Sentiment Analysis of Daily News Topics

ComfyWolf50

2018

Abstract

Since large quantity of text is published every day on the Internet losing track of what is going on can be a frequent complication. Models presented in this paper can be used to process daily news headlines and supply the user with easily comprehensible information. As it turns out this task is not as straightforward as it seems at the first sight and many obstacles must be avoided during the analysis. The main goal of this paper is to create reproducible methods which can be useful not only for researchers but also for practitioners.

1 Introduction

Text mining of news is a useful tool for exploring what events are at the centre of public attention. It is more reliable than simply reading the news as it can combine many sources, and thus reach across several social groups. Certain topics can be under-represented in some sources, as they might not be as important for the subscribers of those media. Lumping news from different media together can show, what topics are important "on average". Our analysis gives a concise visual overview of trending topics in daily news. We collect news headlines from webpages of 15 anglophone news sources (such as ABC, BBC, CBS etc.) via web-scraping.

In the first part, we plot several visualizations of the words and phrases that appear most frequently in the headlines. This helps us to see what the main events of the day or week are (an overwhelming majority of headlines refers to imminent events), at least as viewed by the media.

Not surprisingly, US politics captures a large share of the headlines. More specifically, "trump" is the most mentioned word overall, followed by "shutdown", which refers to a situation when US Government and Congress fail to pass a budget, and, therefore, the Government must discontinue all its non-essential operations until the budget is passed. This happened on Friday, January 19, 2018, and it obviously was a major news reported by all our source newspapers.

In the second part, we apply some elementary bag of words sentiment analysis techniques on the headlines. Even though, bag of words is a naïve approach towards sentiment analysis, it performs reasonably well on the headlines, as they are usually brief and clear (more on this in the Literature review section). A natural consequence of this method is that it sometimes misidentifies the sentiment due to missing context. Interestingly, it more often labels negative statements as positive than vice versa. This probably results from the fact that writers use irony when writing about negative things. Overall, we find that there is approximately two times more negative news than positive. Slightly less than 20 % of the headlines were labelled as positive on the days when we performed this analysis, around 40 % was labelled negative, the rest was labelled neutral.

2 Motivation

Text mining is currently in a trend among researchers. For example Amrita et al. (2017) are using similar techniques as we are using to identify child abuse. This shows that the analysis of text is used also in some of the most serious areas of science. The protection of children is a substantial priority. We do not have to be so careful about imprecisions as they are because in our area of expertise an accidental error cannot have fatal consequences. However, this allows us to use less conservative tools which can have higher ability to describe the story.

Large companies are using text mining to increase the profitability of their business even though such activities can cause them troubles with the law. (Ball 2012) This further shows that text mining is useful and can benefit the society. Even if the equality is not the priority, the total social welfare increases. Processing of texts is already a norm and not an exception.

Nonetheless, in no way we are trying to say that reading original texts becomes obsolete.

The contrary is true. Those events when a person is willing to read a full article or paper mean much more nowadays. That is because in the world where information which is simple to understand is easier to penetrate into brains of consumers, reading a full story means that the person has sincere internal interest in the content and the fact that the full story is read is not a result of aggressive marketing techniques.

Our analysis should help researchers to deeper understand how text processing can be done and practitioners should benefit from unbiased view on a large spectrum of news from different providers. Similar tools exist on the market already, nevertheless the non-commercial dimension of this research should bring added value.

It is important to stress, that the attached R script in which we used for our analysis is constructed so that it scrapes the webpages in real time. The visualizations can therefore be redone with up-to-date data by simply running the script.

3 Literature review

A nice overview of text mining methods and their applications can be found in Lang and Baehr (2012). Given how fast the research in this area moves, we can say that it is already outdated. Nonetheless the concepts barely change, only their implementation advance. An exceptional property of this article is its accent on accountability. This is not something that is commonly viewed in this sector of research and its application.

One of the most similar analysis to our analysis is performed by Chan and hong (2017). Nonetheless, their focus is on financial news while our focus is on general news with slight dominance of politics. Their model is called "sentiment analysis engine" and similarly as in our analysis they used various pre-processing methods to clean the data. Calibration of the model on an English movie review is done as well. Nonetheless, in principle the method they used is bag of words.

Another highly relevant paper is Banks and Said (2006) which was published in a reputable journal. We again have to assert that it is slightly outdate, however the concept introduced in this paper are present in the society until now. The use of text mining for the profit of retail companies is something we stress in several places in this articles. Banks and Said (2006): "[I]t enables better management, new services, lower transaction costs and better customer relations."

