English, French and Spanish Sentiment Analysis on the news documentation of the Paris Novemer 2015 attacks

De Groot, Nicole Louise

Bachelor Communicatie- en Informatiewetenschappen  
Afstudeerrichting Media-analyse  
Minor Digital Humanities

Vrije Universiteit Amsterdam

Opmerking auteur

nicoledegroot1994@hotmail.com

**English, French and Spanish Sentiment Analysis on the news documentation of the Paris Novemer 2015 attacks**

Whenever something bad happens in the world, there is media to cover it and bring it out to the public. We as people immediately feel something about such events, whether it is good or bad. Sometimes, we even openly express this to the rest of the world, to share our opinion or feelings. According to a survey The New York Times Customer Insight Group executed among their customers (as found in Moon, 2014), there are five main reasons why people share things with others: to entertain others, to establish ourselves, to keep track of our relatives and friends, self-fulfillment and promotion of the things and causes we care about. But that is just the public.

What about the news broadcasters? Of course, as a journalist, you have a moral obligation to make the crowd aware of the world around them. You do so in an objective manner, without letting your own beliefs interfere. But is all news always objective? Multiple studies have already shown that there are differences in the way some news is covered (Clayman, 1988; Bendix & Liebler, 1999), which is not such a strange occurrence, because just like we have our reasons for sharing personal things with the world, we have our preferences about our news as well. This has not so much to do with the broadcasters of news, but more with the crowd: we expect certain ‘news values’ from our news. If news does not have these values, we are less likely to read them (Just & Belt, 2007). Take for example, any natural disaster happening in Eastern-Asia. There will probably be less coverage about this than there will be coverage about a semi-large non-fatal attack that happened in your own country. That’s because, according too Just & Belt, the news value local relevance lacks from the story on Eastern-Asia.

But does more distance to a certain event also mean a different kind of coverage, so not just only considering the subject like above, but in particular, the sentiment? Bendix & Liebner (1999) found in their study about geographic variation of newspaper coverage, among other things, that physical distance was in fact a significant predictor when it came to the length and the number of stories about an event. Sadly, there has not been done much research on simply the sentiment, or perhaps, polarity of news articles, also simply because it is not so simple to analyse sentiment in seemingly objective articles.

That is also where the dangers on sentiment analysis lie, there has not yet been found an appropriate method to find the polarity of texts. Much of the work that has already been done on sentiment analysis was experimental, to try out new technical programs and machine learning, but there is no perfect way to go when it comes to technical or computational research. Also, there are many lacunae to be dealt with, like external factors that might be of influence or the target audience that needs to be taken into account. In this research in particular, it turned out exceptionally challenging regard the different languages in which the articles were written and the limitations of the computational tests.

That being said, it is still important to study not only the receiving end of news on sentiment, but also the sending end of news, so the news broadcasters. This is because the more you know about how journalists write news, the better you can write it yourself, but most of all, in the words of Moon (2004): ‘you can better assess your own content, and its ability to result in sustainable growth’. Which is why this study aims to answer the following question: What difference is there between the news media coverage in English, French and Spanish about the Paris November 2015 terrorist attacks as regards to the amount of sentiment expressed in said news articles? It is predicted that the French articles will contain the most sentiment, either positive or negative.

# Method and materials

For this research, two separate datasets were needed that later could be merged together to answer the research question. The first dataset contained the geographical and social data about the terrorist attacks, the second dataset contained the articles that were analysed based on sentiment, time and place. For the actual analysis the focus will lie on the Paris attacks data. These will be compared to the sentiment in articles about those events and later, there will be concluded if there are any differences between those articles.

## The datasets

For the first dataset, the Global Terrorism Database (GTD) was used from the University of Maryland and A Center of Excellence of the U.S. Department of Homeland Security. This database contains all terrorist attacks in the world from 1970 till 2015 and are specified with multiple variables ranging from the amount of fatalities to the (social) group that was the target of the attack to the regions where it happened. From here, it was possible to download a set of terrorism attacks with the help of the advanced search option. To limit the amount of attacks for the research, the search was filtered to: attacks between 2000 and 2015, attacks with 11 to 100 fatalities and injured and attacks that happened in Europe. Therefrom came eleven terrorist attacks that happened about the past 15 years in three major cities and a small island: London, Madrid, Paris and Utøya (Norway). This could next be exported as a .xlsx file.

Before the spatial analysis of the terrorist attacks could begin, the data had to be structured in a distinct manner. This was done by deleting the columns from the dataframe so only the most vital information remained: the event-ID, date, country, city, number of fatalities and number of injured people. Various newsarticles were used to confirm the specific places where the events happened and eventually, the coordinates for these places were added manually.

For the articles about the terrorist attacks that happened in Paris on November 13th, the database from LexisNexis Academic was used. Here, articles were searched under the 'General search' of the site and based on four different languages, that all contained the same parameters.

