

Parallel Worlds

Francesco Prem Solidoro, Michele Salvi

Table of contents

1 Parallel Worlds - a View of Social Media Manipulation and Polarization	1
1.1 Abstract	1
1.2 Introduction	2
1.3 Research Question	3
1.3.a Downloading, Importing, merging and cleaning the dataset	3
1.4 Analysis	6
1.4.a Text analysis	6
1.4.b Bi-gram Analysis	8
1.4.c Sentiment Analysis	10
1.4.d LDA Modelling	13
1.5 Conclusion	20
1.6 Future Work	20
1.6.a Network Analysis	20
1.6.b New data	20
1.6.c LDA opportunities	21
Bibliography	21

1 Parallel Worlds - a View of Social Media Manipulation and Polarization

- Michele Salvi / michele.salvi4@studio.unibo.it / 0001082457
- Francesco Prem Solidoro / francesco.solidoro@studio.unibo.it / 0001071898

1.1 Abstract

What is truth? It can be argued that truth is but the set of facts we intersubjectively agree to be correct as a society. In this view, while voices of specialists hold power sway, it's ultimately public opinion that shapes the shared ground over which political discourse and policy are built. It immediately becomes apparent that choosing one's terrain is a powerful tool in achieving one's objectives. To that end, it becomes essential to:

- win over people to believe in your cause
- make sure people you won over don't leave for other causes.

In comes "Parallel Worlds".

There's no stronger attack against opposing ideas than die-hard fans, and no defense more solid than said die-hard fans *never being exposed to competing ideas*. The pursuit for achieving a parallel

world with a Truth that's centralised is an old one: think of Mussolini and Hitler's use of radio. The renewed threat coming from it is that we're coming off of a new media revolution: just like the printing press, the radio and the television before it, the internet has radically changed the way information flows through society. Analogously, with this revolution, new paradigms of communication have burst into the scene, alongside methods for manipulation of information. The Internet has the unique property of allowing anyone to have a voice, and therefore to represent in an uncanny way the dynamics of social conversation, where not only the loudest, but also the most numerous voices dominate. In this there's an inherent democratic principle of attention being divided equally through society. In this there's also an inherent danger: with bots and technological advancement, it's become easy to artificially boost ideas to favour an agenda through a faux legitimisation from a virtual majority. With the objective of enclosing people who believe they're engaged in an open network into a closed down bubble of ideas, thus spawns the Astroturf. A successful astroturf will appear to be grassroots support for an idea, while driving polarisation to a point where those in support and those in opposition to the target idea will effectively live in different, parallel worlds. With all contact being auspiciously hostile: demonisation of the other side, deification of one's one, discreditation of third party "experts". These are all hallmarks of astroturf campaigns, and they're more and more visible within modern society, as this camouflaged threat rips the promise of the open internet into bubbles floating in the sea of its former potential.

1.2 Introduction

The topic of manipulation has long been studied by scholars in its different facets, and in this paper we'll explore the connection between the political sphere and manipulation when used as an instrument to achieve control.

In the last years social media usage has skyrocketed (Esteban Ortiz-Ospina, n.d.), thus more advanced methods of manipulating public opinion, both in and outside the political sphere, have arisen due to the innovations and reach that social media platforms boast. Historically, the preferred outcome of manipulation has always been polarization, giving rise to strong leaders and divides (Jacob et al., n.d.).

The Merriam-Webster dictionary defines polarization (*Merriam-Webster*, n.d.) as "a state in which the opinions, beliefs, or interests of a group or society no longer range along a continuum but become concentrated at opposing extremes", it's a key factor to view polarization as an outcome rather than a means, since it's always been the objective of different political parties to achieve their goals. Now, in this era of limitless communication, one would see complete polarization as harder to achieve, due to the sheer amount of different opinions available on social media, but there's a new, and arguably the worst ever, means of achieving a polarized society: astroturfing.

Astroturfing is defined by the Merriam-Webster dictionary (*Merriam-Webster*, n.d.) as "organized activity that is intended to create a false impression of a widespread, spontaneously arising, grassroots movement in support of or in opposition to something but that is in reality initiated and controlled by a concealed group or organization". Such a practice is deeply deceptive and thrives on platforms like social media, where identity can be disguised and allegiance faked, to distort and manipulate public opinion.

Astroturfing can be declined into two further classifications, *commercial astroturfing*, centered around the accumulation of profit, and *political astroturfing*, concerned with the attainment of political objectives. Research on political astroturfing has become increasingly present in the last years, and it mainly focuses on analyzing its effects on politics, highlighting the power that this practice represents by exposing its presence in (but not limited to) South Korea's 2012 Presidential Elections (Franziska B. Keller et al., n.d.), the USA's 2016 elections (Ahmed & Anis Rahman, n.d.) and faking Chinese Social media posts (Gary et al., n.d.). These papers highlight the real-world effects and the seriousness of groups and/or individuals rich in economic and social capital, who seek to manipulate public opinion to achieve private objectives, especially in politics, putting them in commanding positions.

The idea behind the "parallel worlds" is that polarization through astroturfing acts silently and unbeknownst to the population, creating fissures and fragmentation in public opinion that effectively alienate and polarize society. The preferred medium for this practice is, as previously stated, through social media, so we focused on the biggest and most documented instance of astroturfing, the 2016 USA elections.

1.3 Research Question

In our pursuit of understanding how parallel worlds are built, we have to understand two things:

- what makes an astroturf?
- what effect does it actually have?
- is it possible to generalize and model an astroturfing from the words used?

To understand this, we will be conducting a sentiment analysis over the datasets of IRA-affiliated tweets utilizing text mining practices such as word frequencies, tf-idf, sentiment analysis through the NRC (*NRC Lexicon*, n.d.), Bing (*Mining and Summarizing Customer Reviews*, n.d.) and AFINN (*AFINN Lexicon*, n.d.) lexicons, and the aid of LDA, latent dirichlet allocation, to model the topics discussed in the tweets.

1.3.a Downloading, Importing, merging and cleaning the dataset

We got the following datasets from 538's github, which were merged into a single dataset.

Since the analysis is quite taxing, we're going to sample 100.000 tweets to perform a preliminary analysis on. In order to conduct the analysis on the full dataset, remove the following block.

