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 automatically detect hyperpartisan news, in this exercise we are going to exploit both corpora and analyze which terms are the most relevant in each of the sets.

We use two different methods for analysing hyperpartisan and non-hyperpartisan documents. First, we calculate log-odd ratios to extract the most relevant words of each category. Then, we use Scattertext (Kessler, 2017) to build an interactive HTML scatter plot. We compare the results of each method and extract some conclusions.

2 Dataset

This dataset contains hyperpartisan and non-hyperpartisan news articles published online. The data is divided into two datasets. One has 1,273 articles, each labeled manually, while the second, larger dataset of 754,000 articles is labeled in a semi-automated manner via distant supervision at the publisher level. Each dataset is further divided into training, validation and testing subsets. The articles and ground-truth information of each subset are contained in separate XML files.

The dataset can be downloaded from Zenodo[1]. You can also use the dataset creation script to create a HuggingFace Dataset automatically with the original and cleaned text. It also contains all the metadata attributes that are present in the original files. This could be used to easily build a classifier using the HuggingFace Transformers library.

2.1 By Publisher Dataset

The first part of the data 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).

2.2 By Article Dataset

The second part of the data 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 balanced test set that contains 628 articles. Again, none of the publishers in this set occurs in the test set.

3 Methods

First, we preprocess the original data to get better results in the analysis. Then, we use two differ-

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

ent 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]

3.1 Preprocessing

As the original files are XML files, we have to preprocess them in order to obtain good insights. First 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, extra whitespaces, 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. We have to divide the news articles contained in the validation XML file, according to their ground truth value.

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.

3.2 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.

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

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.

3.3 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 also gives the terms which occur frequently in all sets of documents being studied (both categories), but relatively infrequent compared to general term frequencies.

Scattertext uses a metric called Scaled F-Score to rank terms. Associated terms have a high category-specific precision and category-specific term frequency. The harmonic mean of precision and frequency is taken to ensure that both are high. Two adjustments are made to come up with the final formulation of Scaled F-Score.

Given a word $w_i \in W$ and a category $c_j \in C$, define the precision of the word w_i with regard to a category as:

$$\text{prec}(i,j) = \frac{\#(w_i, c_j)}{\sum_{c \in C} \#(w_i, c)}.$$

The function $\#(w_i, c_j)$ represents either the number of times w_i occurs in a document labeled with the category c_j or the number of documents labeled c_j which contain w_i.

Similarly, define the frequency a word occurs in the category as:

$$\text{freq}(i,j) = \frac{\#(w_i, c_j)}{\sum_{w \in W} \#(w, c_j)}.$$

The harmonic mean of these two values of these two values is defined as:

$$\mathcal{H}_\beta(i,j) = (1+\beta^2)\frac{\text{prec}(i,j) \cdot \text{freq}(i,j)}{\beta^2 \cdot \text{prec}(i,j) + \text{freq}(i,j)}.$$

$\beta \in \mathcal{R}^+$ is a scaling factor where frequency is favored if $\beta < 1$, precision if $\beta > 1$, and both are equally weighted if $\beta = 1$. F-Score is equivalent to the harmonic mean where $\beta = 1$.

The first problem of the previous formulation is that harmonic means are dominated by precision. To solve this we take the normal CDF of precision and frequency percentage scores, which will be scaled between 0 and 1.

Define the the Normal CDF as:

$$\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(x; \mu, \sigma^2)\, \mathrm{d}x.$$

Where $\mathcal{N}$ is the PDF of the Normal distribution, μ is the mean, and σ^2 is the variance. Φ is used to scale and standardize the precisions and frequencies, and place them on the same scale $[0, 1]$.

Now we can define Scaled F-Score as the harmonic mean of the Normal CDF transformed frequency and precision:

$$\text{S-CAT}_\beta(i, j) = \mathcal{H}_\beta(\Phi(\text{prec}(i, j)), \Phi(\text{freq}(i, j))).$$

μ and σ^2 are defined separately as the mean and variance of precision and frequency. A β of 2 is recommended and is the default value in Scattertext.

Note that any function with the range of $[0, 1]$ can be used in place of Φ. Also, when the precision is very small normalization may be foregone.

The second adjustment consists of making the score fair for negative scoring terms. For this we compute the Scaled F-Score of the negative class. If that score has a higher magnitude than the positive one, we keep that value as a negative score.

