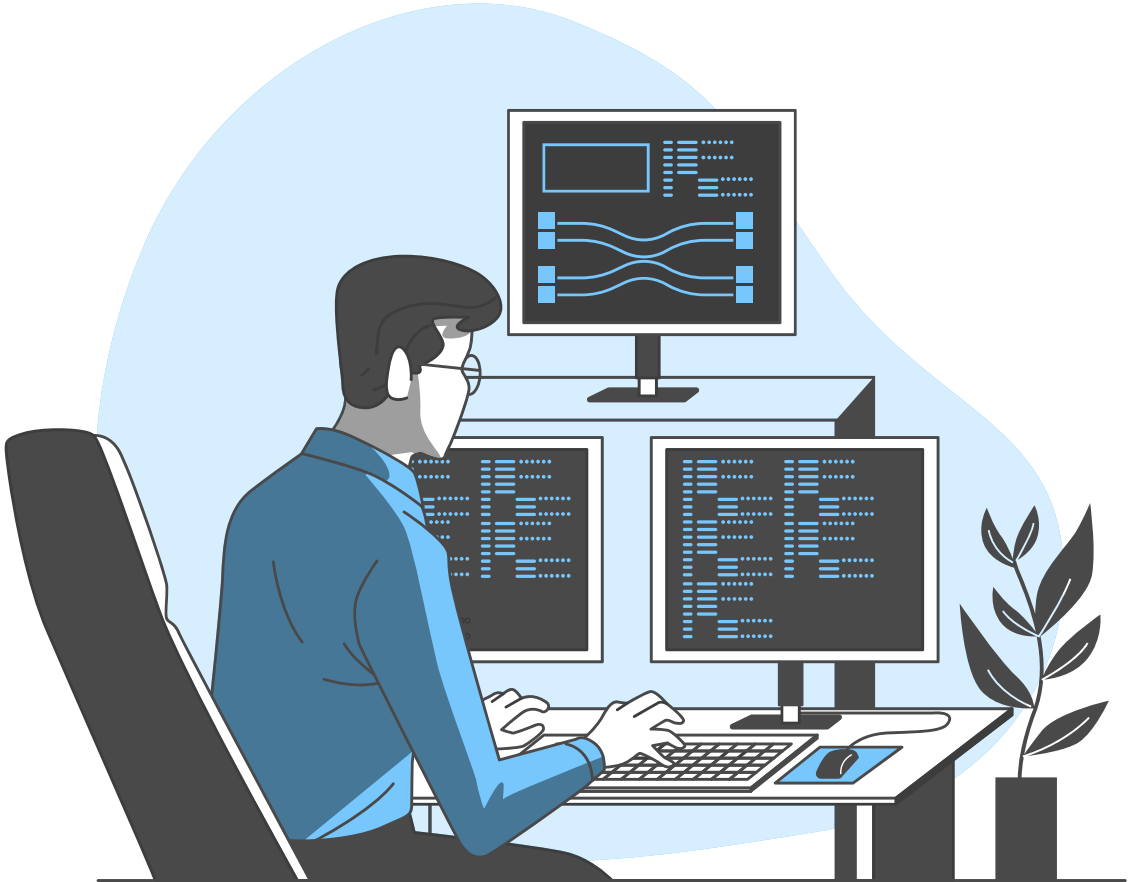
https://colab.research.google.com/drive/1Pm3eFsq1r0B9xuWmNF0Fpd56VaY6-b6I?usp=sharing



Sentiment Analysis

Business & PM Project

Project goal and Research Questions

The research will focus on the topic of **consumer sentiment**, focusing on **sentiment analysis**

First Question:

The research will focus on the following question:

"Can sentiment analysis technology be used in a realistic business environment to understand customer feedback left on social media?"

Then we will analyze some textual reviews and we will label them with our rating model, and we will understand how much the rating of the classified reviews differs from the automatic system, from the rating

of real reviews (assuming that the stars left by users are balanced

with the textual review left).