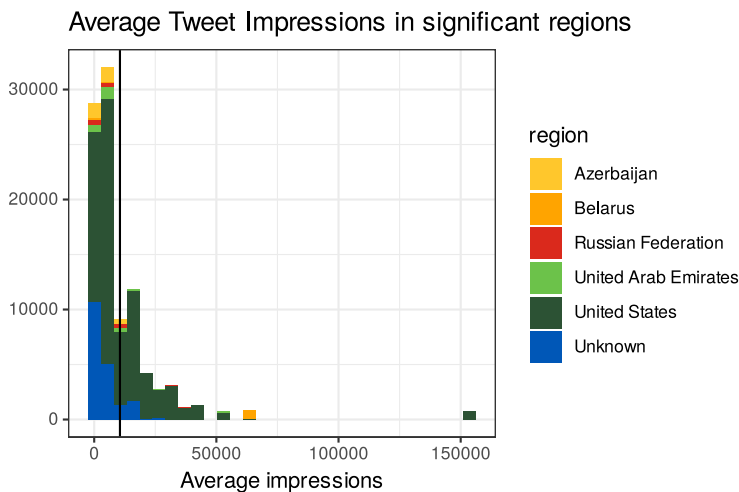
The cleaned and merged dataset contains the following variables:

- **author**: the author of the tweet, categorical
- **content**: the content (text) of the tweet, string
- **region**: the region in which the tweet "originated", if not specified it's marked as "Unknown", categorical
- **language**: the language in which the tweet was written, categorical
- **publish_date**: the date in which the tweet was published, date
- **following**: the number of accounts followed by the tweet author, numerical
- **followers**: the number of accounts that follow the tweet author, numerical

- **updates**: the number of impressions (likes, comments and retweets) summed, numerical
- **post_type**: the type of post (retweet / quote retweet / NA for normal tweets), categorical
- **account_type**: the type of account, previously set, by tweet content, categorical
- **retweet**: weather the tweet was a retweet or not, binary
- **account_category**: the accounts grouped by categories based on tweet content, categorical

To describe the dataset, we plotted a few graphs (some of which are attached, but not essential), showing the proportions of tweet by languages, by account types and by regions (not included in the pdf, code block above)

To represent tweets in proportion to their engagement, so effect on the population, we then plotted tweets by average impressions, according to the regions, highlighting the discrepancies, and different reception of the astroturfing in different regions.



As it can be observed in the graph, despite there being IRA-connected accounts in multiple regions, most of the interactions and impressions come from tweets from the united states.

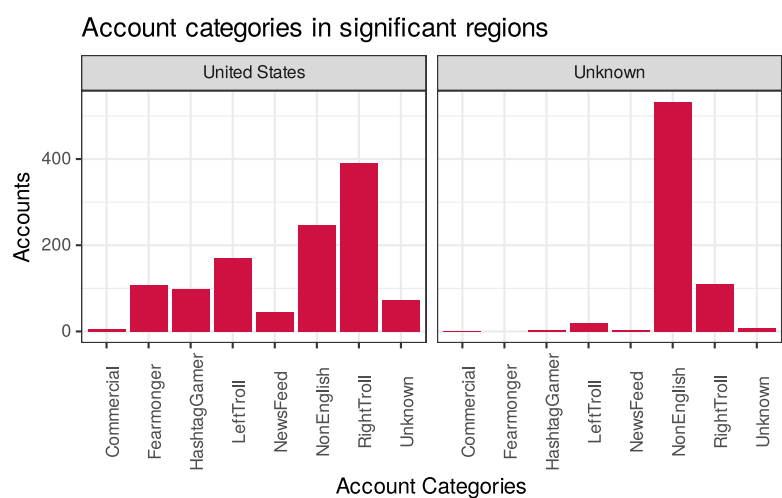
To represent well the most significant regions (United states, Unknown) we categorized the IRA-connected accounts by their type, so the macro-topic they covered in their tweets, along with the retweet rate and descriptive statistics on the follows and followers.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	
X1	1	1e+05	7136.01	14879.11	1313	4089.26	1761.33	0	251245	251245	6.01	
	kurtosis	se										
X1	59.23	47.05										

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	
X1	1	1e+05	3456.67	5564.75	1528	2365.65	2043.02	0	76206	76206	5.12	
	kurtosis	se										
X1	46.52	17.6										

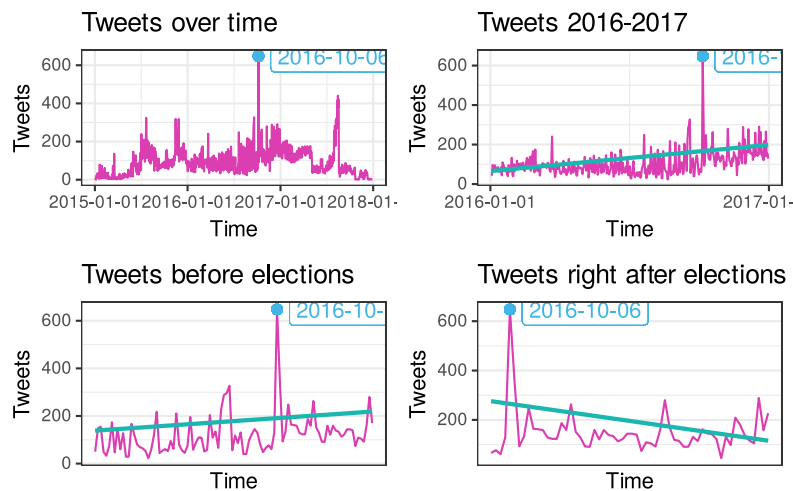
```
[1] 0.44007
```

```
# A tibble: 8 × 2
  account_category retweet_rate
  <chr>            <dbl>
1 LeftTroll        0.820
2 HashtagGamer     0.715
3 NonEnglish       0.548
4 RightTroll       0.437
5 Unknown          0.191
6 Fearmonger       0.148
7 Commercial       0.0638
8 NewsFeed         0.00188
```



Most of the accounts in the tweets can be traced back to either the US or the “Unknown” region, but since most accounts in this last region are “NonEnglish”, we’ll discard those and focus on those composed in English for most of our analysis.

Additionally, since the times at which the tweets were posted are included in the dataset, we plot them, progressively restricting them on the date of the presidential election, to have a look at the patterns.



As it can be observed, tweets steadily increased until a few days after the election, when they started dropping.

1.4 Analysis

1.4.a Text analysis

We began our analysis by focusing on the tweet's text. In this type of "word" centered approach we filtered the tweets to contain only English tweets, removing links and "stop words" (words such as the, a, by...) that would just hinder the process.