Ke et al. (2014) are looking at the text mining from a different perspective, yet their approach is still plausible. Their primary goal is an increase in the efficiency of the algorithms used in this kind of analysis. This can sound unrelated to the pure text analysis. On the other hand, their approach is considerably forward looking as the current level of tools used for text mining is not satisfactory and similar improvements should be even more appreciated by the data science community.

In many areas, most of the available information is unstructured. According to some estimates, more than 70 % of potentially usable in business information is unstructured, often in the form of text. (Kwartler 2017) It is logical to assume that in other areas of human endeavour, excluding science, unstructured text information makes up even larger share of the information pool. Tools and methods for unstructured text analysis enable us to use this valuable resource, which is why they have been developing rapidly in recent years.

Text mining techniques can be divided in two main branches: bag of words being one of them and semantic parsing the other one. bag of words (also referred to as vector space model) is the simplest method of text analysis. As the name suggests, it "puts all words in a document in one bag", i.e. it disregards word order or any other logical relations within and between sentences. (Le and Mikolov 2014) Advanced bag of words techniques (bag-of-n-grams etc.) work with "bags of phrases", which makes up for some of the drawbacks of working with individual words. Nevertheless, bag-of-n-grams model can often lead to a very high dimensional representation that is difficult to generalize. (Le and Mikolov 2014)

In this project, we use a slightly enhanced bag of words method from the *quanteda* package, which compares words in a document to a dictionary *data_dictionary_LSD2015* and labels them based on their sentiment. Furthermore, it can recognize if the positive (or negative) word is negated by the adjacent "not". (Benoit et al. 2017) This is still quite a naïve approach, but for news headlines, which are usually short and clear, it yields reasonable results.

Notwithstanding its aforementioned disadvantages, bag of words is excellent for capturing trending topics. These are usually referred to by a common word or fixed multiword phrase, and bag of words can detect these easily.

Semantic parsing is a more advanced method. It determines meaning of words based

on the context, in which they are used, and therefore is less likely to misrepresent it. It is therefore more reliable approach for sentiment analysis. (Le and Mikolov 2014) We do not use it in this project, as it would make our analysis unnecessarily complicated.

4 Methodology

4.1 Trending topics

In this first part of our analysis, we explore what are the most frequent words (and collocations) mentioned in the news. bag of words method is very useful for this type of analysis, since it parses the text into individual words (or multiword collocations), and makes it easy to count their frequency. This approach disregards context of the sentence, and word order, which, however, does not cause much harm when we care only about the number of occurrences.

It it is possible that some words occur repeatedly in different headlines which refer to a completely different topic, and two separate unimportant topic create a perception of one important (frequently mentioned) topic. This is why we perform the analysis also for collocations for up to three adjacent words. In this way, we can identify such cases more easily.

We can get frequency counts of each word (or collocation) in the text by constructing a document term matrix whose columns correspond to individual words (or collocations), and rows correspond to the input documents (individual headlines). This allows us to visualize the most often used words (and collocations).

In this analysis, we decided to opt for a more advanced version of the document term matrix, so-called document feature matrix, which comes from the *quanteda* package. The function which creates this matrix can simultaneously apply most of the transformation methods, which we would have to perform on a text prior to creating a document term matrix. These transformations include removing punctuation, lowering the case of all letters, and removing words that have low information value (such as articles, or different of forms of the verb "to be").

In the next section, we also demonstrate how the document feature matrix can be used to perform a robust sentiment analysis.

4.2 Sentiment analysis

In this part, we present a simple sentiment analysis of the headlines. As was described in the introduction, we are using bag of words approach.

This method compares words in the individual documents (headlines) with *data_dictionary_LSD2015* included in the *quanteda* package, and labels them as positive, negative, negation of positive, and negation of negative. The whole document is then represented by a matrix with four columns and number of rows corresponding to the number of documents (headlines) in the input text. This matrix holds counts of each of the four types of words in each document (headline).[1] The following example demonstrates decomposition of a sentence.[2]

Original sentence: "This aggressive policy will not win friends."