#1 Question

01

Process Description

how we operated

02

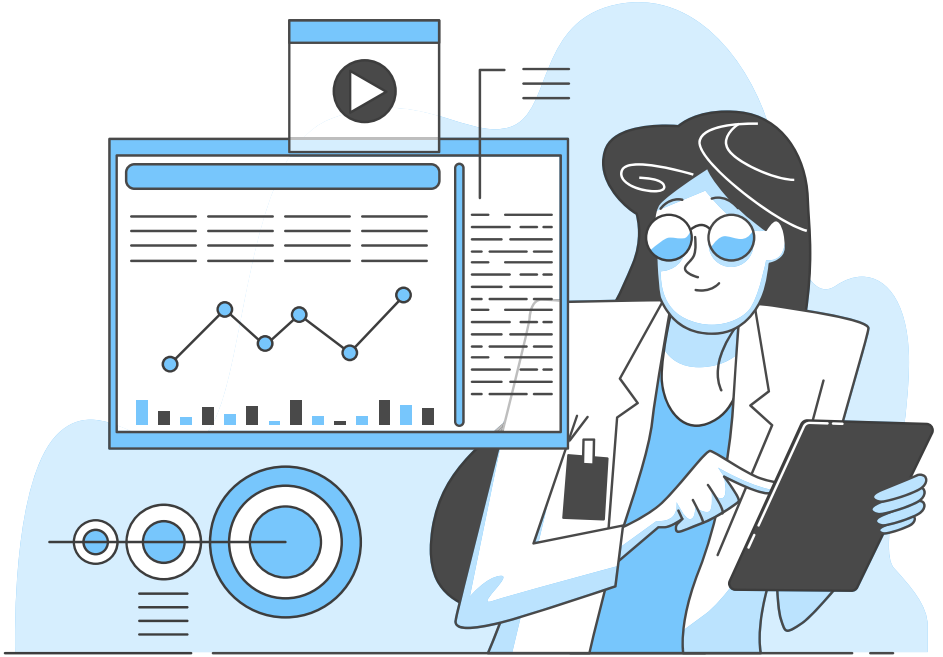
Technical Report

code and implementation

03

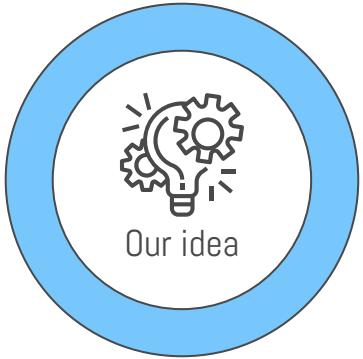
Result Discussion

Analysis on result



01

Process Description



Process:

Our goal is to understand how much the results extracted from the automatic sentiment analysis differ from the real feelings left in the online reviews, in this case hotel reviews.

After **training** several classifiers and **choosing** the best one, the comments will be **automatically classified** with the npl technique in **Good, Neutral, Bad** and then they will be **compared** with the real ratings of the reviews.

All the technical details and the various steps will be shown in the “Technical Report” below.

02

Technical Report

Our Solutions

01 Dataset

Taken from kaggle

02 Dataset Splitting & Label

03 Cleaning Phase

04 Text Elaboration

05 Feature Extraction

tfidf

06 Cross Validation

Choose best

07 Final Classifier

Used in app

1. Dataset

Source: https://www.kaggle.com/datafiniti/hotel-reviews
Attribute : 25
Dimension : 30k review

Attribute selected:
-Hotel Name
-Hotel Country
-Review text
-Review date
-Review Rating

Attribute
Id
dateAdded
dateUpdated
Adress
Categories
primaryCategories
City
Country
Keys
Latitude
Longitude
Name
postalCode
Province
Review.date
Review.rating
Review.link
Review.title
review.userName
review.userCity