We then tokenized the tweets by words, by splitting each word in the content while homogenizing the text.

```
# A tibble: 75,862 × 2
  word      n
  <chr>    <int>
1 news    5176
2 trump   4840
3 police  1875
4 sports  1771
5 people  1746
6 world   1504
7 politics 1500
8 obama   1477
9 workout 1473
10 amp    1366
# i 75,852 more rows
```

```
# A tibble: 75,862 × 7
  word      NewsFeed RightTroll Commercial LeftTroll HashtagGamer Fearmonger
  <chr>    <int>    <int>    <int>    <int>    <int>    <int>
1 news    5176    4840    1875    1771    1746    1504
2 trump   4840    1500    1477    1473    1366    1366
3 police  1875    1366    1366    1366    1366    1366
4 sports  1771    1366    1366    1366    1366    1366
5 people  1746    1366    1366    1366    1366    1366
6 world   1504    1366    1366    1366    1366    1366
7 politics 1500    1366    1366    1366    1366    1366
8 obama   1477    1366    1366    1366    1366    1366
9 workout 1473    1366    1366    1366    1366    1366
10 amp    1366    1366    1366    1366    1366    1366
```

```

1 news      4276      712      2      147      35      4
2 trump     645      3317     1      705     166      6
3 sports    1642      87      3      31      7      1
4 workout   NA        1     1465      4      3     NA
5 politics  1293      146      1      41     19     NA
6 police    1173      310     NA     370     15      7
7 obama     231     1075     NA     141     29      1
8 world     1017      275      7     147     55      3
9 breaking  224     1001      1      65     10     NA
10 hillary   64      931      1      72     52      3
# i 75,852 more rows

```

```

# A tibble: 6 × 5
  account_category word      n tot percent_freq
  <chr>            <chr> <int> <int>      <dbl>
1 NewsFeed        news  4276 145246      2.94
2 RightTroll      trump  3317 180769      1.83
3 NewsFeed        sports  1642 145246      1.13
4 Commercial      workout 1465  35869      4.08
5 NewsFeed        politics 1293 145246      0.890
6 NewsFeed        police  1173 145246      0.808

```

The most common words in the tweets are “news”, “trump” and “police”, by grouping the words by account we can see that those words are fairly common also by controlling for the different account types, further confirmed by the percentage frequencies of the same words.

Then we analyzed the word’s tf-idf, the term frequency-inverse document frequency, which is a measure that assigns lower weight to words that are frequently used and a higher one to the uncommon ones.

The idea is to find a middle point between the two, as some words may be important to describe a text, while not being the most commonly used.

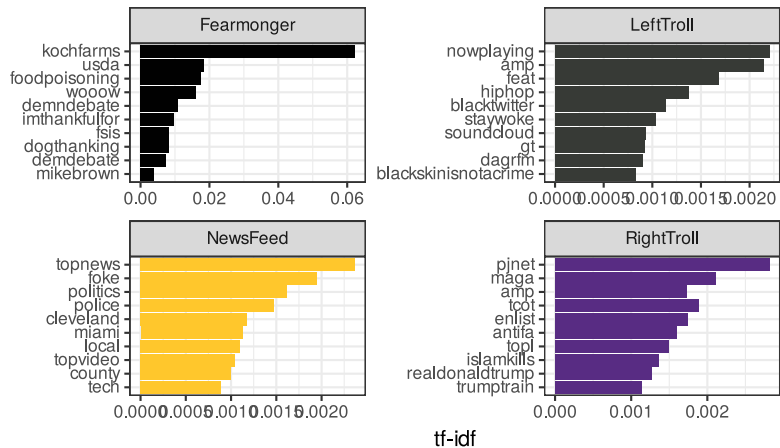
```

# A tibble: 74,759 × 7
  account_category word      n percent_freq      tf      idf      tf_idf
  <chr>            <chr> <int>      <dbl> <dbl> <dbl> <dbl>
1 Fearmonger      kochfarms  127      5.66  0.0566  1.10  0.0622
2 Fearmonger      usda      60      2.67  0.0267  0.693 0.0185
3 Fearmonger      foodpoisoning  36      1.60  0.0160  1.10  0.0176
4 Fearmonger      woow      33      1.47  0.0147  1.10  0.0162
5 Fearmonger      demndebate  35      1.56  0.0156  0.693 0.0108
6 Fearmonger      imthankfulfor  20      0.892 0.00892  1.10  0.00980
7 Fearmonger      dogthanking  17      0.758 0.00758  1.10  0.00833
8 Fearmonger      fsis      17      0.758 0.00758  1.10  0.00833
9 Fearmonger      demdebate  24      1.07  0.0107  0.693 0.00742

```

```
10 Fearmonger      mikebrown      8      0.357 0.00357 1.10  0.00392
# i 74,749 more rows
```

High tf-idf words by account groups



The high tf-idf words, plot by account category, highlight common topics to the 2016 elections and other high impact words (some of which were previously hashtags, but were included to not forego any information) like “islam kills”, “black skin is not a crime”, “food poisoning”, and “stay woke”.

1.4.b Bi-gram Analysis

Tokenization isn’t only possible at the singular words level, we can tokenize according to “n-grams”, which are series of n words.

We decided to focus on Bigrams, series of 2 words, and repeat some of the previous text analysis.

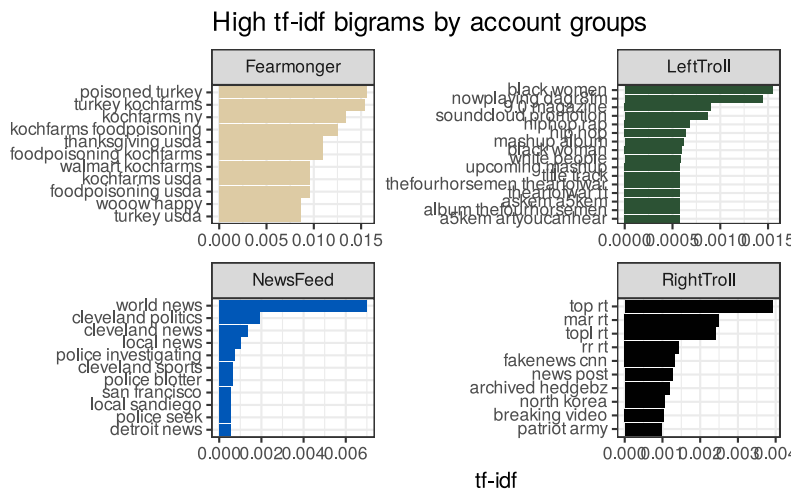
```
# A tibble: 215,978 × 3
  word1 word2 n
  <chr> <chr> <int>
1 world news 928
2 donald trump 462
3 white house 321
4 hillary clinton 270
5 lose weight 235
6 president trump 218
7 top rt 217
8 north korea 197
9 fake news 147
10 local news 143
# i 215,968 more rows
```

