Everyday Citizen Satisfaction

Background

The Office for National Statistics (ONS) is involved in the Big Data ESSNet. This is a programme of work funded by Eurostat, coordinated by CBS Netherlands and involving other 20 NSI partners.
The aim of the Big Data ESSNet is to investigate how a combination of Big Data sources and existing official statistical data can be used to improve current statistics and create new statistics in statistical domains.
There are 9 work packages. Work Package 7 (WP7) focuses on the statistical domains of Population, Tourism/border crossings and Agriculture.
From these domains, ONS has conducted a short pilot study in the "Population" use case.
The purpose of the case study is to examine the level of daily satisfaction by analysing the content of messages for the presence of defined expressions describing emotional states, e.g. joy, sadness, fear, anger. Additionally, emotional states of people can be analysed with respect to public events.
The idea is to explore how we might produce statistics on social sentiment from news sites/blogs/social media towards events/topics and how those can be linked to existing official statistics which annually measure population well-being.
This pilot study was set to provide:

- An in-depth exploration for the chosen source(s)
- An assessment of the API / web scraping conditions of the potential sources
- An exploration and application of lexicon-based sentiment analysis techniques
- An analysis of the results produced between the data sources and within different time units considered
- An assessment of the Facebook users leaving comments to determine distribution of gender and residency, and how this relates to the Guardian readership

The project was split into three stages:

- ❑ **Stage (1) – Data Collection via API/Web Scraping**

Stage (1) focuses on building the tools to gather the data. The two sources targeted are:

- The Guardian Facebook Page: comments made by Facebook users to the Guardian public posts, and reactions counts (Like, Love, Haha, Wow, Sad and Angry)
- The Guardian Website: categories associated with and comments made to the articles on the news portal.

The initial idea was to consider both comments made on the Facebook page of the Guardian and comments made by the readers in the Guardian website. However, the

collection of comments from the Guardian website was later discarded because against the robots.txt[1] rules of the website.

- ❏ **Stage (2) – Text Mining/Sentiment Analysis**

Stage (2) is about evaluating the utility of applying lexicon based sentiment analysis on the comments posted by the users. The evaluation is performed by comparing the scores obtained by the methods against a manually graded gold standard. Posts and comments of daily activity is also analysed over the collection period.

- ❏ **Stage (3) – Quantitative and qualitative assessment of the data**

The data collected in Stage (1) focused over a pre-defined period of time of 33 days (27th February to 31st March 2017). Results are aggregated by different granularities and analysed over time. Relationships between article sentiment/comments sentiment/reactions and word counts are also examined. Overall, this aims at demonstrating whether dates associated with particular events show some evident changes in sentiment.

Stage (1) – Data Collection via API/Web Scraping

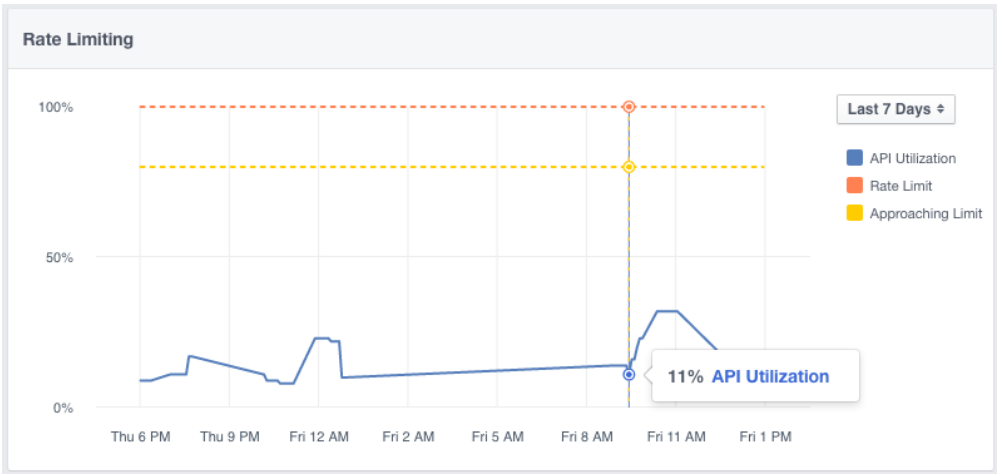
Facebook Graph API

The Graph API is the primary way to get data in and out of Facebook's social graph. It's a low-level HTTP-based API that is used to query data, post new stories, upload photos and a variety of other tasks that an app might need to do.
The Graph API is HTTP based, so it works with any language that has an HTTP library, such as cURL, requests (Python), etc.
facebook-sdk is a Python based client library designed to support the Facebook Graph API.
Most API calls must be signed with an access token.
Different types of access tokens exist, each of which is associated with different rate limits.
The App Dashboard provides an interface to check statistics on the API usage.
The two main sections are:

- API Stats: shows number of API calls made, number of Errors and average request time

[1] Website owners use the robots.txt file to give instructions about their site to web robots (such as web scrapers). robots.txt files use the Robots Exclusion Standard, a protocol with a small set of commands that can be used to indicate what can be accessed on the website and by who.

- Rate Limiting: shows the % of utilisation of the API

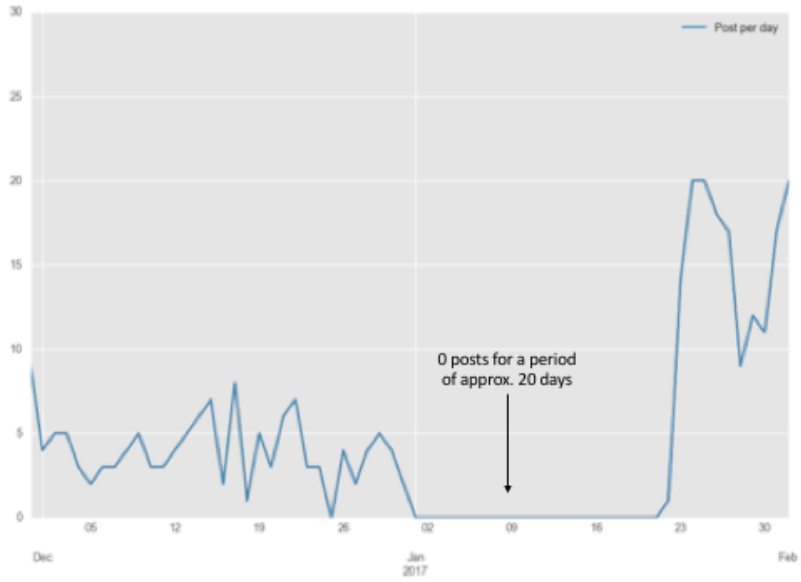


Although the dashboard appears to be a useful tool for understanding API usage, statistics are not shown real-time but rather with a one or even two days lag, which makes impossible to rely on it.

