages (Wang et al., 2019; Pires et al., 2019); and XLM-RoBERTa, which was pretrained on 2.5TB of newly created clean CommonCrawl data in 100 languages (Cañete et al., 2020).

## 3 Experiments

We used the BuzzedFeed-Webis Fake News Corpus 2016 collected by (Potthast et al., 2018) 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 of 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: we cleaned the data excluding articles with empty or bogus texts, duplicates. As a result, we obtained a new dataset with 1555 articles out of 1627.<sup>1</sup> Following the settings of (Potthast et al., 2018), we balanced the training set using random duplicate oversampling.

### 3.1 Masking Content vs. Style in Hyperpartisan News

In this section, we reported the results of the masking technique from two different perspectives. In one setting, we masked *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 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.

<sup>1</sup>The dataset will be available to the research community upon acceptance.

Original text	Masking topic-related words	Masking style-related words
Officers <b>went after</b> Christopher <b>Few after</b> watching <b>an</b> argument <b>between him and his</b> girlfriend outside <b>a</b> bar <b>just before the</b> 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 1: Examples of masking style-related information or topic-related information.

	Left-wing	Mainstream	Right-wing	$\Sigma$
Original data (Potthast et al., 2018)	256	826	545	1627
Cleaned data	252	787	516	1555

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

After the text transformation by the masking process in both the training and test sets, we represented the documents with character  $n$ -grams and compared the results obtained with the *style-based* and the *topic-related models*.

### 3.2 Experimental Setup

**Text Transformation:** We evaluated different values of  $k$  ( $k \in \{100, 200, \dots, 5000\}$ ) for extracting the  $k$  most frequent words from English<sup>2</sup>. For the comparison of the results obtained by each model with the ones of the state of the art, we only showed the results fixing  $k = 500$ .

**Text Representation:** We used a standard bag-of-words representation with  $tf$  weighting and extracted character 5-grams with a frequency lower than 50.

**Classifiers:** We compared the results obtained with Naïve Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF); for the three classifiers we used the versions implemented in *sklearn* with the parameters set by default.

**Transformers:** We approached the hyperparameter tuning by grid search. The best results were obtained with: *learning rate* =  $3e - 5$ ; the *batch size* = 16; and the *adam* optimizer. Moreover, we applied a dropout value of 0.3 to the last dense layer. We have selected a value of 200 for the *max.length* hyperparameter.

**Evaluation:** We performed 3-fold cross-validation with the same configuration

used in (Potthast et al., 2018). Therefore, each fold comprised one publisher from each orientation (the classifiers did not learn a publisher’s style). We used macro  $F_1$  as the evaluation measure since the test set is unbalanced with respect to the three classes. In order to compare our results with those reported in (Potthast et al., 2018), we also used accuracy, precision, and recall.

**Baseline:** Our baseline method is based on the same text representation with the character  $n$ -grams features, but without masking any word.

## 4 Results and Discussion

Table 3 shows the results of the proposed method and the system from (Potthast et al., 2018)<sup>3</sup> in our cleaned dataset (Section 3), both considering topic and style-based methods. In order to compare our results with those reported in (Potthast et al., 2018), 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 (Stamatatos, 2017b) and set  $k = 500$  for this comparison results.

Similar to (Potthast et al., 2018), 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 obtained from the system of (Potthast et al., 2018) (0.63 vs. 0.61). The highest scores of the masking technique were consistently achieved

<sup>2</sup>We use the BNC corpus (<https://www.kilgarriff.co.uk/bnc-readme.html>) for the extraction of the most frequent words as in (Stamatatos, 2017a).

