"What's in a *keyword*?"
An analysis of keywords in social media posts and their implications

Jeeyoon Park
Columbia University
136 W 109th Street
New York, NY 10025
+ 1 312 298 9228
jp3380@columbia.edu

Lleyana Jack
Columbia University
119 W 104th Street
New York, NY 10025
+1 734 635 7023
loj2103@columbia.edu

Rachel Wang
Columbia University
205 W 103rd Street
New York, NY 10025
+1 347 574 1268
sw2854@columbia.edu

ABSTRACT

Text analysis is a relatively modern concept from the late 20th century that stems from data mining. The purpose is to automate the navigation of text usually through statistical pattern matching algorithms and drastically reduce the time and cost of human labor. For this research, we used a dataset that contained Facebook, Twitter, and federal news release information for 2012. A lexicon-based model of sentiment analysis was applied identifying positive, negative, strong and weak emotions in a dataset. In addition, the data was explored using other text analytics techniques to answer research questions, test hypotheses, and reverse engineer information regarding the current events at the time the data was collected. The results of the analyses were visualized. The results showed that: 1) usage patterns for Facebook, Twitter and news are lowest on Mondays and Tuesdays 2) And social media data, Facebook and Twitter, had higher positive, strong and weak word matches than news data 3) There was a 5 hour lag time between news Release updates and Facebook for the same keywords. 4) The trending keywords show a close relationship with the timeline of major events.

Categories and Subject Descriptors

D.2.10 [**Design**]: Methodologies and Representation – *sentiment analysis, keyword visualization.*

General Terms

Algorithms, Measurement, Design.

Keywords

Social media, keywords, sentiment analysis, text data visualization.

1. INTRODUCTION

As Shakespeare asked, "What's in a name? Would a rose by any other name smell as sweet?" The same types of questions can be asked in regards to the keywords of social media posts. Does a keyword represent something more, something deeper than its superficial label? Do they betray users true emotions or portray other relevant information? What can one imply about the situation at hand based on these words? Thanks to technological developments and a new era in data analysis, we are able to attempt to answer these questions and explore the impact of keywords in social media posts.

In this day and age, data is no longer just a quantitative concept limited to numbers alone. We are able to utilize text analytics technique to explore a wealth of qualitative information. This is not just limited to keywords of social media posts, but to the posts in and of themselves, blogs, websites, and entire social media platforms. In fact, the Internet in general serves as a massive collection of information on a range, scale, and granularity never known to mankind before. However, the issue lies in how to process such large amounts of text.

Text analysis is a relatively modern concept from the late 20th century that stems from data mining. The purpose is to automate the navigation of text usually through statistical pattern matching algorithms and drastically reduce the time and cost of human labor. It is estimated that over 80% of information we have today is stored as loose text[1] Therefore, text mining is a powerful tool in making sense of the vast majority of information which is qualitative data.

Text analysis requires a series of data munging and manipulating tasks prior to analysis. The first step is to structure the entirety of text, commonly referred to as corpus, into a suitable format. This process requires parsing the data into base units – for example, by document, sentence, or word. Next, the text data needs to be standardized in format by reducing abnormalities such as special characters, punctuation, or numbers. At this point, user-defined stopwords, or meaningless particles of speech including pronouns, prepositions, and articles, can be removed to make the analysis more efficient and meaningful. The remaining words should be stemmed, shortened by removing gerund endings and conjugations, to be reduced into their most basic forms. Once the data is cleaned and structured into a desired format, the text can be analyzed in ways suitable to the research. This paper is focused on keyword extraction and sentiment analysis.

Sentiment analysis methods can take one of three forms – lexicon based models, rule based models, or machine learning models. Lexicon based models use individual words as the base unit of

[1] http://breakthroughanalysis.com/2008/08/01/unstructured-data-

analysis. These words are matched to a predefined ‘bag of words’, or dictionary, for specific sentiments. Lexicon based models are the simplest form of sentiment analysis. However, due to its simplicity, it cannot detect dual negatives and positives or correctly process complex sentiments, such as sarcasm. Rule based models define sentiments based on preset rules. For example, rule based models work by assigning a particular label to a sentence or document containing specific words or patterns of words. Machine learning based models are based on algorithms that sort text by emotion. These methods are effective in handling commonly repeating patterns such as Twitter hashtags, or emoticons.

For this research, a lexicon-based model of sentiment analysis was applied identifying positive, negative, strong and weak emotions in a dataset. In addition, the data was explored using other text analytics techniques to answer research questions, test hypotheses, and reverse engineer information regarding the current events at the time the data was collected. Lastly, the results of the analyses were visualized.

2. LITERATURE REVIEW

Text mining and sentiment analysis is a convenient way to process large amounts of text data without having to use human labor to manually read through the information. However, with convenience comes with limitations. Text mining is an efficient way to pick up large patterns that might exist in the text, but the researcher does not have a method to test the accuracy of the patterns found in the analysis (Don et. al., 2007). Text mining, especially opinion or sentiment analysis, also depends heavily on the type of documents as text analysis models are built on recurring patterns within text. Once a rich and sufficient size corpus that could identify major patterns is set up, text analysis has a large variety of applications from understanding online reviews to government uses (Pang & Lee, 2008).

As Pang et. al, discuss, a major hurdle in sentiment analysis, as opposed to document sorting, is the small number of generalizable categories, whether it be star-rating or “likes”, applicable across users and platforms (Pang & Lee, 2008). Another challenge is dual meaning in words and multiplicity of emotions within text (Pang & Lee, 2008). It is common for text that contains subjective information such as personal views or opinions to contain multiple, even conflicting emotions.

Liu et. al, tried a more complex approach in their research of analyzing online customer reviews across multiple platforms and products. The researchers wanted to understand what the customers were complaining or liking about the product in their reviews. In order to accurately depict and most importantly, compare reviews, the researchers used two methods. First, they studied the implicit and explicit features of the opinions. (Liu et. al, 2005). Second, they used a defined sequence of words as opinions, rather than simply matching words to sentiment dictionaries. These methods helped researchers accurately analyze one sentence that discusses multiple features and opinions about a product, as well as accurately parse out the negatives and positives in the entire review based on explicit and implicit product features. (Liu et. al, 2005)

3. MOTIVATION

Interests around text analysis are continuously growing. Yet, the barriers of entry into text analysis are quite high due to its firm background in statistics and mathematics. Without a clear understanding of the algorithms, applying complex text analysis models are a daunting and dangerous activity.

