Irish Communication Review

Volume 16 | Issue 1

Article 7

June 2018

Text Analytics Techniques in the Digital World: a Sentiment Analysis Case Study of the Coverage of Climate Change on US News Networks

Jerome Casey
Technological University Dublin, jerome.casey@tudublin.ie

Recommended Citation

Casey, Jerome (2018) "Text Analytics Techniques in the Digital World: a Sentiment Analysis Case Study of the Coverage of Climate Change on US News Networks," *Irish Communication Review*: Vol. 16: Iss. 1, Article 7.
Available at: https://arrow.tudublin.ie/icr/vol16/iss1/7

Text analytics techniques in the digital world: A sentiment analysis case study of the coverage of climate change on US news networks

Jerome Casey

Abstract

This paper analyses nearly 600 news segments relating to climate change broadcast on 3 American news networks over a period of 8 years. The paper demonstrates the typical steps involved in a text analytics solution. It shows how the text data was sourced and imported into a software program. The steps carried out in pre-processing the text data are outlined as well as explaining key terms in the text analytics pipeline. A sentiment analysis is applied using a lexicon and further processing is carried out to answer the original questions posed such as what words drive a particular sentiment category, how the news topic vocabulary varies by news network and how sentiment changes over time. It is argued here that the use of an externally provided lexicon in sentiment analysis is not without its pitfalls. It is also shown how the lexicon can be altered by the implementer and the subsequent effect on the results. The stop words list used also affects the text content downstream which will influence the sentiment score. As such, the integrity of the results output in a sentiment analysis solution can be called into question when the source code itself is not publicly visible and available for inspection.

Introduction

Data science is a discipline that allows raw data to be turned into understanding, insight, and knowledge. One area of such science is text mining or text analytics – the process of distilling actionable insight from text (Kwartler, 2017). Text data are constantly generated and sentiment analysis is a way to measure the attitudes and opinions expressed in text (Silge, 2017). The literature demonstrates the extensive amount of work that has been performed on sentiment analysis. Pang et al. (2008), for example, discuss sentiment analysis applications in areas such as business intelligence in public opinion and how it can affect product sales, government intelligence in public opinion related to government performance, movie and restaurant reviews and speeches. Taboada et al. (2011) describe a lexicon-based approach to extracting sentiment from text. They devise a tool to extract sentiment that shows good performance over various domains. They discuss how the tool handles negations and intensification, both of which constitute challenges encountered in sentiment analysis.

The objective of the study was to demonstrate a typical text analytics pipeline. Although extensive code was developed for this case study to analyse the text dataset, it is not the intention to document the code, rather to describe how the code processes the data, in a stepwise fashion, to produce meaningful visualisations and insights. We start by explaining how the dataset was collated, its structure and some typical text mining preprocessing. We explain the sentiment scoring using a lexicon and then further process the text to answer posed questions such as what words drive a particular sentiment category, how the news topic vocabulary varies by news network, and how sentiment changes over time. In addition, we also discuss external factors that can affect the sentiment score, such as the choice of the sentiment lexicon and the so called stopwords list.

The topic of climate change was chosen as it has been a trending topic in the media for some time. The Economist (2017) and The New Scientist (2017a) cover the topic regularly. In recent times, climate change has received increased coverage from the run-in to the 2016 US presidential election and latterly when President Trump withdrew the United States from the Paris Climate Accord on June 1st 2017. The New Scientist (2017b) made climate change their cover story that month. It is a topic that has continually divided opinion and hence is interesting as a sentiment analysis case study.

Text analytics workflow

A typical workflow used in a text analytics study could have the following steps:

1. Problem definition and specific goals: State the questions you want to examine in the analysis.

   e.g. Which TV news station uses more positive or negative words when discussing climate change?

2. Identify the text to be collected: the text could come from articles, blogs, social media, emails, surveys.

3. Obtain the dataset and import it into a software Independent Development Environment (IDE) such as RStudio.

4. Pre-process the text – this can involve arranging the text into a tidy text format, tokenizing, cleaning the data of errors and removing any stop words.

5. Perform the sentiment analysis.

6. Perform further analysis on the text in response to the posed questions.

7. Reach an insight or conclusion to the original question posed (Grolemund & Wickham, 2017).

A typical text analytics workflow is shown in Figure 1. These steps will be further explained below.

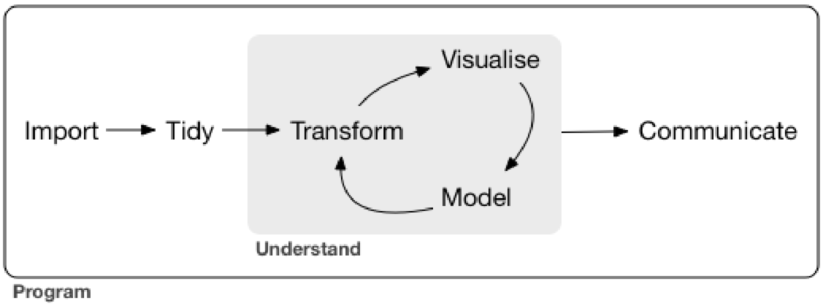


Figure 1: A typical data science workflow (Grolemund & Wickham, 2017).

Sourcing the dataset used in the analysis

The GDELT Analysis Service (GDELT, 2017a) offers a variety of tools and services that allow visualisation, exploration, and exportation of the Global Database of

Events, Language, and Tone (GDELT) project – a real-time database of global society. Closed captioning (CC) text from television news shows are available to download from GDELT's Television Explorer tool (GDELT, 2017b). This web tool uses data from the Internet Archive Television News Archive (2017). It allows entry of keyword searches and returns matching results from the closed captioning streams of up to six years of American television news in the case of some networks.

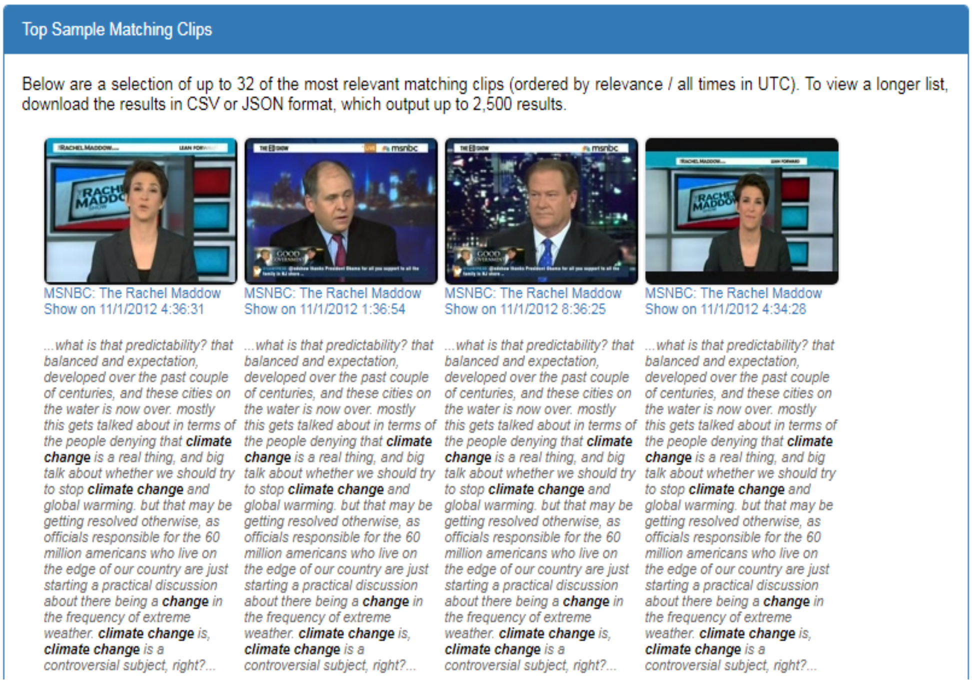
This archive research service is

> … intended to enhance the capabilities of journalists, scholars, teachers, librarians, civic organizations and other engaged citizens. It repurposes closed captioning to enable users to search, quote and borrow U.S. TV news programs. Available at no charge, the public can use the index of searchable text and short streamed clips to explore TV news, discover important resources, understand context, evaluate assertions of fact, engage with others and share insights.