The keywords for all Dutch news articles were: 'Parijs aanval november 2015'. The keywords for all English news articles were: 'Paris attack november 2015'. The keywords for all French news articles were: 'Paris attaque novembre 2015'. The keywords for all Spanish news articles were: 'Paris ataque noviembre 2015'.

Within the search, there was specified to four subjects: 'Government & Public Administration', 'Society, Social Welfare & Lifestyle', 'Crime, Law Enforcement & Corrections' and 'Safety, Accidents & Disasters'.

After this, the search was sorted by relevance. This showed the articles with the most (useful) hits first. To keep the data managable, 33 articles were picked for each language. These articles had to originate from a timeframe of three months after the attack. This timeframe was chosen because during this period, For the Paris attacks this meant that only articles from the 13th of November until the 12th of February the next year were chosen. So, in all, 132 articles about the attacks in four different languages were found.

It was decided not to use the Dutch articles for the analysis, because the package that was going to be used for the analysis was not available for the Dutch language, which would mean another wordlist was needed. This, however, would not be beneficial for the research, so that is why the Dutch articles were dropped. More on this in the discussion.

## Analysis of the terrorism data

The analysis makes use of several computational tests to adhere to the data. These programs are used not only as an analysis tool, but also as an exploration tool and for the presentation of the data. Below, the different techniques will be specified.The terrorist attack data was visualised in QGIS Desktop 2.18.0 'Las Palmas' (2016). The Excel-file was first converted to a CSV-file after which it was loaded into the program. QGIS filtered the coordinates from the dataset to put the separate entries with relative distance from each other. Within QGIS it was possible to use Google Streets as a basemap for the world, so that the entries correspond to the cities where the attacks happened. The result was a simple map of Europe with dots on the cities where the attacks happened.To make deep mapping possible, the data was also put into ArcGIS Online (sd) for exploration purposes. Here, two more variables were added to make a more complete image of the events that happened in Europe, around the four big terrorism cites: the amount of fatalities and the year in which the event happened.

## Analysis of the sentiment data

Following the guidelines of Barr, Golden, Kelly, Stewart & Resende (2001) the method for the analysis of the sentiment data will now follow.

First, the article files needed to be cleaned up and made ready for the analysis. This was done by manually removing the metadata from the articles in the textfiles. The metadata contained information about copyrights, which paper the article originated from, the date it was published, the journalist who wrote it, in which section of the paper the article was found and the number of words in the article.

The package that would be used for the analysis (more information on that below) would analyse the files on sentiment per sentence or group. Therefore, the textfiles were edited with an online tool to replace every newline with a semicolon. This meant that the package would run the sentiment analysis over paragraphs which contained about 35 words per paragraph.

Next, because R i386 3.3.1 (2016) and Rstudio 0.99.903 (2016) were going to be used for the analysis of the sentiment in the articles, the textfiles had to be converted to Comma Delimited Texfiles, known as CSV-files. The link to these files can be found in the appendix.

After searching for the right package for sentiment analysis, it was decided to use a combination of packages found on GitHub . This package, sentimentr, makes use of two other packages, namely qdap and pacman. As the site states ‘sentimentr is designed to quickly calculate text polarity sentiment at the sentence level and optionally aggregate by rows or grouping variable(s)’ (trinker, 2016). It makes use of the dictionary made by Hu & Liu (2004), as mentioned in the introduction.

Other packages that were needed for the analysis were tm, SnowballC, wordcloud, RColorBrewer, fpc and ggplot2. These were all loaded in a script.R file and next, the CSV-files were loaded into R. The packages were used, for example, to create a corpus and to create barplots.

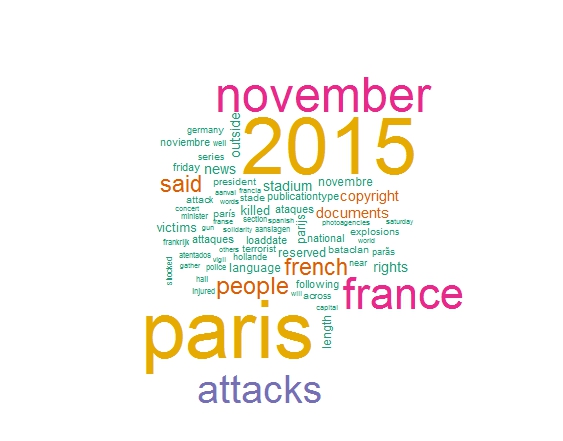
## General inquiry

Before the actual sentiment analysis, a general inquiry of the data was made using the packages mentioned above.

First, a worcloud was made from all article files to find out what words were used the most and to see if the assumptions were right about the keywords.