Note that the range of $\mathcal{S}$ is now $[-1, 1]$, where $\mathcal{S} < 0$ indicates a term less associated with the category is question than average, and a positive score being more associated.

4 Results

The results we get by extracting the words relevant words using log-odd ratios and using Scattertext are different.

4.1 Using log-odd ratios

We get relevant words and bigrams from hyperpartisan and non-hyperpartisan articles by calculating their log-odd ratio.

4.1.1 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.

4.1.2 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. Table 1 shows all words and bigrams.

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)

4.2 Scattertext

We visualized the differences between hyperpartisan and non-hyperpartisan of the original text and cleaned text (before and after the preprocessing of the text mentionned in section 3.1).

4.2.1 Original text

Most frequent words in both hyperpartisan articles are stopwords, and we can also see that in both cases some there are some non-words character sequences (= twsrc%5etfw, type="external"¿august). We can see all the words and which of them appear most in hyperpartisan and non-hyperpartisan articles in Figure 1.

Relevant words of hyperpartisan articles:

- Names of people and words of the political domain (donald trump, politics, republican, second amendment)

- Names of countries (iranian, afghanistan)

- Stopwords (it's, he's, does, that's)

Relevant words of non-hyperpartisan articles:

- Words that doesn't seem to have any political connotation (dental, tooth, tooth regeneration). These words probably appear in the same articles, because they are all from the same field

- Characters that don't form words (", — a, —)

Overall we see that this corpus needs to be cleaned, as there are a lot of stopwords and character sequences that doesn't form words (from URLs, for example).

4.2.2 Clean text

By cleaning the text, we get better results, as the all the words we get for both top-hyperpartisan and non-hyperpartisan terms exists. We can see all the

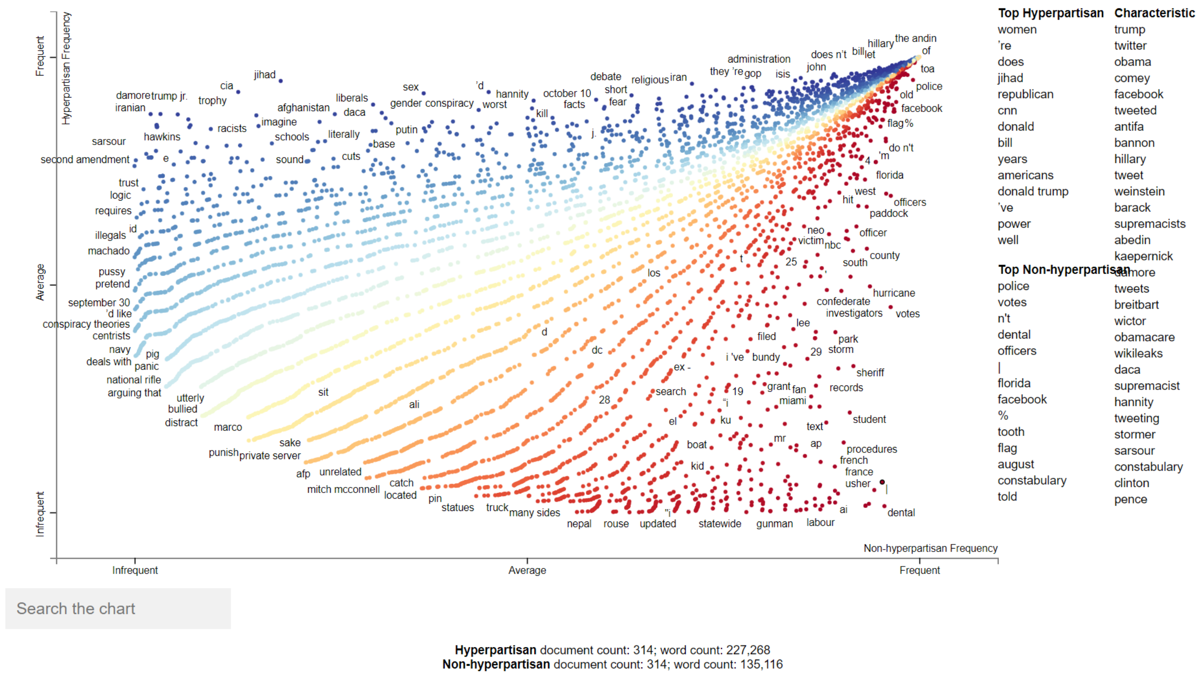


Figure 1: Scattertext plot and top words for the original test set of the By Article dataset.

words and which of them appear most in hyperpartisan and non-hyperpartisan articles in Figure 3.