Using the Search Options area of the site a search for segments was configured as follows:

> Primary Keyword/Phrase:climate change
>
> Television Network:MSNBC
>
> Output Format: Web Visualizations

The results returned are as shown in Figure 2.



Top Sample Matching Clips

Below are a selection of up to 32 of the most relevant matching clips (ordered by relevance / all times in UTC). To view a longer list, download the results in CSV or JSON format, which output up to 2,500 results.

MSNBC: The Rachel Maddow Show on 11/1/2012 4:36:31

...what is that predictability? that balanced and expectation, developed over the past couple of centuries, and these cities on the water is now over. mostly this gets talked about in terms of the people denying that ***climate change*** *is a real thing, and big talk about whether we should try to stop* ***climate change*** *and global warming. but that may be getting resolved otherwise, as officials responsible for the 60 million americans who live on the edge of our country are just starting a practical discussion about there being a* ***change*** *in the frequency of extreme weather.* ***climate change*** *is,* ***climate change*** *is a controversial subject, right?...*

MSNBC: The Rachel Maddow Show on 11/1/2012 1:36:54

...what is that predictability? that balanced and expectation, developed over the past couple of centuries, and these cities on the water is now over. mostly this gets talked about in terms of the people denying that ***climate change*** *is a real thing, and big talk about whether we should try to stop* ***climate change*** *and global warming. but that may be getting resolved otherwise, as officials responsible for the 60 million americans who live on the edge of our country are just starting a practical discussion about there being a* ***change*** *in the frequency of extreme weather.* ***climate change*** *is,* ***climate change*** *is a controversial subject, right?...*

MSNBC: The Rachel Maddow Show on 11/1/2012 8:36:25

...what is that predictability? that balanced and expectation, developed over the past couple of centuries, and these cities on the water is now over. mostly this gets talked about in terms of the people denying that ***climate change*** *is a real thing, and big talk about whether we should try to stop* ***climate change*** *and global warming. but that may be getting resolved otherwise, as officials responsible for the 60 million americans who live on the edge of our country are just starting a practical discussion about there being a* ***change*** *in the frequency of extreme weather.* ***climate change*** *is,* ***climate change*** *is a controversial subject, right?...*

MSNBC: The Rachel Maddow Show on 11/1/2012 4:34:28

...what is that predictability? that balanced and expectation, developed over the past couple of centuries, and these cities on the water is now over. mostly this gets talked about in terms of the people denying that ***climate change*** *is a real thing, and big talk about whether we should try to stop* ***climate change*** *and global warming. but that may be getting resolved otherwise, as officials responsible for the 60 million americans who live on the edge of our country are just starting a practical discussion about there being a* ***change*** *in the frequency of extreme weather.* ***climate change*** *is,* ***climate change*** *is a controversial subject, right?...*

Figure 2: Visualisation of the closed captioning text segments returned by the search using GDELT's Television Explorer tool (GDELT, 2017b).

The search output format can be set to comma-separated values (CSV) allowing the text to be downloaded as a csv file and the search can be repeated for other US national networks such as CNN and Fox News. The data for both MSNBC and CNN extend back to 2009, whereas that for Fox News extends back to 2011. The data can then be imported into a software program such as RStudio (2017) for text analysis using the R programming language.

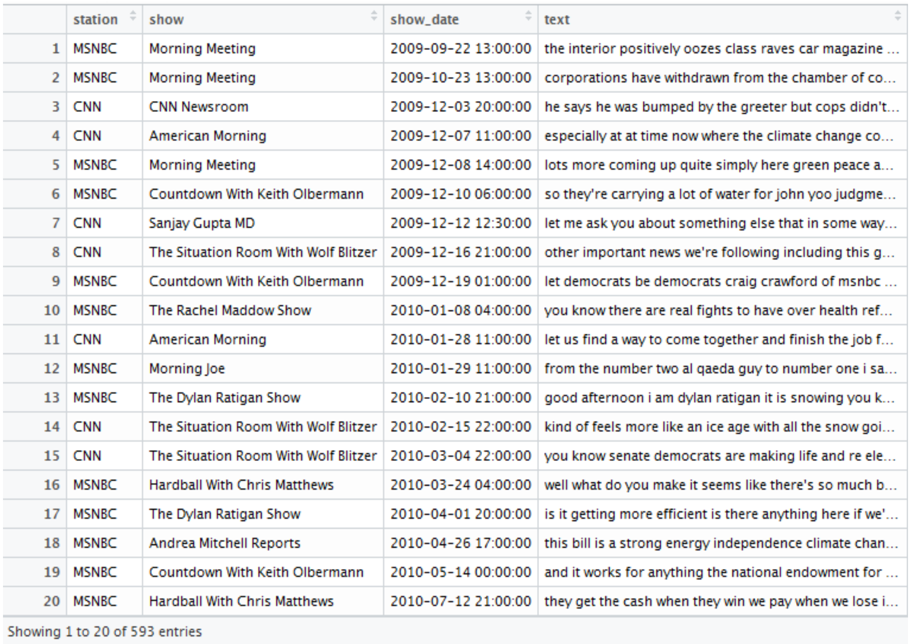
RStudio and R

RStudio (2017) is a free and open-source integrated development environment (IDE) that can be used to program in R, a language for statistical computing and graphics. A number of R library packages will be used in this paper to process the text data. Library packages contain functions that we can apply to the text data to process it in some way, such as calculating word frequencies, filtering out rows that match a condition, sorting rows or producing plots. They can also contain data such as the lexicons used in the sentiment analysis. It is not the intention to document the code used in the study, rather to describe the typical processes involved in a text to pipeline and what is being done by the code at each stage.

Description of the Dataset Used in the Analysis

The text dataset is shown below in Figure 3. after being imported into RStudio. Each of the 593 rows represent an individual news segment containing 4 columns or fields, namely:

> station: the TV news station from which the text comes, CNN, MSNBC, Fox News
>
> show: the show on that station where the text was spoken
>
> show_date: the broadcast date of the spoken text
>
> text: the actual text spoken on TV