Image 1

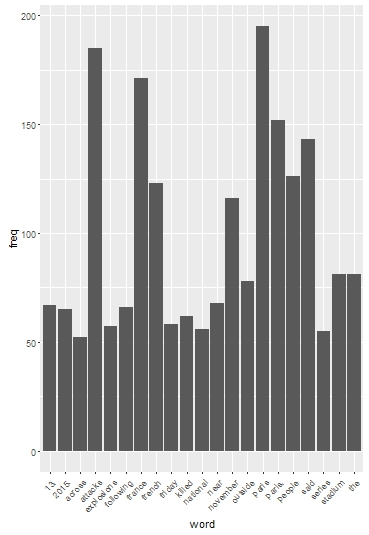
*Wordcloud showing words with a frequency over 150 from four artcile-files.*

  
Note that the keywords that were used in the query also return in the wordcloud. The words with the same colours fall under the name frequency category.

Next, the english, french and spanish article-files were viewed more closely. To get a deeper understanding of the data barplots were made showing the frequency of a sample of words, reigning from a frequency of 10 and higher to 50 and higher.

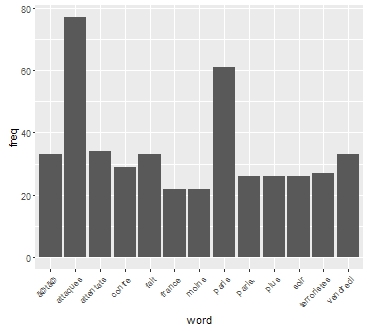
Barplot 1

*Barplot for english words with a frequency over 50*

  
It seemed the english articles had more words that were more frequent than the other languages.

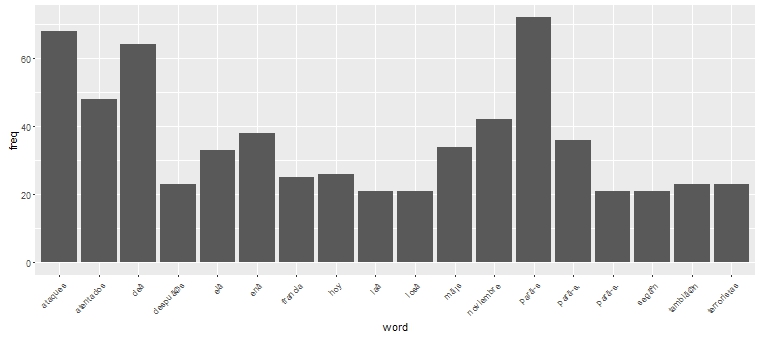
Barplot 2

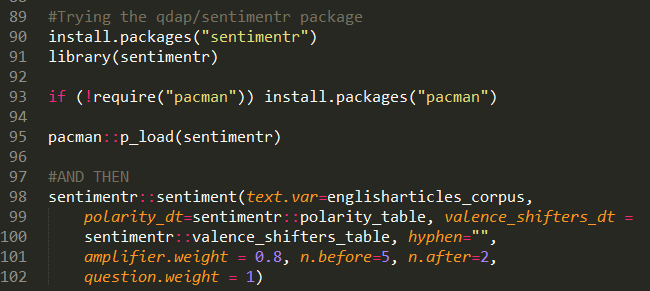
*Barplot for french words with a frequency over 20*

**It is interesting to note that in all barplots, the days of the week on which the terrorist attacks happened also returned.

Barplot 3

*Barplot for spanish words with a frequency over 20*

  
In all barplots, the words attack(s) and november come back.

Next, the sentiment analysis was conducted on the separate language files. For more information on the formula that was used to create the algorythm in the sentimentr package, please consult the GitHub page. Below is a snippet of code that was used to get the sentiment data for the files. In this case, it is a snippet from the english data script.

# Results

Like with the general inquiry, the data set with the articles was first summarised to look at the extend of the data.

Table 1

*Descriptive statistics from the sentiment analysis on the English, French and Spanish articles respectively*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | English | French | Spanish | Total |
| N | 236 | 116 | 1511 | 1863 |
| Total sentiment | -34.28 | -120.51 | -52.66 | -207.45 |
| Mean (SD) | -0.15 (0.78) | -1.04 (4.84) | -0.03 (0.38) | -0.11 (1.30) |
| Sum of Squares | 146.23 | 2815.95 | 215.93 | 3178.11 |
| Variance | 0.60 | 23.40 | 0.14 | 1.69 |
| Std. Error | 0.05 | 0.45 | 0.01 | 0.03 |

N = amount of paragraphs

To determine whether or not the differences in means from the sentiment analysis were significant or not, a one-way indepedent-samples ANOVA was conducted. The results can be viewed in the table below.