Relevant words of hyperpartisan articles:

- Words related to politics and organizations (republican, cia, isis, jihad)
- People's names (damore, wictor)
- Names of countries (iranian, afghanistan)

Relevant words of non-hyperpartisan articles:

- Words related to the institutions (votes, flag, police, officers, constabulary)
- As in the original text, words that doesn't seem to have any political connotation (dental, tooth, tooth regeneration, august)

If we compare the results of Figures 1 and 3), we can see that words from the cleaned corpus also appear in the original ones, and for most of them in the same place. And if we look at the more characteristic words of the whole corpus, we see that they are almost the same, and they usually have political connotation (trump, obama, antifa, supremacist, neonazi...).

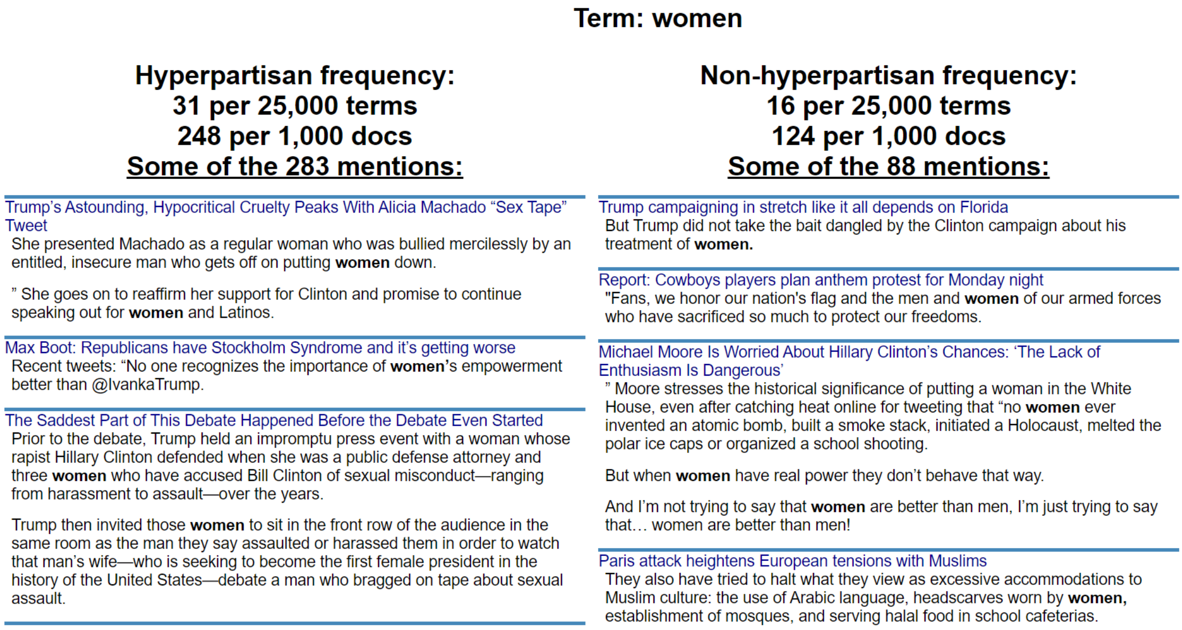
5 Conclusions

We saw that with both methods, we got similar results, we find almost the same words in the hyperpartisan and non-hyperpartisan, but the visualization with Scattertext makes the results more readable.

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.



Term: women

Hyperpartisan frequency:	Non-hyperpartisan frequency:
31 per 25,000 terms	16 per 25,000 terms
248 per 1,000 docs	124 per 1,000 docs
Some of the 283 mentions:	Some of the 88 mentions:

Hyperpartisan mentions:

Trump's Astounding, Hypocritical Cruelty Peaks With Alicia Machado "Sex Tape" Tweet

She presented Machado as a regular woman who was bullied mercilessly by an entitled, insecure man who gets off on putting **women** down.

" She goes on to reaffirm her support for Clinton and promise to continue speaking out for **women** and Latinos.

Max Boot: Republicans have Stockholm Syndrome and it's getting worse

Recent tweets: "No one recognizes the importance of **women**'s empowerment better than @IvankaTrump.