	station	show	show_date	text
1	MSNBC	Morning Meeting	2009-09-22 13:00:00	the interior positively oozes class raves car magazine ...
2	MSNBC	Morning Meeting	2009-10-23 13:00:00	corporations have withdrawn from the chamber of co...
3	CNN	CNN Newsroom	2009-12-03 20:00:00	he says he was bumped by the greeter but cops didn't...
4	CNN	American Morning	2009-12-07 11:00:00	especially at at time now where the climate change co...
5	MSNBC	Morning Meeting	2009-12-08 14:00:00	lots more coming up quite simply here green peace a...
6	MSNBC	Countdown With Keith Olbermann	2009-12-10 06:00:00	so they're carrying a lot of water for john yoo judgme...
7	CNN	Sanjay Gupta MD	2009-12-12 12:30:00	let me ask you about something else that in some way...
8	CNN	The Situation Room With Wolf Blitzer	2009-12-16 21:00:00	other important news we're following including this g...
9	MSNBC	Countdown With Keith Olbermann	2009-12-19 01:00:00	let democrats be democrats craig crawford of msnbc ...
10	MSNBC	The Rachel Maddow Show	2010-01-08 04:00:00	you know there are real fights to have over health ref...
11	CNN	American Morning	2010-01-28 11:00:00	let us find a way to come together and finish the job f...
12	MSNBC	Morning Joe	2010-01-29 11:00:00	from the number two al qaeda guy to number one i sa...
13	MSNBC	The Dylan Ratigan Show	2010-02-10 21:00:00	good afternoon i am dylan ratigan it is snowing you k...
14	CNN	The Situation Room With Wolf Blitzer	2010-02-15 22:00:00	kind of feels more like an ice age with all the snow goi...
15	CNN	The Situation Room With Wolf Blitzer	2010-03-04 22:00:00	you know senate democrats are making life and re ele...
16	MSNBC	Hardball With Chris Matthews	2010-03-24 04:00:00	well what do you make it seems like there's so much b...
17	MSNBC	The Dylan Ratigan Show	2010-04-01 20:00:00	is it getting more efficient is there anything here if we'...
18	MSNBC	Andrea Mitchell Reports	2010-04-26 17:00:00	this bill is a strong energy independence climate chan...
19	MSNBC	Countdown With Keith Olbermann	2010-05-14 00:00:00	and it works for anything the national endowment for ...
20	MSNBC	Hardball With Chris Matthews	2010-07-12 21:00:00	they get the cash when they win we pay when we lose i...

Showing 1 to 20 of 593 entries

Figure 3: The original text data imported (20 records shown of the 593 news segments)

Two sample news segments are shown below to better understand the data. Punctuation has been removed and there are some typos emanating from the closed captioning process:

> 1.’there was a big climate change conference and the president removed most climate change information from its website to quote reflect the approach of new leadership as you know president trump has wrongly referred to climate change as a chinese hoax as being a man who believes in climate change does this concern you the thing that does concern me is that my friends in the environmental community and their refusal to even consider it is something that has been san impedestriiment moving forw yes i think that climate change is real’

> 2.'jesse let's pick it up no littering on day what do you think is a bigger threat to the united states climate change or terrorism climate change no littering on day what do you think is a bigger threat to the united states climate change or terrorism climate change climate change at the moment what do you think is a bigger threat to the united states climate change or terrorism climate change climate change at the moment i would go with climate change no jesse let's pick it up no littering on day what do you think is a bigger threat to the united states climate change or terrorism'

A preliminary analysis of the data shows there are:

> 148 articles from CNN from Dec 2009- Apr 2017
>
> 262 articles from MSNBC from Sep 2009- Apr 2017
>
> 183 articles from Fox NEWS from July 2011- Apr 2017

At this point, the text is in a raw format. However some initial preprocessing is required before reaching a point of applying the sentiment analysis and the additional processing needed to answer the questions posed. These questions are outlined below.

Preprocessing the text

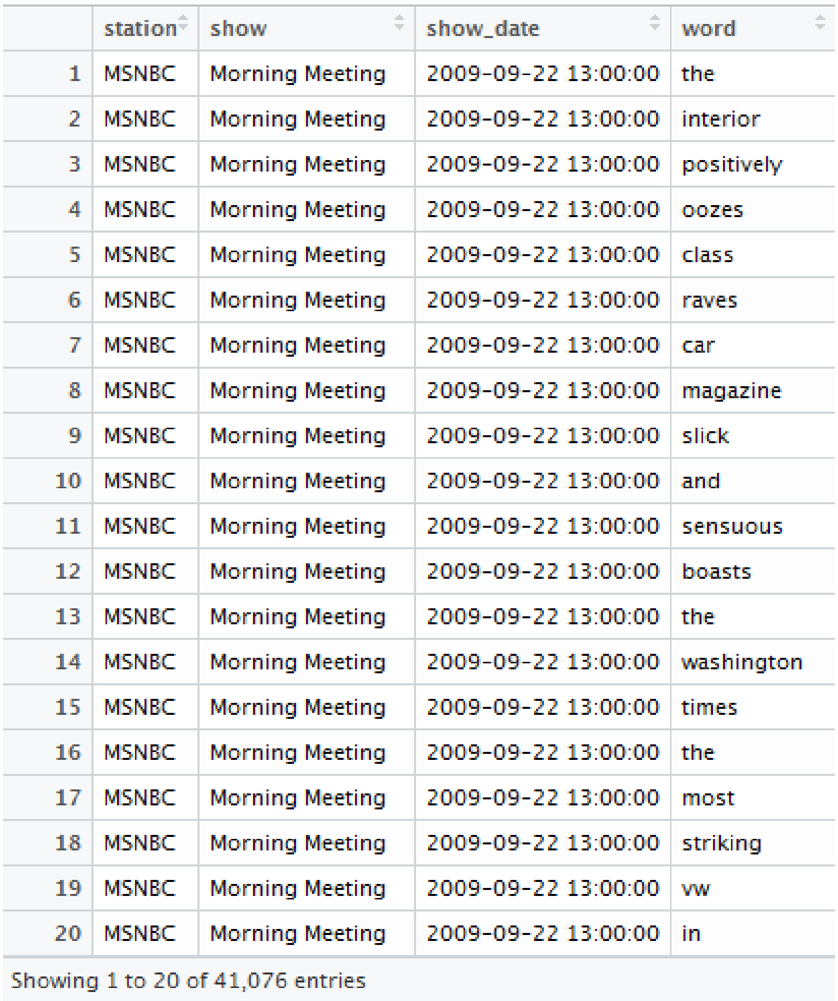
Possible steps that could be involved at this stage are:

1. Arrange the text into a **tidy format**. Tidy datasets have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This structure makes it easier subsequently to apply data analysis functions to manipulate, model and visualise the data. (Wickham, 2014 and Silge & Robinson, 2017).
2. **Tokenizing:** reducing a large body of text to individual tokens. A token is a meaningful unit of text for analysis – it can be individual paragraphs, individual sentences, or groups of words denoted as n-grams (Mullen, 2016). Two words occurring together are bi-grams whereas a single word is a unigram. This process will simultaneously leave the text in a tidy data structure with one token per row.
3. **C**reating a **document-term matrix** (or DTM). This is a matrix where each row represents one document (such as a book or article or news

segment), each column represents one term (a word), and each value typically contains the number of appearances of that term in that document. In this case study we work with data in a tidy format and do not create a DTM. By applying a count or a group_by/summarize that contains counts or another statistic for each combination of a term and document, this is comparable to creating a DTM, but with the advantage that we have data in a tidy format. As stated above this allows us to apply a large variety of functions to produce meaningful visualisations and insights, as we shall see.

4. Remove **stop words** from the text. Often there are words in a text that occur frequently, but provide little information. So it may be desirable to remove these so-called stop words. Some common English stop words include 'I', 'she'll', 'the', etc. 'of', 'a'.
5. **Stemming** and stem completion – segment words to their base e.g. 'complicated' and 'complication' would be transformed to their stem 'complicate'.
6. Remove tabs and whitespaces from the text.
7. Remove punctuation (., !) – this step may or may not be carried out depending on the text data under analysis. For example, if the text data are tweets, this step is skipped as the tweets may contain text emoticons which convey meaning.
8. Convert all text to lowercase.
9. Map symbols to a word e.g. $-> dollar, % -> percent...and more.

Returning to the text dataset we see in Figure 3 above, that the text is already in a tidy format. The next step is to apply the tokenization – in this case we choose the token to be an individual word or unigram. The resulting output is as shown in Figure 4. with the textfield being replaced by the wordfield in the process. As a result, the dataset goes from having 593 rows of the individual news segments to 41,076 rows with the individual words. Running a little programming code, we determine that this is made up of 19,487 rows relating to MSNBC, 10,713 relating to CNN, and 10,876 relating to Fox News, as will be seen later in Figure 8. It is important to keep account of these values to calculate the sentiment proportions later on, as by that stage we will have altered the number of rows after applying the steps of stop-word removal and sentiment analysis.

