

INTRODUCTION

Fake news is never out of today's headlines. From profit, power and politics to prejudice, paranoia and propaganda, whatever is its motivation, Fake news has thrived through history. From the flying moon-bats in The Sun in 1835, to the more recent 'pizzagate' story, the chequered history of fake news never lacked imagination. With advent of 'deep fake', this problem has become now more than just a menace. But how do we separate the wheat from the chaff, fact from the fiction? In this project, I have attempted to study the anatomy of 'fake-news'. This modest attempt will in no way stem the ever-growing weed of falsehood, nor will it detect every lie told, but will throw some light on some techniques on studying the fake news systematically.

EXECUTIVE SUMMARY

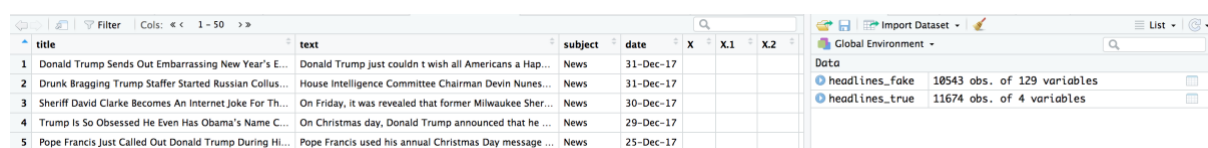
In this project, I have analysed two datasets real.csv and fake.csv, which have been curated from several thousand news headlines from around the world. The titles are labelled as real (indicating that they are valid news articles) and fake (representing fake and propaganda news articles) from spurious internet sources. My primary goal here is to use Text Mining as a tool to identify the differences that can help someone understand how real news headlines differ from fake news headlines. Examine the data given to you (given dataset) using multiple Text mining / opinion mining techniques. I have used UDPIPE and RAKE methods of unsupervised analysis to study the actual words that in the headlines, to develop the foundations of building a classification model out of it.

THE DATA

The dataset True.csv contains 11,674 observations with 4 variables.

	title	text	subject	date
1	As U.S. budget fight looms, Republicans flip their fisc...	WASHINGTON (Reuters) – The head of a conservative ...	politicsNews	31-Dec-17
2	U.S. military to accept transgender recruits on Monda...	WASHINGTON (Reuters) – Transgender people will be ...	politicsNews	29-Dec-17
3	Senior U.S. Republican senator: 'Let Mr. Mueller do his...	WASHINGTON (Reuters) – The special counsel investig...	politicsNews	31-Dec-17
4	FBI Russia probe helped by Australian diplomat tip-o...	WASHINGTON (Reuters) – Trump campaign adviser Ge...	politicsNews	30-Dec-17
5	Trump wants Postal Service to charge 'much more' for...	SEATTLE/WASHINGTON (Reuters) – President Donald ...	politicsNews	29-Dec-17
6	White House, Congress prepare for talks on spending...	WEST PALM BEACH, Fla./WASHINGTON (Reuters) – Th...	politicsNews	29-Dec-17

The Fake.csv is slightly messier with 10,543 observations, and 129 variables.



title	text	subject	date
1 Donald Trump Sends Out Embarrassing New Year's E...	Donald Trump just couldn't wish all Americans a Hap...	News	31-Dec-17
2 Drunk Bragging Trump Staffer Started Russian Collus...	House Intelligence Committee Chairman Devin Nunes...	News	31-Dec-17
3 Sheriff David Clarke Becomes An Internet Joke For Th...	On Friday, it was revealed that former Milwaukee Sher...	News	30-Dec-17
4 Trump Is So Obsessed He Even Has Obama's Name C...	On Christmas day, Donald Trump announced that he ...	News	29-Dec-17
5 Pope Francis Just Called Out Donald Trump During Hi...	Pope Francis used his annual Christmas Day message ...	News	25-Dec-17

We begin by selecting the columns we need.

```
headlines_fake <- headlines_fake[,1:4]
```

We then introduce 3 more columns for Year, Month, and Day.

	title	text	subject	date	year	month	day
1	As U.S. budget fight looms, Republicans flip their fisc...	WASHINGTON (Reuters) – The head of a conservative ...	politicsNews	31-Dec-17	2017	12	31
2	U.S. military to accept transgender recruits on Monda...	WASHINGTON (Reuters) – Transgender people will be ...	politicsNews	29-Dec-17	2017	12	29
3	Senior U.S. Republican senator: 'Let Mr. Mueller do his...	WASHINGTON (Reuters) – The special counsel investig...	politicsNews	31-Dec-17	2017	12	31
4	FBI Russia probe helped by Australian diplomat tip-o...	WASHINGTON (Reuters) – Trump campaign adviser Ge...	politicsNews	30-Dec-17	2017	12	30
5	Trump wants Postal Service to charge 'much more' for...	SEATTLE/WASHINGTON (Reuters) – President Donald ...	politicsNews	29-Dec-17	2017	12	29

We then look at the number of headlines by year for TRUE and FAKE:

TRUE		FAKE	
year	n	year	n
<dbl>	<int>	<dbl>	<int>
2016	4716	2016	5841
2017	6958	2017	4666
		2018	35
		NA	1

For this project, I have included only the years 2016 & 2017. And we can see TRUE has more headlines in 2017, and Fake in 2016. Since there is not much of a difference in the count of headlines, we see if there is any glaring difference in the count by months.

2016 TRUE		2016 FAKE	
month	n	month	n
1	246	1	644
2	432	2	635
3	490	3	597
4	383	4	560
5	394	5	527
6	419	6	427
7	338	7	423
8	265	8	412
9	351	9	410
10	336	10	427
11	637	11	367
12	425	12	412

Except for the last two months in 2016, Fake data has more headlines than True data.