Decomposed version: "This" "NEGATIVE" "policy" "will" "NEG_POSITIVE" "POSITIVE"

This is an example from the original project documentation. However, even on that, we can see some limitations of the method. The word "friends" is labelled positive, even though, the whole phrase "will not win friends" is rather negative. We determine sentiment of the whole headline by the following formula:

$$\begin{aligned}\text{Sentiment score} = {} & \text{sum of positive words} + \text{sum of negated negative words} \\ & - \text{sum of negative words–sum of negated positive words.}\end{aligned}$$

For the aforementioned sentence, this would mean:

$$\text{Sentiment score} = 1 + 0 - 1 - 1 = -1.$$

This gives us a measure that overall performs quite well, even though it is not flawless. Sentences with negative score tend to be negative, whereas sentences with positive score tend to be positive; there is also many sentences which are labelled as neutral, since their sentiment score is 0.

[1] This is similar to the creation of matrix with frequency counts in the previous parts; where columns correspond to each word. Here they correspond to word sentiment types.

[2] https://cran.r-project.org/web/packages/quanteda/quanteda.pdf

4.3 Data

Our data are download from public news providers. The list of news providers is based on Engel (2014) and it was carefully crafted to include the whole ideological spectrum. It includes Washington Post, The Wall Street Journal, USA TODAY, Slate, The New Yorker, POLITICO, The New York Times, Huffington Post, NBC News, The Guardian, Fox News, CBS News, Bloomberg, ABC News and BBC. The effect is that our results are balanced. We were trying to always select just the main headlines. Opinions, sport news and longer readings should not be included. Substantial reverse engineering of news websites was required to guarantee that data are correctly download and clean even after some time.

Our research is reproducible and the provided source code can be used to download new data from news providers in any time. Nonetheless, for the purpose of interpreting results we use data retrieved on the project deadline day January 21, 2018.

5 Results

5.1 Trending topics

Figures 1, 2, 3 and 5 show visualizations of the most frequently debated (written about) topics according to the news headlines.

First, Figure 1 examines individual words that have the most occurrences in the headlines. Not surprisingly, many of the most frequent words refer to the US president. To be more precise, "trump" refers to his last name, "shutdown" refers to a situation when the "u.s." "government" does not manage to pass budget in the Congress, and therefore it must limit its activity to essential tasks until the budget is agreed. "trump" has recently celebrated his "first" "year" in the office. Other notable events included "women"'s march, which is a global protest.

Next, Figure 2 presents most frequent terms in a form of a word cloud. Word clouds are extremely useful since they can visualize the words more vividly than a bar chart. Furthermore, they enable us to capture more topics in one picture. A bar chart for thirty words would not be legible, whereas word cloud captures them easily.

Naturally, word cloud cannot show precise number of occurrences for individual words.

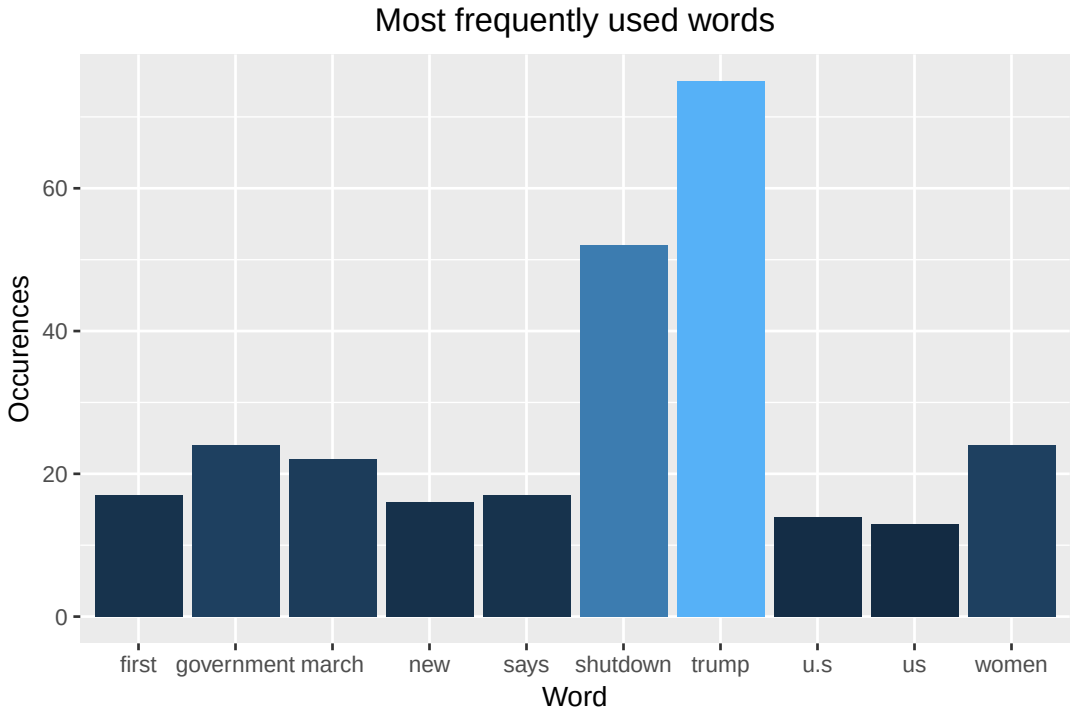

Figure 1: Most frequent words

The word frequency is represented by their size and frequency groups are represented by the same colour.