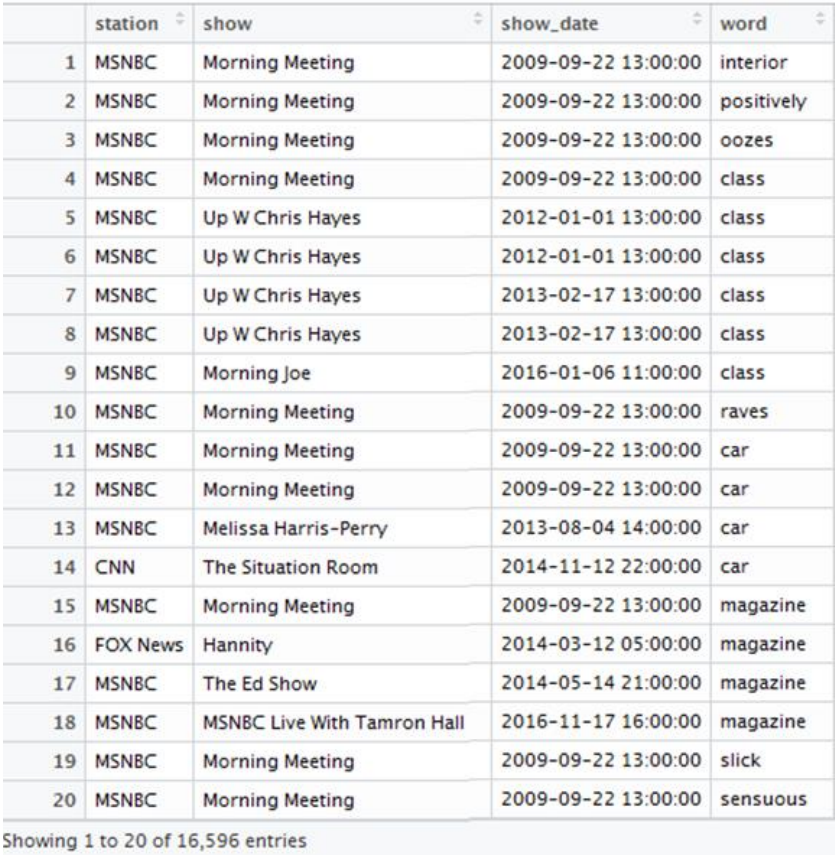


	station	show	show_date	word
1	MSNBC	Morning Meeting	2009-09-22 13:00:00	the
2	MSNBC	Morning Meeting	2009-09-22 13:00:00	interior
3	MSNBC	Morning Meeting	2009-09-22 13:00:00	positively
4	MSNBC	Morning Meeting	2009-09-22 13:00:00	oozes
5	MSNBC	Morning Meeting	2009-09-22 13:00:00	class
6	MSNBC	Morning Meeting	2009-09-22 13:00:00	raves
7	MSNBC	Morning Meeting	2009-09-22 13:00:00	car
8	MSNBC	Morning Meeting	2009-09-22 13:00:00	magazine
9	MSNBC	Morning Meeting	2009-09-22 13:00:00	slick
10	MSNBC	Morning Meeting	2009-09-22 13:00:00	and
11	MSNBC	Morning Meeting	2009-09-22 13:00:00	sensuous
12	MSNBC	Morning Meeting	2009-09-22 13:00:00	boasts
13	MSNBC	Morning Meeting	2009-09-22 13:00:00	the
14	MSNBC	Morning Meeting	2009-09-22 13:00:00	washington
15	MSNBC	Morning Meeting	2009-09-22 13:00:00	times
16	MSNBC	Morning Meeting	2009-09-22 13:00:00	the
17	MSNBC	Morning Meeting	2009-09-22 13:00:00	most
18	MSNBC	Morning Meeting	2009-09-22 13:00:00	striking
19	MSNBC	Morning Meeting	2009-09-22 13:00:00	vw
20	MSNBC	Morning Meeting	2009-09-22 13:00:00	in

Showing 1 to 20 of 41,076 entries

Figure 4: The tokenized dataset showing a new field called word containing a single word.

The next step is to remove the stopwords using a built-in list available in the commercial R*tidytext* package. The stopwords come from 3 separate lexicons namely ‘SMART’, ‘snowball’ and ‘onix’ (Silge & Robinson, 2017). As a result of this step, the dataset reduces from 41,076 rows to 16,596 rows as shown in Figure 5. It should be noted that using different stopword lists affects the downstream analysis of the text, because of the words that are removed in the process.



	station	show	show_date	word
1	MSNBC	Morning Meeting	2009-09-22 13:00:00	interior
2	MSNBC	Morning Meeting	2009-09-22 13:00:00	positively
3	MSNBC	Morning Meeting	2009-09-22 13:00:00	oozes
4	MSNBC	Morning Meeting	2009-09-22 13:00:00	class
5	MSNBC	Up W Chris Hayes	2012-01-01 13:00:00	class
6	MSNBC	Up W Chris Hayes	2012-01-01 13:00:00	class
7	MSNBC	Up W Chris Hayes	2013-02-17 13:00:00	class
8	MSNBC	Up W Chris Hayes	2013-02-17 13:00:00	class
9	MSNBC	Morning Joe	2016-01-06 11:00:00	class
10	MSNBC	Morning Meeting	2009-09-22 13:00:00	raves
11	MSNBC	Morning Meeting	2009-09-22 13:00:00	car
12	MSNBC	Morning Meeting	2009-09-22 13:00:00	car
13	MSNBC	Melissa Harris-Perry	2013-08-04 14:00:00	car
14	CNN	The Situation Room	2014-11-12 22:00:00	car
15	MSNBC	Morning Meeting	2009-09-22 13:00:00	magazine
16	FOX News	Hannity	2014-03-12 05:00:00	magazine
17	MSNBC	The Ed Show	2014-05-14 21:00:00	magazine
18	MSNBC	MSNBC Live With Tamron Hall	2016-11-17 16:00:00	magazine
19	MSNBC	Morning Meeting	2009-09-22 13:00:00	slick
20	MSNBC	Morning Meeting	2009-09-22 13:00:00	sensuous

Showing 1 to 20 of 16,596 entries

Figure 5: The tokenized dataset with the stop words removed from the word field.