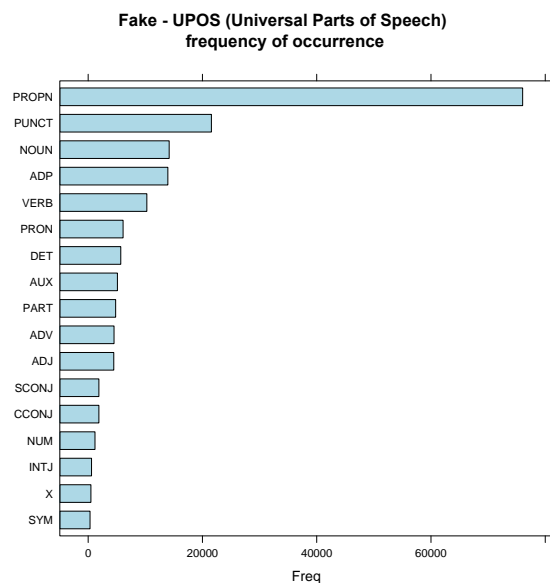
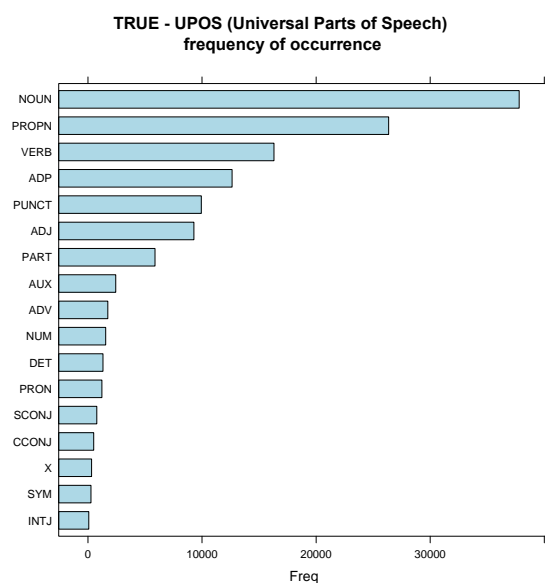
2017 – TRUE

2017 - FAKE

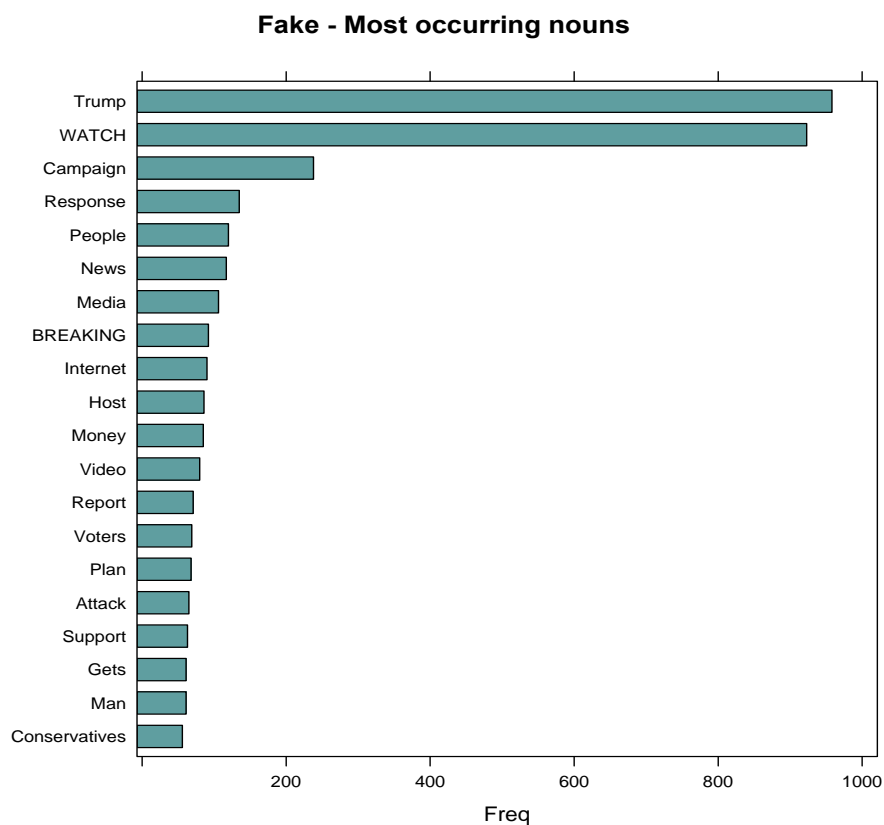
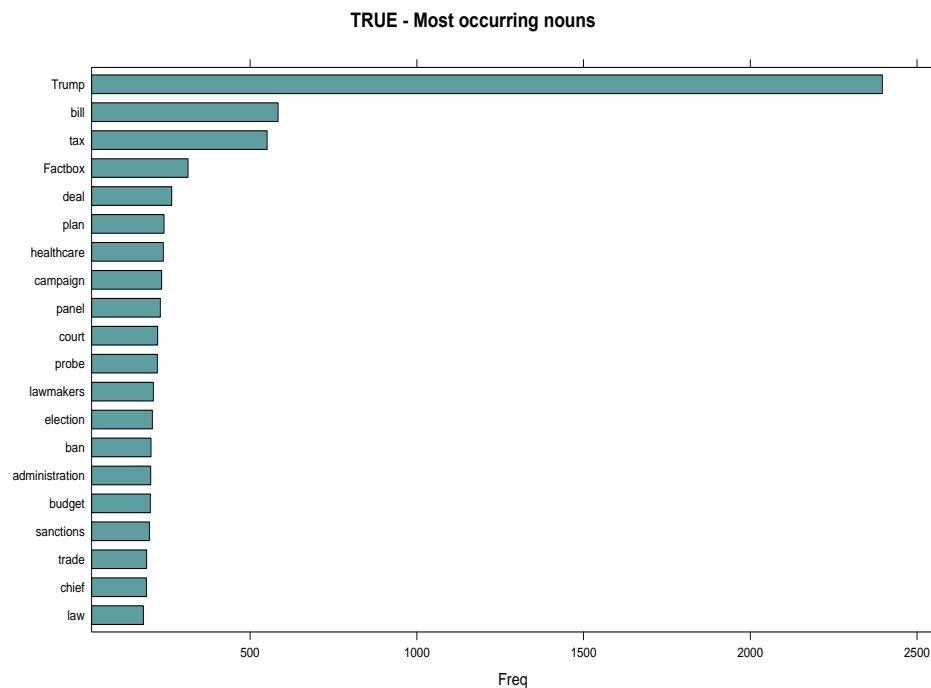
month	n	month	n
1	750	1	460
2	591	2	393
3	684	3	425
4	543	4	308
5	505	5	313
6	556	6	394
7	540	7	506
8	435	8	491
9	497	9	402
10	528	10	354
11	511	11	346
12	818	12	274

