Hyperpartisan News Analysis With Scattertext

Julen Etxaniz
University of the Basque Country
jetxaniz007@ikasle.ehu.eus

Oihane Cantero
University of the Basque Country
ocantero003@ikasle.ehu.eus

Abstract

1 Introduction

Hyperpartisan news are those that take an extreme left-wing or right-wing standpoint. Detecting hyperpartisan news automatically can be useful to tag them and inform readers. This was the goal of the SemEval 2019 Task 4 (Kiesel et al., 2019).

The purpose of this work is to analyze the usage of words in documents which are hyperpartisan and non-hyperpartisan. Hyperpartisan news are those that exhibit blind, prejudiced, or unreasoning allegiance to one party, faction, cause, or person.

Whereas the task on semeval was to design a system to auto- matically detect hyperpartisan news, in this exercise we are going to exploit both corpora (hyperpartisan and not hyperpartisan news), and analyze which terms are the most relevant in each of the sets.

2 Related Work

3 Dataset

The data is split into multiple files. The articles are contained in the files with names starting with ”articles-” (which validate against the XML schema article.xsd). The ground-truth information is contained in the files with names starting with ”ground-truth-” (which validate against the XML schema ground-truth.xsd).

The dataset can be downloaded from Zenodo[1]. You can also use the dataset creation script to create a HuggingFace Dataset automatically.

3.1 By Publisher Dataset

The first part of the data (filename contains ”by-publisher”) is labeled by the overall bias of the publisher as provided by BuzzFeed journalists or MediaBiasFactCheck.com. It contains a total of 750,000 articles, half of which (375,000) are hyperpartisan and half of which are not. Half of the articles that are hyperpartisan (187,500) are on the left side of the political spectrum, half are on the right side. This data is split into a training set (80%, 600,000 articles) and a validation set (20%, 150,000 articles), where no publisher that occurs in the training set also occurs in the validation set. Similarly, none of the publishers in those sets occurs in the test set (4,000 articles).

3.2 By Article Dataset

The second part of the data (filename contains ”byarticle”) is labeled through crowdsourcing on an article basis. The data contains only articles for which a consensus among the crowdsourcing workers existed. It contains a total of 645 articles. Of these, 238 (37%) are hyperpartisan and 407 (63%) are not, We will use a similar (but balanced!) test set that contains 628 articles. Again, none of the publishers in this set occurs in the test set.

4 Methods

First, we preprocess the original data to get better results in the analysis. Then, we use two different methods for analysing hyperpartisan and non-hyperpartisan documents. On the one hand, we calculate log-odd ratios to extract the most relevant words of each category. On the other hand, we use scattertext (Kessler, 2017) to build an interactive HTML scatter plot. Code is available at GitHub: [2]

4.1 Preprocessing

As the original files are XML files, we have to preprocess them in order to obtain good insights. First

[1] https://zenodo.org/record/5776081

[2] https://github.com/juletx/hyperpartisan-news-detection

we use the `lxml` library in python to analyze the XML documents and extract the necessary information. Preprocessing also includes tokenizing, converting words to lowercase, removing punctuation, numbers, stop words, XML entities and image tags.

To calculate the log-odd ratios, we select the validation set of the By Publisher dataset that contains 150,000 articles. The first step is to generate two text files for hyperpartisan and non-hyperpartisan news articles, respectively. You have to divide the News articles contained in articles-validation-bypublisher-20181122.xml.zip into two text files (`hyperpartisan.txt` and `non-hyperpartisan.txt`), according to their ground truth value in ground-truth-validation-bypublisher-20181122.xml.zip.

For scattertext, we select the test set of the By Article Dataset, which contains 628 articles. Scattertext needs a smaller number of articles because otherwise the interactive site takes very long to load. This size is big enough to extract the most relevant words.

4.2 Log-odd ratio

After preprocessing text files, we extract the log-odd ratios of each word. Because the log-odd ratio is sensitive to infrequent words, we discard words that appear less than 20 times in the corpus. We also extract the log-odd ratios of the bigrams in the corpus. Having the log-odd ratios, we can extract the most relevant 50 words and bigrams in hyperpartisan and non-hyperpartisan documents. If we analyze these words, we can draw some conclusions about hyperpartisan news.