This gives us additional context on the most common words, like “world news” and “donald trump”

We can also repeat the `tf_idf` analysis with bigrams to capture additional context that might have been discarded by single-word analysis.

```
# A tibble: 226,201 × 6
  account_category bigram          n      tf    idf  tf_idf
  <chr>            <chr>      <int>  <dbl> <dbl>  <dbl>
1 Fearmonger      poisoned turkey    10 0.00873 1.79 0.0156
2 Fearmonger      turkey kochfarms   16 0.0140  1.10 0.0153
3 Fearmonger      kochfarms ny       14 0.0122  1.10 0.0134
4 Fearmonger      kochfarms foodpoisoning 8 0.00698 1.79 0.0125
5 Fearmonger      foodpoisoning kochfarms 7 0.00611 1.79 0.0109
6 Fearmonger      thanksgiving usda    7 0.00611 1.79 0.0109
7 Fearmonger      foodpoisoning usda   10 0.00873 1.10 0.00959
8 Fearmonger      kochfarms usda      10 0.00873 1.10 0.00959
9 Fearmonger      walmart kochfarms   10 0.00873 1.10 0.00959
10 Fearmonger     turkey usda         9 0.00785 1.10 0.00863
# i 226,191 more rows
```

Again, bigrams like “poisoned turkey”, “isis accounts”, reveal patterns that weren’t captured by the previous `tf-idf` plot

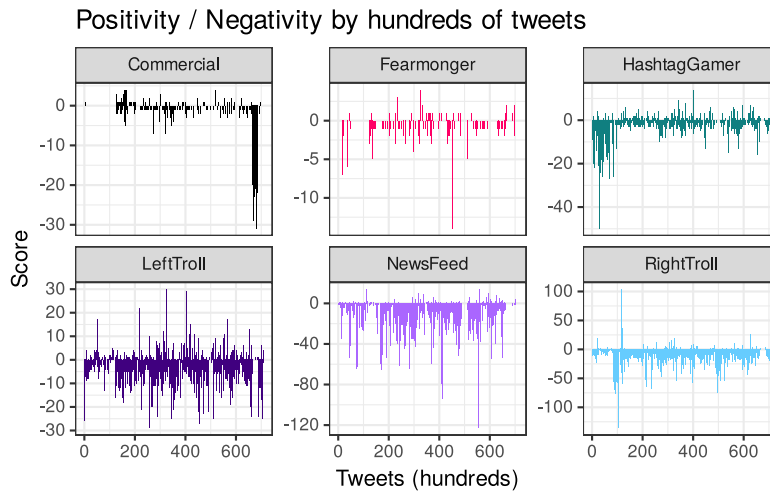
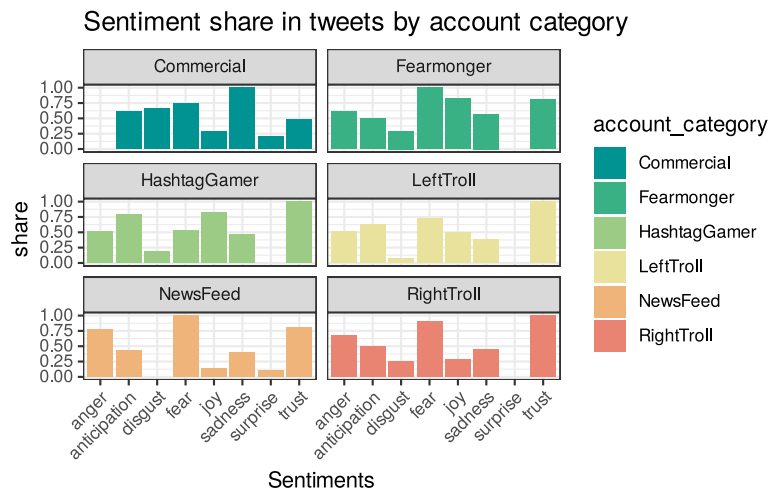
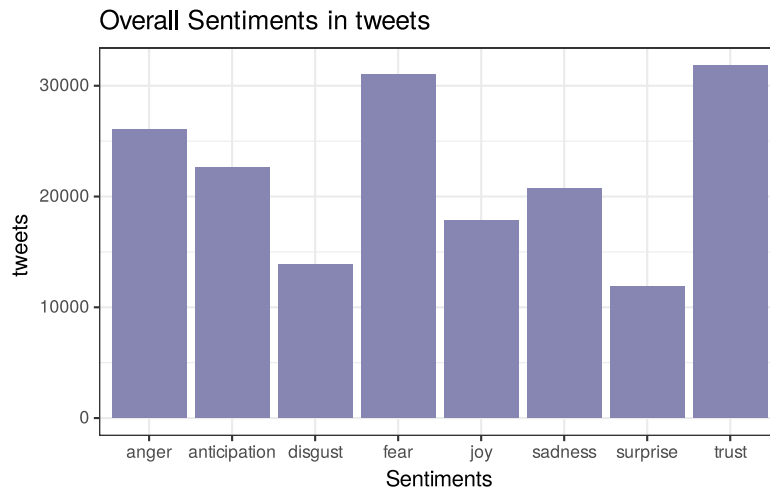


Additionally, the bigram combinations can be plot in a network graph:

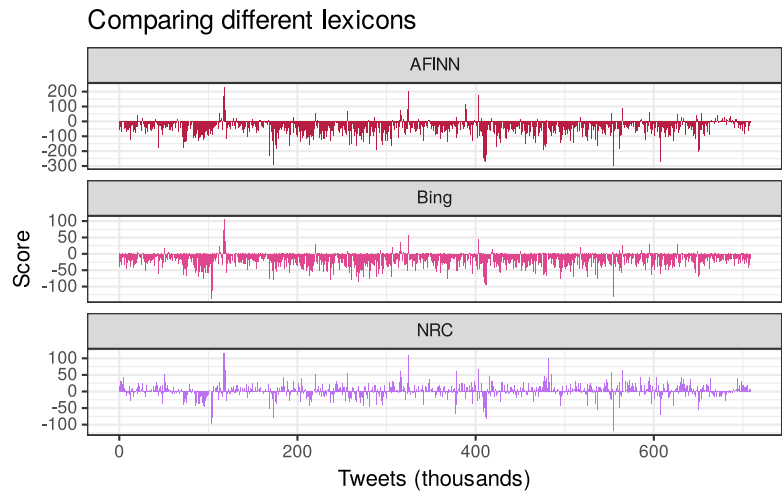
1.4.c Sentiment Analysis

This block filters out the word “trump” (to avoid misinterpretation in sentiment analysis) and joins the dataset with the NRC sentiment lexicon to classify words into sentiment categories. The result is a dataset of words with assigned sentiments.

```
# i 26,755 more rows
```

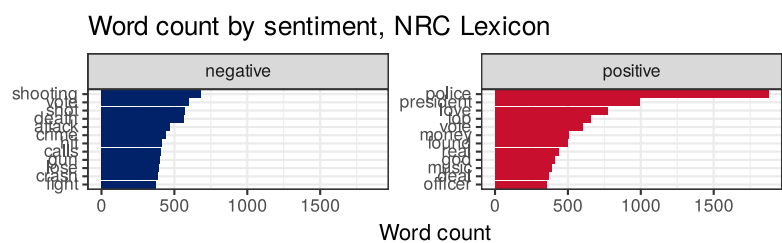
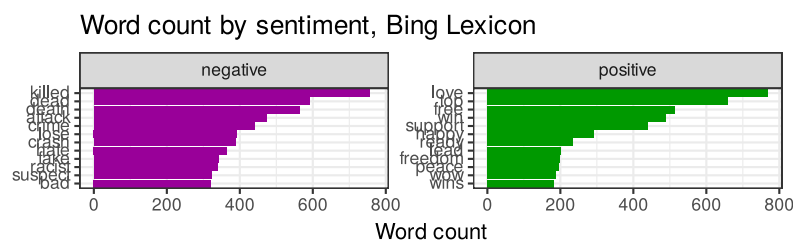


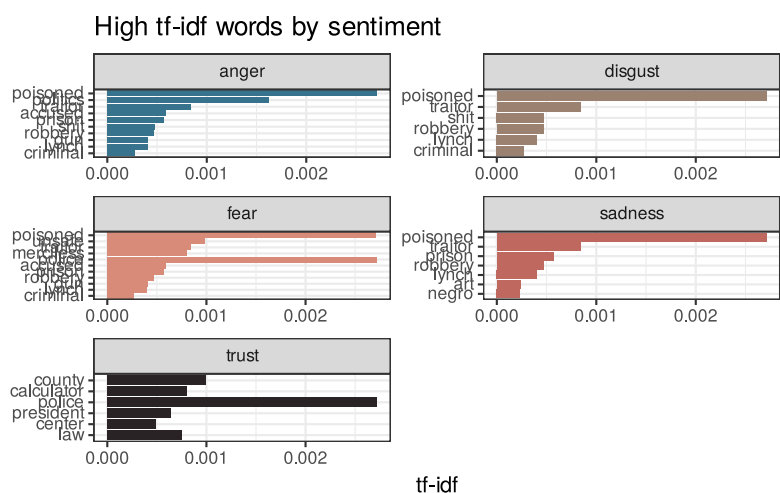
Now we compare the 3 lexicons.



```
# A tibble: 2 × 2
  sentiment      n
  <chr>      <int>
1 negative   3316
2 positive   2308
```

```
# A tibble: 2 × 2
  sentiment      n
  <chr>      <int>
1 negative   4781
2 positive   2005
```





We can observe that NRC is more positive than the other lexicons, even after correcting the issue with Trump being interpreted as a word of “surprise”.

1.4.d LDA Modelling

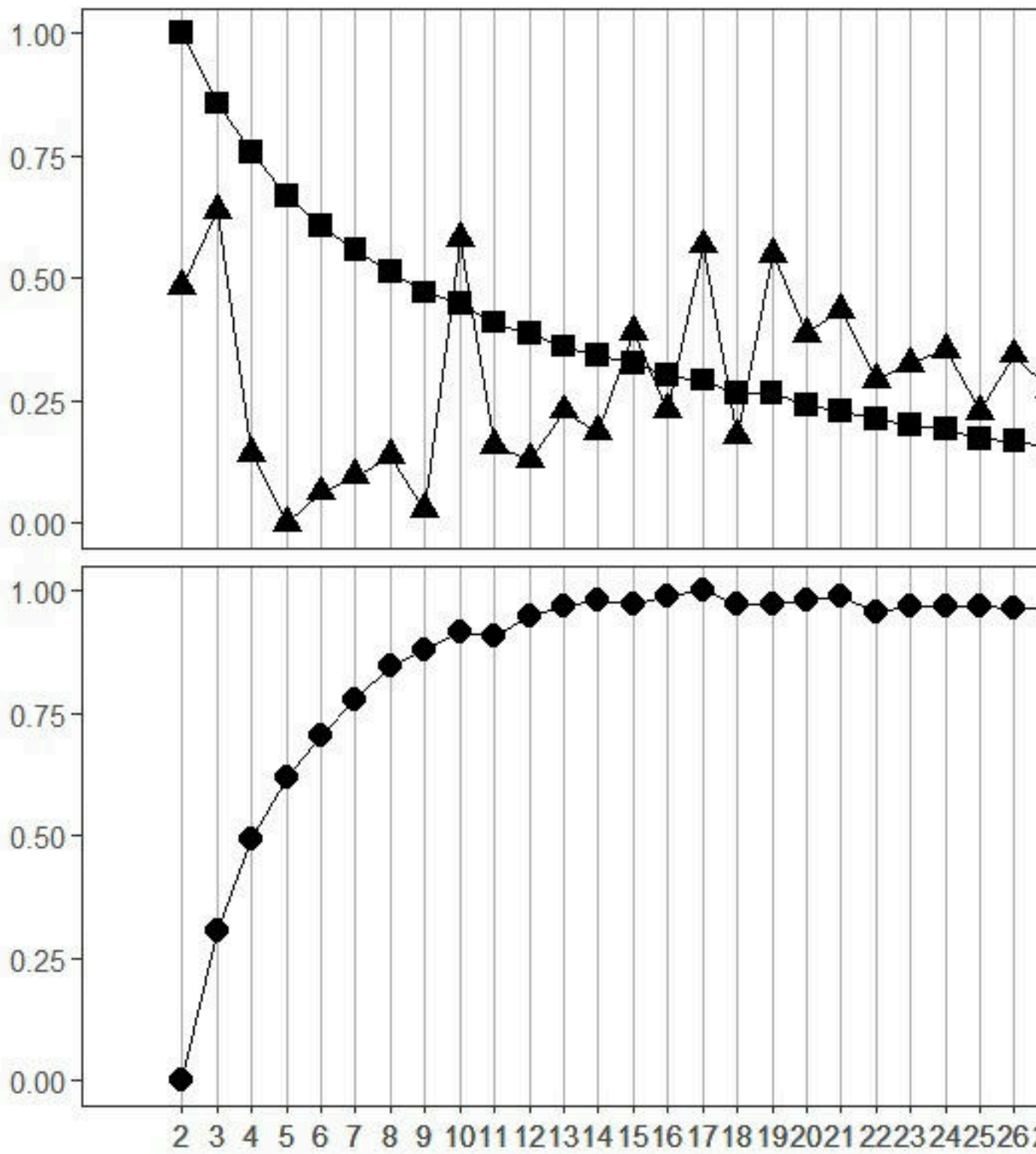
We proceed to use LDA analysis (David M. Blei et al., n.d.) to create a model that splits the tweets we found into topic. We take each tweet as a document and its text as a corpus. The tweets are then aggregated through LDA into topics, which returns us a model for how likely each word is to come out each topic.