Whereas with the exception of the month of August, True dataset has more headlines than Fake. Again, there is no glaring spike in any of the months.

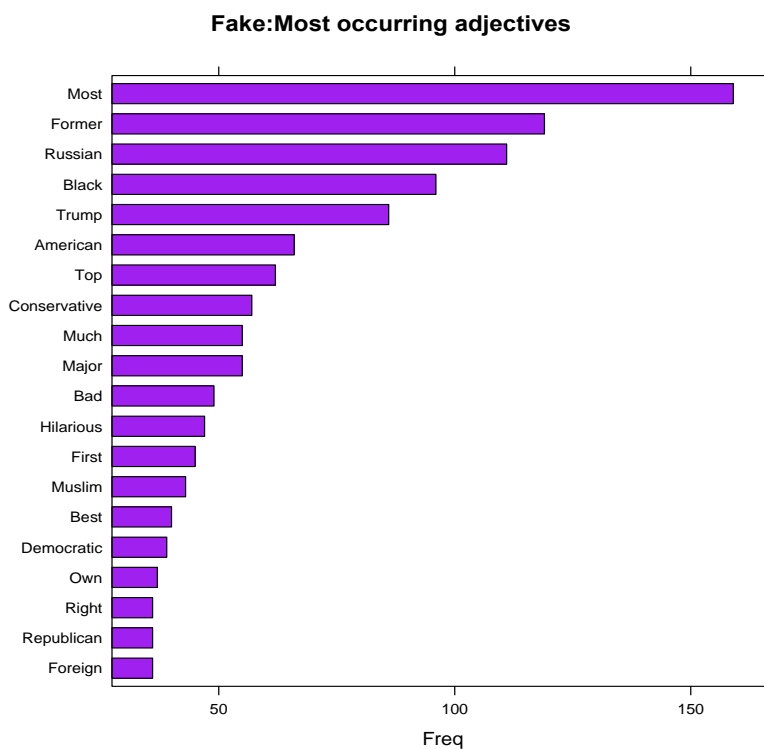
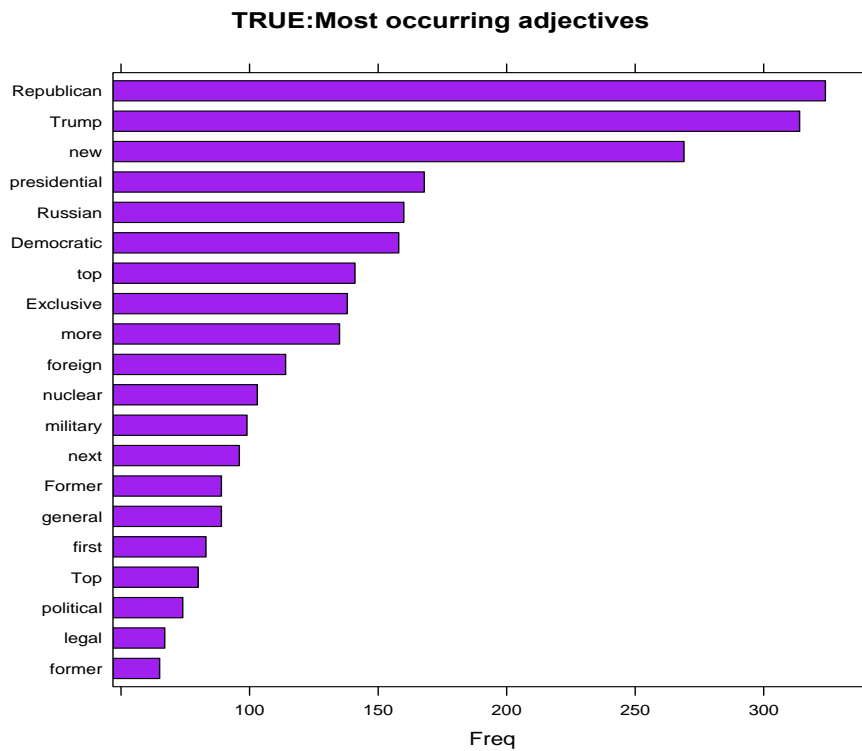
Parts of speech tags or Universal Parts of Speech (UPOS) can reveal a lot about the way a text is structured.