Table 2

*Summary of the one-way ANOVA-test*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SS | Df | MS | F | P |
| Between groups | 108.91 | 2 | 54.45 | 33.25 | <.0001 |
| Within groups | 3046.10 | 1860 | 1.64 |  |  |
| Total | 3155.01 | 1862 |  |  |  |

From the one-way ANOVA-test we can conclude that one or more of the means are significantly different from each other, because the F-statistic is lower than .05. Following is the Tukey HSD test to identify which of the pairs are significantly different.

Table 3

*Summary of the Tukey HSD post-hoc test*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tukey HSD Q statistic | Tukey HSD p-value | Tukey HSD inference |
| English vs French | 8.71 | 0.0010053 | p<0.01 |
| English vs Spanish | 1.74 | 0.4358962 | Insignificant |
| French vs Spanish | 11.52 | 0.0010053 | p<0.01 |

On average, as predicted, sentiment in the French texts was more negative (M=-1.04, SE=0.45), than the sentiment in the English (M=-0.15, SE=0.05) and the Spanish texts (M=  
-0.03, SE=0.01). The difference in sentiment between the English and French, 0.89, and between the French and Spanish texts, -1.01, were significant, F(2, 1860)=33.25, p<.0001. However, the difference in sentiment between the English and the Spanish texts, -0.12, was not significant. Also, the Tukey HSD post-hoc test revealed that the French texts differed significantly from the other two texts.

**Conclusion and discussion**

Coming back to the research question in the beginning: What difference is there between the news media coverage in English, French and Spanish about the Paris November 2015 terrorist attacks as regards to the amount of sentiment expressed in said news articles? In this case, the data showed that indeed, the French data contained more sentiment about the November 2015 Paris attacks than the English or the Spanish data.

This study, however, was nowhere near perfect and still leaves multiple questions unanswered. In the future, it would interesting, as well as smart, as implementing the following thoughts. First, it would be better for generalisation to do the same study for multiple events in multiple places, so that you have more comparative data. Second, the limitations of the R-packages that were used were primarily language-based. Therefore, it would be interesting to do the same study with more languages than this one. Third, during this study, there was a problem with the gathering of the data. Should we choose certain English newspapers and limit the database or search all English news (that did not just originate from the United Kingdom) with the risk of not being able to generalise? Fourth, there was the issue with the Dutch articles. There does not exist yet a Dutch sentiment analysis package, which made sentiment analysis useful for those articles. Unfortunately, those had to be dropped. In all fairness, there was no Dutch terrorism attack that could later on be studied anyway.

It is interesting to see how far we have already come in technical and linguistically computing, but like this research showed, there is still much to uncover in the field computational linguistics.

# Bibliografie

*ArcGIS Online*. (n.d.). Retrieved from https://www.arcgis.com/home/index.html

Barr, R., Golden, B., Kelly, J., Stewart, W., & Resende, M. (2001). Guidelines for designing and reporting on computational experiments with heuristic methods. *Proceedings of International Conference on Metaheuristics for Optimization*, 15.

Bendix, J., & Liebler, C. M. (1999). Place, Distance, and Environmental News: Geographic Variation in Newspaper Coverage of the Spotted Owl Conflict. *Annals of the Association of American Geographers, 89*(4).

Clayman, S. E. (1988). Displaying neutrality in television news interviews. *Social problems, 35*(4), 474-492.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 10.

Just, M., & Belt, T. (2007). The Magic Formula: How to Make TV that Viewers Will Watch. In T. Rosenstiel, M. Just, T. Belt, A. Pertilla, W. Dean, & D. Chinni, *We interrupt this newscast: How to improve local news and win ratings, too.* (pp. 94-125). Cambridge University Press.

University of Maryland & A Center of Excellence of the U.S. Department of Homeland Security. (2009-2016). *Global Terrorism Database*. Retrieved December 6th, 2016 from https://www.start.umd.edu/gtd/

Moon, G. (2014, 10 2). *Why People Share: The Psychology of Social Sharing*. Retrieved 12 21, 2016, from http://coschedule.com/blog/why-people-share/

QGIS Desktop 2.18.0 'Las Palmas'. (2016, 11 25). The Noun Project collection. Retrieved from https://www.qgis.org/nl/site/forusers/download.html

Schumaker, R., Zhang, Y., Huang, C., & Chen, H. (2012). Evaluating Sentiment in Financial News Articles. *Decision Support Systems, 53*(3), 458-464.

*The Facebook: not just for students*. (n.d.). Retrieved 12 15, 2016, from http://www.browndailyherald.com/media/paper472/news/2005/11/03/CampusWatch/The-Facebook.Not.Just.For.Students-1044229.shtml

trinker. (2016, 11). *trinker/sentimentr*. Retrieved 12 2016, from https://github.com/trinker/sentimentr

# Appendix

The appendix can be found online under:

<https://github.com/nicolelouise/vth2016>