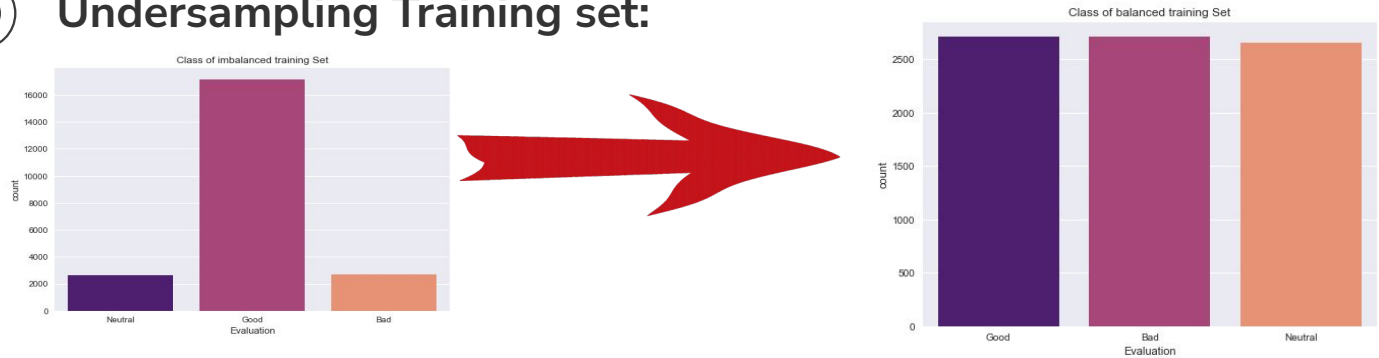
2. Dataset Splitting & Label Phase

Split Phase:
Training set = 75%
Test set = 25%

Label Phase:

- 4-5 stars→ Good
- 3 stars → Neutral
- 1-2 stars → Bad

Undersampling Training set:



3. Cleaning Phase

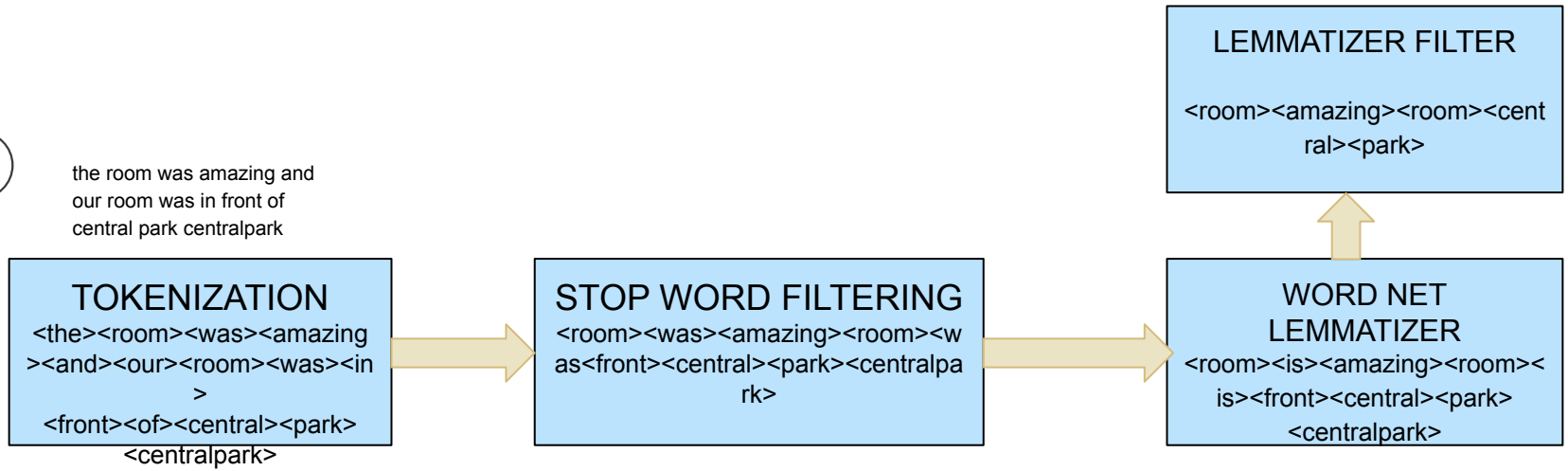
1. Lower case words
2. Remove punctuation from sentences
3. Remove mention @ or Hashtags from sentences
4. Remove link http from sentences
5. Remove emoji from sentences

```
#cleaning
review = review.lower()
review = review.translate(str.maketrans('', '', string.punctuation)) #remove puntuaction
review = re.sub('@[^\s]+','',review) #remove mention
review = re.sub('http[^\s]+','',review) #remove link
review=emoji_pattern.sub(r'', review) # remove emoji
```

4. Text Elaboration

In this phase we took the cleaned text coming from the previous one and we did:

1. Tokenization
2. Stop word filtering (english dictionary)
3. WordNet Lemmatizer
4. Lemmatizer filtering (reduced number)



5. Feature extraction technique - TFIDF

For our proposal we used a **Term Frequency - Inverse Document Frequency** in order to get the relevance of the single token of the sentence.

The following sparse matrix is a compressed view of the matrix .

