PAGE DE GARDE

TABLE OF CONTENT

[Introduction to Social Listening 3](#_Toc10628092)

[Internet and Traditional Marketing 3](#_Toc10628093)

[Lectures 5](#_Toc10628094)

[Methodology 6](#_Toc10628095)

[Data Preparation 7](#_Toc10628096)

[Data Collection 7](#_Toc10628097)

[Data Cleaning 8](#_Toc10628098)

[Data quality 9](#_Toc10628099)

[Analysis 11](#_Toc10628100)

[Exploratory Analysis 11](#_Toc10628101)

[Sentiment Analysis 14](#_Toc10628102)

[Clustering 14](#_Toc10628103)

[Twitter Network 14](#_Toc10628104)

[Conclusion 16](#_Toc10628105)

# Introduction to Social Listening

The launch of Facebook in **DATE** and Twitter in **DATE** marked the beginning of a direct access to consumer’s minds. From Internet to traditional, marketing will be shaken by these new insights coming right of the consumers’ fingertips. The strength of these companies does not only lie into their capacity to connect people together but to a certain extent in the possibility for anyone who listens to learn more about the opinion of a particular group of people.

Chiffres sur le contenu créée en ligne (tweets, posts, reveix etc)

Augmentation ? etc

What is social listening

What insisgts can be taken

limitations

The heart of any social listening strategy lies into the quality and quantity of the data collected. There are several options to collect from social networks everything they expose. The simplest one is to use their dedicated APIs that allows through paid plans to access everything, for example Twitter, or part, for example Facebook, of their network. This can be rapidly a cost if the social listening strategy requires a daily update and that the topic monitored generates a large quantity of data.

The second option is to go for some pure players like Talkwalker, Sprout Social or Hootsuite that allow through a paid plan to access everything they expose. This tools can aggregate several social media sources, for example Twitter and Instagram

## Internet and Traditional Marketing

Marketing, meaning all the activities a company must accomplish to promote and sell their product, has been done since the Antiquity with mosaic design where an artist has spent time to create advertisement campaign for fish sauce[[1]](#footnote-1). But marketing has done quite the leap since Antiquity and nowadays, it gathers four major methods being printing, broadcasting, direct mailing and phoning. Almost every company in the world use at least one of these techniques to advertise its product.

I

# Lectures

# Methodology

Data Collection

Data cleaning

Sampling

Exploratory analysis

Wordcloud

Bigrams and trigrams

Sentiment analysis

Labeled word and clustering

LDA

Naïve Bayes

Emoji?

# Data Preparation

## Data Collection

Of all methods evoked in the Introduction to gather data, we decided here to go for a simpler method. To collect tweets from around the time Volkswagen announced that the used cheating software for their emissions tests, we used a classic web scraping package.

Web scraping allows to gather the content of a web page via a programming language, here Python, and retrieve from the web page every content that is displayed. The trick to build a good web scraper is to handle queries, exceptions and to respect the limitations asked from the website. To avoid building reliable web scrapers, we will use an open source package.

We used the twitterscraper[[2]](#footnote-2) Python package. It has been chosen because it is reliable and often updated and because it allows to go around Twitter Search API limitations of only 7 days historical data for the free plan. Since we need tweets that dates back to 2015, it was very important.

This web scraping method allows us to gather the username and the full name of the person that writes a tweet, the tweet-id, its URL, the text, the HTML, the tweet timestamp and finally the number of likes, of replies and of retweets the tweet has.

To have the tweets that we want, we used the following query:

twitterscraper volkswagen -bd 2015-09-01 -ed 2015-10-15

It means that the web scraper will gather all tweets that match three criteria. That the text of tweet contains Volkswagen, that it has a timestamp between the September 1st,2015 and October 10th, 2015.

This query has allowed us to gather 914 274 tweets.