Data Collection: Comments and Posts

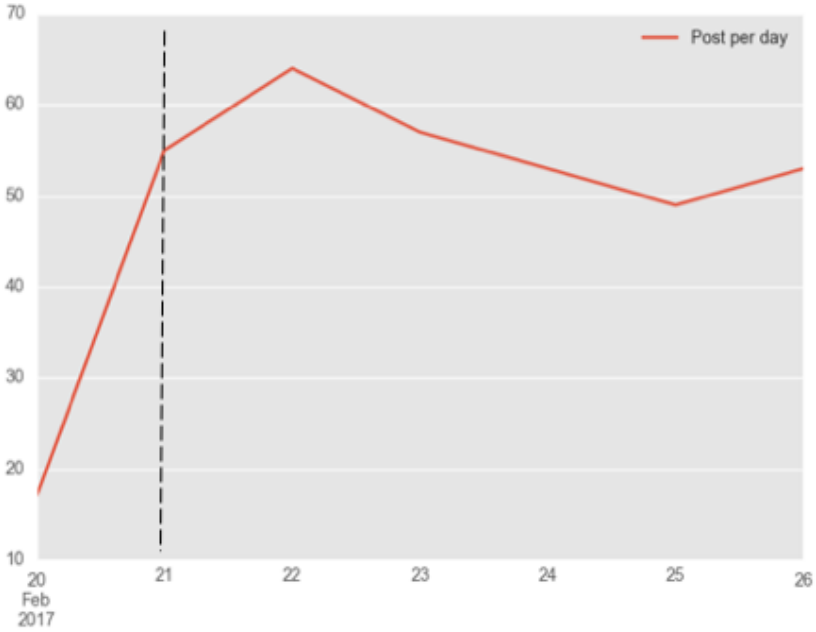
While collecting the data from the Facebook API for the initial agreed period (December 2016 – January 2017) it was noted that the number of posts collected per day is not consistent and suffers from days with no posts returned from the API (Figure 1). It appears that going back in time is not feasible as the Facebook API tend to return only a subset or no data at all with no clear criteria for the selection.
Figure 1 shows the number of posts per day collected through the API on the 1st of March 2017, while querying the 1st December 2016 – 1st February 2017 period of time.

Figure 1 - Number of posts per day collected for the period 1st December 2016 – 1st February 2017. Time of querying the data is Wednesday 1st of March 2017.



Querying more recent data shows a different picture.

Figure 2 - Number of posts per day collected for the period 20th February 2017 – 26th February 2017. Time of query is Wednesday 1st of March 2017.



In this case, the query was run the 1st of March to retrieve the number of posts per day published by the Guardian Facebook page on the week commencing 20th February (Monday) to 26th February (Sunday). The number of posts retrieved per day is much higher (mostly between 50 to 60 posts per day). Figure 2 shows also a clear difference in the number of posts returned on the 20th (data queried more than 8 days old) vs. number of posts returned from the 21st onwards (data queried up to 8 days old).

Following this assessment, the assumption is that it is better to query the API within one week from the time the query is done.
Data from the Facebook Graph API were collected daily. On each day, data were collected leaving a six-day gap between the collection day and the reference day queried.
Overall, the collection covered a period of 33 days, from the 28th of February to the 31st of March.
1,704 posts were collected from the Guardian Facebook page, an average of approximately 52 posts per day, and a total of 561,748 comments, an average of 330 comments per post and 17,000 comments per day.
The posts data were also integrated with information scraped from the Guardian website to add article information about categories, tags, authors, etc. The data collected were stored in a MongoDB database in two separate collections:

- The *posts* collection which contains the posts published by the Facebook page. Any valid access token can read posts on a public Page. The data associated with each post is:
    - Reactions: in 2015 Facebook extended the original Like button with five[2] additional reactions to give users more ways to share their feelings towards a post in a quick and easy way. Along with the "like", the new set of reactions includes "love", "angry", "haha", "wow" and "sad". The Facebook API allow to collect reaction counts to a post for each type of reaction.
    - Post_id: The post ID
    - Comment_count: Number of comments to this post
    - Article_url: The link attached to this post. Most of the times this url links to a Guardian article in the Guardian website (this is the case for 1584 out 1704 posts), but it could also link to another media content published by the Guardian Facebook page (such as a video, etc)
    - Created_time: The time the post was published
    - Message: The status message in the post

  If the post links to a Guardian article, then the following additional fields are scraped from the Guardian website and attached to the post object.
    - Article_title: the title of the article on the Guardian website
    - Main_category: the main category associated used by the Guardian to categorise articles.
    - Categories: sub-categories scraped from the article page
    - Tags: tags scraped from the article page. Tags goes in more detail and can often be person's name, places, companies cited in the article
    - Authors: the authors of the article

[2] An additional temporary reaction called "thankful" rolled out by Facebook in 2016 in occasion of Mother's Day. Currently not visible by the users in the Facebook website but still accessible via the API.

Figure 3 - An example of post object.


```
{
        "_id" : ObjectId("58bd41d59ff1026d58d7f05f"),
        "article_title" : "SpaceX to send two people around the moon who paid for a 2018 private mission",
        "reactions" : {
                "thankful" : 0,
                "love" : 45,
                "like" : 862,
                "total_count" : 996,
                "angry" : 3,
                "haha" : 8,
                "wow" : 77,
                "sad" : 1
        },
        "tags" : [
                "SpaceX",
                "Space",
                "Elon Musk",
                "The moon",
                "news"
        ],
        "main_category" : "science",
        "post_id" : "10513336322_10155128653556323",
        "comment_count" : 104,
        "article_url" : "https://www.theguardian.com/science/2017/feb/27/spacex-moon-private-mission-2018-elon-musk?CMP=fb_gu",
        "authors" : [
                "Alan Yuhas"
        ],
        "created_time" : "2017-02-27T22:31:58+0000",
        "message" : "SpaceX CEO Elon Musk declined to name the two people or what they had paid, though he said the \"private citizens\" know each other and are “very serious” about the flight.",
        "categories" : [
                "spacex",
                "home",
                "science"
        ]
}
```


- The *comments* collection which contains the comments: comments are linked to the *posts* collection by the post_id field. A comment can be made on various types of content on Facebook. Most Graph API nodes have a /comments edge that lists all the comments on that object. It is possible for comment objects to have a /comments edge too, which is called 'comment replies'. The structure is the same for these.
  The fields attached to each comment are:
  - Comment_id: The comment ID
  - Post_id: the post_id to which the comment refers to
  - Comment_count: Number of replies to this comment
  - Like_count: Number of times this comment was liked
  - User: The person that made this comment which is composed of the ID and the person's full name
  - Created_time: The time this comment was made
  - Message: The comment text
  - Parent_id: For comment replies only, the comment that this is a reply to

Figure 4 - An example of comment object:


```
{
        "_id" : ObjectId("58bc2c6d9ff1026d58d7adfe"),
        "Comment_id" : "10155128653556323_10155128655371323",
        "post_id" : "10513336322_10155128653556323",
        "comment_count" : 2,
        "like_count" : 59,
        "user" : {
                "name" : "Jack Marshall",
                "id" : "10212458974738919"
        },
        "created_time" : "2017-02-27T22:33:14+0000",
        "message" : "Hopefully Donald Trump and Piers Morgan.\n\nEDIT: Actually, I don't care who it is as long as it's a couple of wankers and it's a one way ticket."
}
```

Data Collection: Volumes

Figure 5 and 6 show the posts and comments daily activity over the period of the 33 days, with articles causing a sudden urge[3] of comments highlighted in Figure 6 (with number of comments in brackets).

Figure 5 - Daily Posts volume over time.