The Saddest Part of This Debate Happened Before the Debate Even Started

Prior to the debate, Trump held an impromptu press event with a woman whose rapist Hillary Clinton defended when she was a public defense attorney and three **women** who have accused Bill Clinton of sexual misconduct—ranging from harassment to assault—over the years.

Trump then invited those **women** to sit in the front row of the audience in the same room as the man they say assaulted or harassed them in order to watch that man's wife—who is seeking to become the first female president in the history of the United States—debate a man who bragged on tape about sexual assault.

Non-hyperpartisan mentions:

Trump campaigning in stretch like it all depends on Florida

But Trump did not take the bait dangled by the Clinton campaign about his treatment of **women.**

Report: Cowboys players plan anthem protest for Monday night

"Fans, we honor our nation's flag and the men and **women** of our armed forces who have sacrificed so much to protect our freedoms.

Michael Moore Is Worried About Hillary Clinton's Chances: 'The Lack of Enthusiasm Is Dangerous'

" Moore stresses the historical significance of putting a woman in the White House, even after catching heat online for tweeting that "no **women** ever invented an atomic bomb, built a smoke stack, initiated a Holocaust, melted the polar ice caps or organized a school shooting.

But when **women** have real power they don't behave that way.

And I'm not trying to say that **women** are better than men, I'm just trying to say that… women are better than men!

Paris attack heightens European tensions with Muslims

They also have tried to halt what they view as excessive accommodations to Muslim culture: the use of Arabic language, headscarves worn by **women,** establishment of mosques, and serving halal food in school cafeterias.

Figure 2: Scattertext search example of the term women in hyperpartisan and non-hyperpartisan documents.

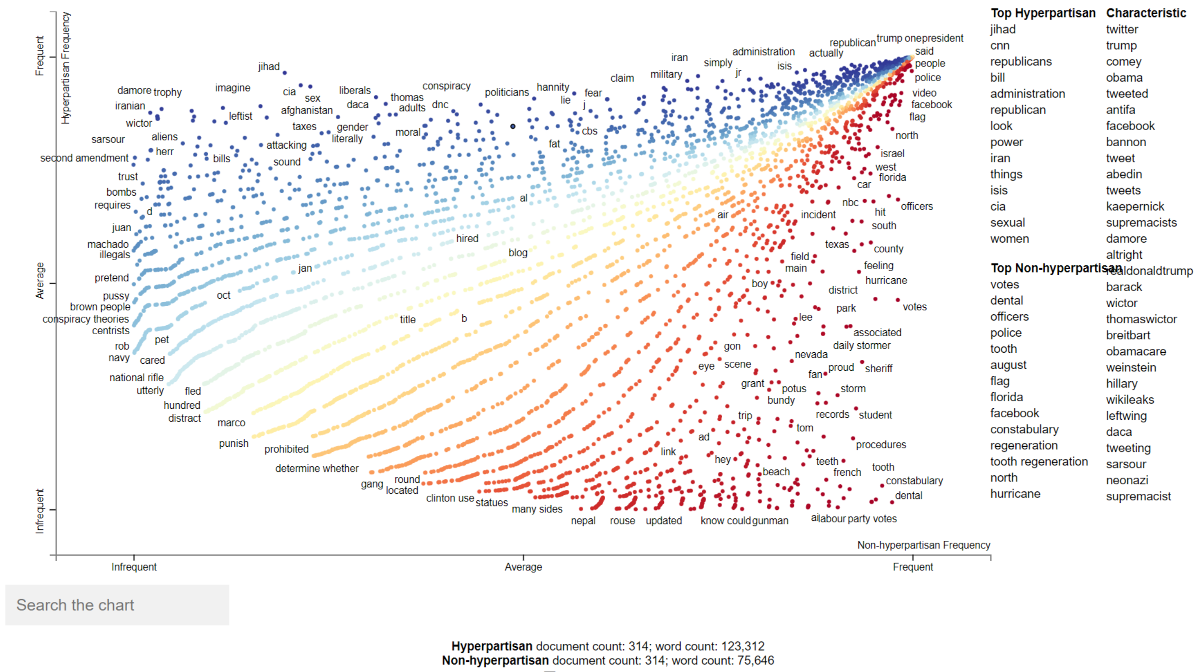


Figure 3: Scattertext plot and top words for the cleaned test set of the By Article dataset.

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.