Apart from the major topics, which we already discussed above, we see that many newspapers wrote, among others, about sexual harassment, terrorist attack in Kabul and North Korea.

The word cloud function can cluster the words which have similar frequencies. This clustering is then depicted by the colour groups in the word cloud. We try to modify this by assigning whole headlines to clusters prior to the word cloud creation. We perform k-means clustering with 15 centres. This assigns headlines to to 15 groups based on the similarity among their words. Figure 3 displays the result. We can see that most of the words end up in the same colour. This word cloud therefore does not provide new information.

The last word cloud which is presented in Figure 5 in Appendix is made of n-grams (polygrams, multigrams), i.e. not only single words but also collocations of up to three words. We decided to use this length, because collocations of three words can still make a meaningful fixed phrase, which would be repeated often enough for the word cloud to capture it.

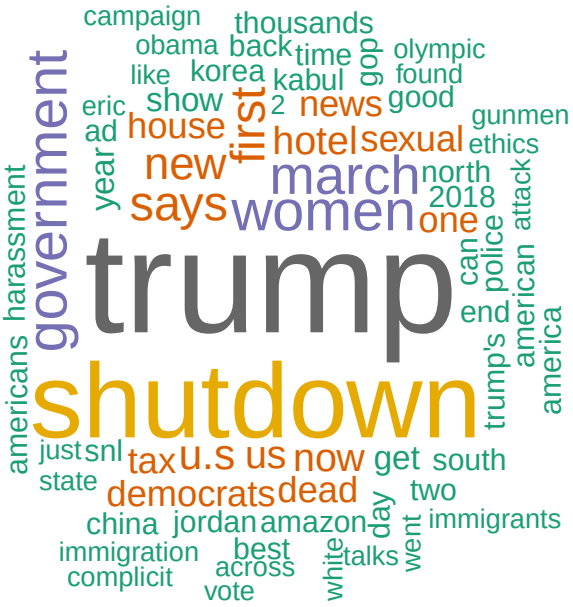

Figure 2: Word cloud from news topics based on unigrams

Document feature matrix for n-grams is constructed in a similar fashion to the one for unigrams. In this case, however, the columns represent all possible combinations of up to n adjacent words in the document. The matrix can therefore be quite large, and this is why context analysis with bag of words is difficult.

We see that the words which dominated the unigram word cloud, dominate also this one. This is logical since every word which is included in a document in a specific collocation must also be included as individual word, but not vice versa. This helps us to see which collocations are, in fact, the main topics of the day. Individual words can be mentioned in different context, and thus appear to be important, by coincidence. For collocations of several words, it is less likely.

The word cloud tells us that the "government shutdown", and "women s march" were, indeed, important topics at least according to the media attention. "hotel siege", i.e. the terrorist attack in Kabul, was also reported. On the other hand, phrase "first year" does not appear in the word cloud which suggests that media were not so much interested in the anniversary of Donald Trump's government; or, more precisely, they did not refer to it by using this phrase.

Figure 3: Word cloud from news topics created by manual k-means clustering

5.2 Sentiment analysis

Below is a list of ten sentences with the lowest sentiment score. The list is called "TEN MOST NEGATIVE HEADLINES". Nevertheless, we must always bear in mind that the labelling does not account for degrees of sentiment.[3] (Please note that we excluded duplicate headlines. Quite often, several newspapers use identical headline).

TEN MOST NEGATIVE HEADLINES

1. Dylan Farrow details her sexual assault allegations against Woody Allen
2. Family Tom Petty died of accidental drug overdose
3. One year on Spanish rangers fear for lives after double murder
4. Stormy Daniels Stops By Weekend Update to Remind Us That the World Is Terrible and Its Our Fault
5. Tens of thousands protest against corruption
6. The Road Movie Exploits Pain and Death

[3]I.e. "lethal chemical attack" has the same sentiment score as "unlucky traffic accident".