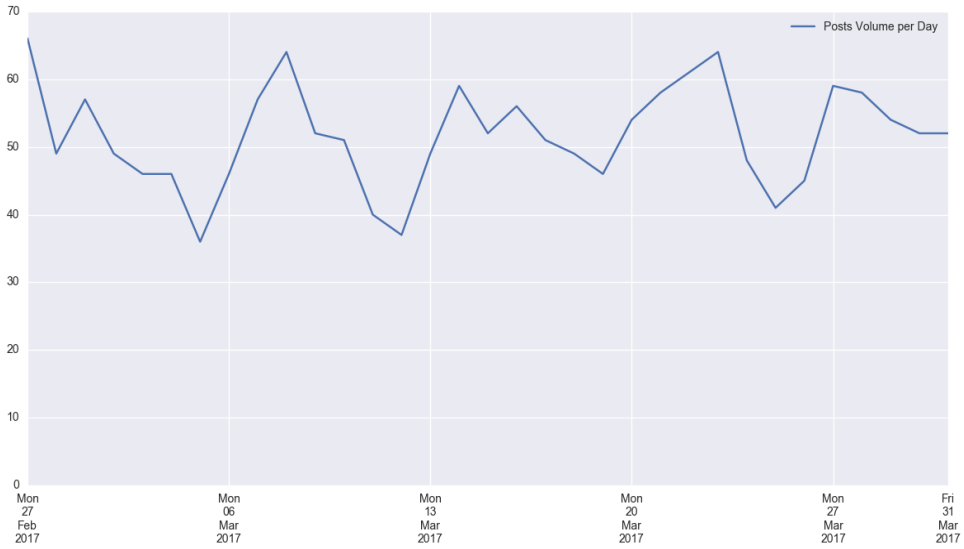
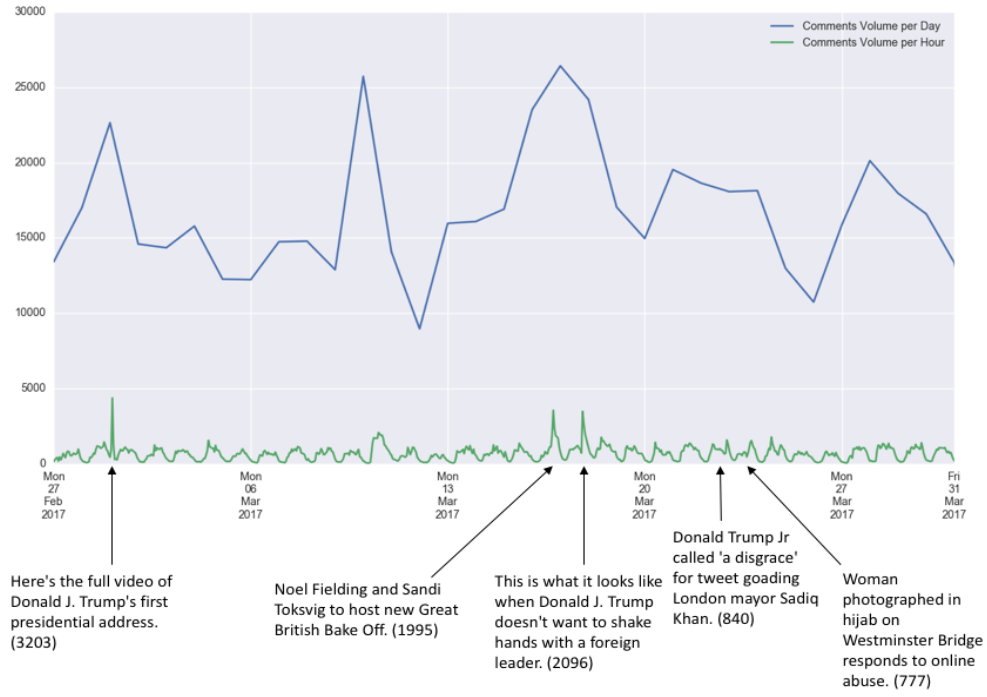


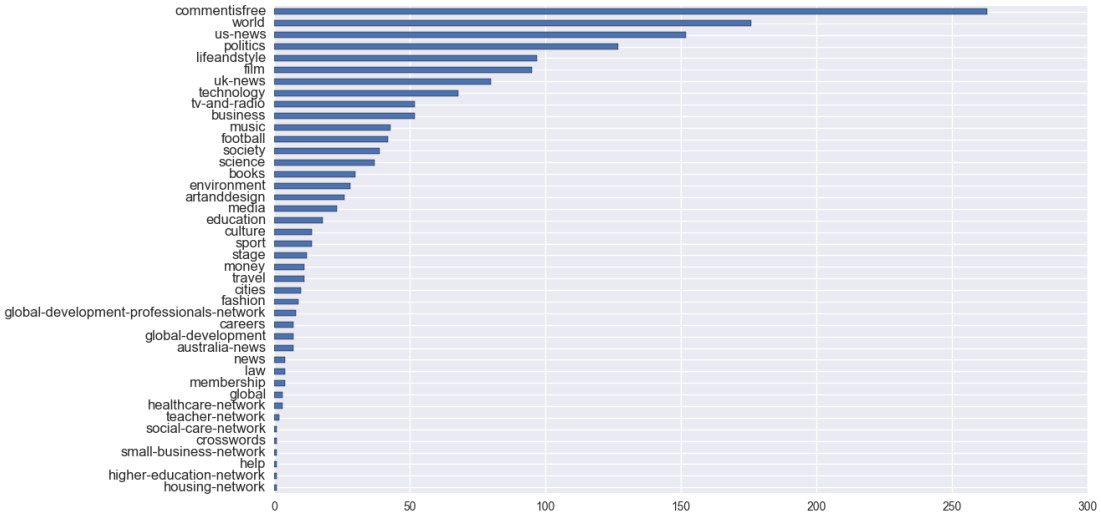
Figure 6 - Daily/Hourly Comments volume over time. Highlighted posts with n. of comments in brackets.



[3] A sudden urge of comments is identified by calculating the total number of comments per hour and per post.

Figure 7 shows the count of post per category (obtained from the Guardian website). Most of the posts are under the ‘commentisfree’ category which are opinion articles.

Figure 7 - Number of Posts by category.



Stage (2) – Text Mining/Sentiment Analysis

The primary focus of the sentiment analysis is to determine a writer’s feeling from a given text. The feeling might be his/her attitude, emotion, or opinion. The most important step of this analysis is to classify the polarity of the given text as positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification tools, look, for instance, at the valence of the sentiment, such as positive sentiment ranging from 1 to 5, or at emotional states such as "angry", "sad", and "happy".

There exist two main approaches to the problem of extracting sentiment automatically.
The lexicon-based approach involves calculating orientation for a document from the semantic orientation of words or phrases in the document. The text classification approach involves building classifiers from labeled instances of texts or sentences, essentially a supervised classification task. The latter approach could also be described as a statistical or machine-learning approach.

❏ Lexicon-based sentiment analysis

A substantial number of sentiment analysis approaches rely greatly on an underlying sentiment (or opinion) lexicon. A sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative, or a number ranging from e.g -1 to 1. A number of manually constructed lexicons exists and will be described in more detail in the next section.

❏ Machine learning sentiment analysis

Because manually creating and validating a comprehensive sentiment lexicon is labor and time intensive, machine learning based sentiment analysis practices "learn" the sentiment-relevant features of text. Most popular methods include Naive Bayes, Maximum Entropy or Support Vector Machines. Machine learning approaches are not without drawbacks:

1. they require (often extensive) training data which are, as with validated sentiment lexicons, sometimes troublesome to acquire
2. they depend on the training set to represent as many features as possible
3. they are often more computationally expensive in terms of CPU processing, memory requirements, and training/classification time
4. they often derive features "behind the scenes" inside of a black box that is not (easily) human interpretable and are therefore more difficult to either generalize, modify, or extend (e.g., to other domains).

For this case study, we follow the first method, in which we use dictionaries of words annotated with the word's semantic orientation, or polarity.

Lexicons assessment: Vader, NRC, Syuzhet, AFinn, Bing

Several lexicons are available. Here we examine 5 of them, VADER available on the Python NLTK library, and the other four included in the R package called Syuzhet that interfaces with the Stanford CoreNLP library. We migrated the last four lexicons from R to Python to create a consistency with our approach.

- ❏ VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. In addition to a list of words, VADER incorporates a full list of Western-style emoticons, sentiment-related acronyms and initialisms (e.g., LOL and WTF) and commonly used slang with sentiment value (e.g., nah, meh and giggly). The valence of the sentiment it is also adjusted for the impact of word-order sensitive relationships between terms. For example, degree modifiers (such as intensifiers, booster words, or degree adverbs) impact sentiment intensity by either increasing or decreasing the intensity. Finally, negators are also considered to reverse the sentiment from one polarity to the opposite.
- ❏ The Syuzhet package was originally created as an attempt to reveal the latent structure of narrative by means of sentiment analysis. It includes four different lexicons:
    - "syuzhet" is a custom sentiment dictionary developed in the Nebraska Literary Lab
    - "afinn" developed by Finn rup Nielsen
    - "bing" developed by Minqing Hu and Bing Liu
    - "nrc" developed by Mohammad, Saif M. and Turney, Peter D.