The log-odd ratio is a measure of words compared on two sets of documents (i and j), which in our case corresponds to hyperpartisan and non-hyperpartisan documents, respectively. Each word then can be associated with its log-odd ratio r_w, which is a number that can be positive or negative: positive numbers are associated with set i, and negative numbers with set j.

The log-odd ratio r_w is defined as:

$$p_w^{(i)} = \frac{f_w^{(i)}}{N^{(i)}}; p_w^{(j)} = \frac{f_w^{(j)}}{N^{(j)}}$$

$$o_w^{(i)} = \frac{p_w^{(i)}}{1 - p_w^{(i)}}; o_w^{(j)} = \frac{p_w^{(j)}}{1 - p_w^{(j)}}$$

$$r_w = \log o_w^{(i)} - \log o_w^{(j)}$$

where $f_w^{(i)}$ is the frequency of word w in group i (hyperpartisan or non-hyperpartisan), and $N^{(i)}$ is the number of words in group i.

For example, suppose that the word $gold$ appears $2,500$ times on hyperpartisan documents ($f_{gold}^{(i)} = 2500$), and 760 times on non-hyperpartisan documents ($f_{gold}^{(j)} = 760$). Furthermore, suppose that there are $25,000$ words in hyperpartisan documents ($N^i = 25000$), and $17,500$ on non-hyperpartisan documents ($N^j = 17000$). Then:

$$p_{gold}^{(i)} = \frac{2500}{25000} = 0.1; p_{gold}^{(j)} = \frac{760}{17500} = 0.045$$

$$o_{gold}^{(i)} = \frac{0.1}{1 - 0.1} = 0.11; o_{gold}^{(j)} = \frac{0.045}{1 - 0.045} = 0.047$$

$$r_{gold} = \log 0.11 - \log 0.047 = 0.369$$

and therefore the log odd ratio of $gold$ is 0.369.

4.3 Scattertext

Scattertext (Kessler, 2017) is a tool for finding distinguishing terms in corpora and displaying them in an interactive HTML scatter plot. It is intended for visualizing what words and phrases are more characteristic of a category than others. We can use it to compare hyperpartisan and non-hyperpartisan news. It could also be used to compare news with left and right bias.

5 Results

5.1 Relevant Words

There are many differences betweeen the 50 most relevant words of hyperpartisan and non-hyperpartisan news. Here are the main findings of each class. Table 1 shows all words and bigrams.

Relevant words of hyperpartisan articles:

- Hyperpartisan articles contain words ending in -ist/-ism/-ity(anarchist, anarchism, globalist, globalists, individualist, anarchists, zionists, vulgarity, profanity). These do not appear in non-hyperpartisan words.
- Other hyperpartisan words that describe people (slager, teabagger, shep, lgbtq, courteous). Similar terms do not appear in non-hyperpartisan words.
- Bad words in hyperpartisan articles (fucking, trolling, fuck, fck). There are no bad words in non-hyperpartisan.

- News sites or webs in hyperpartisan articles (wonkette, realclearpolitics, newsbusters, vox, gofundme, newsmax, foxnewscom). This suggest that these news sites are commonly associated with hyperpartisan news. Most correspond to news agencies in the US. No news agencies appear in non-hyperpartisan words.

- Other organizations in hyperpartisan news (usmc (United States Marine Corps), splc (Southern Poverty Law Center), emmys). They are US organizations.

- People in hyperpartisan articles (oreilly, kilmeade, chomsky, cavuto, grahamcassidy, omalley, madsen, beyoncé, willard, odonell, kliff, machado, watters, susteren). They correspond to politicians, journalists and famous people.

Relevant words of non-hyperpartisan articles:

- Demonyms in non-hyperpartisan articles (subsaharan, nigerians, thai, israelpalestinian, nigerian). They correspond to people from other countries. No demonyms appear in hyperpartisan words.

- Places in non-hyperpartisan articles (bangkok, myanmar, rakhine, nigeria, thailand, tribune, kyoto, lima). Many places appear in non-hyperpartsan words, none in hyperpartisan words. They correspond to other countries and cities.

- People in non-hyperpartisan articles (straus, zuma, newsom, hun, schwarzenegger, hu (Hu Jintao)). They correspond to politicians and famous people.

- Organizations in non-hyperpartisan articles (treasuries, boko, haram, tic, utaustin (The University of Texas at Austin), anc (African National Congress, pri (Partido Revolucionario Institucional), nld (National League for Democracy)). Unlike hyperpartisan organizations, many organizations are from other countries different from the US.