A typical tweet collected looks like this:

|  |  |
| --- | --- |
| **Column** | **Output** |
| fullname | *Rhondas Romance* |
| html | *<p class="TweetTextSize js-tweet-text tweet-text" data-aria-label-part="0" lang="en">I liked a <a class="twitter-atreply pretty-link js-nav" data-mentioned-user-id="10228272" dir="ltr" href="/YouTube"><s>@</s><b>YouTube</b></a> video <a class="twitter-timeline-link" data-expanded-url="http://youtu.be/pmzZbUioFAQ?a" dir="ltr" href="http://t.co/L6oksJ9rUO" rel="nofollow noopener" target="\_blank" title="http://youtu.be/pmzZbUioFAQ?a"><span class="tco-ellipsis"></span><span class="invisible">http://</span><span class="js-display-url">youtu.be/pmzZbUioFAQ?a</span><span class="invisible"></span><span class="tco-ellipsis"><span class="invisible">\xa0</span></span></a> 2015 <strong>Volkswagen</strong> Sales Event | “Model Rear End” Passat Commercial</p>* |
| id | *640675779253280768* |
| likes | *0* |
| replies | *0* |
| retweets | *0* |
| text | *I liked a @YouTube video http://youtu.be/pmzZbUioFAQ?a\xa0 2015 Volkswagen Sales Event | “Model Rear End” Passat Commercial* |
| timestamp | *2015-09-06T23:59:43* |
| url | */RhosBookReviews/status/640675779253280768* |
| user | *RhosBookReviews* |

## Data Cleaning

To effectively retrieve insights from the collected tweets it is very important to process a cleaning of the texts. A text as such “*I liked a @YouTube video http://youtu.be/pmzZbUioFAQ?a\xa0 2015 Volkswagen Sales Event | “Model Rear End*” Passat Commercial” has a lot of unnecessary noise that would make the analysis harder.

The first step is to remove all the unnecessary **noise** in the textual data. The first noise is the links in a tweet. Most of the time they embed an image and the link does not provide any information on the text. We also remove the # and @ because it does not add much. Later on, we will only focus on the mention and the hashtag but for pure textual analysis and sentiment analysis it does not add any value. We then remove all numbers because of the low information on sentiment they provide and then of course the punctuation. Finally, everything is put to lower case. One could argue that the tweets in capital letters could bear more information or more vigor in the tone, but it drastically increases the number of different words where in reality, *Foo*, *FOO* and *foo* are the same word.

The second step is to keep only English tweets. This is an arbitrary choice because we could have kept all tweets and perform the analysis on multi language tweets. However, working on only one language is easier. It is easier to draw insights from a known language, the stemming process is, as of today, better performed on the English language and it is a way to reduce, at least for training the size of the dataset. Meaning that after this phase, we actually lost 62% of the database, with only **339 367 tweets remaining**.

The third step is to remove **stop words**. Stop words are very common words of a language that, in a sentence, do not create meaning. For example, in the phrase “I am a student.”, “I” and “a” are stop words. Removing these words allow faster computation and better relevance in the analysis. We will use the default stop words list of the NLTK package. However here we will add one stop word which is “volkswagen” since it is on every tweet, it does not add any useful information to the tweet.

The fourth step is **lemmatization**. The process of lemmatization allows to transform all words in third person to first person and verbs that have past tense or future tense to be put pack to present. The fifth and last step is **stemming**. Stemming is the process of transforming a word back to its stem, or root. Once more, it allows to reduce the number of words and yet does not reduce the accuracy and the information of the tweet. Thanks to both these steps, it reduces the number of different versions of one verb or word to keep the most information out of every tweet.

## Data quality

Data quality is at the center of every problematic linked to data. At the start in relational database, a huge part of the database management is to make sure that the right format is entered, there is no missing values or no duplicates across the system. For unstructured data such as text, it is even harder to monitor the overall quality of the database.

To assess to quality of the tweets we have collected we will use the Text Quality Dimensions of Daniel Sonntag[[3]](#footnote-3). Note that we will not compute a score but only using Sonnatg’s dimensions as framework to think about the quality of our tweets.

The **intrinsic text quality**, as stated in the paper, can yet not be assessed automatically. We can only infer that as most of the tweets are emitted by non-competent authorities or persona or bots, the intrinsic text quality is not particularly good. The **accessibility of texts** is quite good since this is public data once it is one Twitter’s platform. It is a bit diminish by the fact that we used a particular software to retrieve them. The **contextual text quality**, meaning that the quality “*must be considered within the context of the task at hand*” is very good. Since we want to analyze the impact of the Diesel Gate on Volkswagen digital representation each tweet speaking or referring to Volkswagen is relevant in the analysis. And last but not least, **representational text quality and deficit resources**. Sonntag describe two problem. Single source problems such as “*wrong formulated data values, typing errors, different spellings of same word, co-reference problems, and lexical ambiguity*”. And multi-source problems “*homonym name conflicts and document duplicates*”. For Twitter in general, these problems are the scourge of any good analysis. We will find slang, a lot of typing errors, insults, irony and a lot of duplicates with some bots posting a lot of time the same tweet.