Existing research is focused on intricate and complicated text analyses models that perform well across various types of text. However, this paper is focused on applying a simple lexicon based sentiment analysis and word frequency counts to test whether these rudimentary models can still outline the major patterns in large amounts of text data.

Our motivation in this research was to make text mining more approachable and easy to understand through the use of simple models and visualizations.

4. RESEARCH QUESTIONS

This paper analyzes the relationship between social media posts and time. More specifically, it looks at the relationship between time and trending keywords from both narrow and wide scopes and perspectives.

4.1 Daily Frequencies

Among the social media and news release data, we wanted to find out what the frequency of keyword represented for each day across the entire year and if there were any recurring patterns. *Hypothesis 1*: This is an exploratory analysis technique and therefore does not have a typical hypothesis. However, we expect to find a relationship between the most frequent keywords and the state of the world on the day being analyzed.

4.2 Usage Patterns

What is the pattern of social media use over days of the week and months of the year, compared to news release activity? Do social media updates outpace news release in number and frequency? *Hypothesis 2*: We hypothesize that social media update activity would be higher across major holidays and weekends.

4.3 Sentiments

Are there any patterns in the sentiment of the keywords in social media posts that are different than that of news release information?
Hypothesis 3: We hypothesize that news data will have the least amount of sentiments expressed across the three media platforms, while Facebook and Twitter will score higher than news but similar to one another due to the similarity in the products’ function.

4.4 Facebook Timeline

We want to explore when Facebook posts are published online. In present times, Facebook is used by millions of users every day. The ability to crowd source information in real time has the potential to surpass the ability of more traditional news outlets. In this age of technology, is it possible that Facebook users are able to distribute the news faster than traditional news media sites?

Hypothesis 4: We hypothesize that Facebook posts, on average, will be published before news updates.

4.5 Top Keywords and Events

We are interested in how the top keywords are distributed over time. How does the change in distribution tells us about what was happening at those times? We wanted to explore whether we can recreate the stories about what was happening in the world at that time by looking at the data.

Hypothesis 5: Similar to Hypothesis 1, this is an exploratory analysis. Therefore, we are looking more for the patterns. This exploration however seeks to compare the change in the distribution of the frequency of words to current events instead of focusing on just one day.

4.6 Keyword Social Network

We are interested in how the keywords are connected to and influencing each other in the social network. As information in a social network pass on node by node, it would be interesting to see how the top keywords are related to each other in graphics.

Also, the long tail theory assumes some distributions of numbers is the portion of the distribution having a large number of occurrences far from the "head" or central part of the distribution. We are intrigued by the “keyword long tail” and want to test its validity.

Hypothesis 6: There is a long tail effect in the long tail theory, and the top keywords should be in the abstract center of keywords social network.

5. DATASET

The data we used was from Voxgov, an organization dedicated to providing information on US federal government through news and document search as well as alerts. The information contained Facebook, Twitter and federal news release information (from here on out just called Release) for the calendar year of 2012 formatted as JSON (JavaScript Object Notation) files. The analyses used the entire data set, aggregating data from all three sources for all 366 days of 2012 as it was a leap year .

A function was constructed in R to automatically collect data from each day for Facebook, Twitter, and Release data respectively. The function used the `rjson` package to convert the JSON files into a more easily parsed format in R. Then it iterated through all the files in a category (Facebook, Twitter, or Release) to extract selected subcategories within a single JSON file. We selected keywords and date from each file which corresponded to a day in the year 2012 and combined the entire year’s results back into a single csv for each selection.

6. VISUALIZATION DESIGN

We constructed three types of visualizations to answer our research questions and test the hypotheses. To understand aggregate activity and sentiment level by the months of the year and days of the week (Research Questions 4.2 & 4.3), we used Tableau to create static visualizations and the results were compared side by side. We also used static visualization to show the results of Research Question 4.4, looking at the publishing time differences between Facebook and news Release posts. In order to examine top key words by date (Research Question 4.1), we used the `shiny` package in R to design an interactive word cloud that produces the top key words by user defined date, frequency, and number of words. Lastly, to understand the relationship between trending key words and events, we mapped the top trending key words in December 2012 by date using an interactive data driven document created using JavaScript (Research Question 4.5).

6.1 Static Visualization

The first set of static visualization, Figures 1 and 2, attempt to answer Research Question 4.2 regarding usage patterns on social media. In this chart, the number of updates for each medium, Twitter, Facebook and News were collected for each day of the year. Then, this information was plotted by months of the year and days of the week to understand the activity level of each medium. The results showed that Twitter had the highest number of activity measured in number of updates among the three media. This is not surprising given the nature of Twitter in comparison to news or Facebook. Twitter is a platform where users can share short from (140 characters or less) updates. Because the length of each update is limited, often times, users share multiple updates a day. Also, the majority of publishers partner with Twitter to distribute information on trending articles and breaking news. Facebook also partners with Twitter to enable users to link their Twitter and Facebook accounts together. As we did not check for unique users by ID or select for updates unique to each media, it is entirely possible that the high activity on Twitter is an endogenous value that accounts for both Facebook and news releases.

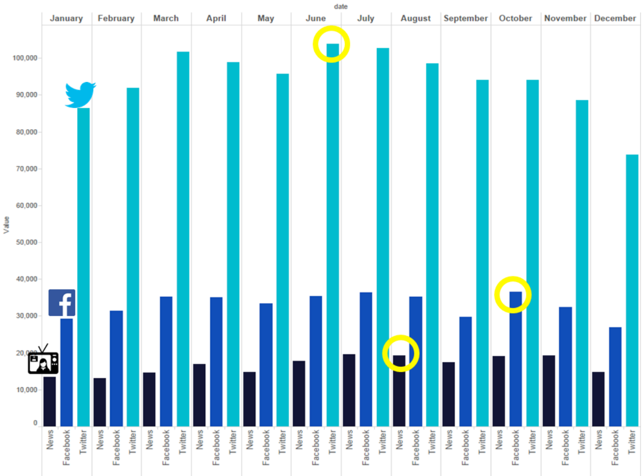


Figure 1 Activity Level by Months of the Year
Twitter activity is highest in June and lowest in December. Facebook activity is highest in October and lowest in December. News activity was highest in August and lowest in January of 2012.