Using UDPIPE model to reveal the Parts of Speech on the True Vs Fake, we see the glaring difference between the two sets, straight off the bat. There are way more nouns (almost 20,000) used in the True set than in Fake. Whereas the Fake headlines contains a lot more names of People and places in terms of Proper Noun. This Noun Vs Proper noun ratio could be a key identifier. There are also more Adjectives and Verbs in True than in Fake.

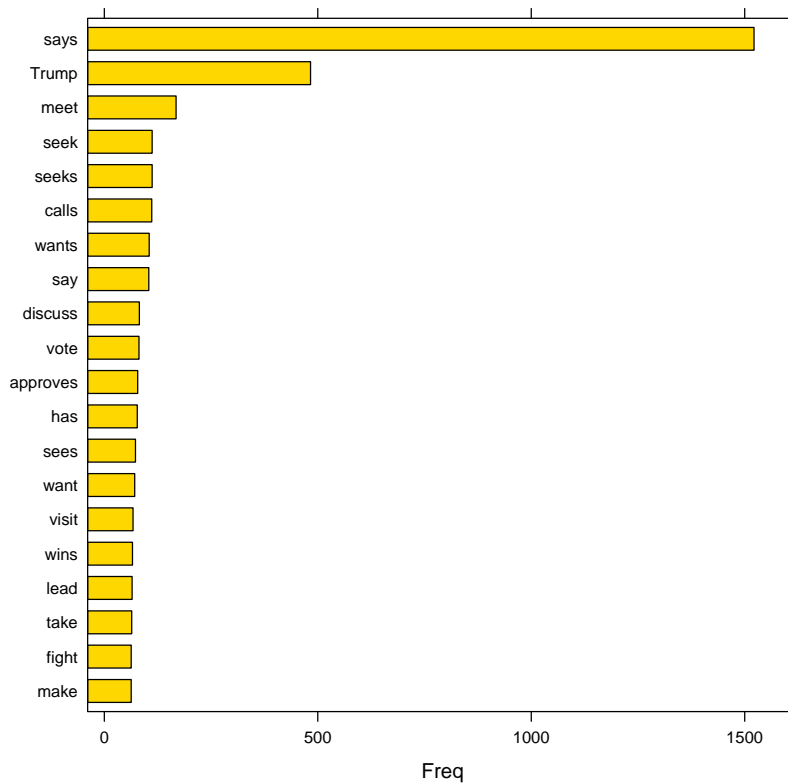


When we compare the most occurring nouns between the True Vs Fake, we see that except for 'Trump' and 'campaign' there is no relation between the datasets. Again, a potential classifier. Though 'Trump', 'Russian', 'Democratic' appear on both the sets, there are enough differences in the way adjectives are used to qualify as a classifier.

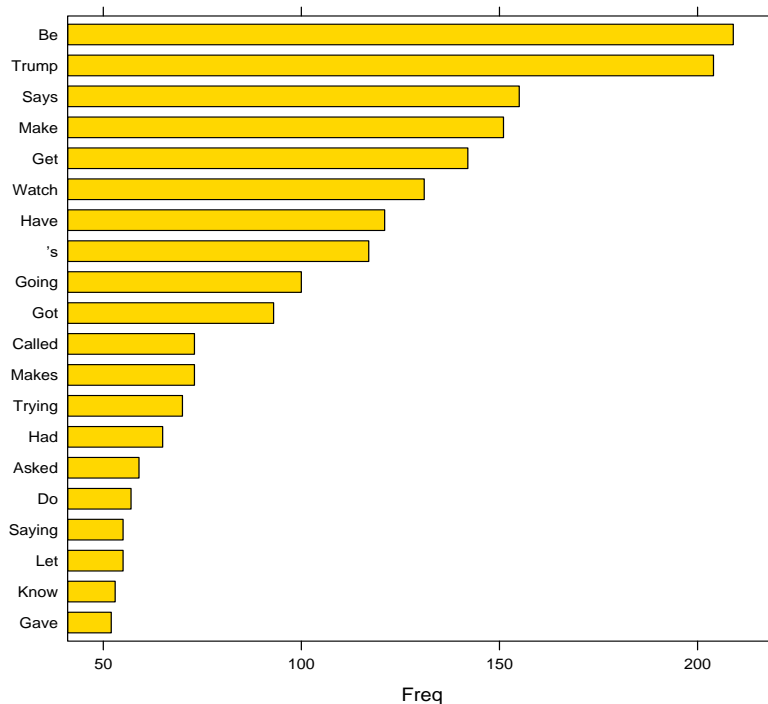


What about verbs? The usage of verbs is not very dissimilar between True Vs Fake, except for the fact, a lot of 'capitalized verbs' turn up in the Fake set.