Sentiment lexicons

To conduct sentiment analysis we need to employ a lexicon. Lexicons are lists of words that are scored in some way according to the emotion or opinion content of the words. The words can be scored:

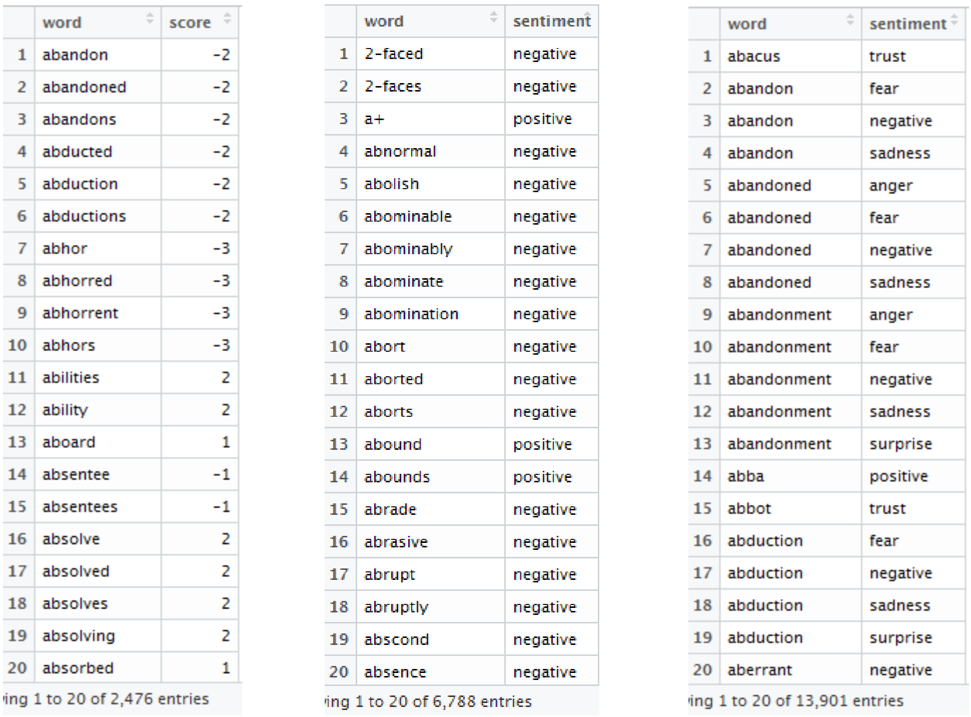
1. In a binary fashion (positive/negative) such as the *bing*lexicon constructed byLiu, *et al.* (2004)

2. Numerically from negative to positive sentiment (-5, +5) such as the *afinn*lexicon constructed by Årup Nielsen (2011).

3. Specific emotion categories such as the *nrc*lexicon constructed by Mohammad & Turney, (2011). This latter lexicon will be employed in this study due to its breadth of sentiment. It contains the following 10 categories:

positive, *negative*, *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust*.

Some of these words, due to their meaning, can appear in more than one category. For example as shown in Figure 6, ‘abandon’ can covey both a negative sentiment and sadness. Neutral words do not appear in the lexicons. The three different lexicons are shown below each with its own sentiment score. All three sentiment lexicons are contained in the *sentiments* dataset within the R *tidytext* package.



	word	score
1	abandon	-2
2	abandoned	-2
3	abandons	-2
4	abducted	-2
5	abduction	-2
6	abductions	-2
7	abhor	-3
8	abhorred	-3
9	abhorrent	-3
10	abhors	-3
11	abilities	2
12	ability	2
13	aboard	1
14	absentee	-1
15	absentees	-1
16	absolve	2
17	absolved	2
18	absolves	2
19	absolving	2
20	absorbed	1

ing 1 to 20 of 2,476 entries

	word	sentiment
1	2-faced	negative
2	2-faces	negative
3	a+	positive
4	abnormal	negative
5	abolish	negative
6	abominable	negative
7	abominably	negative
8	abominate	negative
9	abomination	negative
10	abort	negative
11	aborted	negative
12	aborts	negative
13	abound	positive
14	abounds	positive
15	abrade	negative
16	abrasive	negative
17	abrupt	negative
18	abruptly	negative
19	abscond	negative
20	absence	negative

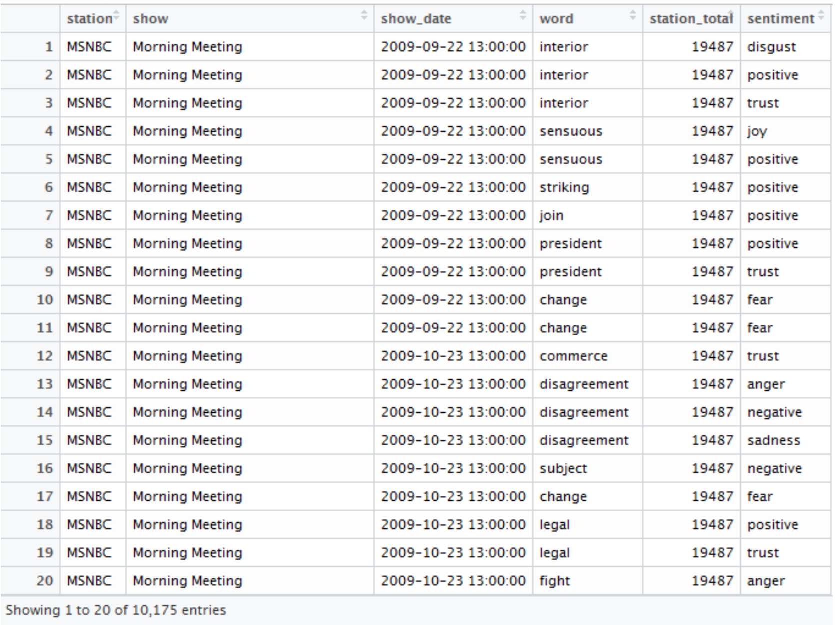
ing 1 to 20 of 6,788 entries

	word	sentiment
1	abacus	trust
2	abandon	fear
3	abandon	negative
4	abandon	sadness
5	abandoned	anger
6	abandoned	fear
7	abandoned	negative
8	abandoned	sadness
9	abandonment	anger
10	abandonment	fear
11	abandonment	negative
12	abandonment	sadness
13	abandonment	surprise
14	abba	positive
15	abbot	trust
16	abduction	fear
17	abduction	negative
18	abduction	sadness
19	abduction	surprise
20	aberrant	negative

ing 1 to 20 of 13,901 entries

Figure 6: Three different lexicons are shown each with its own sentiment score. Words repeat in the nrc lexicon when a number of sentiment categories apply. Neutral words do not appear in the lexicon. left: afinn, middle: bing, right: nrc.

Applying the *nrc*lexicon to our text will result in a new sentiment field appearing as shown in Figure 7.



	station	show	show_date	word	station_total	sentiment
1	MSNBC	Morning Meeting	2009-09-22 13:00:00	interior	19487	disgust
2	MSNBC	Morning Meeting	2009-09-22 13:00:00	interior	19487	positive
3	MSNBC	Morning Meeting	2009-09-22 13:00:00	interior	19487	trust
4	MSNBC	Morning Meeting	2009-09-22 13:00:00	sensuous	19487	joy
5	MSNBC	Morning Meeting	2009-09-22 13:00:00	sensuous	19487	positive
6	MSNBC	Morning Meeting	2009-09-22 13:00:00	striking	19487	positive
7	MSNBC	Morning Meeting	2009-09-22 13:00:00	join	19487	positive
8	MSNBC	Morning Meeting	2009-09-22 13:00:00	president	19487	positive
9	MSNBC	Morning Meeting	2009-09-22 13:00:00	president	19487	trust
10	MSNBC	Morning Meeting	2009-09-22 13:00:00	change	19487	fear
11	MSNBC	Morning Meeting	2009-09-22 13:00:00	change	19487	fear
12	MSNBC	Morning Meeting	2009-10-23 13:00:00	commerce	19487	trust
13	MSNBC	Morning Meeting	2009-10-23 13:00:00	disagreement	19487	anger
14	MSNBC	Morning Meeting	2009-10-23 13:00:00	disagreement	19487	negative
15	MSNBC	Morning Meeting	2009-10-23 13:00:00	disagreement	19487	sadness
16	MSNBC	Morning Meeting	2009-10-23 13:00:00	subject	19487	negative
17	MSNBC	Morning Meeting	2009-10-23 13:00:00	change	19487	fear
18	MSNBC	Morning Meeting	2009-10-23 13:00:00	legal	19487	positive
19	MSNBC	Morning Meeting	2009-10-23 13:00:00	legal	19487	trust
20	MSNBC	Morning Meeting	2009-10-23 13:00:00	fight	19487	anger

Showing 1 to 20 of 10,175 entries

Figure 7: The tokenized dataset after applying the sentiment analysis, as denoted by the new field sentiment. An additional field station total gives the total number of words used by the station over the entire period.