Given the possible endogeneity in Twitter data, it is still interesting to observe that activity for all media, including news release, was lowest on Mondays and Tuesdays. While news activity increased from the beginning of the week and peaked on Thursdays, social media activity on Twitter and Facebook were highest on Fridays.

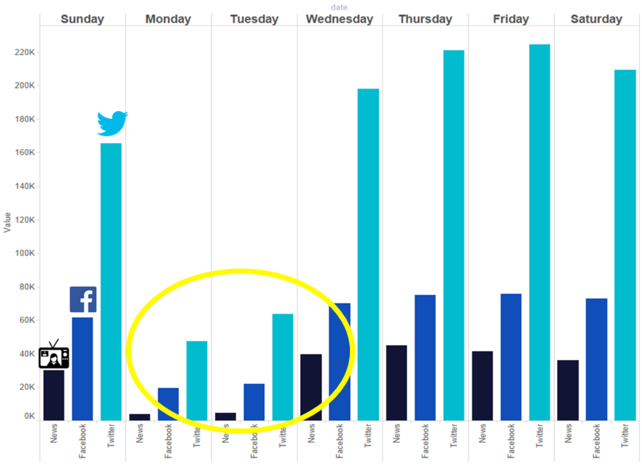


Figure 2 Activity Level by Days of the Week

Activities for Twitter, Facebook and news were lowest on Mondays and Tuesdays but increased rapidly starting half way through the week on Wednesdays.

The next set of static visualizations, Figures 3-6 and Table 1, address Research Question 4.3 regarding the sentiment of the keywords. A lexicon based sentiment analysis was used on the keywords from Twitter, Facebook and news release data. The parsed JSON keywords were combined by date and then again by source. Regular expressions were used to remove special characters and to change all the text to lower case. The dictionary for the lexicon based sentiment analysis was selected from two sources. One was from the opinion lexicon dictionary by Bing Liu at the University of Illinois in Chicago (UIUC).[2] The positive and negative word dictionary from this opinion mining, sentiment analysis and opinion spam detection word base contains an extensive number of words that had a higher match rate than the Harvard lexicon dictionary.[3] In other words, more words in our corpus corresponded to the words included in this positive and negative word dictionary than that of Harvard's.

The other source was the Harvard lexicon dictionary, which contained a larger variety of sentiment word dictionary than the one from Liu at UIUC. From this database, dictionaries for strong and weak words were selected. Strong words contain terms that indicate power, strength or intense emotions such as president, energy and love. Negative words contain terms that refer to submission or lack of resources such as debt or deficit. Keywords were matched to the four sentiment dictionaries for positive, negative, strong and weak.

The matching process used a function in R that counted the number of corresponding words in each dictionary and outputted the words within a given dictionary and the frequencies of each word's appearance in our corpus. This resulted in a table of words found in the four sentiment dictionaries and number of times each word appeared. A higher frequency indicated more frequent appearances of this word.

For all four sentiments displayed in Figure 3 through Figure 6, we observe that all three media show strong fluctuations between high and low emotion on a day-to-day basis. On a scatter plot, it seems like all three media have two streams – high match updates and low match updates.

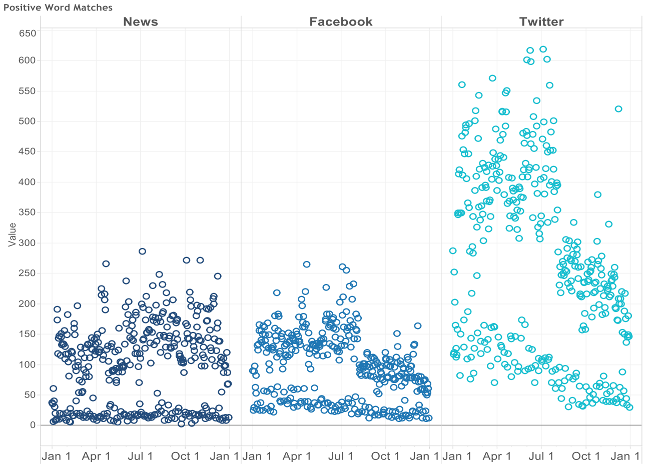


Figure 3 Positive Word Matches by Media

Positive word matches are significantly higher for Twitter showing that more words in the Twitter corpus matched words in the positive sentiment dictionary. Facebook and News matched at similar levels in positive sentiments. Positive sentiments on Twitter and Facebook dropped around late July of 2012.

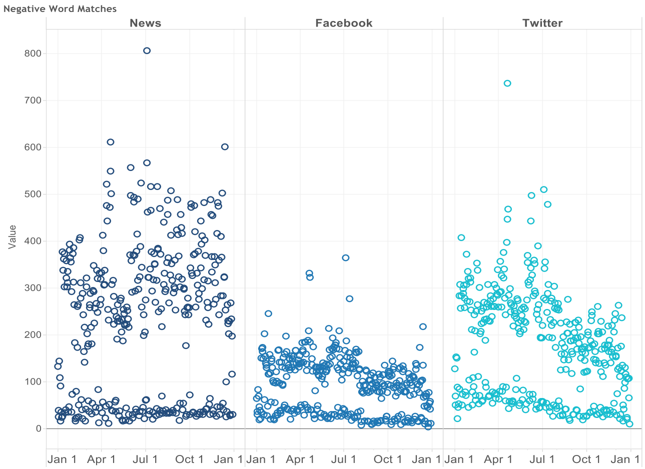


Figure 4 Negative Word Matches by Media

Facebook had the lowest negative word matches compared to Twitter and news. For both Facebook and Twitter, negativity dropped during the second half of the 2012. News release had a higher number of negative word matches throughout the year and match frequency increased during the second half of the year.

Twitter has more positive, strong, and weak word matches in keywords among the three media. Interestingly, Twitter's high match scores for positive, strong and weak words all fell sharply around late July of 2012. Although not as drastic, Facebook showed a similar pattern to the same three sentiments as well. News release had a lower positive match score, but higher

[2] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[3] http://www.wjh.harvard.edu/~inquirer/Strong.html

negative latch score compared to the other two media. Strong and weak word match scores were somewhat consistent for news throughout 2012.