We proceed to use LDA analysis to create a model that splits the tweets we found into topic. We take each tweet as a document and its text as a corpus. The tweets are then aggregated through LDA into topics, which returns us a model for how likely each word is to come out each topic.

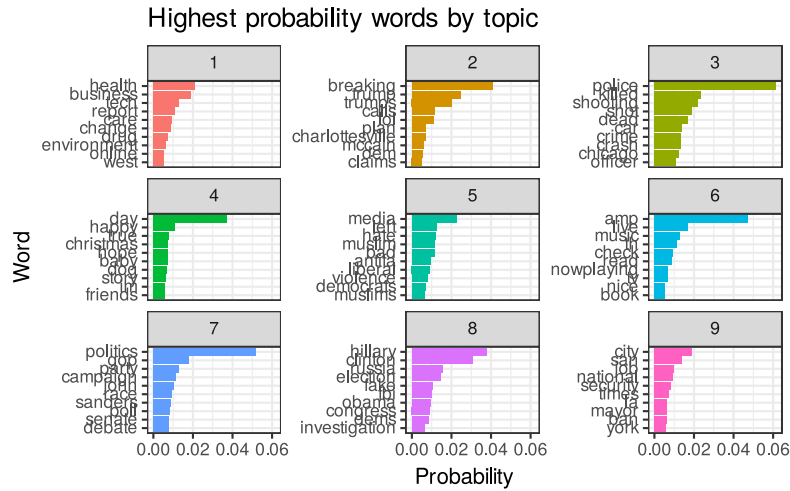
First we have to train the model, so we’re going to divide the tweets into training and testing data with an 80/20 split.

We then run the model for different numbers of topic, and evaluate its accuracy (which is measured by the model’s perplexity). We toggle this codeblock’s execution off because it takes a long time to run, but the model is most accurate with 25 topics (as shown by the plot).

It’s worth noting that perplexity remains high, despite the singular iteration, but that’s to be expected: even in human interpretation it’s recognised that Tweets, by their very nature, are often ambiguous. Astroturf-related tweets, which by their own objective try to be even more ambiguous, are bound to share that quality.



That being said, we can clearly observe patterns of topic forming within the model



```
# A tibble: 248 × 26
  term    topic1 topic2 topic3 topic4 topic5 topic6 topic7 topic8 topic9 topic10
<chr>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 west    0.00538      NA      NA      NA      NA      NA      NA      NA      NA      NA
2 onli... 0.00542      NA      NA      NA      NA      NA      NA      NA      NA      NA
3 envi... 0.00618      NA      NA      NA      NA      NA      NA      NA      NA      NA
4 drug    0.00752      NA      NA      NA      NA      NA      NA      NA      NA      NA
5 chan... 0.00896      NA      NA      NA      NA      NA      NA      NA      NA      NA
6 care    0.00936      NA      NA      NA      NA      NA      NA      NA      NA      NA
7 repo... 0.0111       NA      NA      NA      NA      NA      NA      NA      NA      NA
8 tech    0.0127       NA      NA      NA      NA      NA      NA      NA      NA      NA
9 busi... 0.0190       NA      NA      NA      NA      NA      NA      NA      NA      NA
10 heal... 0.0212       NA      NA      NA      NA      NA      NA      NA      NA      NA
# i 238 more rows
# i 15 more variables: topic11 <dbl>, topic12 <dbl>, topic13 <dbl>,
#   topic14 <dbl>, topic15 <dbl>, topic16 <dbl>, topic17 <dbl>, topic18 <dbl>,
#   topic19 <dbl>, topic20 <dbl>, topic21 <dbl>, topic22 <dbl>, topic23 <dbl>,
#   topic24 <dbl>, topic25 <dbl>
```

```
# A tibble: 1,756,575 × 3
  document topic  gamma
<chr>      <int> <dbl>
1 1          1 0.0370
2 2          1 0.0392
3 3          1 0.0385
4 4          1 0.0377
5 5          1 0.0392
6 6          1 0.0357
7 7          1 0.0741
```