<sup>3</sup><https://github.com/webis-de/ACL-18>



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
<b>System from (Potthast et al., 2018) (applied to our cleaned dataset)</b>												
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
<b>Transformer-based models</b>												
M-BERT		0.76	0.83	0.65	0.75	0.93	0.49	0.86	0.92	0.56	<b>0.93</b>	0.80
XLNet		0.80	0.86	<b>0.80</b>	0.76	<b>0.95</b>	0.50	<b>0.91</b>	0.94	0.61	0.83	<b>0.95</b>
BERT		<b>0.86</b>	<b>0.89</b>	0.77	<b>0.87</b>	0.94	<b>0.75</b>	0.86	<b>0.96</b>	<b>0.76</b>	0.87	<b>0.95</b>

Table 3: Results of the proposed masking technique ( $k = 500$  and  $n = 5$ ) 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 (Potthast et al., 2018) to our cleaned dataset (Section 3).

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 (Potthast et al., 2018), the left-wing orientation is harder to predict, possibly because this class is represented with fewer examples in the dataset.

Another reason why our masking approach achieves better results than the system from (Potthast et al., 2018), could be that we use a higher length of character n-grams. In fact, comparing their results against our baseline model, it is possible to note that even without masking any word, the classifier obtains better results. This suggests that the good results are due to the length of the character n-grams rather than the use of the masking technique.

The last three rows of Table 3 show the results of the transformer-based models. As we can see, these models achieved the highest results, 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 models at predicting the correct orientation using just the beginning of the news ( $max\_length = 200$ ). This is aligned to the work of (Ghanem et al., 2021) that focused on analyzing the initial part of false news articles. The authors as-

Baseline model			Style-based model			Topic-based model		
left	main	right	left	main	right	left	main	right
_imag	_cnn_	e_are	but_*	n_thi	y_**_	ant_*	_cnn_	hilla
_that	said_	lary_	out_w	s_**s	out_a	lies_	ics_*	als_*
e_tru	_said	_your	t_**_	_how_	as_to	_*_ex	sday_	_*_le
e_don	y_con	n_pla	you_h	at_he	o_you	etty_	ed_be	_dail
_here	ry_co	e_thi	t_and	m_**t	ell_*	donal	_cnn_	_*_te
s_of_	_cnn_	s_to_	_is_a	**_**u	and_n	n_**c	day_*	**_ame
for_h	said_	illar	h_**_a	e_#_**	hat_w	onald	cs_**_	_*_am
donal	_said	hilla	_of_#	and_*	**_#_#	ying_	ics_*	illar
racis	ore_t	llary	or_hi	**_**_	_it_t	thing	**_**e	llary
_kill	story	_hill	for_h	t_the	e_of_	â_**_*	ed_be	_*_le
_that	_said	_let_	**_**_	and_*	o_you	eâ_**_	y_con	_*_ri
_trum	tory_	_comm	_in_o	_**_tw	n_it_	nâ_**_	tory_	_hill
trump	ed_be	lary_	hat_*	**_two	and_n	tâ_**_	story	_bomb

Table 4: Most relevant features to each class.

sumption 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.

#### 4.1 Relevant Features

Table 4 shows the features with the highest weights from the SVM (we used `scikit-learn`’s method to collect feature weights). It is possible to note that the mention of *cnn* was learned as a discriminative feature when the news from that publisher were used in the training (in the topic-based model). However, this feature is infrequent in the test set where no news from CNN publisher was included.

The features related to Donald Trump (*donal* and *onal*), and Hillary Clinton (*llary* and *illar*) are more frequent in one of the hyperpartisan orientation, and none of them occurs frequently in the