It is quite a common approach in text analytics to operate on the individual word level (Grolemund & Wickham, 2017, Silge & Robinson, 2017). When using this approach with sentiment analysis, the sentiment content of the whole text is taken as the sum of the sentiment content of the individual words (Silge & Robinson, 2017). This is the approach followed in this work and is applicable due to the low word count of the news segments. If, however, one was interested in sentiment content at a finer level, like the sentence level, then negations need to be considered. Negations (not, isn't, wasn't, and aren't) alter the sentiment at the point where they occur, but are not considered here. Holtzman et al. (2011) analysed television transcripts from cable news during the 12-month period of 2008. They included a facility in their software to translate common negations into unique tokens. They gave the example 'He is **not** liberal' which would be reduced to the novel token 'notliberal', without a space, thereafter treated as its own word. Pang *et al.*(2008) also describe negations adding that 'not all appearances of explicit negation terms reverse the polarity of the enclosing sentence' and 'negation can often be expressed in rather subtle ways'. They discuss other methods to handle negations. Silge and Robinson (2017) describe a methodology to handle negations by using bigrams. When a negation is encountered in the text the sentiment was 'flipped' in order to correct the sentiment score.

Questions to pose the dataset

We will now use text analytics to answer the following questions in turn:

1. Which TV news station uses more positive or more negative words for this topic?
2. Which words contribute to the 10 types of sentiment scores? i.e. which words specifically are driving sentiment scores?
3. Which negative words did each news station use when talking about climate change?
4. How does sentiment change over time? Is the proportion of positive and negative words increasing or decreasing over time?

Often, as we process a question it will present new lines of discovery to be investigated, as we will see with Q2 below.

Q1. Which TV news station uses more positive or more negative words for this topic?

Methodology

As a first step we need to calculate the total words used by each station. This necessitates taking the tidy text data before the stop words were removed and the sentiment analysis was applied. The dataset is processed as follows:

1. Group the text data on the station column. This creates a copy of the data in 3 separate groups, that being the total number of stations. Any functions called subsequently will manipulate each group separately.

2. Count the total number of words used by each station and store in a variable station_total.

3. Apply the sentiment analysis using the nrc lexicon – this will add the sentiment column to the dataset.

4. Filter the data to display only records where the sentiment column has the value ‘negative’.

5. Count the total number of ‘negative’ words used by a station and store a variable n.

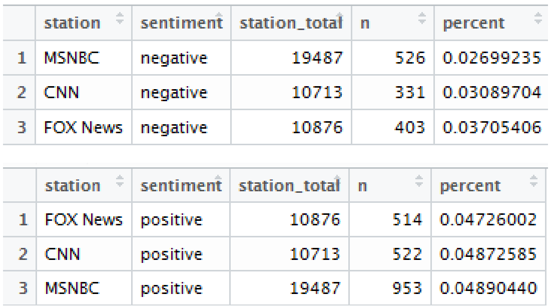
6. Calculate the percentage of negative sentiment for each station as divided by station_total and store in a variable percent.

7. Sort the data in ascending order of the percent column.

8. Repeat the process filtering for ‘positive’ sentiment.

Q1. results & discussion

The results shown in Figure 8. indicate that MSNBC used a low proportion of negative words (2.7%) but a high proportion of positive words (4.9%) when covering this topic. This would seem to agree with the liberal bias associated with the network as attributed by Holtzman et al. (2011). The opposite is the case for Fox News which tends to have a more conservative bias while CNN lies in between. However, since we are looking at sentiment at an aggregate macro level here, we need to carry out a more fine-grained level of analysis (e.g. topic-oriented sentiment analysis, aspect-oriented sentiment analysis) to fully support these hypotheses.



	station	sentiment	station_total	n	percent
1	MSNBC	negative	19487	526	0.02699235
2	CNN	negative	10713	331	0.03089704
3	FOX News	negative	10876	403	0.03705406

	station	sentiment	station_total	n	percent
1	FOX News	positive	10876	514	0.04726002
2	CNN	positive	10713	522	0.04872585
3	MSNBC	positive	19487	953	0.04890440

Figure 8: Results showing the total number of words used for the 3 news stations scored for (top) Negative sentiment and (bottom) Positive sentiment.

Q2. Which words contribute to the 10 types of sentiment scores? i.e. which words specifically are driving sentiment scores?

Methodology

The dataset is processed as follows:

1. Count by word and sentiment which creates a new column n, representing the number of instances of a word within its associated sentiment category.

2. Group by sentiment. This creates a copy of the data in 10 separate groups, that being the total number of sentiment categories. Any functions called subsequently will manipulate each group separately.

3. Filter out the top 10 words for each sentiment category by operating the top_n function on the value in column n, and then order the sentiment field alphabetically.

4. Ungroup the data – this returns the data to the ungrouped state but retains the results of the previous processing, namely the word field filtered to show only the top 10 with a corresponding word count field, n, and all the sentiment categories ordered alphabetically.

5. Reorder the data in descending order of the word count field, n.

6. Create bar plots with the word field on the y-axis, the number of uses n on the x-axis, and the sentiment category from the sentiment field appearing in its own pane.

Q2. results & discussion

The results in Figure 9 demonstrate which words contribute to the sentiment scores. We see the words 'terrorism' and 'abortion' feature in some news segments when discussing the topic of climate change. We also see that proper names like Gore and Trump, which should be treated as neutral, were scored for sentiment because of the common English meaning of those words. Gore appears in the anger, disgust, fear, negative and sadness sentiment categories, whilst Trump appears in the surprise sentiment category. It is important to see which words contribute to the sentiment scores so the sentiment lexicons can be adjusted if appropriate.

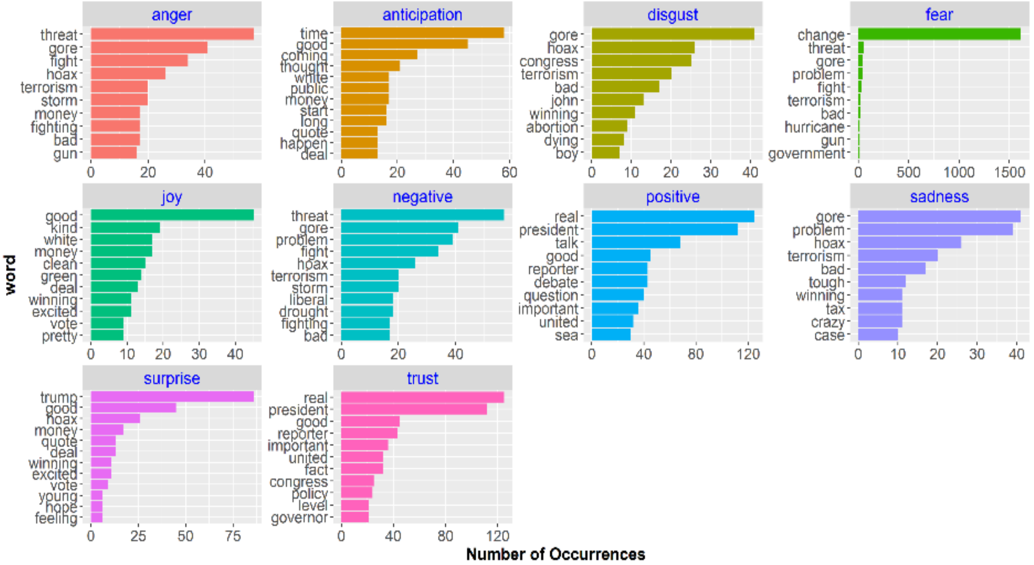


Figure 9: The top 10 words driving each type of sentiment category.