  Compared to Vader, these are larger lexicons but no attempt is made to take in consideration things like negators, intensifiers or booster words. Additionally, other than word's associations with positive polarity negative polarity, the NRC lexicon

has also manual annotations of words' associations with eight emotions: joy, sadness, anger, fear, trust, disgust, surprise and anticipation. These emotions have been argued to be the eight basic and prototypical emotions on Plutchik's wheel of emotions theory.

Figure 8 - Plutchik's wheel of emotions.

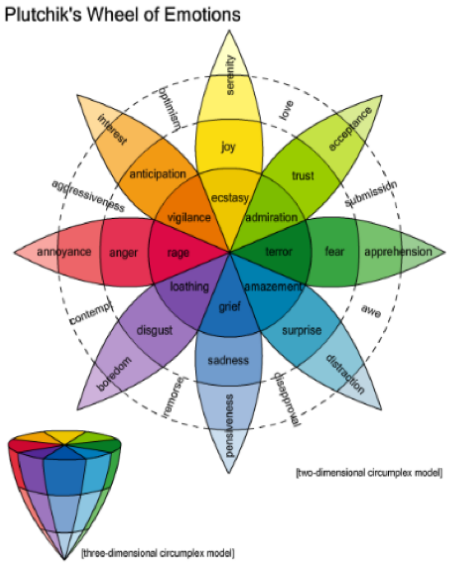


FIGURE 1. Plutchik's wheel of emotions. Similar emotions are placed next to each other. Contrasting emotions are placed diametrically opposite to each other. Radius indicates intensity. White spaces in between the basic emotions represent primary dyads—complex emotions that are combinations of adjacent basic emotions. (The image file is taken from Wikimedia Commons.)

Plutchik suggested 8 primary bipolar emotions: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation.
In the NRC lexicon, each word is associated with positive polarity, negative polarity and eight different emotions and their corresponding valence.

In Table 1 the sizes of the different lexicons are compared, distinguishing between positives (sentiment > 0), negatives (sentiment < 0) and neutral words (sentiment = 0). The NRC lexicons appear to have a large number of neutral words, where the sentiment equals 0. This might due to the fact that these words are sentiment neutral, but are still included in the lexicon for the emotions they carry (any between joy, sadness, anger, fear, trust, disgust, surprise and anticipation).

Table 1 - Comparison of lexicons' size

	VADER	Syuzhet	Afinn	Bing	NRC
N° of words	7502	10747	2477	6786	14182
- Positives	3333	7160	878	2006	2231
- Negatives	4169	3587	1598	4780	3243
- Neutral	0	0	1	0	8708

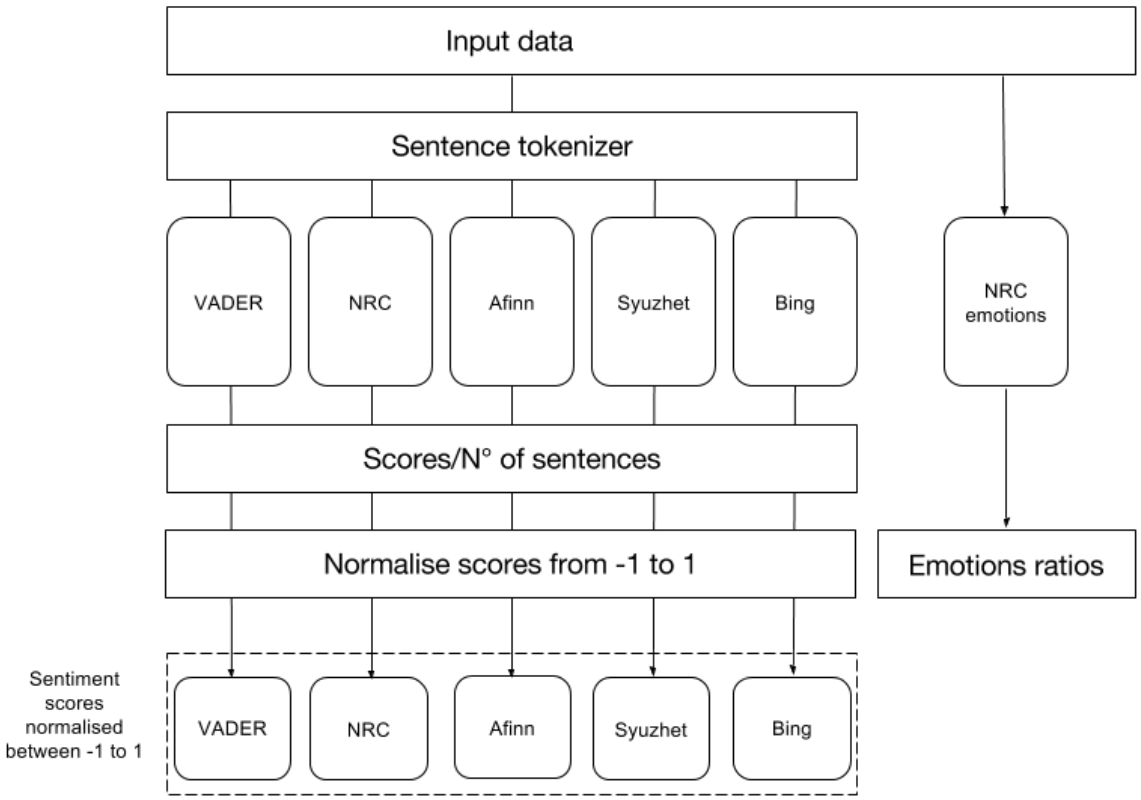
Model overview

Initially the block of text (e.g. a Facebook comment) is parsed into sentences that are then passed into the sentiment analysis step that is based on the five different lexicons. This step produced a number for each sentence composing a single comment and ranges from the negative polarity (number is negative) to the positive (number is positive).
To obtain a unique number per comment, the average of all sentences scores is taken.
Finally, to make possible the comparison between the different lexicons, the scores are normalised to range between -1 and 1. The normalisation formula is the same used by the VADER library:

$$Normalised\ score\ = \frac{score}{\sqrt{score^2 + \alpha}}$$

Where we keep the default alpha value used by the Vader library K = 15. This number was derived empirically by the authors of the library and approximates the maximum expected value of sentiment words can be found in a sentence. If the resulting normalised score is outside the range of -1 and 1, the normalised score is rounded to -1 and 1 respectively.
Separately, the block of text is also used to extract the emotions using the NRC lexicon and calculate emotions ratios.
Figure 9 summarises the steps just described.

Figure 9 - Model overview.



Performance Evaluation

To assess lexicons performance, 696 Facebook comments were manually graded between -1, very negative, and 1, very positive.
Table 2 below shows the correlations between the scores computed by the different methods. It is possible to observe how the Vader scores are the ones correlating the most with the manually annotated scores.

Table 2 - Correlations between lexicon sentiment and manually annotated sentiment.

	Manual	Afinn	Bing	NRC	Syuzhet
Afinn	**0.40**				
Bing	**0.34**	0.66			
NRC	**0.17**	0.40	0.43		
Syuzhet	**0.34**	0.71	0.77	0.65	
Vader	**0.43**	0.78	0.64	0.40	0.67

The accuracy of a sentiment analysis system is, in principle, how well it agrees with human judgments. This is usually measured by precision and recall.
The sentiment scores were therefore converted to a multinomial classification problem where a score:

- > 0.2corresponds to a Positive sentiment (**P**)
- < -0.2 corresponds to a Negative sentiment (**N**)
- $-0.2 \leq score \geq 0.2$ corresponds to a Neutral sentiment (**X**)

Table 3 reports for each method the averaged precision, recall and F1-score among all classes weighted by the support (the number of true instances for each label).