Topic-related model	
left	(...)which his son <b>pretty</b> much confirmed in a foolish statement. The content of those tax returns has been the subject of much speculation, but given Trump’s long history of tax evasion and political bribery, it doesn’t take much <b>imagination</b> to assume he’s committing some kind of fraud * * son <b>pretty</b> * confirmed * foolish statement * content * * tax returns * * * subject * * speculation * * Trump * * * tax evasion * * bribery * doesn * * <b>imagination</b> * assume * * committing * * fraud
main	Obama proved <b>beyond</b> a shadow of a doubt in 2011 when he released his long-form birth certificate (...) <b>CNN</b> and Fox News cut away at points in the presentation. Networks spent the day talking about Trump’s history as a birther (...) Before Friday, the campaign’s most recent deception came <b>Wednesday</b> when campaign advisers told reporters that Trump would not be releasing results of his latest medical exam Obama proved <b>beyond</b> shadow * doubt * 2011 * * released * * * birth certificate (...) <b>CNN</b> * Fox News cut * * points * * presentation Networks spent * * talking * Trump * * birther (...) * Friday * campaign * recent deception * <b>Wednesday</b> * campaign advisers told reporters * Trump * * releasing results * * latest medical exam
right	The email, which <b>was</b> dated March 17, 2008, and shared with POLITICO, reads: Jim, on Kenya your person in the field might look into the impact there of Obama’s public comments about his father, I’m told by State Dept officials <b>that</b> Obama publicly derided his father on (...) Blumenthal, a longtime confidant of both Bill and <b>Hillary</b> Clinton, emerged as a frequent correspondent in the former secretary of (...) * email * <b>dated</b> March 17 2008 * shared * POLITICO reads Jim * Kenya * * * field * * * * impact * * Obama * comments * * told * * Dept officials * Obama publicly derided * * (...) Blumenthal longtime confidant * * Bill * <b>Hillary</b> Clinton emerged * frequent correspondent * * former secretary * (...)

Table 5: Fragments of original texts and their transformation by masking the  $k$  most frequent terms. Some of the features from Table 4 using the topic-related model are highlighted.

mainstream orientation. On the other hand, the relevant features from the style-based model involve function words that are frequent in the three classes (e.g., *out*, *you*, *and*, *of*) even if the combination between function words and other characters can lightly differ in different orientations.

#### 4.2 Features with the Highest Attention Scores

Transformer-based models allow us to visualize different parts of the news according to the scores they received to obtain the final prediction. In Figure 1 we show examples of news predicted correctly by BERT (the model with the highest  $F_1$  score). Due to space limitations, we provide fragments of six news, two per orientation. The more intense the color, the greater is the weight of attention given by the model.

In the examples from 1a, the left-wing orientation remarks the names of the opposite politicians, and it is possible to see which of them 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, examples 1c 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*. However, examples from 1c offer a comparison in which keep the reader in a neutral position. Moreover, in the second 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).

We can conclude that the attention mechanism of the transformers not only help in doing effective predictions, but offer some extra information that could be useful to understand some insights about hyperpartisanship. For example, the words with the highest scores can be used in other strategies to confirm the previous results that topic-based models outperform a style-based one at distinguishing left, right and mainstream orientations (Potthast et al., 2018).

## 5 Conclusions

In this paper we presented initial experiments on the task of hyperpartisan news detection. In particular, we aimed to explore the trade-off between performance and transparency, and proposed a comparison of two different approaches. First, we explored the use of masking techniques to boost the performance of a lexicalized 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 comparing our results with the state of the art. We evaluated different values of the parameters and see that finally our baseline model, in which we extract character 5-grams without applying any masking process, achieves the better results. This seems to indicate a strong lexical overlap between different sources with the same orientation, which, in turn, calls for more challenging datasets and task formulations to encourage the development of models covering more subtle, i.e., implicit, forms of bias. Future datasets could consider more topics and different time spans to avoid the models learn from the topic, rather than the target classes.

Second, we used three transformer-based models (BERT, M-BERT, and XLM-RoBERTa) that are resource-hungrier than the masking technique, and achieved the highest results. We also presented some examples of how these models, through their attention scores, provide additional information about the relevant parts of the text for distinguishing their political orientation. Considering the high effectiveness of these models, and that they only observe the first part of the news, we will evaluate as future work how necessary is to use all the

on the topic of climate change , hillary clinton seems more knowledgeable of donald trump ' s words than he does , earlier in monday night ' s presidential debate at hofstra university , democratic nominee hillary clinton pointed out that the gop nominee previously said that

once again , donald trump and the republican party ' s fear - mongering about immigrants is proven false , ever since an improvised explosive device injured 29 in chelsea , new york city , trump and his goons have revived one of their favorite talking points vilifying syrian refugees