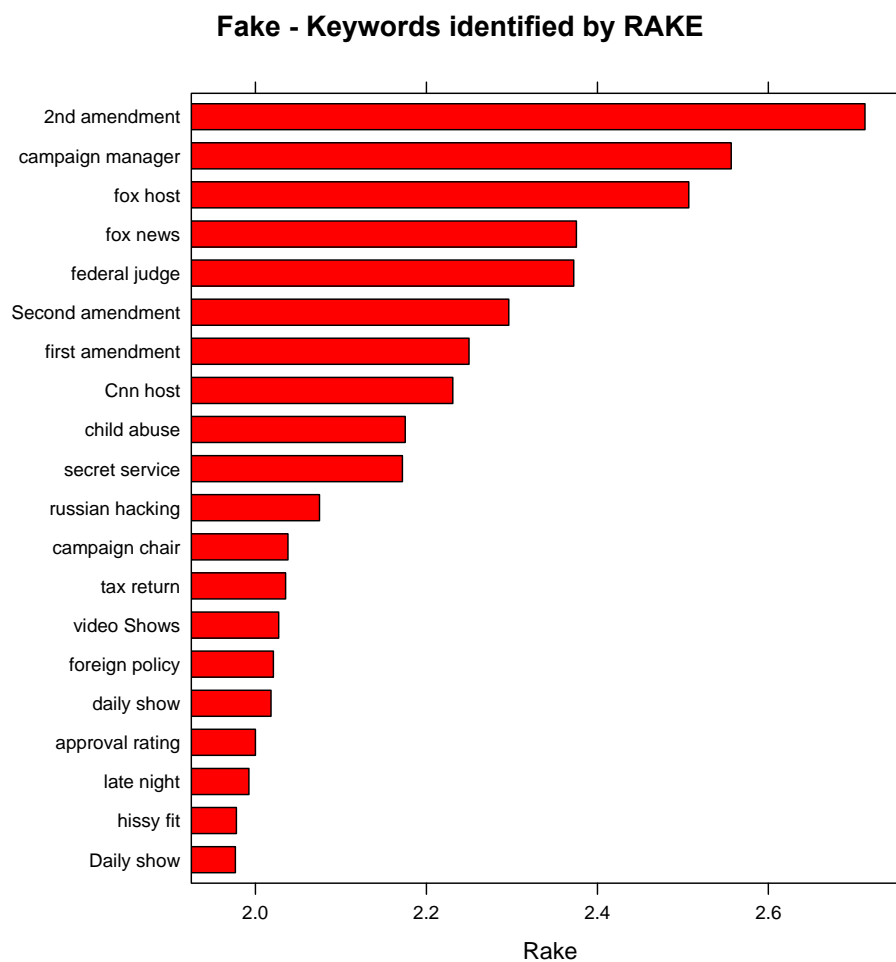
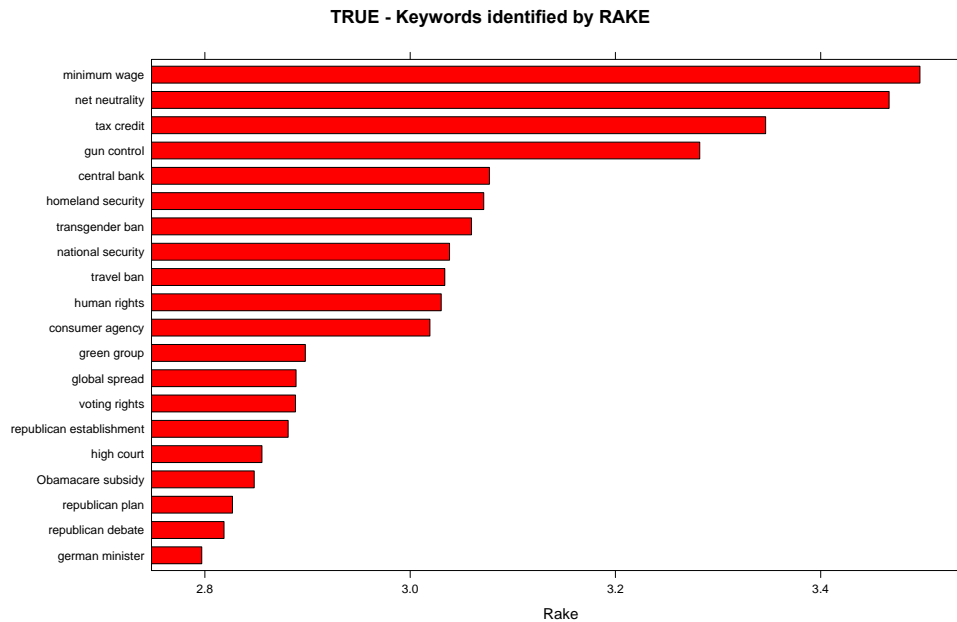
TRUE - Most occurring Verbs