Table 3 - Precision, Recall and F1-score for each method.

	Afinn	Bing	NRC	Syuzhet	Vader
Precision	0.54	0.52	0.44	0.53	0.55
Recall	0.52	0.49	0.44	0.46	0.52
F1-score	0.51	0.46	0.39	0.38	0.5

The detailed scores are also shown here. Again, Vader is one the best performing along with the Afinn method, which surprisingly has also the smallest lexicon. This could be because a smaller lexicon means less words that offset each other and thus push the sentiment towards either the negative or positive side.
Also interesting to note, the recall for the negatives class is in every case very low which makes them hard to detect in terms of quantity. Only a small portion of them are currently detected by the methods, which might due to the fact that sometimes the negativity of the comment is influenced by the context rather than the exact words used in it.

Table 4 - Afinn lexicon classification report.

Afinn				
	Precision	**Recall**	**F1-score**	**Support**
N	0.6	0.32	0.42	233
P	0.41	0.58	0.48	149
X	0.56	0.64	0.6	314
avg / total	**0.54**	**0.52**	**0.51**	**696**

Table 5 - Bing lexicon classification report.

Bing				
	Precision	**Recall**	**F1-score**	**Support**
N	0.64	0.23	0.34	233
P	0.39	0.4	0.39	149
X	0.49	0.72	0.58	314
avg / total	**0.52**	**0.49**	**0.46**	**696**

Table 6 - NRC lexicon classification report.

NRC				
	Precision	**Recall**	**F1-score**	**Support**
N	0.47	0.1	0.17	233
P	0.33	0.42	0.37	149
X	0.48	0.69	0.57	314
avg / total	**0.44**	**0.44**	**0.39**	**696**

Table 7 - Syuzhet lexicon classification report.

Syuzhet				
	Precision	**Recall**	**F1-score**	**Support**
N	0.72	0.09	0.16	233
P	0.35	0.24	0.29	149
X	0.46	0.83	0.59	314
avg / total	**0.53**	**0.46**	**0.38**	**696**

Table 8 - Vader lexicon classification report.

Vader				
	Precision	**Recall**	**F1-score**	**Support**
N	0.67	0.27	0.39	233
P	0.43	0.61	0.5	149
X	0.53	0.66	0.59	314
avg / total	**0.55**	**0.52**	**0.50**	**696**

Overall, the lexicon based sentiment classification exercise reach at best a 47.7% classification error against the baseline value of 54.9%, that is if we were to predict all comments as the most occurring class.

Stage (3) – Quantitative and qualitative assessment of the data

In Stage (3) results are aggregated by different granularities and analysed over time. Relationships between article sentiment/comments sentiment/reactions and word counts are also examined.

Analysis over time

Looking at the sentiment produced by the different lexicons over the 33 days period, Figure 10 and 11 show that they tend to follow more or less the same patterns. In the first chart all comments are considered, whereas in the second plot only parent comments to a post are used (i.e. comment replies are excluded). Considering parents comments only shows that the patterns over time are very similar but amplified.
Duplicate comments (same message, posted for the same post by the same user, but different times) are removed as well as comments were all the lexicons give score 0 (this has the intent to exclude empty comments and comments where the user is just mentioning another user).
The following charts are produced by averaging the sentiment every hour and applying a moving average on a window of 24 hours. While a moving average is useful to remove noise, data on the edges is lost and thus the sentiment tend to level off. Nevertheless, such smoothing can be useful for getting a sense of the emotional trajectory.
Table 9 and 10 report the correlations between lexicons sentiment over time for all and parents only comments.

Figure 10 - Sentiment of all comments over time.

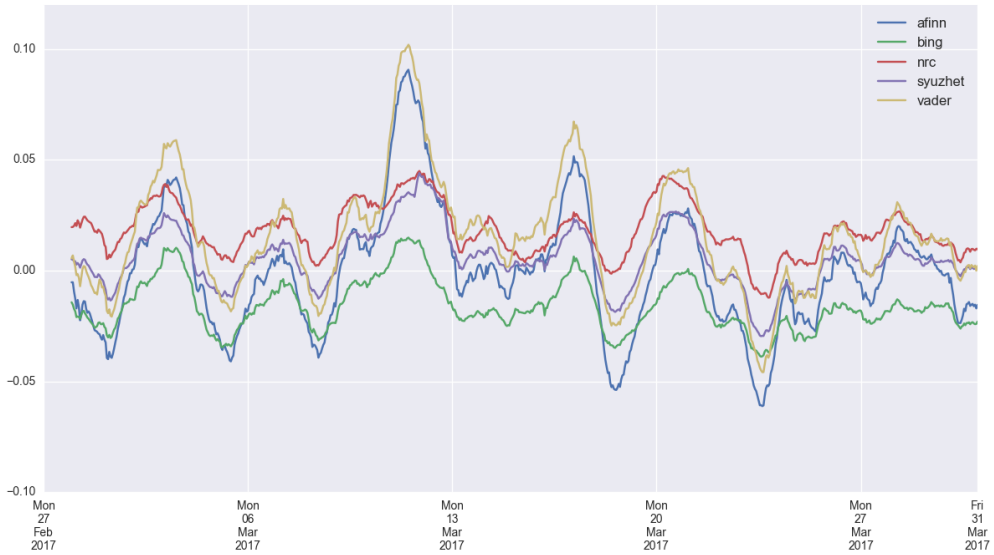


Table 9 - Correlations between lexicon sentiment over time for all comments.

	Afinn	Bing	NRC	Syuzhet
Bing	0.67			
NRC	0.49	0.53		
Syuzhet	0.72	0.77	0.70	
Vader	0.77	0.61	0.46	0.67

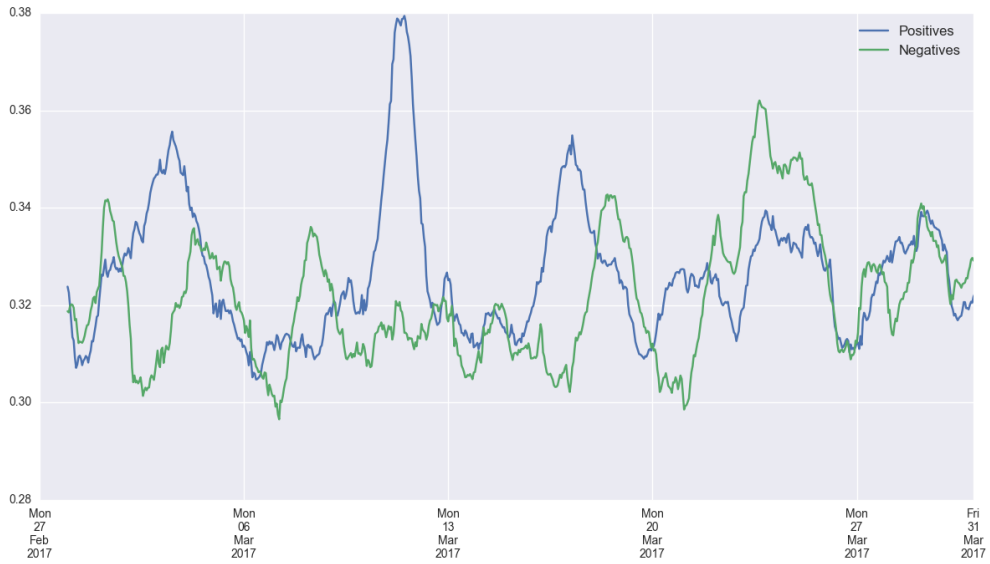
Figure 11 - Sentiment of parent comments only over time.

Table 10 - Correlations between lexicon sentiment over time for parents comments only.