- There are many economics terms in non-hyperpartisan articles (renminbi, yen, rebalance, cfr, depreciation, exporters, aggregator, outflow). Some correspond to currencies and other to actions or people.

5.2 Relevant Bigrams

There are many differences betweeen the 50 most relevant bigrams of hyperpartisan and non-hyperpartisan news. Here are the main findings of each class.

Relevant bigrams of hyperpartisan articles:

- More negative words than on non-hyperpartisan news (threats violence, hate group, divestment sanctions, overdose deaths, illegal alien)

- People (obama, trump, bill oreilly, romney, darren wilson, mr comey, van susteren). They correspond to politicians or famous people

- Media related terms (media research, independent journalism, media matters, corporate media, associate editor)

- Politics related terms (trump obama, legislature, obamacare, america health, basic income, ruling class)

Relevant bigrams of non-hyperpartisan articles:

- Demonyms in non-hyperpartisan articles (southeast asian, african).

- Places in non-hyperpartisan news (texas, us, china, southeast asia, travis county, austin). Some places are repeated a lot in different bigrams: china, us and texas are the most repeated ones.

- Many economics related bigrams also appear a lot (emerging economies, exchange rate, emerging markets, direct investment, private investors...).

- Organizations (boko haram, international institutions, china gorvernment...)

- People are also mentioned (dan patrick, jacob zuma, suu kyi, president jacob, david dewhurst)

5.3 Scattertext

6 Conclusions

References

Jason S. Kessler. 2017. Scattertext: a browser-based tool for visualizing how corpora differ.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

Hyp Words	Non Words	Hyp Bigrams		Non Bigrams	
wonkette	subsaharan	repeat	offenders	texas	tribune
vulgarity	straus	state	shall	may	subject
realclearpolitics	treasuries	media	keep	emerging	economies
newsbusters	zuma	reserve	right	complete	list
oreilly	boko	media	research	exchange	rate
profanity	tic	trump	nt	us	assets
kilmeade	renminbi	white	privilege	texas	house
vox	haram	illegal	alien	boko	haram
chomsky	bangkok	like	college	emerging	markets
courteous	checker	law	shall	dan	patrick
gofundme	nigerians	person	shall	southeast	asian
rcp	newsom	legislature	may	direct	investment
newsmax	utaustin	without	warning	sovereign	wealth
anarchism	heremore	threats	violence	us	current
trolling	tribune	monday	friday	guest	post
cavuto	cfr	agree	terms	members	may
fucking	custodial	obama	nt	us	exports
foxnewscom	myanmar	us	maintain	china	trade
anarchist	rakhine	news	hour	growth	china
susteren	hun	bill	oreilly	us	firms
grahamcassidy	grist	obamacare	repeal	net	exports
newsletter	thai	media	matters	research	associate
chez	exporters	news	team	exchange	rates
usmc	nigeria	black	panther	international	institutions
splc	denuclearization	hate	group	travis	county
omalley	crossposted	independent	journalism	global	governance
fuck	anc	van	susteren	private	investors
banter	yen	false	flag	balance	sheet
madsen	depreciation	season	two	story	updated
watters	rebalance	research	team	jacob	zuma
beyoncé	schwarzenegger	happening	world	china	central
willard	thailand	corporate	media	china	government
jerk	aggregator	romney	leads	east	north
lgbtq	sponsors	officer	darren	texas	austin
globalist	kyoto	privately	owned	development	goals
odonnell	lima	overdose	deaths	suu	kyi
emmys	pri	shall	made	think	worth
mises	israelpalestinian	darren	wilson	advanced	economies
globalists	outflow	big	league	bretton	woods
kliff	suu	show	today	president	jacob
individualist	nigerian	associate	editor	news	views
fck	nld	game	thrones	north	texas
machado	uschina	author	necessarily	texas	senate
anarchists	cyberspace	basic	income	states	china
painkillers	rebalancing	ruling	class	david	dewhurst
slager	inaudible	divestment	sanctions	chinese	state
shall	ph	support	continue	think	china
zionists	odinga	america	health	southeast	asia
teabaggers	bretton	mr	comey	fort	worth
shep	hu	romney	tax	african	national

Table 1: Most relevant hyperpartisan and non-hyperpartisan words and bigrams according to log-odd ratios.