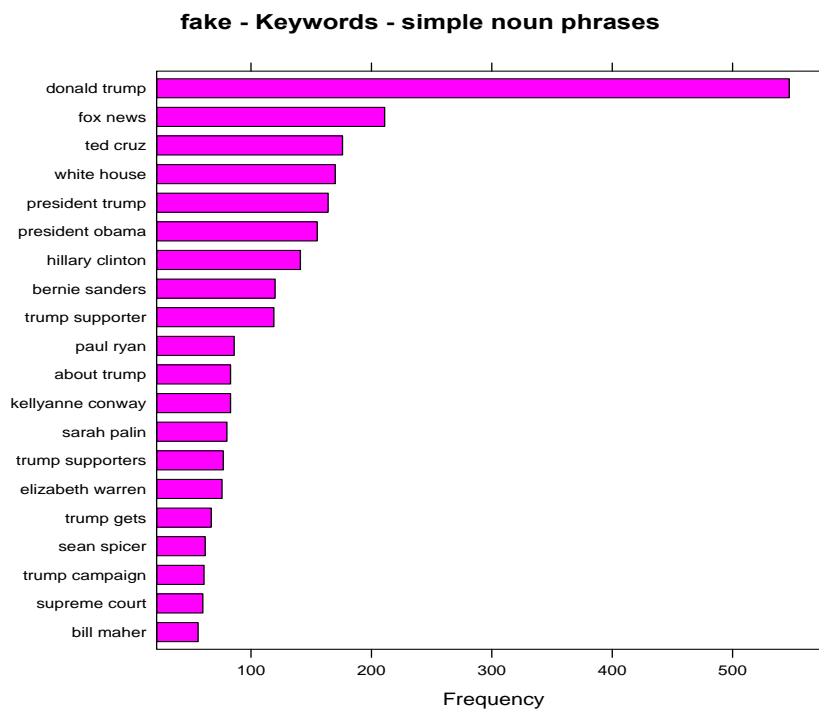
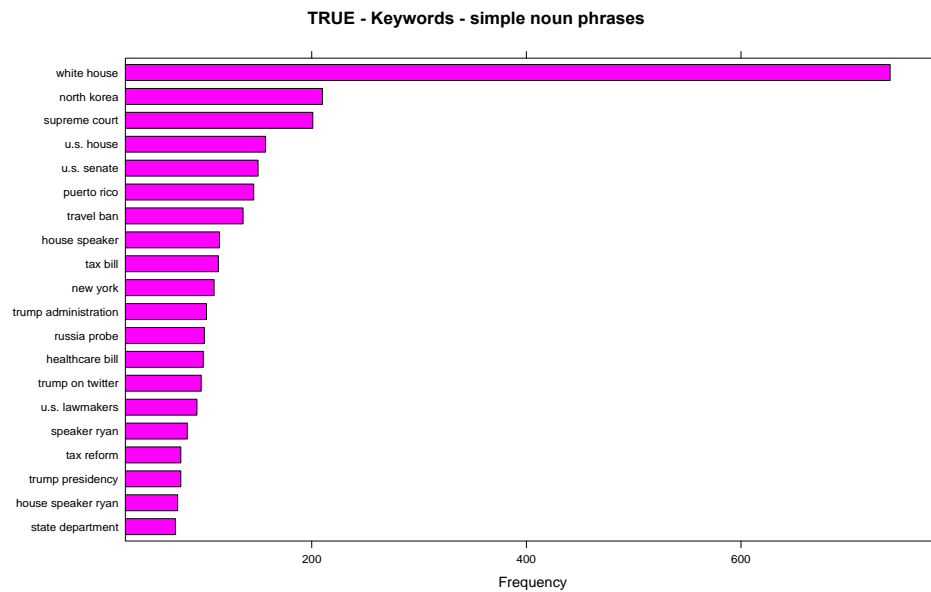
Fake - Most occurring Verbs



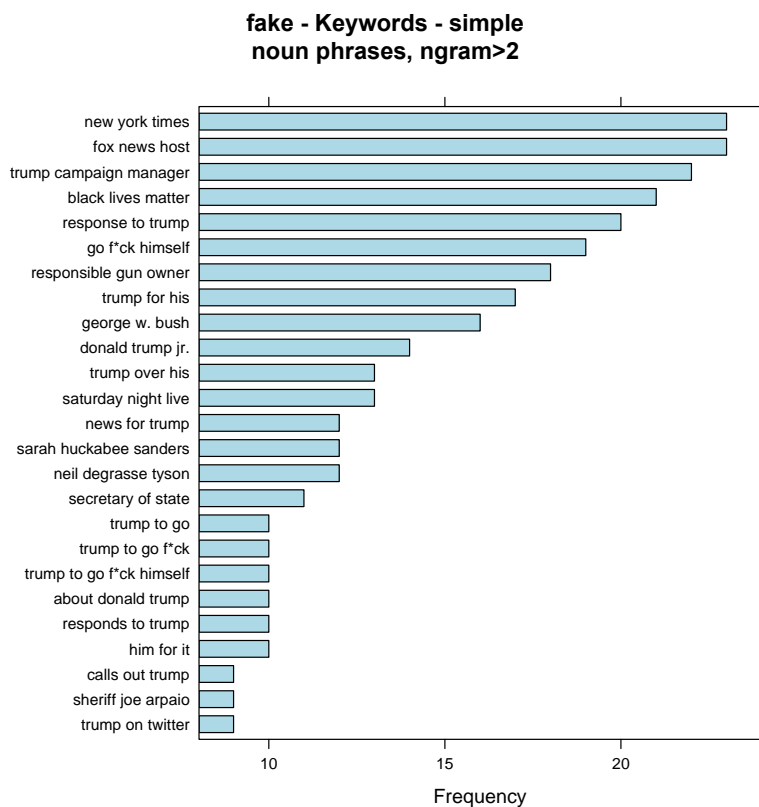
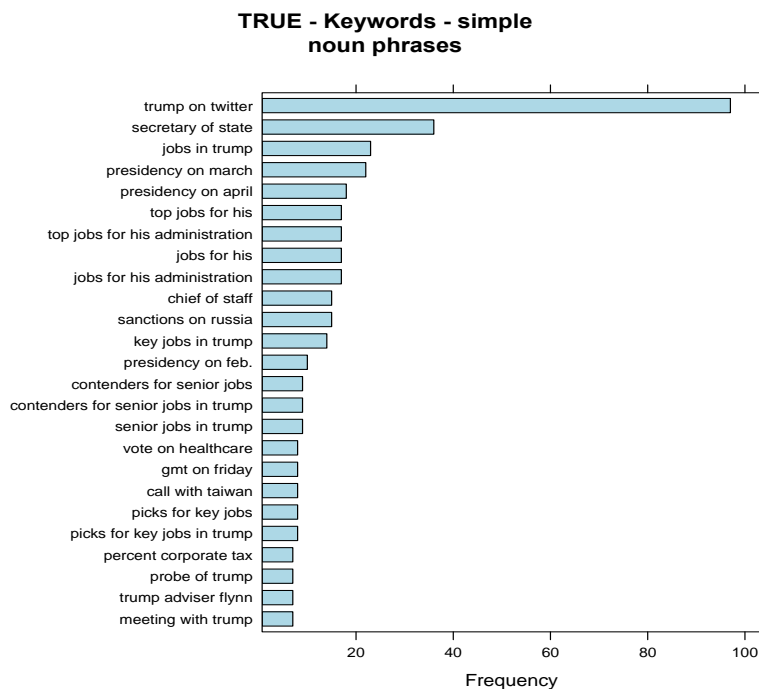
The scene changes dramatically when we start extracting the keywords using Rapid Automatic Keyword Extraction (RAKE) method. Suddenly the difference between the True and Fake comes out as chalk and cheese. None of the True keywords are there in the Fake set.