(a) Hyperpartisan (left-wing) news.

donald trump feels like a man half his age , and hillary clinton is " quite delighted " that the topic of the septua - and sexagenarian ' s ages haven ' t been an issue throughout their presidential campaigns , both candidates responded to aarp bulletin for the cover story of its

when donald trump took his campaign to high point , north carolina , tuesday , his topics ranged broadly from trade to immigration to terrorism , in other words , none of the hot - button issues that are currently roiling the political landscape in the battleground state that

(b) Non-hyperpartisan (mainstream orientation) news.

there shouldn ' t be an estate tax period , right now the rate stands at 40 % , if hillary clinton gets her way , she ' ll raise it to a whopping 65 % and i would not be surprised to see it go even higher , we are being taxed to death and hillary loves taxes , taxation equals slavery ... it is

hillary is without a doubt , the worst and most despicable liar to ever run for the office of president of the united states ... hillary is a sociopathic liar , a sociopath is typically defined as someone who lies incessantly to get their way and does so with little concern for others , a sociopath

(c) Hyperpartisan (right-wing) news.

Figure 1: Fragments of news (two for each political orientation) with the visualization of the attention learned by BERT. The more intense the color, the greater the weight of attention.

news (and not only the beginning), e.g. with the Transformer-XL model (Dai et al., 2019). Moreover, we are motivated to take advantage of the attention scores to study in more detail the style used in hyperpartisan news in order to improve the predictions.

### Acknowledgments

The work of the authors from the Universitat Politècnica de València was funded by the Spanish Ministry of Science and Innovation under the research project MIS-MIS-FAKEnHATE on MIS-information and MIScommunication in social media: FAKE news and HATE speech (PGC2018-

096212-B-C31). Experiments were carried out on the GPU cluster at PRHLT thanks to the PROM-ETEO/2019/121 (DeepPattern) research project funded by the Generalitat Valenciana.

### References

Clio Andris, David Lee, Marcus J. Hamilton, Mauro Martino, Christian E. Gunning, and John Armistead Selden. 2015. The rise of partisanship and Super-Cooperators in the U.S. house of representatives. *PLoS ONE*, 10(4):e0123507.

Talita Anthonio. 2019. Robust document representations for hyperpartisan and fake news detection.

- Master's thesis, University of the Basque Country UPV/EHU.
- Ricardo Baeza-Yates. 2018. *Bias on the web*. *Communications of the ACM*, 61(6):54–61.
- José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PMLADC at ICLR 2020*.
- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. *Transformer-XL: Attentive language models beyond a fixed-length context*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Bilal Ghanem, Simone Paolo Ponzetto, Paolo Rosso, and Francisco Rangel. 2021. *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*, pages 679–689, Online. Association for Computational Linguistics.
- Ana Granados, Manuel Cebrian, David Camacho, and Francisco de Borja Rodriguez. 2011. Reducing the loss of information through annealing text distortion. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1090–1102.
- L. Hemphill, A. Culotta, and M. Heston. 2016. #polar scores: Measuring partisanship using social media content. *Journal of Information Technology & Politics*, 13(4):365–377.
- Marjan Hosseinia. 2020. *Content and Stylistic Models for Authorship, Stance, and Hyperpartisan Detection*. Ph.D. thesis, University of Houston.
- 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*. pages 829–839.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.
- Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylistic inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240.
- J. Sánchez-Junquera. 2018. Adaptación de dominio para la detección automática de textos engañosos. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica. [Available here](#) (in Spanish).
- Efstathios Stamatatos. 2017a. 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*, volume 1, pages 1138–1149.
- Efstathios Stamatatos. 2017b. 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*, pages 1138–1149. Association for Computational Linguistics.
- Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.