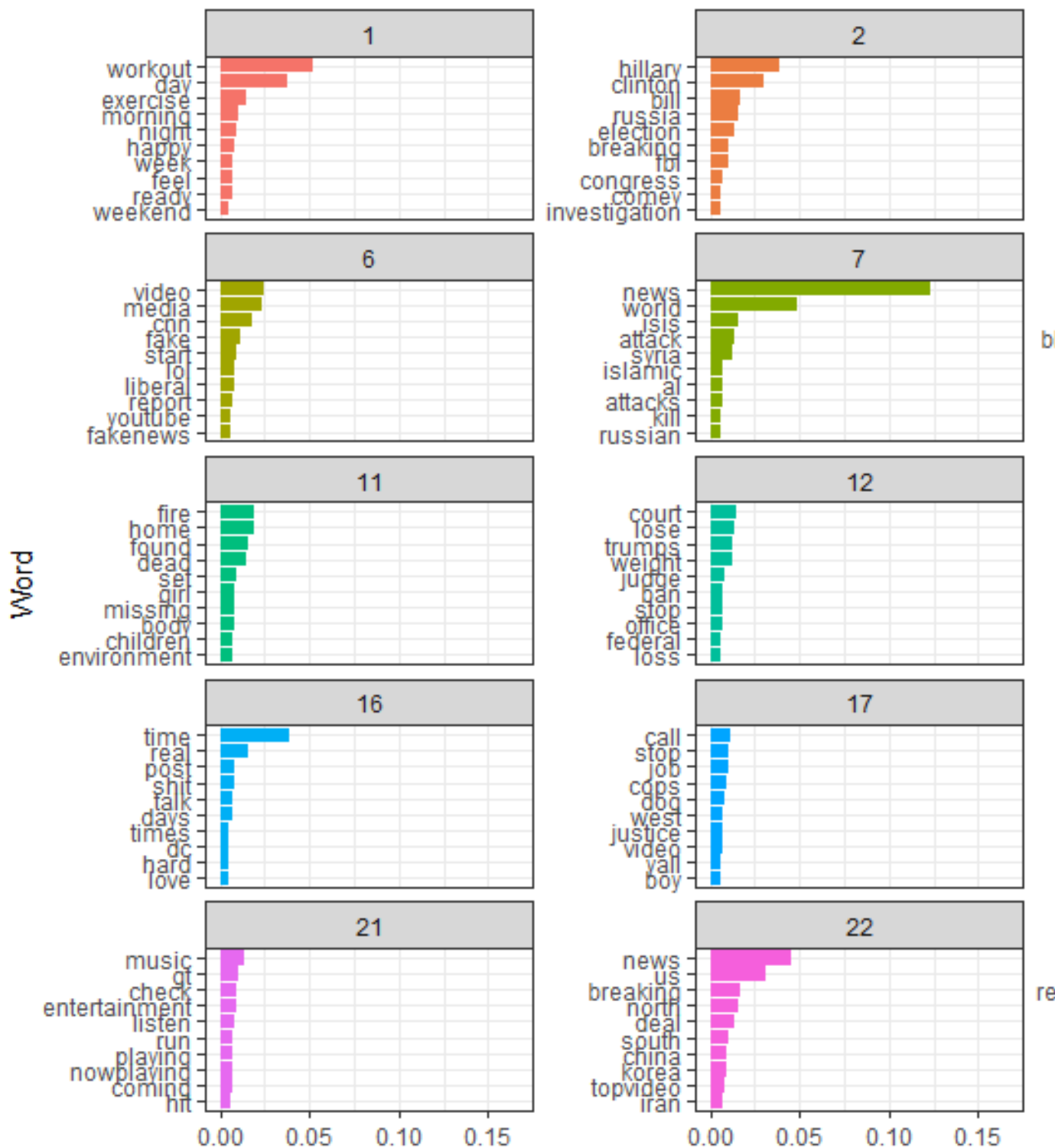
```
8 8          1 0.0345
9 9          1 0.0714
10 10        1 0.0385
# i 1,756,565 more rows
```

```
[1] 9470.8
```

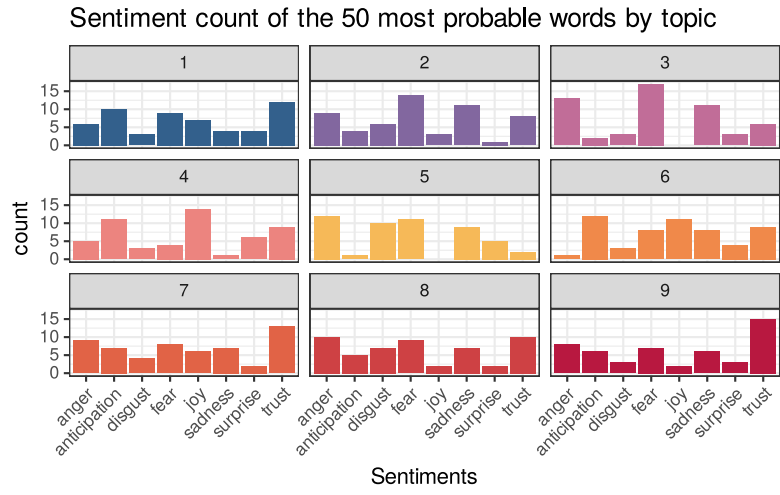
```
[1] 41050.34
```

We can observe several clear topic being distinguished, some of which along political lines, like the Obama presidency, Hillary Clinton and Donald Trump. Some of which along cultural topics, like workouts or sports.