A lot of uncapitalized proper nouns tend to dominate the Fake set as Noun phrases

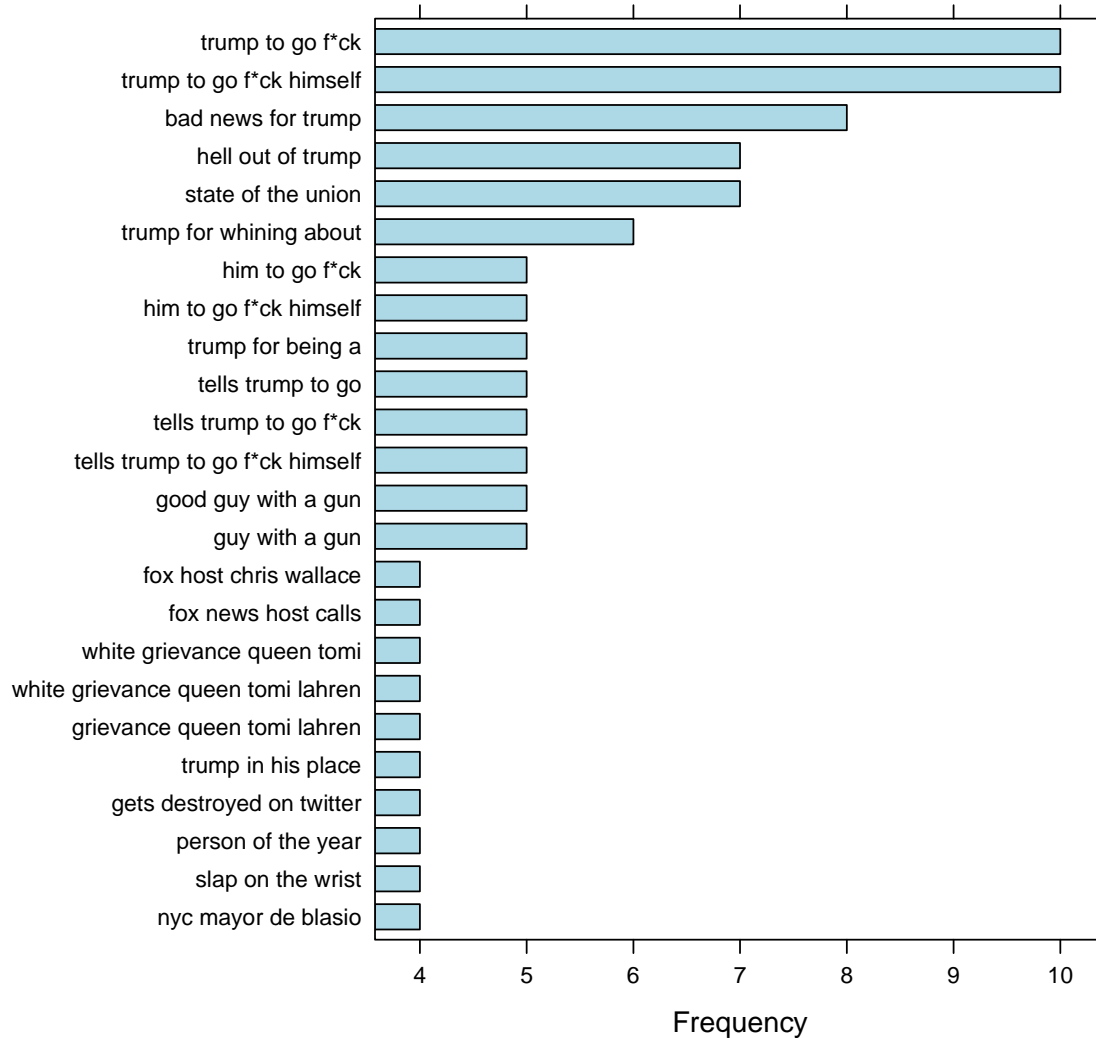


Noun phrases can be extracted by specifying the ngrams to be studied. If we increase the ngrams to more than 2, a whole new picture start to emerge.