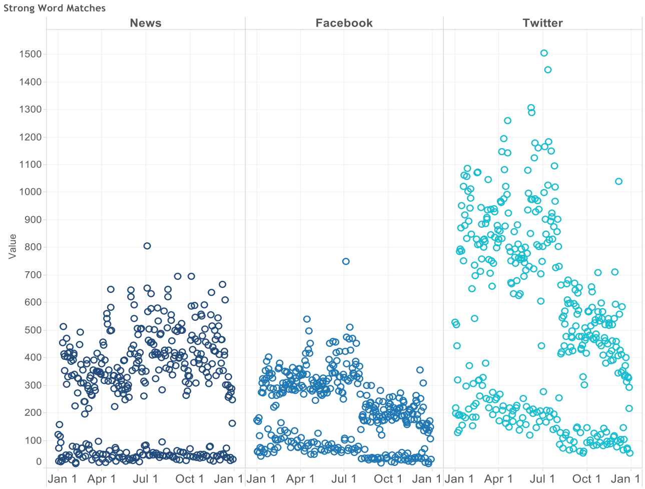


Figure 5 Strong Word Matches by Media

Twitter scored the highest strong word match. Strong word matches for Facebook and Twitter decrease around the same time positive word matches decrease on these two platforms.

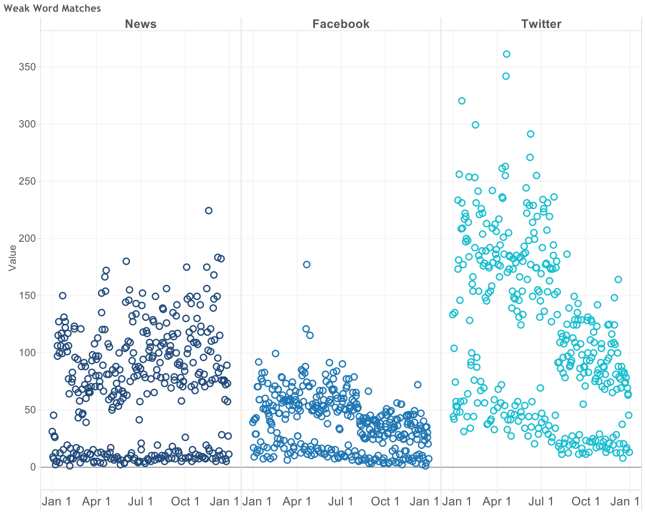


Figure 6 Weak Word Matches

Weak word matches closely mimic the pattern of strong words match in Figure 5. Again in Facebook and Twitter, sentiment drops around the end of July. News had a higher negative word than positive word match but did not show particular differences in strong or weak word match.

Next, we examined the top 10 most frequent words in each emotion. The matched words were stemmed to decrease redundancy. Same words of different tense or grammatical form (noun versus verb, for example) were collapsed into one word. It is here where the caveat of lexicon based word matching became clear. There was no way to tell whether certain words such as "fall" were referring to the season or action. While certain words were not clear on their meaning, the majority of the words that appeared on the top list of each sentiment were sensible.

Table 1. Top 10 Key Words by Sentiment

Strong	Positive	Weak	Negative
Energy	Congratulation	Help	Cancer
Join	Support	Support	Miss
Headquarters	Glad	Miss	Debt
President	Welcome	Debt	Fraud
Love	Love	Fraud	Prison
Support	Great	Disaster	Drought
Great	Winner	Deficit	Disaster
Violence	Work	Pass	Terrorism
Fort	Honor	Wish	Attack
Ready	Ready	Depression	Conspiracy

The last set of static visualizations, Figures 7 and 8, pertain to Research Question 4.4, regarding whether, on average, Facebook or traditional new sites publish new stories first. Based on earlier findings during our analyses regarding the behaviour of Facebook and news Release keyword data, it has been established that the keyword behaviour for Facebook and the Release data are very similar. Keeping that in mind, we decided to compare the difference between the posting times for Facebook and Release data to see which, on average, was published first. To begin, we set our function to select for dates in addition to keywords. The dates were encoded as character strings and were in JSON format, which meant they were counted as milliseconds from midnight January 1st, 1970 and preceded by the word "Date" along with some other random punctuation marks. Regular expressions and scripting was used to clean the data and convert it to numeric values that could then be averaged. Then, the average posting times for each day for the Facebook data was subtracted from the average posting time for each day for the Release data. Afterwards, these numbers were converted into hours since the original times were in milliseconds, and plotted in Figure 7.

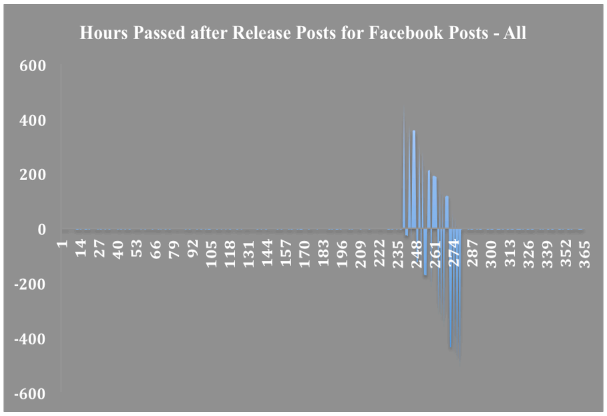


Figure 7 Published Time: Release vs. Facebook

This graph shows the time differences in hours on the y-axis. The x-axis shows the days in 2012 in order. There are 366. When the blue is above the x-axis, the positive range of Y, it means the posts were published on Facebook before the news. When the blue area is below the x-axis, it means the keywords were present on the news Release before appearing on Facebook.

The range of the y values in the graph in Figure 7 is from positive to negative 600. However, in a single day, no post could be more than 24 hours later or earlier than any other. That indicates some obvious some flaws in the data

Therefore, another visualization was created, Figure 8, which restricted the data to only include days where the average difference between times of the Facebook posts and news Release posts were 24 hours or less.