Overall, we could divide our data quality assessment in two categories. The relevance for the subject would be the first one. As a matter of fact, Twitter listening always brings a lot of noise data but most of the time it is consistent in quantity or content. Thus, the real content which is interesting is always found and relevant to the listened subject, here the reaction on the Diesel scandal. The second category would be the raw quality of the textual data collected which is quite poor.

# Analysis

## Exploratory Analysis

As a reminder, we have an exploitable set of tweets made of 339 367 tweets, the remaining 38% of the original dataset. After the data collection and processing we will now proceed with the analysis of these tweets. The first step is to understand the overall structure of these tweets.

The base ground of Twitter textual analysis is to know what people are talking about and in what sense. The simplest approach, and yet efficient, is to know which are the **most frequent words** used by the users.

|  |  |  |
| --- | --- | --- |
|  | **Word** | **Count** |
| N°1 | emiss | 90 254 |
| N°2 | scandal | 89 313 |
| N°3 | car | 57 082 |
| N°4 | diesel | 37 303 |
| N°5 | cheat | 36 875 |
| N°6 | ceo | 35 819 |
| N°7 | new | 29 424 |
| N°8 | say | 27 598 |
| N°9 | test | 23 192 |
| N°10 | via | 20 499 |

We can see where the interest of most tweet is. Out of the top 10 most frequent words, 4 words (emiss, scandal, cheat, test) and possibly two others (diesel, ceo) refers to the Volkswagen scandal.

To strengthen the analysis, we will not only focus on the 10 most frequent words, but we are going to go look for the 150 most frequent. In order to keep it understandable we are going to draw word clouds to help us grasp the main topics discussed in the tweets. We used the Python package wordcloud[[4]](#footnote-4) to generate the word clouds from the frequency distribution of every word.

Word clouds is a smart way of displaying the most frequent, or important word in a corpus. In a business point of view, it can help identify customers’ pain points. As the words vary in size according to their frequency the bigger words mean that they bear a lot of importance. In our context, bigger words would mean the most tweeted. Of course, we will find again the same words as in the top 10, but the others are as significant.

In the word cloud below (**Figure 1**), we notice a lot of words around the former Volkswagen CEO Martin Winterkorn with words as *chief*, *boss*, *ceo*, *martin*, *winterkorn*, *board*, *resign*, *execut* (for executive for example). We can also notice reference to the stock price of Volkswagen with *price*, *action*, *market*, *stock*. We find among all the tweets about the scandal some more common with words like *sale*, *audi*, *golf*, *buy*, *beetl* (for beetle one of their car), *porsch* (for Porsche the brand).

**Figure 1**



We will also look at the hashtags used in the corpus of tweet because it is a highly relevant information on Twitter. However, on that corpus, it appears that hashtags are more of a noise than anything else. In **Figure 2**, we can see things like *#freedomforkesha*, *#worldrhinoday* or *#madeintheam* that bear no sense and that actually are bots polluting the space (see **Figure 3**). You can still see some relevant though with *#fraud*, *#dieselgate* or *#vwscandal*.

In fact, bots are a large part of Twitter’s traffic, especially in 2015 when Twitter had not yet strengthened its automation rules[[5]](#footnote-5). These bots use a large quantity of hashtags to try to stay relevant and have a proportion to post the same post all over again. For example, the tweet in **Figure 3** has been posted by the account 893 times!

**Figure 2 – Word Cloud by hashtags**



**Figure 3 – Screenshot of a bot**



Finally, for the word clouds, we are going to split our corpus of tweets in 7 parts from the most recent to the oldest. We choose 7 because it allows to have the same number of tweets in all splits. The output will be 7 different word clouds that will represent the evolution of the words used through our studied period, from September 1st, 2015 to October 15th, 2015. At the beginning of the period, the scandal of Volkswagen was already pretty advanced and words like *emiss*, *cheat* and *recall*, which is one of the most frequent word in the first word cloud. The more time passes the more words like *scandal*, *winterkorn*, *ceo* and *resign* appear.