7. Tom Petty Died of Accidental Drug Overdose
8. Quebec Jury Finds Three Not Guilty of Negligence in Deadly 2013 Derailment
9. Report Congressman leading charge against assault settles own misconduct case
10. Sick to death of hearing Brigitte Bardots bigoted views Me too

We can see that these events usually include death or some form of criminal activity. Furthermore, all these headlines are labelled correctly. Follows a list of ten sentences with the highest sentiment score.

TEN MOST POSITIVE HEADLINES

1. Amazons HQ2 Losers Hold Out Hope for Consolation Prize
2. Moderates on both sides offer private glimmers of a breakthrough
3. Your Emotional Support Duck Is Not Welcome in Seat 15C
4. Amazon Is Totally Putting Its Second Headquarters in DC Right
5. Buttler inspires England to seriesclinching ODI win over Australia
6. CBS News honored with duPontColumbia Awards
7. Ed Sheeran is very happy and in love with new fianc<e9>e
8. If Vikings win historic homefield Super Bowl could require steepest ticket ever
9. Im Pretty Sure This SNL Sketch Isnt How the Credits to The Fresh Prince of Bel Air Usually Went
10. In Trumps inauguration crowd members of Russias elite anticipated a thaw between Moscow Washington

Our simple scoring method performs quite successfully, even though, for example, the ninth sentence is mislabelled. Words "Fresh" and Credits earned it a positive score.

Based on several attempts on different days (i.e. with different news headlines), we can claim that the situation where non-positive headline is labelled as positive is more common than non-negative sentence being labelled as negative. This would suggest that

our sentiment evaluation might overestimate the share of positive sentences. Below is a graph depicting shares of headlines with different sentiment.

Figure 4 shows the distribution of negative, neutral and positive sentiments in the news headlines. Interestingly, even though the share of positive headlines might be overestimated, there is still twice as many negative headlines. Due to a strange human tendency, people like to read about sad events, and newspapers adjust to the demand. For example, the article about Tom Petty's drug overdose resulting in his death appeared in most of the newspapers.

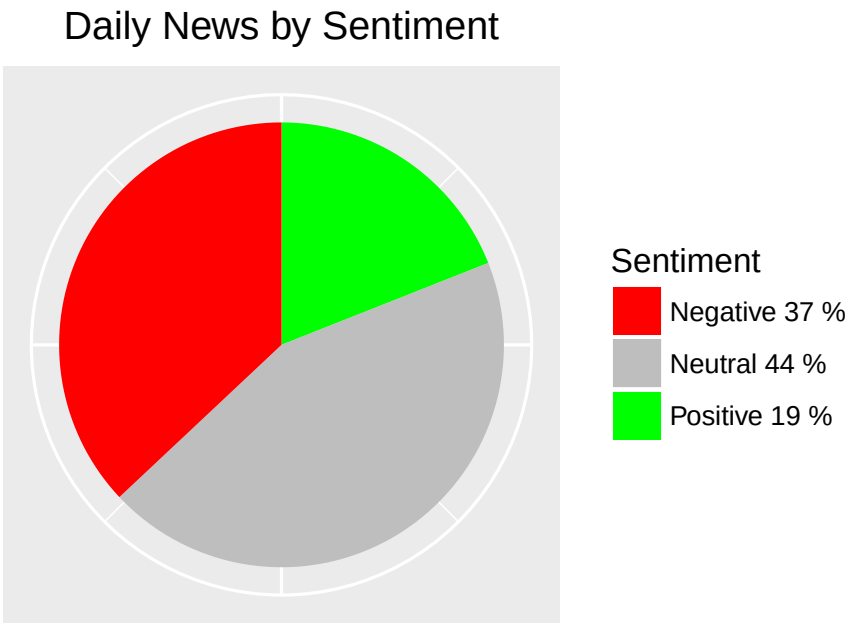

Figure 4: Pie chart showing sentiments in news headlines

6 Discussion

One of the goals of our analysis is also to help practitioners to navigate in the complicated system of world news. For example, Frunzeanu (2015) is explaining in the paper how word clouds can be used as a learning tool in primary schools . Given the robustness of our research, we can expect that the word clouds created in our analysis can be used by professionals working in industries who do not have time to read all newspapers, yet they care about what is going on in the society and they are interested in public affairs. Fast life does not always give them the opportunity to read whole articles on the Internet. However, with the word cloud produced during our research only relevant words show up

in the figure. The algorithm behind our word cloud and the unique dataset are the key components.