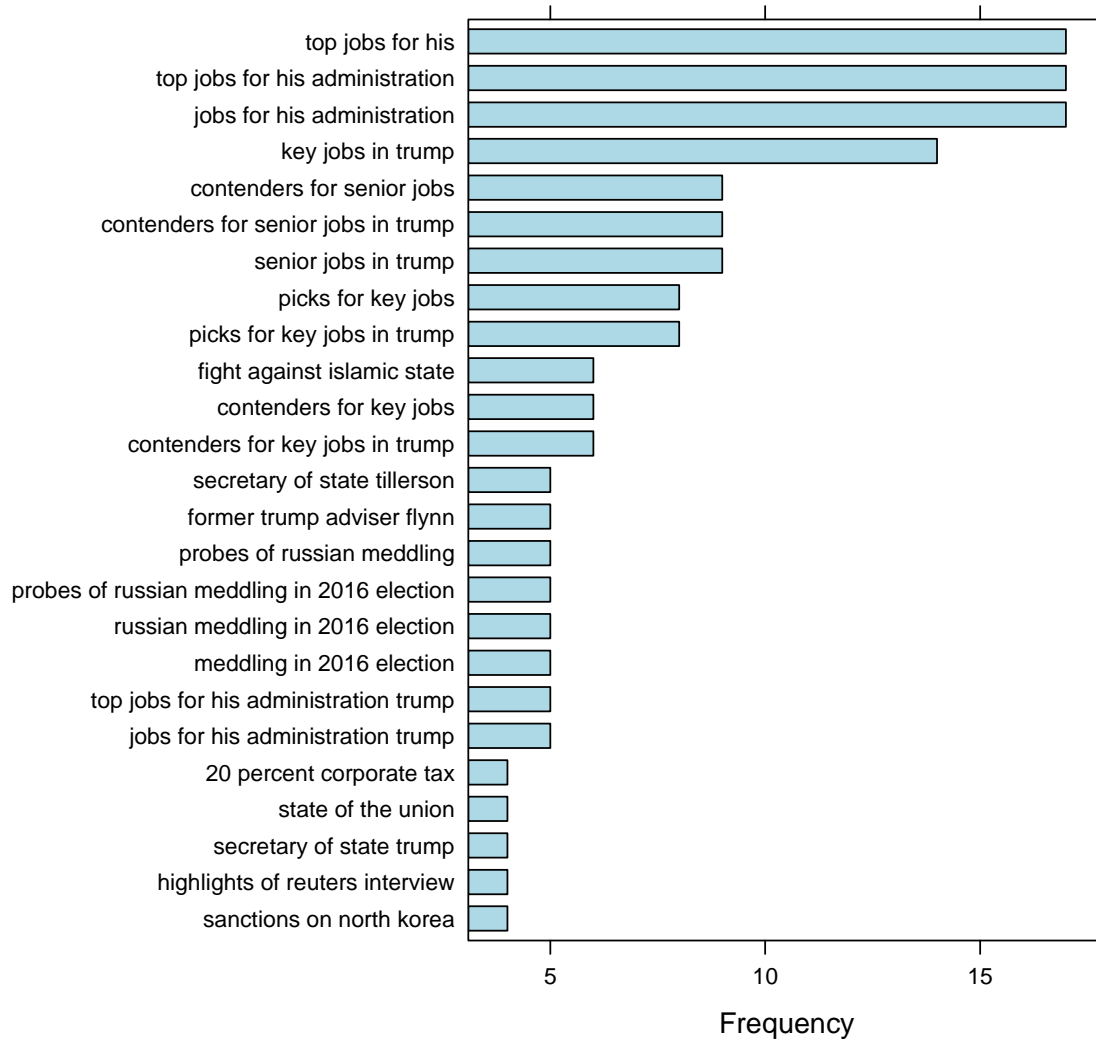


The true colour of the fake reviews show up when we dial up the ngrams...

fake - Keywords - simple noun phrases,ngrams>3



TRUE - Keywords - simple noun phrases



Co-occurrence of words is another indicator to tell the fact from fiction.

CO-OCCURRENCE -TRUE Vs FAKE

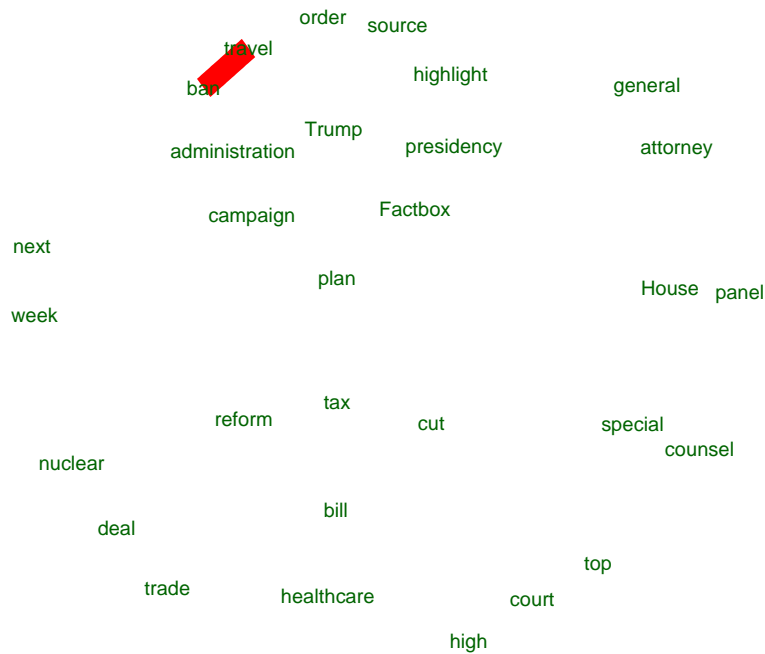
```
> head(stats)
```

	term1	term2	cooc		term1	term2	cooc
1	travel	ban	138	1	Trump	campaign	48
2	Factbox	Trump	120	2	campaign	manager	48
3	tax	bill	118	3	Trump	supporter	37
4	Trump administration	bill	102	4	watch	Trump	32
5	healthcare	bill	100	5	news	host	26
6	tax	reform	83	6	watch	fox	25

Lastly the clinching evidence is the WordNetwork map. The bolder the colour indicates the stronger the relationship. And the difference is self explanatory.

TRUE - Co-occurrences within 3 words distance

Nouns & Adjectives



fake -Co-occurrences within 3 words distance

Nouns & Adjectives



CONCLUSION

I started looking at two datasets, one true, other fake. By studying the Parts of Speech, Cooccurrence of words, noun phrases and ngrams – all of these prove to be efficient classifiers when it comes to separating the truth from lies. These classifiers can help us build a model, based on which a body of text, can be determined as Truth or manufactured with considerable certainty.