**Figure 3 – 7 Word Clouds throughout September 1st, 2015 to October 15th, 2015**

The individual word clouds are in Appendices x

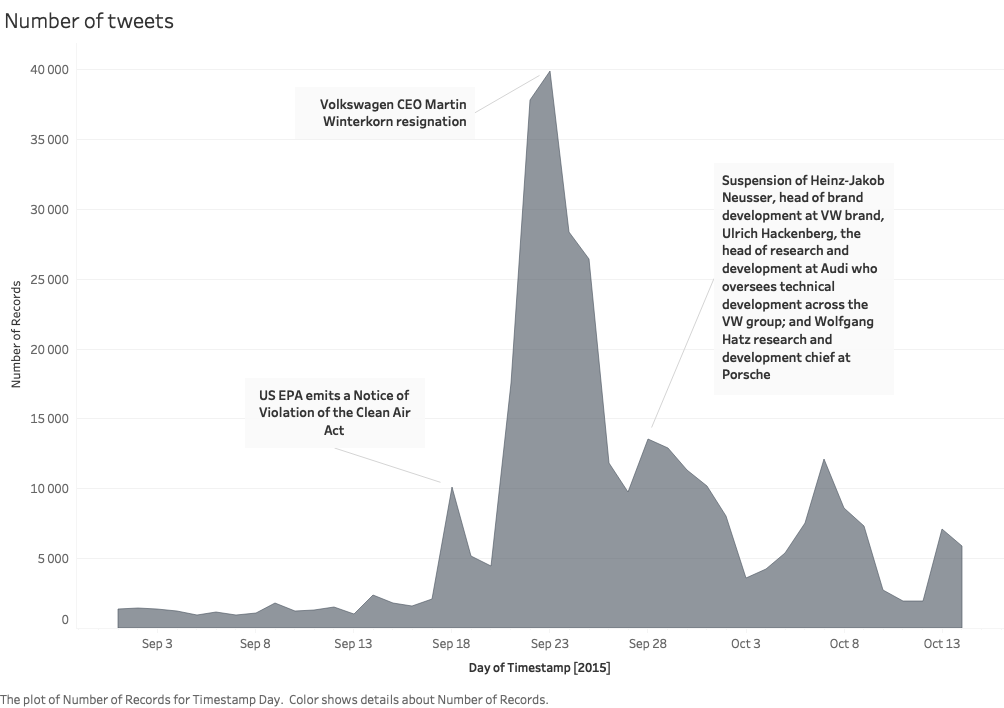


This interest is also seen in the number of tweets. During the monitoring of a Twitter hashtag of word (here volkswagen), one of the most interesting things to follow is the number of tweets emitted by day. The rise and fall of interest in a subject in a social network like Twitter is highly correlated to the number of tweets written. The more tweets the more likely it is for a user to see at least one about it and react. Most of the time, increase in volume comes from an event that triggers reactions all across the network.

The following figure (**Figure 4**) is the number of tweets emitted per day during the studied period. We can see a net increase of the number of tweets on September 18th, following the emission of a Notice of Violation (NOV) of the Clean Air Act on Volkswagen Group by the US Environnemental Protection Agency (EPA)[[6]](#footnote-6).

We also can see the immense increase in volume between September 22th and 23th following among other the resignation of Volkswagen CEO Martin Winterkorn *Volkswagen* CEO Martin Winterkorn[[7]](#footnote-7). It is also remarkable that these picks in the number of tweets and reactions to the subject is parallel to a fall of 40% of the action of Volkswagen at the Frankfurt Stock Exchange[[8]](#footnote-8).