	Afinn	Bing	NRC	Syuzhet
Bing	0.68			
NRC	0.51	0.55		
Syuzhet	0.73	0.78	0.71	
Vader	0.78	0.63	0.48	0.69

The Vader sentiment was then split into Positive comments (sentiment > 0) and Negative comments (sentiment < 0). The two trajectories (with the negatives in absolute values) are here compared over time. Figure 12 shows as before the hourly value with a moving average over a window of 24 hours.

Figure 12 - Positive vs. Negative trajectories with a 24 hours moving average



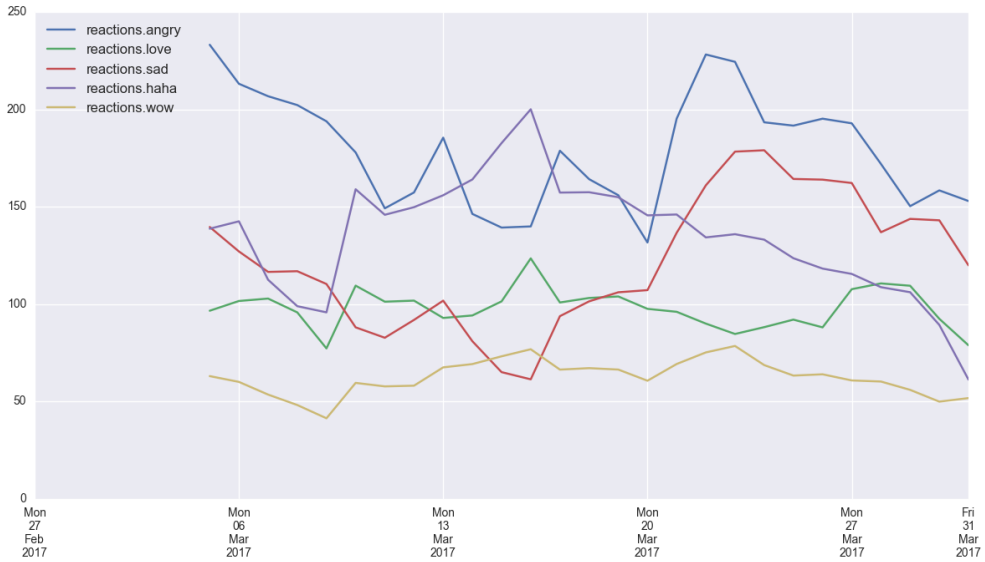
It appears to be a pattern of a positive spike followed by a negative spike a few days later. In the fourth big spike, the positive and negative spikes seem to coincide. In the case of the 2nd big spike instead, the spike is only positive with no change in the negative sentiment. An analysis by category indicates the *media* category as being one of the main influencer of the positive spike around the 11 of March and the *uk-news* category the major influencer for the negative spike around the 22 March, day of the terrorist attack in London. A more in depth analysis should aim at identifying whether this is just coincidence or whether there might be a relationship between positive and negative sentiment referring to a specific event.

Figure 13 looks at the daily mean with a moving average over a window of 7 days.

Figure 13 - Positives vs. Negatives trajectories with a 7 days moving average

Figure 14 shows the 7 days moving average window applied to the reactions count of the Facebook posts. Reactions counts over time also shows an increase of *angry* and *sad* reactions the week of the terrorist attack.

Figure 14 - Reactions counts with a 7 day moving average



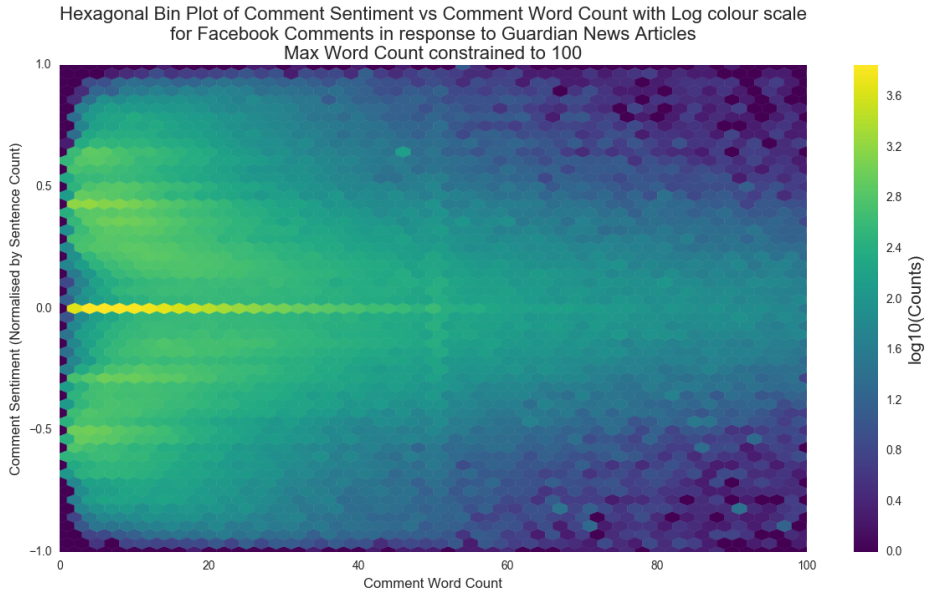
Relationships between Comments/Articles and Word Counts

Relationships between various elements are analysed here using the scores produced by the Vader sentiment tool.

The relationship between Facebook comment sentiment and comment word count for the articles featured on The Guardian Facebook page suggests that the majority of comment data is focussed around the 0. It appears that there is much more variation in sentiment at low word counts. As the word count increases the variation between sentiment extremes become less and less pronounced.

For comparison, a hexbin plot is used in Figure 15: this is a 2d histogram density plot, where the colour intensity indicates the frequency of data entries occurring in that particular region of the plot.

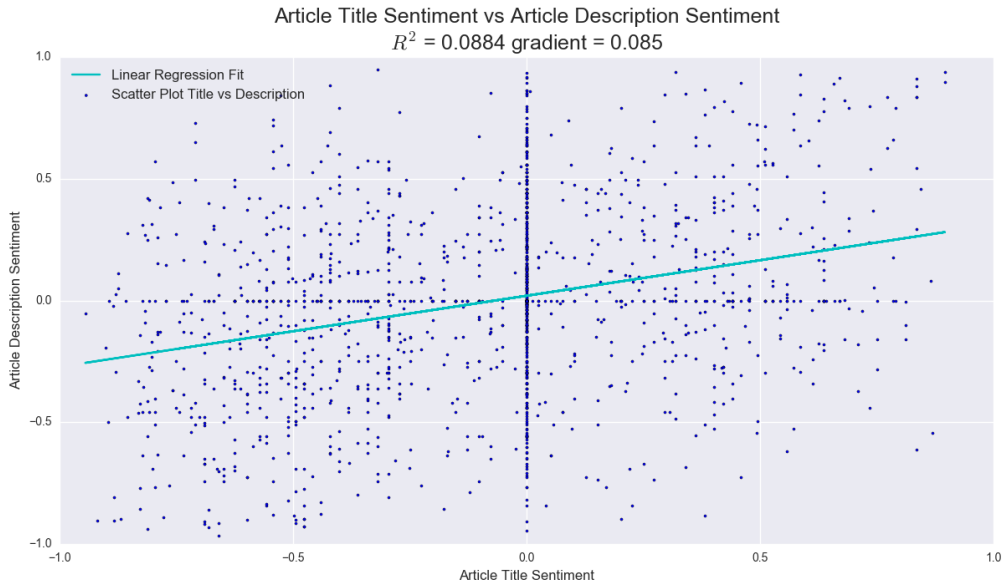
Figure 15 - Hexagonal bin plot Comment Sentiment (Overall score) vs. Comment word count for all comments.