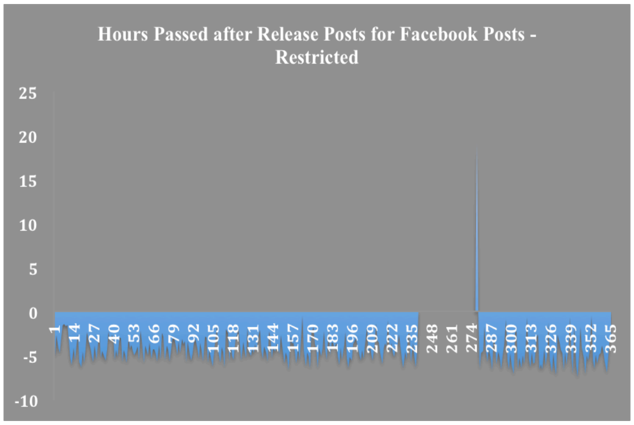


Figure 8 Published Time: Release vs. Facebook (restricted)

Once again, time differences in hours are on the y-axis. The 366 days of 2012 are on the x-axis. [4]

In Figure 8, we can see most of the activity in blue is in the negative region of the y-axis, around -5, which represents 5 hours later. This indicates that. on average, Facebook published the posts of the same keywords 5 hours after the news release published these stories.

6.2 Interactive Visualization Using Shiny

`Shiny` was used to explore Research Question 4.1, regarding the daily word frequencies. `Shiny` is a package in R that allows users to build custom interactive web applications. It consists of two main R script components, a ui.R script, which contains the code that defines for what the users see when interacting with the interface, and a server.R script, which has the code for the behind-the-scenes processes that make the visualization work. Those are the two main components, but there may also be additional scripts, for example, a global.R script, which dictates extra functions or information that is accessible by both the server and ui scripts. In addition to `shiny`, an add on package called `shinyIncubator` was also used to create this visualization.

For this visualization, all 3 of these scripts were used. In the ui (user interface) script, control widgets for date, word frequency, and total number of words were used. These dictated the input the user could enter and in turn that input was passed onto the server script. `Shiny` enables interactive use by having reactive capabilities; it allows the builder to link the user input to the behind the scenes calculations and update what the user sees accordingly.

On the back end, the server and global scripts contained code that performed text analyses on the data. The global script contains a function using the `tm` package that takes the keywords documents that were created earlier and formats them into corpus, where they are then stemmed, cleaned, stripped of stopwords[5], and converted into a term document matrix.

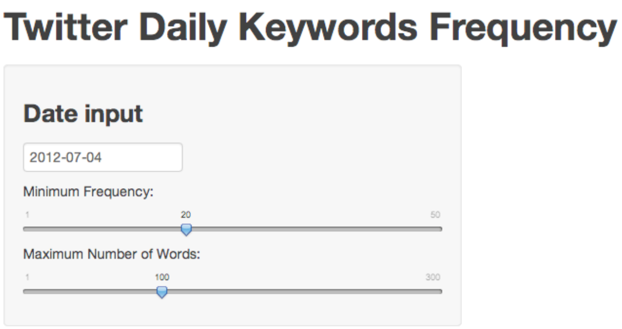
The server script used the term matrix from the global script to create a wordcloud using the `wordcloud` package. The function was adjusted to make some usually static aspects reactive so they would be sensitive to user input. Since the keywords were listed in a data frame and indexed by numbers 1-366, some code was written to transform the date inputs in the format "yyyy-mm-dd" into a single integer. Lastly, a qualitative colour palate[6] from the `RColorBrewer` package was used for

[4] At this moment in time, there is no explanation for the spike in the data around day 274, September 30th, 2012.

[5] We also added our own stopwords to the package. For example, for each day, the original vectors of keywords were all different lengths. However, they needed to be uniform and formatted in a data frame for the functions to iterate through them correctly. So, one of the author's very unique names was imputed into the blank cells to create vectors of uniform lengths. However, that name ended up being the most frequent word per day! Therefore, the name was added to the stopword list.

[6] "Qualitative palettes, do not imply magnitude differences between legend classes, and hues are used to create the primary visual differences between classes. Qualitative schemes are best suited to representing nominal or categorical data" (Harrower & Brewer, 2011).

the words in the word cloud. Figures 9 and 10 show the user input and resulting word cloud, respectively.



Twitter Daily Keywords Frequency

Date input

2012-07-04

Minimum Frequency:

1 20 50

Maximum Number of Words:

1 100 300

Figure 9 Shiny Data Visualization input

This tool enables the user to display the most frequent key words for any given date within the dataset. In this example, the input date is set to July 4th, 2012.

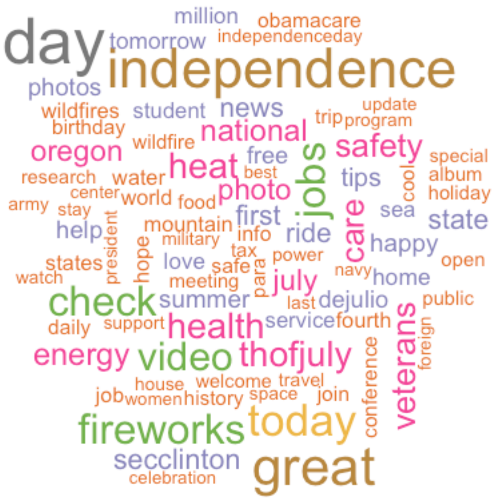


Figure 10 Shiny Data Visualization Result

The resulting word cloud based on the input for July 4th, 2012. The size of the word corresponds to its frequency. Even without knowing that it was Independence day, it would be easy to reconstruct what was happening on that day by the popularity of the keywords such as 'independence', 'day', 'fireworks', and 'thofjuly'.

6.3 Interactive Visualization Using D3

A data driven document, was created using JavaScript to explore Research Question 4.5, how are the top keywords distributed over time. For this visualization we focused on just the Facebook keywords in December 2012 to enable us to get depth and dig deeper into the meaning behind the data. After running text analyses, the top 10 most frequent words of the month were selected and their distributions were visualized in a streamgraph (Figure 11).