**Figure 4 – Number of tweets**

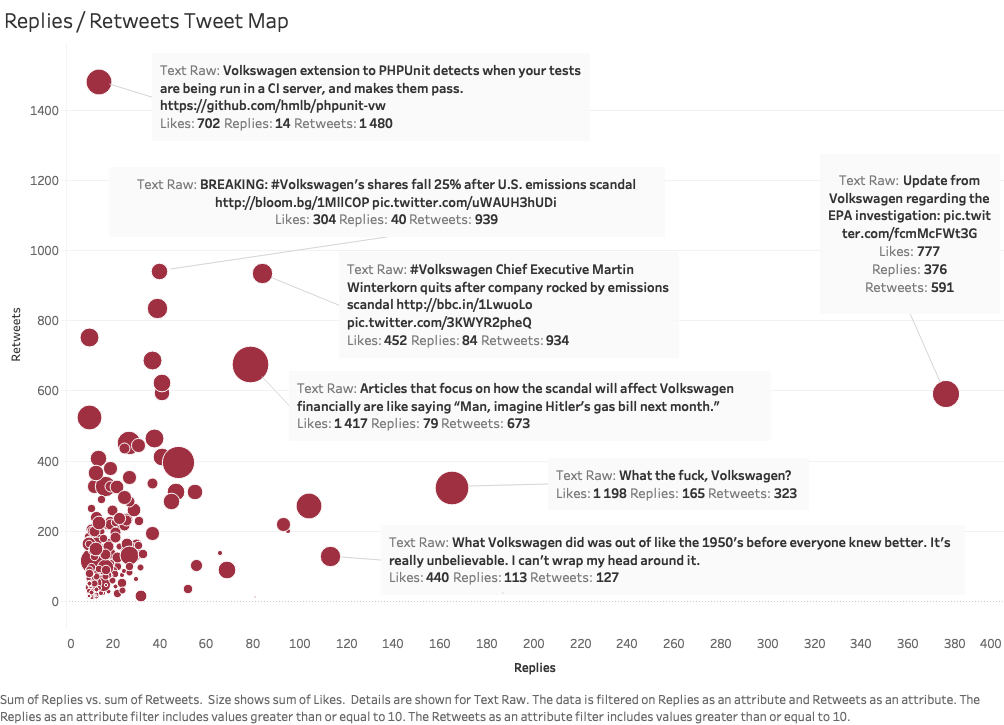


However, just relying on the most frequent words and the assumption that the increase in volume are due to Diesel Gate related events is not enough. To see with more accuracy what are the topics that the users care about, we are going to look at a Replies/Retweets Map (**Figure 5**). This map acknowledges that the two most important metrics to measure the virality of tweet is the number of replies it gets, and the number of times it is retweeted. The map allows us to see that, along to the size of each tweet that is proportionate to the number of likes it received. Note that on that graph, for better computation and visibility, all tweets with less than 10 replies and 10 retweets have not been drawn. That leaves us with 213 tweets that match these criteria.

We have highlighted 7 tweets, with among them those which have been the most retweeted and with the most replies. We observe with no surprise that these tweets all talk about the scandal but in a very different way.

For example, the most retweet tweet “*Volkswagen extension to PHPUnit detects when tour tests are being run in a CI server, and makes them pass.*” is a joke from developers that developed packages in several programming language that allowed to pass all unit tests if the test were conduct in a test environment, which is always the case. We observe also that news tweets take a major part with 3 tweets out of 7. The 3 remaining tweets express the irritation of users in reaction to the announcement of the scandal.

**Figure 5 – Impact of tweets map**



## Sentiment Analysis

## Clustering

Explain what I am trying to achieve 🡪 Monitor given categories, perform analysis inside, allow to track opinions on different topics in the same network of people

Explain what I did

The algorithms and their differences

The outputs

## Twitter Network

Understand who is important in the network. The different classes inside. Identify potential threats and opportunity

Edges and nodes concept

Viz

# Conclusion

# Appendices

1. A personalized floor mosaic from Pompeii, Robert I. Curtis. https://www.jstor.org/stable/504744?seq=1#page\_scan\_tab\_contents [↑](#footnote-ref-1)
2. Twitterscraper Package by Ahmet Taspinar, MIT License, <https://github.com/taspinar/twitterscraper> [↑](#footnote-ref-2)
3. Daniel Sonntag, in *Assessing the quality of natural language text data*, <http://www.dfki.de/~sonntag/text_quality_short.pdf> [↑](#footnote-ref-3)
4. Wordcloud Package by Andreas Mueller, MIT License, <https://github.com/amueller/word_cloud> [↑](#footnote-ref-4)
5. Twitter Automation Rules from Twitter website, <https://help.twitter.com/en/rules-and-policies/twitter-automation> [↑](#footnote-ref-5)
6. US Environmental Protection Agency. 18 September 2015, <https://web.archive.org/web/20150919011334/http://yosemite.epa.gov/opa/admpress.nsf/d0cf6618525a9efb85257359003fb69d/dfc8e33b5ab162b985257ec40057813b!OpenDocument> [↑](#footnote-ref-6)
7. Volkswagen Group. <https://web.archive.org/web/20150925115541/http://www.volkswagenag.com/content/vwcorp/info_center/en/news/2015/09/Statement.html> [↑](#footnote-ref-7)
8. Volkswagen perd en Bourse près de 25 milliards en trois jours, Le Figaro, 22/09/2015

   <http://bourse.lefigaro.fr/indices-actions/actu-conseils/volkswagen-perd-en-bourse-pres-de-25-milliards-en-trois-jours-4632836> [↑](#footnote-ref-8)