Altering the lexicon

These are just two word examples that stand out from the results. This demonstrates that a word that is actually a proper name may be incorrectly labelled as a particular sentiment, thereby altering the overall sentiment score.

We correct the scoring by removing these proper names from the scoring lexicon and running the code again to see what new words are driving the sentiment, as shown in Figure 10. This also demonstrates a possible limitation of using a lexicon. Although the publicly available *nrc* lexicon was used, the implementer can easily add, remove or alter the lexicon via code. This has obvious benefits to achieve a more precise sentiment analysis, but what if the intention of the implementer is to alter the scored sentiment for some advantage? This can be done just as easily, but is harder to detect if the code is not visible to an observer.

Additionally a word could refer to a proper name in one context but in another context actually belong to a sentiment category, the above cases of Gore/gore and Trump/trump being examples. This may occur very infrequently but nevertheless demonstrates a challenge when scoring a sentiment analysis implementation using a lexicon and one more difficult to address.

A final comment– when a word occurs in the text dataset but does not appear in the supplied lexicon, it will be treated as neutral even though it may have a sentiment that fits one of the 10 categories. This could be considered a limitation –the lexicon used may not be appropriate for the text being scored; or the lexicon may be appropriate in scoring one body of text but not another due to the vocabulary used for the topic; or no lexicon is appropriate due to some characteristic of the text being scored such as the writing style (sarcasm) or word usage (early or modern English).

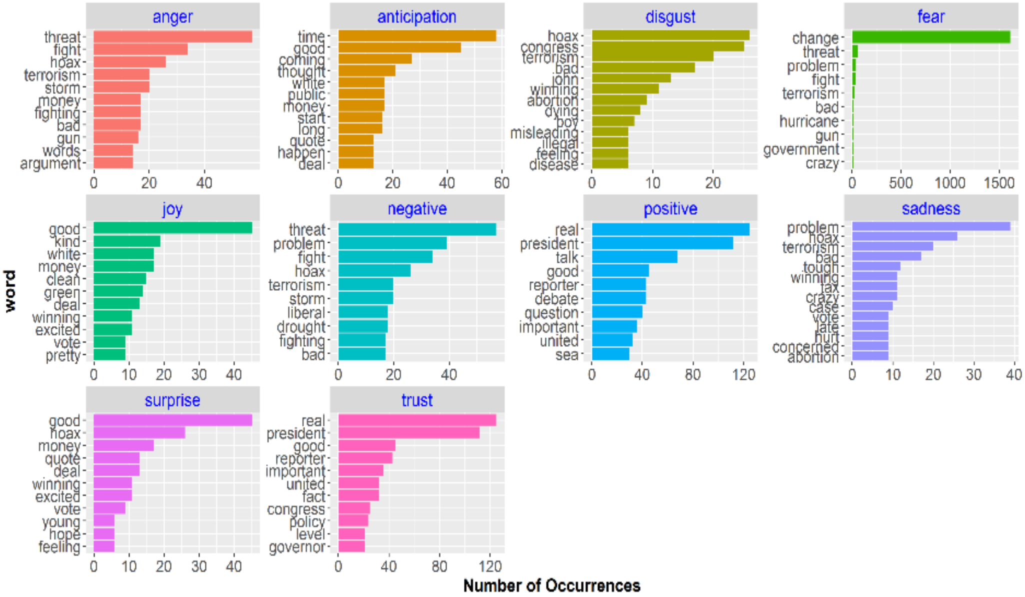


Figure 10: The top 10 words driving each type of sentiment category after filtering out Gore and Trump.

Q3. Which negative words did each news station use when talking about climate change?

Methodology

The dataset is processed as follows:

1. Apply the sentiment analysis and filter the sentiment column for only 'negative' words and also, filter out proper names as above.

2. Count by word and station which creates a new column n, representing the number of occurrences of a negative word, on a per station basis. This advances the goal of determining which words are contributing most overall to the negative sentiment scores.

3. Group the text data on the station column. This creates a copy of the data in 3 separate groups, that being the total number of stations. Any functions called subsequently will manipulate each group separately

4. Filter out the top 10 words for each station by operating the top_n function on the value in column n.

5. Ungroup the data – this returns the data to the ungrouped state but retains the results of the previous processing.

6. Reorder the data in descending order of the word count field, n.

7. Create bar plots with the wordfield on the y-axis, the number of uses n on the x-axis, and the station field appearing in its own pane.

Q3. results & discussion

The results in Figure 11 display the top 10 word choices used by each station when conveying a negative sentiment. Some words, like ‘threat’ are used by all three stations but some word choices are quite different. CNN has ‘hoax’ as its top choice. Curiously, the more conservative Fox News includes ‘terrorism’ in discussing the topic.

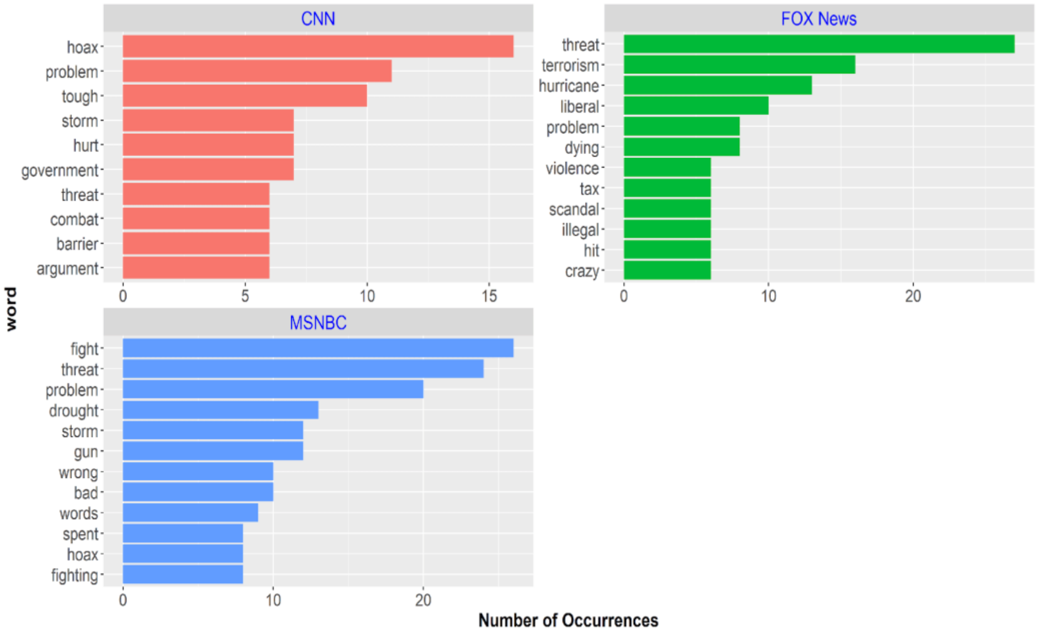


Figure 11: Word choices used by TV station to express negative sentiment.

Q4. How does sentiment change over time? Is the proportion of positive and negative words increasing or decreasing over time?

Methodology

The dataset is processed as follows:

1. Define a new column called date and assign it the value of the date in the show_datefield rounded down to the nearest 6-month unit.

2. Group by the new date column so that now all data is grouped into its corresponding 6 month period.

3. Count the total number of words used in each 6 month period –assign the value calculated to the field total_words.

4. Ungroup the data and implement the sentiment analysis– this will add the sentiment column to the dataset.

5. Filter the sentiment field for 'positive' and 'negative' words in order to plot both and also filter out Proper names as above.