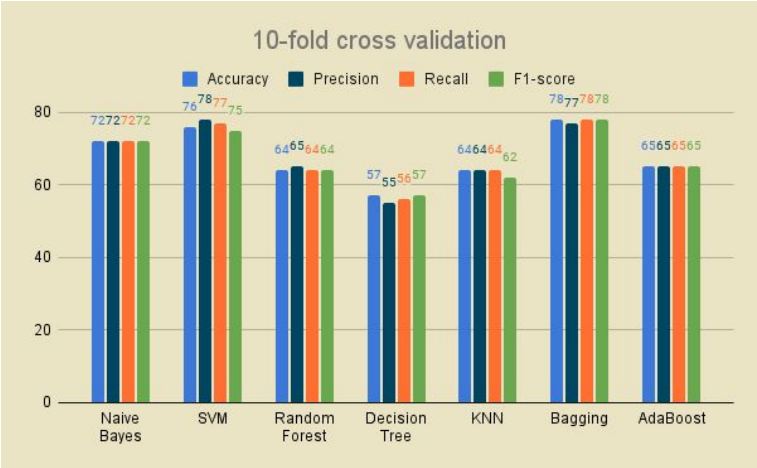
	zealand	zephyr	zero	zion	zip	zone	zoo	zorro	zu
0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
...	...	...	...	...	...	...	...	...	...
8070	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
8071	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
8072	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
8073	0.0	0.0	0.0	0.0	0.0	0.0	0.268154	0.0	0.0
8074	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0

6. Cross Validation of Classifier

Classifiers that have been **used** are:

- Multinomial Naive Bayes;
- Support Vector Machine;
- Random Forest;
- Decision Tree;
- KNN;
- Bagging;
- AdaBoost.

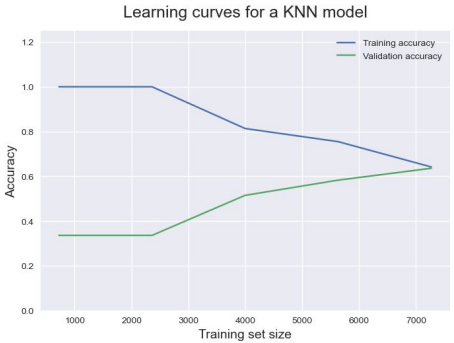
We used the K-FOLD to **evaluate** the accuracy of different Classifier

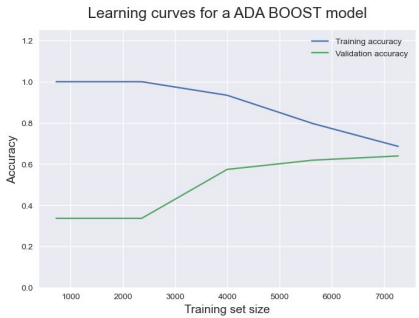


best classifiers that are not overfitted are **Bagging** and **SVM**

6. Cross Validation of Classifier

We also analyzed the learning curves to understand if they were **overfitted**





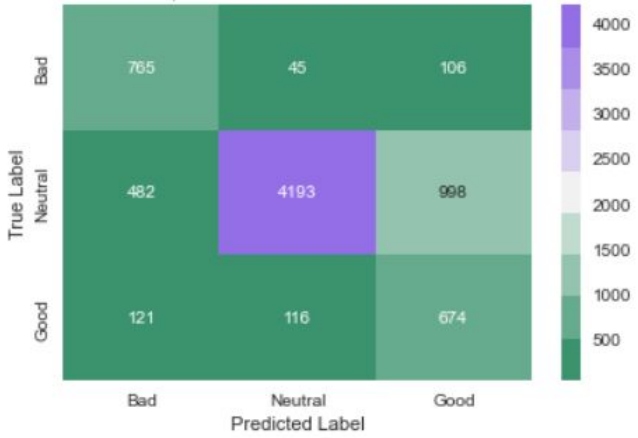
We used the t test to compare the two best classifiers obtained from the k fold. And these are the results:

```
from mlxtend.evaluate import paired_ttest_kfold_cv

t, p = paired_ttest_kfold_cv(estimator1=clf,
                             estimator2=bgclassifier,
                             X=X_train_tf,
                             y=np.ravel(df_train_y.astype(int)),
                             random_seed=1)

print('p value: %.6f' % p)
```

```
p value: 0.000498
```

Since $p<\alpha$ (=0.05), we can accept the Alternative Hypothesis. This means that we take the best classifier between the two to build our application. **Bagging Classifier**

