PAGE DE GARDE

TABLE OF CONTENT

[Introduction to Social Listening 3](#_Toc10102723)

[Internet and Traditional Marketing 3](#_Toc10102724)

[Lectures 4](#_Toc10102725)

[Methodology 5](#_Toc10102726)

[Data Preparation 6](#_Toc10102727)

[Analysis 7](#_Toc10102728)

[Conclusion 8](#_Toc10102729)

# 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

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 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. For example, “I liked fishes” and “I like fish” should be considered synonyms since it is the same sentiment. Stemming will output “I like fish” for both tweets. It suppresses the tense mark, the plural mark and reduce to their root adverbs and such.

At the end of the all cleaning process, our tweet “I liked a @YouTube video http://youtu.be/pmzZbUioFAQ?a\xa0 2015 Volkswagen Sales Event | “Model Rear End” Passat Commercial” has been transformed to “like youtub video sale event rear passat commerci”

## Data quality

# Analysis

# Conclusion

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)