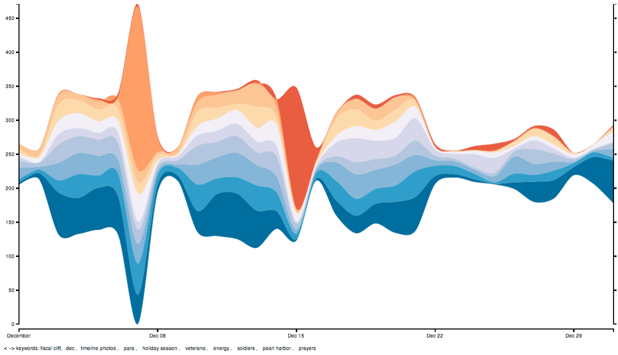


Figure 11 Streamgraph - December 2012 : Top 10 keywords

Across the x-axis are the dates in December 2012. The y-axis represents the frequency of the word. Each word is represented by a different color. The top 10 words (fiscal cliff, dec, timeline photos, para, holiday season, veterans, energy, soldiers, pearl harbor, prayers) can be found at the bottom of the screen

Table 2. Top Ten Keywords and Their Frequency

keywords	fiscal cliff	dec	timeline photos	para	holiday season
frequency	931	583	573	440	406
keywords	veterans	energy	soldiers	pearl harbor	prayers
frequency	401	336	325	311	299

The streamgraph enables users to highlight a single point on the graph or observe how the distribution changes over time. In Figure 12, the highlighted keyword is 'prayers'.

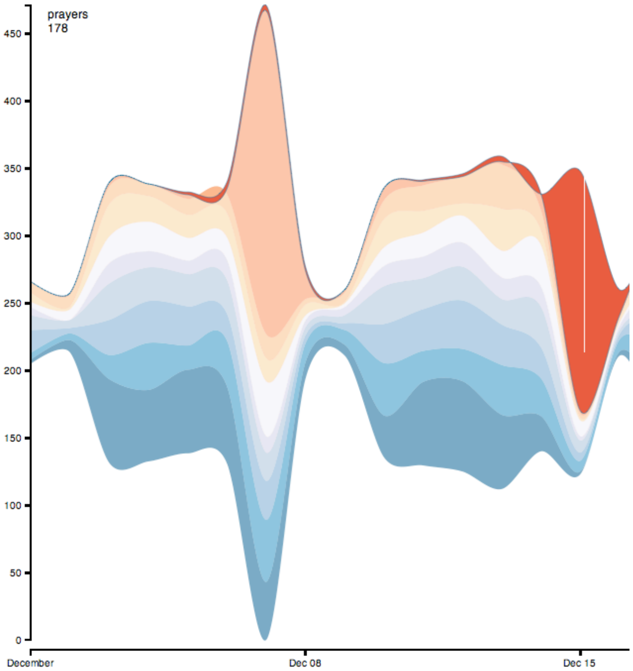


Figure 12 Streamgraph - Close-up: "prayers" on Dec 14th

This close up of the streamgraph showcases a user highlighting the date December 14th and the keyword 'prayers,' which was the top word was for that day. In the upper left hand corner, we can see that the word appeared 176 times in the keywords for that day.

However if the user were to mouse over other areas of the graph, they would see that distribution of the word 'prayers' changes dramatically from appearing 0 times on December 13th to appearing 176 times on December 14th. Linking this to the current events at that time, there was the now infamous Sandy Hook elementary school shooting on December 14th, 2012 in Connecticut where 20 children were killed.

Another one of the major keywords of December 2012 was fiscal cliff, which appears as the darkest blue stream at the bottom of the graph. The interest was sparked by then Federal Reserve Chairman Ben Bernanke's speech on the fiscal cliff regarding the coming year 2013. In early December, Pearl Harbor was a trending keyword in recognition of the attack on Pearl Harbor on December 7th 1941. Trending keyword Pearl Harbor appears as the salmon colored spike at the top of the graph in early December.

By focusing in on this subset of data, we can easily observe that keywords are often driven by major events. This visualization confirms that social media is closely connected, if not reflective of our reality and we can uncover major happenings and stories by looking at trending keywords in a given time frame.

6.4 Keywords Social Network Analysis Using Gephi

The cleaning of this dataset is very different from the measures mentioned above, as we need to preserve the information of links between different keywords. By seeing the keywords that appeared in the same Facebook post as linked together (known as edges in Gephi), we can explore the "keyword network". By looking at the basic statistics in Gephi, we can have an overall understanding of the features of keyword network. Take the last day of 2012 Facebook posts as an example.

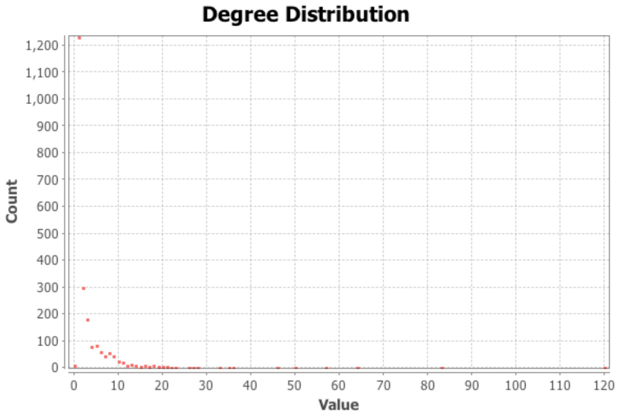


Figure 13 Degree Distribution of Keyword Network

This graph perfectly shows the long tail effect in social network—top ten keywords got more than 50% of the total frequency in the whole day. It actually makes great sense if we think about social network pattern that can be compared as a spider network.

Knowing that there is a long tail, our focus shift on to the relationship between the top keywords. The size distribution chart helps us view the modularity of different top keywords in a nutshell.

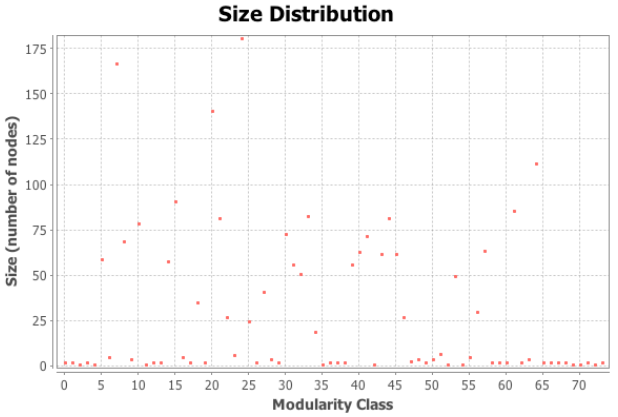


Figure 14 Size Distribution of Keyword Network

The size distribution shows several peaks along the modularity class axis, showing no cluster of the top keywords, which is interesting as it shows different people's attention on various affairs and topics.