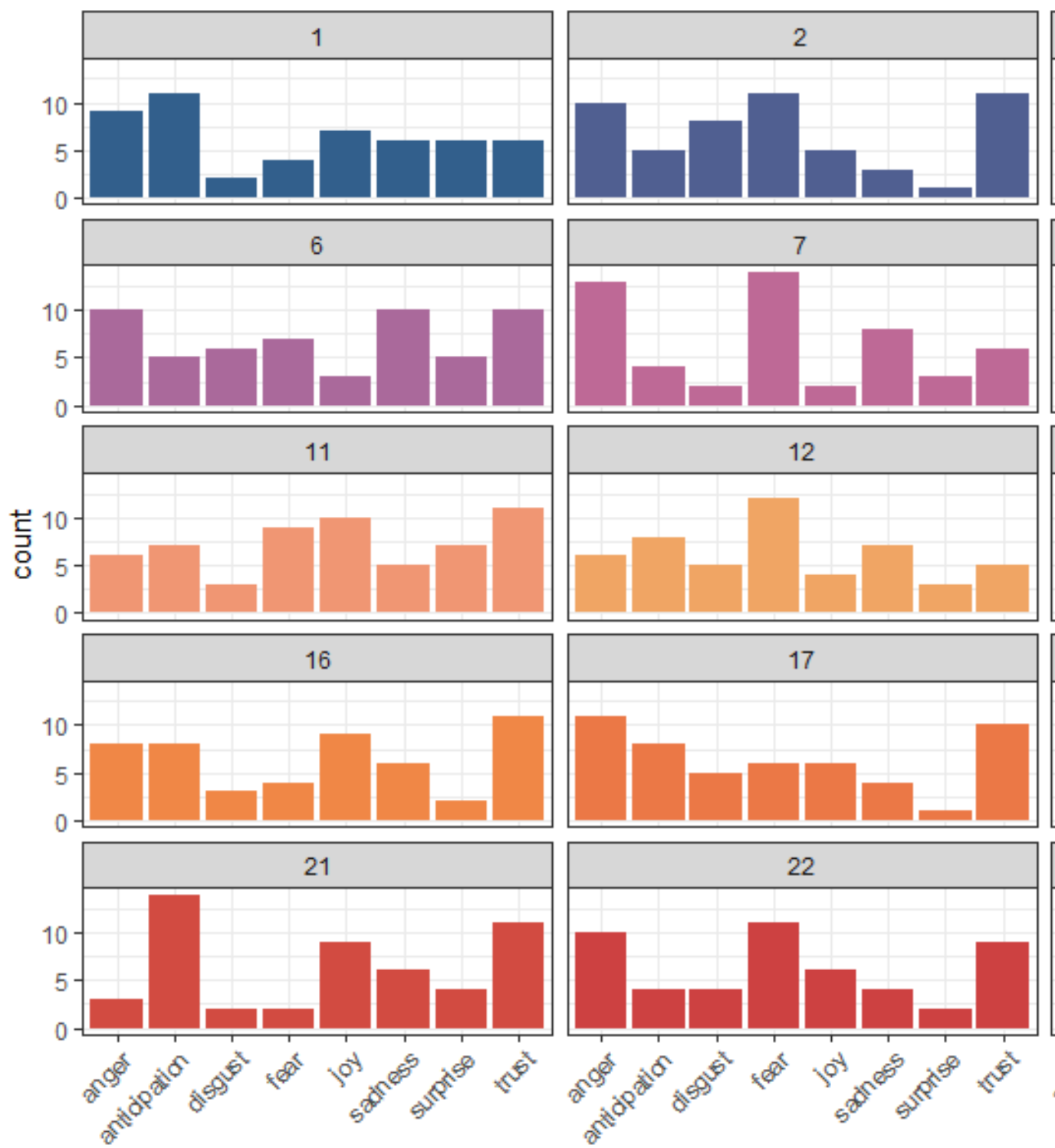
Highest probability words by topic



We then evaluate sentiment by topic



Sentiment count of the 50 most probable words by topic



We can observe that topic Trump is correlated with high feelings of trust, hope. This is in stark contrast with the feelings around other political candidates. This is far and away the strongest visible pattern of this analysis, showing clearly the intent behind the operation: through subtle and hard-to-detect political messaging, to associate Trump with hope for the future and positivity, and his opposition with negativity.

Additionally high values of trust and fear oriented words can be observed around multiple topics, but generally fear, trust and anger are the most shared sentiments.

1.5 Conclusion

In this paper we observed that during the 2016 election season, there's clear patterns that we can observe through sentiment analysis that characterize this specific astroturf operation orchestrated by the IRA. Through text analysis, we could observe specific sets of words, through sentiment analysis we could observe the choice in sentiments to be boosted, and through LDA analysis we could observe the distribution of both of these across specific topics. The findings of such tactics were supplemented by the use of more common text-mining approaches like word frequencies and tf-idf.

For example, we noticed that Trump-adjacent topics were commonly associated with positive sentiment, like "trust", and by far right-leaning words and ideas were by far the most shared on our dataset.

From the combination of these analyses and our interpretative work, we can corroborate the assumption that IRA's intention was to aid President-Elect Donald in winning the presidency in 2016. This tracks with the findings of the Mueller report and of most other literature on the subject. What's interesting is the possible application in iterating on these findings to look for consistent patterns across many known astroturf operations, to be finally able to model their occurrence, and finally be able to tackle them before the consequences of their presence become apparent, as has been the case so far.

1.6 Future Work

1.6.a Network Analysis

It's opportune to verify the hypothesis that astroturf operations such as this one have effects on polarization, observed through network betweenness. We hope to do so by checking for differences between data gathered before and after the election

1.6.b New data

It'd be interesting to observe the evolution of astroturf given the recent developments in generative technology. To do so, we'd need to access new, large amounts of data. This can't cheaply be done off of Twitter (now X), so we hope to do so through fediverse or bluseky (or AT-based) platforms.

With this new data, we could then run a similar kind of analysis to observe if the patterns we found in 2016 remain and fully develop a recognition model. At the same time, we could better develop a model using the help of network analysis: should our hypothesis about polarization

hold, increases in network polarization would be possible markers of astroturf campaigns, and therefore “hot spots” to analyse.

1.6.c LDA opportunities

Due to computational constraints on the datasets of such a size, we were able to run LDA with only one iteration, leading to fairly significant results. This leaves us optimistic on the possible results that could be achieved by running the algorithm with at least 1000 iterations, potentially optimizing astroturf recognition and detection.

Bibliography

AFINN lexicon. <https://arxiv.org/abs/1103.2903>

Ahmed, A.-R., & Anis Rahman. *Manufacturing Rage, The Russian Internet Research Agency's political astroturfing on social media*. <https://firstmonday.org/ojs/index.php/fm/article/download/10801/9723?inline=1>

David M. Blei, Andrew Y. Ng, & Micheal I. Jordan. *Latent dirichlet allocation*. <https://dl.acm.org/doi/10.5555/944919.944937>

Esteban Ortiz-Ospina. *The rise of social media*. <https://ourworldindata.org/rise-of-social-media>

Franziska B. Keller, David Schoch, Sebastian Stier, & JungHwan Yang. *How to Manipulate Social Media Analyzing Political Astroturfing Using Ground Truth Data from South Korea*. https://www.researchgate.net/publication/317290047_How_to_Manipulate_Social_Media_Analyzing_Political_Astroturfing_Using_Ground_Truth_Data_from_South_Korea

Gary, k., Jennifer, P., & Margaret E, R. *How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument*. <https://doi.org/https://doi.org/10.1017/S0003055417000144>

Jacob, J., Suresh, N., Ethan, K., & Laurence, W.-S. *Political Polarization and the Dynamics of Political Language Evidence from 130 Years of Partisan Speech*. https://www.columbia.edu/~lhw2110/2012b_Jensen.pdf

Merriam-Webster. <https://www.merriam-webster.com/dictionary/polarization>

Merriam-Webster. <https://www.merriam-webster.com/dictionary/astroturfing>

Mining and summarizing customer reviews. <https://dl.acm.org/doi/10.1145/1014052.1014073>

NRC lexicon. <https://doi.org/10.4224/21270984>