If we compare our results with results achieved by Chan and hong (2017), we can say that our method yields easier understandable outcome. However, even though our method can be considered substantially robust to outliers, the method of Chan and hong (2017) is even more robust. Nonetheless, at this point we have to say that their method requires to build a "language parser for sentiment analysis" which needs constantly update dictionary and their method shows large sensitivity to the quality of the dictionary. Our method while being relatively simpler does not suffer so much from the low quality dictionary. Even if dictionary used in the sentiment analysis were broken, we can still rely on the clustering embedded in the word cloud method.

7 Conclusion

This paper shows how news headlines can be processed in order to get useful insights from them. We are using a large portion of text processing methods covered during the course to build a robust model for visualizing daily news topics. There is a comprehensive overview of methods used by other researchers. Various manuals covering broad range of text mining techniques are discussed as well.

In the empirical part we are using data from trusted news providers to build a large spectrum of text visualisation tools. The analysis requires usage of custom function R specifically prepared for this project. The text mining is divided into two parts. First, we explore trending topics using word clouds and k-mean clustering. There are several word cloud models including polygram word cloud model which according to our knowledge has not been used previously in other research of similar scope. Second, we conduct a sentiment analysis. In this part we find ten most positive and 10 most negative headlines. Consequently, we are able to divide headlines into positive and negative with surprisingly high accuracy. Finally, we compare our results with other researchers and show evidence that similar tools as we created are used by practitioners.

References

Amrita, Chintan et al. (2017). "Identifying child abuse through text mining and machine learning". In: *Expert Systems with Applications* 88, pp. 402–418.

Ball, James (2012). *Me and my data: how much do the internet giants really know?* URL: https://www.theguardian.com/technology/2012/apr/22/me-and-my-data-internet-giants.

Banks, David L. and Yasmin H. Said (2006). "Data Mining in Electronic Commerce". In: *Statistical Science* 21.2, pp. 234–246.

Benoit, Kenneth et al. (2017). *Package 'quanteda'*.

Chan, Samuel W.K. and Mickey W.C.C hong (2017). "Sentiment analysis in financial texts". In: *Decision Support Systems* 94, pp. 53–64.

Engel, Pamela (2014). *THere's How Liberal Or Conservative Major News Sources Really Are*. URL: http://www.businessinsider.com/what-your-preferred-news-outlet-says-about-your-political-ideology-2014-10.

Frunzeanu, Mirela (2015). "Using Wikis, Word Clouds and Web Collaboration in Romanian Primary Schools". In: *Procedia - Social and Behavioral Sciences* 94.5, pp. 580–585.

Ke, Xiaohua et al. (2014). "Complex dynamics of text analysis". In: *Physica A: Statistical Mechanics and its Applications* 415.1, pp. 307–314.

Kwartler, Ted (2017). *Text Mining: Bag of Words*. URL: https://www.datacamp.com/courses/intro-to-text-mining-bag-of-words.

Lang, Susan and Craig Baehr (2012). "Data Mining: A Hybrid Methodology for Complex and Dynamic Research". In: *College Composition and Communication* 64.1, pp. 172–194.

Le, Quoc and Tomas Mikolov (2014). *Distributed Representations of Sentences and Documents*.

8 Appendix

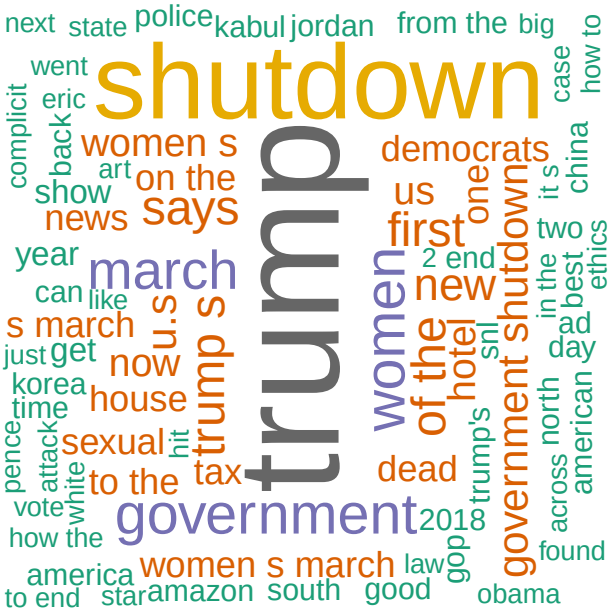

Figure 5: Word cloud from news topics based on polygrams