6. Final Classifier



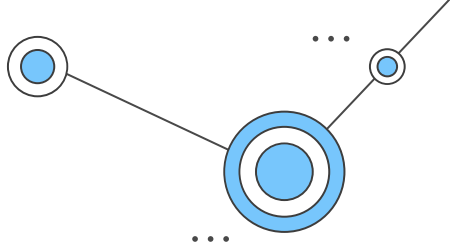
The Classification report:

	precision	recall	f1-score	support
Bad	0.56	0.84	0.67	916
Good	0.96	0.74	0.84	5673
Neutral	0.38	0.74	0.50	911
accuracy			0.75	7500
macro avg	0.63	0.77	0.67	7500
weighted avg	0.84	0.75	0.78	7500

Accuracy: 0.75

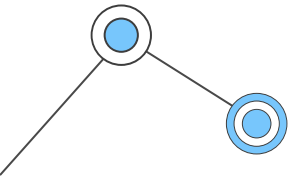
The **bagging classifier** was used for our app

Neural Network Classifier



Neural networks usually **require much more data** than traditional machine learning algorithms, as in at least thousands if not millions of labeled samples. This isn't an easy problem to deal with and many machine learning problems can be solved well with less data if you use other algorithms.

Usually, neural networks are also more **computationally expensive** than traditional algorithms

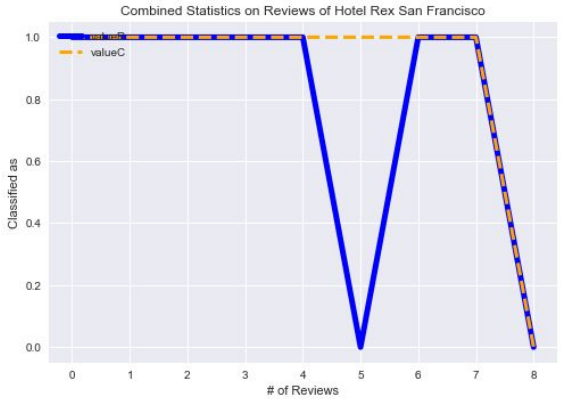


03

Result Discussion

Differences between REAL vs AUTO rating

Assuming the graph given by the rating (BLUE) as a real trend, the trend calculated with our classifier (ORANGE) doesn't differs so much.



YES!

this software could be used in real life

We compared our software with software currently on the market and it was found that the achievable accuracy is ~ 85%.

Various Purpose for our Software:

- Have an **overview** of the reviews of their products, aggregating the results of different sites;
- Use sentiment analysis to instantly know when the **brand is mentioned** in a negative way, in order to understand the origin of that data and not ruin the reputation.
- Use this tool to understand in automatic way , the **satisfaction of the clients** after using customer service of the company
- preserve the online **reputation** of the company

software improvements:

- See positive and negative mentions volume and reach to see how your **reputation changes over time**;
- Track reach and learn how many people **talked about your new campaign** or a press release. To understand which hut has the best performance;
- Track the behavior of the entire sector or of a direct competitor to understand if there is a possibility of improvement and **diversification** of the product. Perhaps by modifying the software you could also understand the negative and positive aspects mentioned in the review so that you can more accurately extract the data.
- By analyzing the reviews of the company and the sector, it would be possible to create a "**Perceptual Map**", selecting the attributes that most interest us for our analysis.

Analyzing the feedback is very important, because it could increase the experience and implement the “**learning curve effect**” , furthermore the characteristic of Ai companies is to learn much faster than traditional companies..

Our Consideration

- we must avoid having a few large buyers, so as not to give them too much “Power of Buyer”
- we must create a service that is difficult to replace
- being a growing sector we should have less aggressive rivals
- We aim for differentiation, therefore with unique features different from competitors. Such as an after-sales team that analyzes customer data once a month.
- being a data analytics company, we could easily do a market segmentation
- our services are mainly offered and b2b, so our strategy will have to be adequate, having experienced customers
- Positioning : the company must be credible, clear and consistent in the advertising of the services offered
- being a service, distribution is simpler, in fact we can have a direct channel.
- For communication it might be ideal to choose online communication. Like for example email marketing and social media, a content site on the topic of AI could also be created, to attract new buyers and industry experts, as Bitdefender did.
- Being a service it can be sold internationally, but in this case you have to be careful of the "price escalation" due to various factors.
- If we have to launch a new service, a good pretotyping strategy could be the "Fake Door", as for example did Airb & b

Project goal and Research Questions

The research will focus on the topic of **consumer sentiment**, focusing on **sentiment analysis**

Second Question:

The second research focuses on the following question:

“Which are the main topics of the sentiment analysis application in the management field?”