It is possible that this is an effect of the Vader analysis tool, as it is ideally meant for shorter comment analysis, rather than long passages of text.

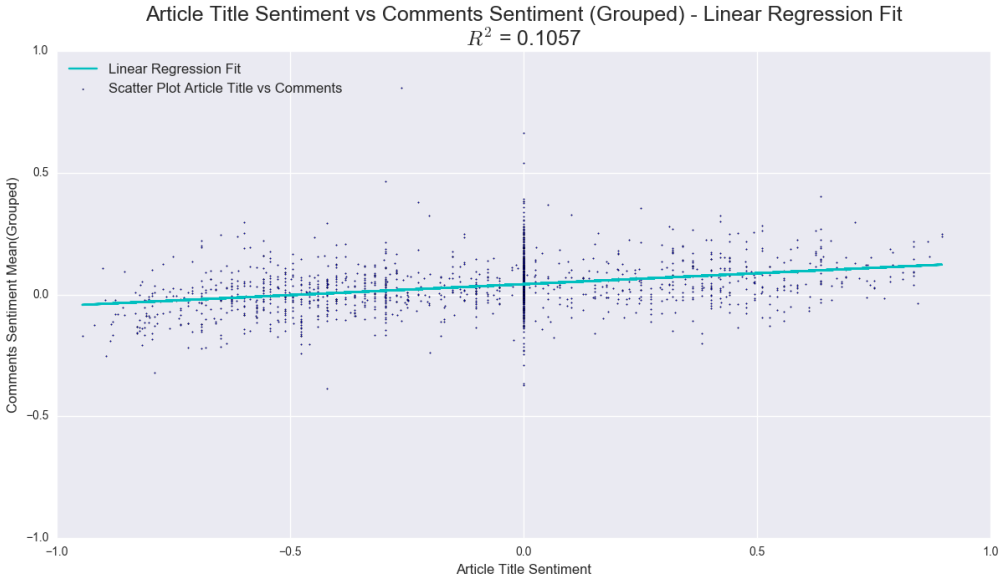
Figures 16, 17 and 18 look at the different levels of relationships: at first an analysis of the relationship between the sentiment of the article title with that of the sentiment of the article message is presented, followed by an analysis of the relationship between parent level comments and child level comments (comment replies) and finally a comparison between the sentiment of Facebook comments posted on an article and the sentiment of the article title.

Figure 16 - Comparison between the sentiment of the article title with that of the sentiment of the article message.



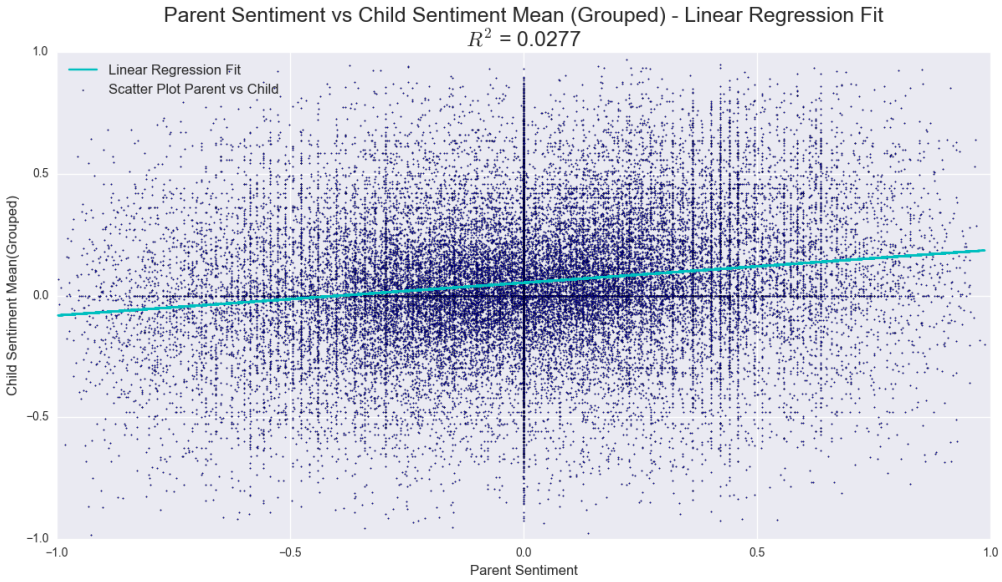
In the next Figure, the goal was to see if the sentiment of an article title influences the sentiment of the comments made in response to this title.

Figure 16 - Comparison between the sentiment of Facebook comments posted on an article and the sentiment of the article title



In the final plot, the correlation between parent level comments and child level comments is analysed to see if the sentiment for a particular parent comment influences the sentiment of the set of associated child comments (i.e. replies to that parent comment).

Figure 17 - Comparison between the sentiment of parent level comments and child level comments.



These relationships show weak or no relationships whatsoever between the various elements, although this apper difficult to understand due to the high presence of noise and neutral sentiments.

Emotions

The NRC lexicon is a large word–emotion association lexicon, and can be used to use it to extract emotions from a piece of text.
Given a target text, such as a Facebook comment, the model determines which of the words exist in the NRC emotion lexicon and calculates the emotions ratios. These are given by the number of words associated with an emotion to the total number of emotion words in the text. This simple approach may not be reliable in determining if a particular sentence is expressing a certain emotion, but it is reliable in determining if a piece of text has more emotional expressions compared to others in a corpus.

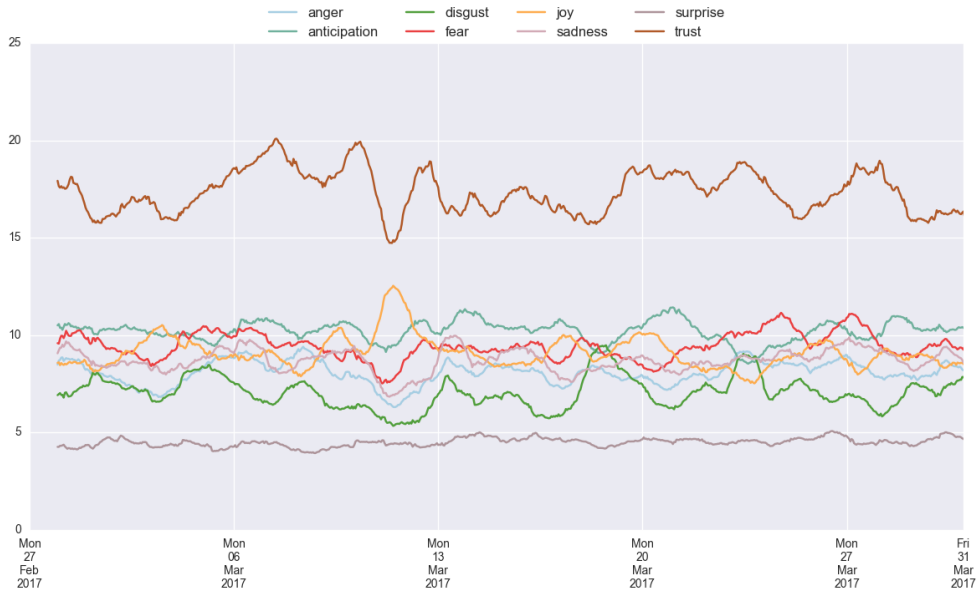
Due to a disambiguation, we first removed from the lexicon the following words: don, trust, and john.
The following Table shows the 10 most frequent words associated with each emotion:

Table 11 - Top 10 most frequent words associated with each NRC emotion.