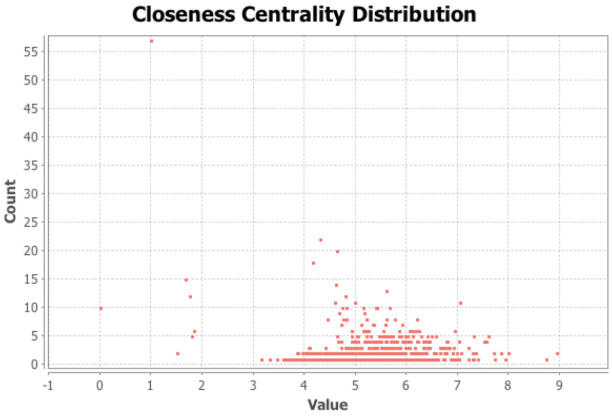


Figure 15 Closeness Centrality Distribution of Keywords

Closeness centrality shows a certain amount of cluster, which is unseen in the hypothesis. One possible explanation is that most links are averagely close, and their average value is about 5 to 6.

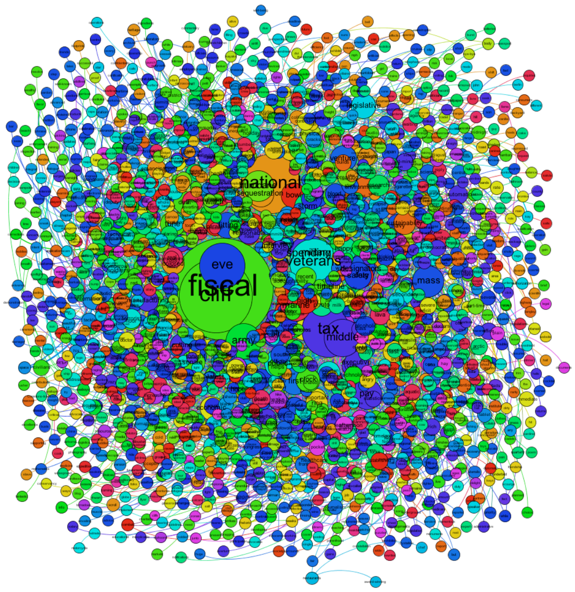


Figure 15 Closeness Centrality Distribution of Keywords

After having a general image of feature numbers, we can finally dig into the micro level keyword distribution and network in two-dimension space. From the Gephi graphic, we can see that the top keywords are all located in the center of the network and relative close (but separated) from each other. Taking a closer look, we can see that fiscal is overlapping with cliff and eve, which refers to Ben Bernanke's forecast that after the New Years Eve of 2012, America will face a fiscal cliff.

This method can be further used to see the clustered topics, as related top keywords are more likely to appear with an edge in Gephi. This again, proved our hypothesis of long tail theory, by presenting the thousands of small bubbles and several big bubbles.

7. SUMMARY OF VISUALIZATION TOOLS

On the coming end of this paper, we think it would be helpful to do a summary for contemporary data visualization tools. Data visualization, meaning creating and analyzing data and information in visual representation, has evolved from a supportive tool of data analysis to an individual research field related to information graphics, information visualization, and statistical graphics in the past two decades. Other than conveying ideas and communicating information, modern data visualization also requires sophisticated, eye-catching designs to stimulate engagement of viewers. By transforming massive amount of vector/matrix-formatted data to interactive graphics, a good piece of data visualization work should connect isolated members and facts and erase out the unseen patterns for audience.

The common models of data visualization include bar chart, streamgraph, treemap, scatter plot, Gantt chart, calendar view, chord diagram, choropleth, word cloud and so on. These models all approach data from their own visual dimensions, like color, time, position, size, length and interactiveness to utilize various features of a certain dataset.

As of data visualization tools, there are many different software /language/API to take advantage of. These tools vary at user-friendliness, complex level, and suitable objects.

Some of the most popular online data visualization tools include Google Chart API, Flot, Raphael, D3(Data Driven Documents), Visually. These tools provide relatively easy online access for all users to customize their own data visualizations. Raphael and D3 are both javascript based to provide gorgeous graphics and interactiveness, while one need to keep in mind when to stay simple while using them.

GUI (Graphical User Interface) is also an innovative way to build highly interactive data visualization. Tools like Crossfilter, Tangle make it possible for people to get according data by adjusting the input range on a graph or chart. Buttons, pull-down lists, and slide guides are all factors that contribute to more complicated interface.

Creating maps used to be one of the hardest tasks in web-building, until Google Maps changed the landscape. Many datasets with geographic factors look great in map visualizations, while Google Maps API let people build maps on their own websites. Data Visualization Tools like Modest Maps, Leaflet, PloyMaps, OpenLayers, Kartograph, CartoDB help people create geographical visualizations.

Besides the online data visualizations tools stated above that are majority-friendly, some expert data visualization tools provide programmers more space for processing and customizing data. Processing, NodeBox, R, Weka, and Gephi are among the best free ones.

The Tools we used in this paper covered at least one tool from each category, so that the readers can have a rather clear vision of the data from various approaches.

8. EVALUATION

In evaluating our data, processes, and visualizations, we do realize that there are some limitations.

For example, upon extracting the data, we did not identify user ids or unique updates. Without doing so it is difficult to say whether or not we oversampled some persons or populations. There is also cross over between the information posted on social media outlets. Tweets and Facebook posts often contain links, perhaps to news Release articles. Facebook also allows users to link their Twitter accounts, which means that their posts would appear on both websites. Therefore, we acknowledge and assume a certain level of endogeneity in these variables.

On the other hand, when calculating the keyword differences for Facebook and news Release data for Research Question 4.4, we acted under the (educated) assumed that the two mediums are very similar. This assumption in the other direction, that the users and data is very similar instead of very unique, also has it drawbacks and said over arching of assumption could lead to calculation errors.

In addition regarding the same Research Question 4.4, looking specifically at Figures 7 and 8 there may also be some underlying issue with the data collection. Almost all of the data stating that Facebook published stories earlier than news Release in Figure 7 was dropped in Figure 8 because it was outside of the sensible time range of being posted more than 24 hours before another post on the same day. Also, all of the outlier and flawed times and dates were in one 5-week section centered around the month of September. This may indicate a larger data collection flaw, at least in regards to the data and time that may be worth noting.