6. Count by using the three fields date, sentiment and total_words. This creates a new column n, representing the counts of a word within each 6 month period.

7. Ungroup the data – this returns the data to the ungrouped state but retains the results of the previous processing.

8. Calculate both the negative and positive sentiment percentage for each 6 month period as n divided by total_words and store in a variable percent.

9. Set up your plot with date on the x-axis, percent on the y-axis, and have colour correspond to sentiment.

10. Fit dashed trend lines through both sets of data.

Q4. results & discussion

The results in Figure 12 chart the positive and negative sentiment trend over time. The periodic trend fluctuates with notable spikes at certain periods. The overall proportion of positive words appears to be relatively unchanged since 2009 at around 5% of the total word usage for the topic. The proportion of negative words seems to be increasing, having had large increases in the first halves of 2014 and 2015, possibly related to the run-in to the 2016 presidential election.

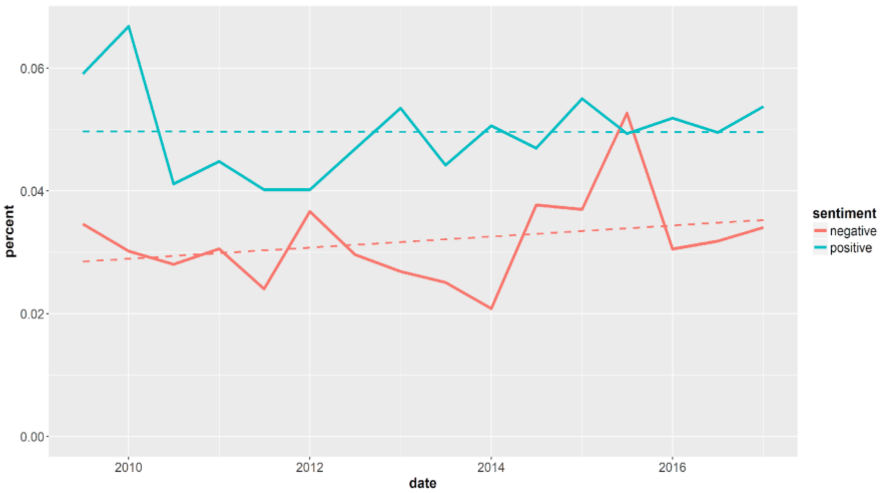


Figure 12: Overall positive and negative sentiment over time.

Conclusion

Sentiment analysis, as an added tool in the process of information mining, is used by different industries in order to identify trends and opinions that are inferred from language usage. These text analytics tools are being incorporated into different decision making mechanisms, in some cases, as black boxes or algorithmic codes that are difficult to interpret by the final user of the tool.

In this paper, we presented a step by step guide of the sentiment analysis process that will help the final user understand the process and the dangers and possible bias (deliberate or not) included in the results.

The sources from which the information is obtained should be identified and referenced. This allows the final user of the tool to know the possible sources of bias. For example, a news summary product using different websites may want to limit the sources to those that follow a specific political line, or on the other hand, a more balanced view might be sought and sources from different sides of the political spectrum may be used. An example of how to identify the political line of a newspaper can be obtained using sentiment analysis for a particular topic of enquiry such as climate change, as inferred in experiment 1. In this experiment a lexicon is used to identify the positive or negative sentiment of the news associated with climate change.

In addition to the sources of information a lexicon is another external resource used in sentiment analysis. Lexicons are compilation of words with a sentiment attached to each of them. Lexicons are commonly available; however, they can be easily modified to customize the results obtained. For example, the sentiment of certain words can be modified and changed from positive to negative. This could help the customization of the system if the final user is aware and in control of the process. On the other hand, if the final user has no control or knowledge about the lexicon and the way it is defined, the results could be biased, or intentionally used to deceive or suppress negative opinion. The stop words list was also identified as influential in the processing pipeline, as it is externally generated and alterable and affects the text available for sentiment scoring downstream.

The paper demonstrated the steps carried out in sourcing, importing and pre-processing some text data using the RStudio software and the R programming language. Techniques were demonstrated as well as explaining key terms in the text analytics pipeline. After the sentiment analysis was carried out using a lexicon, we further processed the text to answer posed questions such as what words drive a particular sentiment category, how the news topic vocabulary varies by news network, and how sentiment changes over time.

We saw that although we may begin with a large unstructured dataset, we can perform a series of text analytics techniques to tidy and process that data to lead to interesting visualisations and insights about the data.

References

Årup Nielsen, Finn (2011) AFINN Sentiment Lexicon [Online] Available at:http://corpustext.com/reference/sentiment_afinn.html
(Accessed: 04 November 2017).

Liu, Bing et al. (2004) Opinion Mining, Sentiment Analysis, and Opinion Spam Detection[Online] Available at:https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon(Accessed: 04 November 2017).

Economist, (2017) Climate-Change Articles, The Economist Newspaper Limited. [Online] Available at: https://www.economist.com/topics/climate-change (Accessed: 04 November 2017

GDELT, (2017a) The GDELT Analysis Service offers a variety of tools and services that allow you to visualize, explore, and export the Global Database of Events, Language, and Tone (GDELT) project - a real-time database of global society. [Online] Available at:
http://analysis.gdeltproject.org
(Accessed: 04 November 2017).

GDELT, (2017b) The GDELT Television Explorer tool.[Online] Available at: http://television.gdeltproject.org/cgi-bin/iatv_ftxtsearch/iatv_ftxtsearch
(Accessed: 04 November 2017).

Grolemund, Garrett & Wickham, Hadley (2017) R for Data Science - Import, Tidy, Transform, Visualize, and Model Data, O'Reilly Media, Inc. [Online] Available at: http://r4ds.had.co.nz
(Accessed: 04 November 2017).

Holtzman, Nicholas S., Schott, John Paul, Jones, Michael N., Balota, David A. & Yarkoni, Tal (2011) Exploring media bias with semantic analysis tools: Validation of the Contrast Analysis of Semantic Similarity (CASS) Behavior Research Methods 43:193–200.

Internet Archive's TV News Archive (2017).[Online] Available at: https://archive.org/details/tv
(Accessed: 04 November 2017).

Kwartler, Ted (2017) Text Mining: Bag of Words.[Online] Available at: https://www.datacamp.com/courses/intro-to-text-mining-bag-of-words
(Accessed: 04 November 2017).

Mohammad, Saif and Turney, Peter (2011) NRC Word-Emotion Association Lexicon (aka EmoLex), National Research Council Canada (NRC). [Online] Available at:
http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm (Accessed: 04 November 2017).

Mullen (2016) R Package Tokenizers. [Online] Available at: https://cran.r-project.org/web/packages/tokenizers/tokenizers.pdf
(Accessed: 04 November 2017).

New Scientist, (2017a) Climate Change Articles, [Online] Available at https://www.newscientist.com/article-topic/climate-change (Accessed: 04 November 2017

New Scientist, (2017b) Special Report – How to live with climate change and how to beat it, 24 June 2017, No 3131, pages 28-35.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1-135.

RStudio(2017) an integrated software development environment. You can download and install it from http://www.rstudio.com/download

Silge, J. (2017) Sentiment Analysis in R: The Tidy Way. [Online] Available at: https://www.datacamp.com/courses/sentiment-analysis-in-r-the-tidy-way (Accessed: 04 November 2017).

Silge,J. &Robinson,D. (2017) Text Mining with R - A Tidy Approach, O'Reilly Media, Inc.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

Wickham,H. (2014) Tidy Data, Journal of Statistical Software, Volume 59, Issue 10. [Online] Available at: https://www.jstatsoft.org/article/view/v059i10 (Accessed: 04 November 2017).