Anger	Anticipation	Disgust	Fear
money	good	bad	government
vote	time	hate	bad
bad	money	blame	problem
court	white	idiot	god
hate	vote	terrorist	change
politics	pay	hell	war
blame	long	terrorism	military
words	thought	shame	court
terrorist	god	ridiculous	hate
hell	coming	death	case
Joy	**Sadness**	**Surprise**	**Trust**
good	vote	good	good
love	leave	money	system
money	bad	vote	president
white	problem	leave	money
vote	hate	hope	white
pay	case	deal	vote
god	black	young	guardian
true	tax	terrorist	real
hope	lost	expect	fact
majority	terrorist	guess	pay

When plotted and smoothed over time, a sensible difference between the trust emotion vs the others is evident, as shown in Figure 18. Trust is constantly the largest proportion in the comments, whereas on the other end the surprise emotion is constantly the smallest.
No clear interpretation seems possible from the graph.

Figure 18 - NRC Emotions over time.



Users

Each comment collected from the Facebook API is associated with a user. The only information returned by the API about the person that made the comment are the ID and the person's full name. However, it was found that the ID of the person is unique only to each app/API and cannot be compared across different apps. Given that two different APIs were used for this use case due to an initial inexperience with the types and rate limits, the result is that it is not possible to track the behaviour of the same user across time since different IDs might be present in the data for the same user.
The lack of demographic information along with lack of information of where the user live are two big limitations of the Facebook API compared to others social network such as Twitter.
A 2010 article on the Demographic profile of Guardian readers suggests an evenly distributed readership across age groups, more or less compatible with the UK adult distribution, and with a higher presence of men than women across gender.

Guardian reader profile				
		All Adults %	Guardian reader	Guardian reader %
Solus			733000	80%
Social Grade	A	5%	158,000	14%
	AB	26%	696,000	62%
	ABC1	54%	994,000	89%
Age Group	15-24	16%	200,000	18%
	25-34	16%	177,000	16%
	35-44	17%	180,000	16%
	45-54	17%	194,000	17%
	55-64	14%	190,000	17%
	65+	20%	178,000	16%
Gender	Male	49%	597,000	53%
	Female	51%	522,000	47%
	ABC1 men	27%	528,000	47%
	ABC1 women	28%	460,000	41%
Education	TEA 18+	37%	813	73%
	TEA 21+	22%	698	62%
Working status	Full Time	42%	525,000	47%
	Full time 35+	27%	363,000	32%
	Full time 45+	17%	245,000	25%

Source: NRS Oct 2010 - Sept 2011

Additional information about Guardian online readership are provided here, where is possible to assess the detailed demographic information about the Guardian's site users and how they compare to the average UK internet user.

The Guardian site user profile vs. average UK Internet user		
	Average UK Internet User	Average Guardian Site User
Gender		
Male	50%	52%
Female	50%	48%
Age		
2 to 17 years old	13%	5%
18 to 24 years old	11%	11%
25 to 34 years old	15%	15%
35 to 49 years old	27%	30%
50 to 64 years old	27%	30%
65 years old or older	7%	9%
Average age	39.3	42.8

To understand the demographics of the Guardian readers on Facebook instead, one possibility could be to use the Facebook Ads Create page. This Facebook service allows any user to create an ad while specifying a target population according to a combination of factors, such as locations, age, gender, interests and also users behaviours, derived from the Facebook users profiles data and external sources.
By selecting *The Guardian* as interest, we can see different combination of demographic characteristics for the users which have *The Guardian* Facebook page in their interest.

As an example, restricting the location to the people who live in the United Kingdom and have *The Guardian* in their Facebook interests the distributions are the following compared to the UK 2015 estimates:

- By Gender

	Facebook numbers	**Facebook %**	**UK % (2015 estimates)**
All	3,100,000	-	-
Male	1,500,000	48	51
Female	1,600,000	52	49

- By Age Groups

	Facebook numbers	**Facebook %**	**UK % (2015 estimates)**
Up to 17	64,000	2	21
18-24	610,000	20	9
25-34	820,000	26	14
35-49	920,000	29	20
50-64	500,000	16	18
65+	210,000	7	18

- By Income

	Facebook numbers	Facebook %
£20,000-£24,999	210,000	18
£25,000-£29,999	190,000	16
£30,000-£34,999	170,000	15
£35,000-£39,999	160,000	14
£40,000-£49,999	210,000	18
£50,000-£74,999	220,000	19
£75,000+	4,800	0

The categories to which it is possible to refine these statistics include a very detailed range of possibilities derived from information provided by the users themselves in their Facebook profiles or derived from their likes in conjunction with information provided to Facebook by Oracle Data Cloud on UK consumer data (where consumers shop, how they shop, what products and brands they purchase, the publications they read, and their demographic and psychographic attributes).
Although, this may technically be the demographics of people who list *The Guardian* as an interest, rather than the exact page likes, those statistics could provide a good approximation of its demographics.
Finally, to understand who is actually commenting in the Guardian Facebook page one possible option to consider could be to create ad hoc automated data collection tools for users demographics, with Facebook consent.

Conclusions

There are several limitations that arose from using lexicon-based sentiment analysis methods for analysing Facebook comments. We try to summarise them below:

- Lexicon-based sentiment analysis is known to work better with short text from social network, such as tweets from Twitter, which are short and thus usually straight to the point. In those occasions, methods such as Vader have demonstrated to perform better than machine learning models. However, sentiment analysis for discussions, comments, and blogs tend to be a much harder task, since they generally involve multiple entities, multiple opinions, comparisons, noise, sarcasm, etc. The evaluation of the lexicon based methods against manually evaluated comments shows a performance just above the baseline model. It should be noted however that the manual annotation involved just one person, which might introduce some subjectivity in the scores.
- Context in which the comment is expressed is not included in the scores produced by the lexicons but it is considered to be an important influence of it. This appear to be particularly relevant for the Negatives comments which are indeed with all methods underestimated.
- Keyword-based approach is totally based on the set of keywords. Therefore, sentences without any keyword would imply that they do not carry any sentiment at all.
- Meanings of keywords could be multiple and vague, as most words could change their meanings according to different usages and contexts. This appears particularly relevant in the emotions analysis where three keywords had to be removed from the NRC lexicon.

Overall, the most salient points of the analysis can be summarised in:

- The collection of data from the Facebook API highlighted that is not feasible as a means of collecting historical data, but it may be possible for collecting data up to a week old
- The evaluation showed that Vader is the best performing method along with the Afinn method, which surprisingly has also the smallest lexicon of all. Considering that could be due to the fact that a smaller lexicon means less words that offset each other, Vader should be preferred its completeness
- Looking at the sentiment produced by the different lexicons over time, it appears that that they tend to follow more or less the same sentiment trajectory. When considering only parent comments (i.e. excluding comment replies), extremes appeared more pronounced.
- Despite a method for identifying the most influencing tags or posts of the positive vs. negative sentiment over time was not developed, an analysis by category indicated the *media* category as being one of the main influencer of the positive spike around the 11 of March and the *uk-news* category the major influencer for the negative spike around the 22 March, day of the terrorist attack in London.

- In conjunction with that, reactions counts over time also showed an increased in of *angry* and *sad* reactions the week of the terrorist attack, although it might require some more analysis in order to confirm the two are related
- Looking at relationships between various elements using the sentiment scores produced by the Vader sentiment tool, it appears that there is much more variation in sentiment at low word counts. As the word count increases the variation between sentiment extremes become less and less pronounced. It is possible that this is an effect of the Vader analysis tool, as it is ideally meant for shorter comment analysis, rather than long passages of text. Other relationships showed weak or no relationships whatsoever between the various elements
- Finally, it is not clear the outcome of the emotions analysis based on the emotions extracted using the NRC lexicon.
- The Facebook API does not provide any other information than person's full name and ID. However, it was found that users' IDs are meaningful only if the same API/app is used for collection. Additional information about users might be collected using Facebook Ads, although in this short study it was not possible to link those to the commenters information.
- Overall, this study aimed at understanding whether dates associated with particular events show some evident changes over the sentiment. From a superficial analysis, it appears changes in the sentiment trajectory could be detected, although more work is required.