Static visualizations also have its disadvantages. It is limited in its capacity to present layers of information. For example, we could not produce a clear visual that could convey the top 10 words of four different sentiments. Also, the results in Table 1 are only the top words that matched to the dictionaries we selected. It is possible we lost some words that occurred in our data but not in the dictionaries.

The benefit of using static visualization was the ability to hold the data still for the reader or user to observe. However, static visualizations were not the best tool to convey repetitive information, for example, positive sentiment for Facebook, Twitter and news. Presenting similar looking graphs across different measures for the same data was not a visually engaging task.

Unlike matching the corpus to an external dictionary, the interactive visualization in `shiny` approached finding top key words more organically. The function behind this visualization found the most frequent words within the data. The results of this interactive visualization revealed trending keywords that the sentiment analysis could not find. Most of the cases that got lost within in the sentiment analysis static visualizations consisted of combined words such as “obamacare” or “independenceday”.

The benefit of the `shiny` interactive visualization is that the user does not have to standardize the input text format as the backend code handles cleaning, stemming and removing of stopwords.

The interactive visualization in D3 was a useful way to show when trending keywords happen and how it relates to real life events. However, the gradient color scheme used in this visualization was not ideal. The diverging color palette inadvertently suggests that each category of color has a specific value assigned to it. Therefore, it would have been a better choice to use different colors for each of the top 10 keywords in December (a qualitative palette).

For Research Quesntion 4.6, we only to a one-day data subset as an example, while the whole year data can be more interesting and show some clear topics. Moreover, it is still debatable that if “appearing in the same post” can be seen as the only edge in the keyword network.

Lastly, when looking at the data set in its entirety, we worked with was a nuanced data of political news and public social media data of 2012. We suspect that much of the trending keywords of 2012 in our data were heavily driven by the presidential reelection. Therefore, it is possible that the conclusions we draw from this research are not generalizable to more broader and diverse text data of different time frames. Also, we imagine that public versus private social media data may have dissimilar characteristics in the nature of the text.

9. RESULTS

We used a mix of static and interactive visualizations in `shiny` and D3 to answer five research questions discussed in Section 4.

Answering Research Question 4.2, we did indeed find a pattern in the static plots that revealed that update amounts for Facebook, Twitter and news are lowest on Mondays and Tuesdays. The activity patterns over the months of the year were more even but in general showed more activity in the warmer months than colder months.

However for Research Question 4.3, the static visualization on sentiment analysis did not show a clear pattern. The trending keywords varied on the sentiment scale depending on time of the year. This variance could be driven by seasonality changes or major events that occurred around specific times in 2012. We could not draw meaningful conclusions based on one year of data.

In the static sentiment match visualizations, we detected that Facebook and Twitter had higher positive, strong and weak word matches than news data. Twitter especially had high matches for all three emotions but scored low with negative words. It is possible that social media for public figures in the political sphere is used a public relations tool, which might explain the high positive and strong word match and low negative word match. Another interesting pattern we observed was the sudden drop in sentiment matches at the end of July. The word match frequencies to the four sentiment dictionaries for social media noticeably dropped simultaneously at the end of July. However, news release data did not show any sudden changes. We did not delve into the potential cause behind this.

Looking at Research Question 4.4, Facebook Timeline, we hypothesized that Facebook posts would precede news Release data. However, we were incorrect and the opposite was true; there was a lag time of, on average, 5 hours between Facebook updates and their news Release counterparts containing the same keywords. In hindsight, this outcome seems sensible as many of the posts regarding current events and news information on Facebook do link to a web published article, which means that the article must come before the Facebook post. Generally, it also makes sense that to effectively crowdsource information, the information must still come from a reliable source.

Our two exploratory interactive visualizations that addressed Research Questions 4.1 and 4.5, were useful in uncovering the relationship between keywords and events. In accordance with our exploratory hypotheses, we did find patterns in the words present on a given day and a relationship with current events. The shiny interactive visualization enabled users to look up trending keywords on any given day and we further explored this idea in our D3 visualization, showing that keywords and real life events could be closely related, with the highlighted trending keywords fitting very closely to a timeline of major events.

As for Research Quesntion 4.6, we confirmed the hypothesis that the social network keywords have a clear long tail effect. And the top keywords are located relatively centered of the keyword network. If two keywords are highly related, they may appear overlapped in the graphic, which in return, can help us with the topic organizing.

10. FUTURE WORK

Future research could entail subdividing this data further into blue, red, and swing states. Comparing the trending keywords in these three respective types of states prior to and following the presidential debate and Election Day could show different patterns. While social media data is relatively new, repeated text analysis on more election years or specific dates could reveal patterns that can be used in predictive analysis or future campaign management.

Overall, the techniques applied in our analysis – activity calculations, lexicon based sentiment analysis, and keyword visualizations based on word frequency – are applicable to a wide range of text data. However, these techniques are not sufficient for more complex models that are able to capture subtleties in the text such as sarcasm or correctly process misspelled words. In the future, it would be useful to be able to improve on these technique to recognize these complexities of language to improve our analyses of text data.

11. REFERENCES

[1] Don, A., Zhlelva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., Plaisant, C. 2007-2008. Discovering interesting usage patterns in text collections: Integrating text mining with visualization. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 213-222. DOI= http://dl.acm.org/citation.cfm?id=1321440.1321473.

[2] Pang, B., Lee, L. 2008. Opinion Mining and Sentiment Anaysis. *Foundation and Trends in Information Retrieval.* 2, 1-2 (2008), 1-135. DOI= http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf.

[3] Liu, B., Hu, M., Cheng, J., Opinion Observer: Analyzing and Comparing Opinions on the Web, *International World Wide Web Conference Committee,* (May 2005), 342-351. DOI= http://people.cs.pitt.edu/~huynv/research/aspect-sentiment/Opinion%20observer%20analyzing%20and%20comparing%20opinions%20on%20the%20web.pdf.

[4] Indurkhya, N., Damerau, F.J., The Handbook of Natural Language Processing (Second Edition), CRC Press, 2010

[5] Harrower, M. and Brewer, C. A.2011. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps, in The Map Reader: Theories of Mapping Practice and Cartographic Representation doi: 10.1002/9780470979587.ch34