A company must always keep in touch and grow along with the evolution of technology. We focused our attention in the detection of which are the main uses and possibilities of the application of the sentiment analysis through the analysis,conducted by using Natural Language Processing techniques, of papers taken from MIT Sloan school related to the management and business field.

#2 Question

01

Process Description

how we operated

02

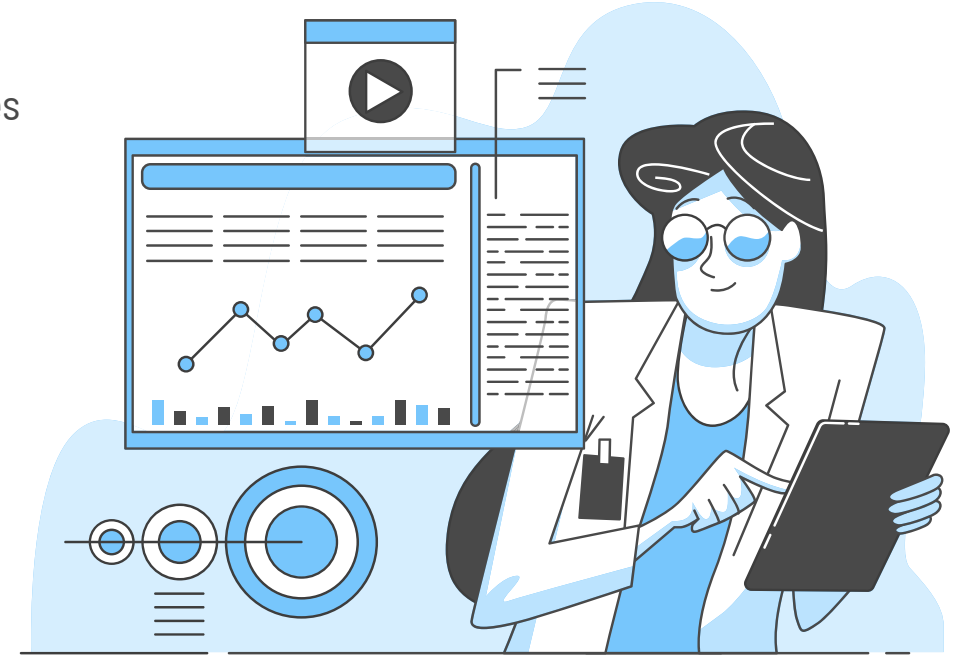
Technical Report

Description of the processes actuated

03

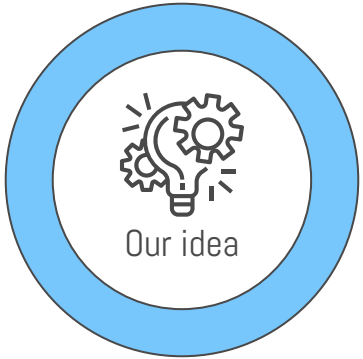
Result Discussion

Analysis on result



01

Process Description



Process:

The process of the second question consists in the use of the natural processing language on a series of papers in the field of management taken from MIT's management school.

After a first conversion of the papers, we select only those containing information about data mining and sentiment analysis, then,after a pre-processing phase, NLP techniques were used to obtain readable diagrams containing the main topics inside the documents (at first considering the papers one at a time and then all them joined together.

02

Technical Report

Our Solutions

01 Dataset

A series of papers scraped from the MIT sloan website.

02 Conversion & filtering

03 Text PreProcessing

04 Topic Extraction

05 Topic Analysis

06 Final considerations

1. Dataset

The Dataset used for this section consists in 128 papers taken from the MIT Sloan management school covering the years from 1990 to 2022. For each year there are four papers, one for each season (spring,summer,fall and winter).

Inside each paper there are multiple articles related to many different topics about business and economy divided as:

Link	Full page	Title	Author	Year	Text	References	About

2. Conversion & Filtering

The files scraped were in the RDS format so before any other operation we had to convert them into CSV format.
In order to do that we used a script in R which took the whole folder containing the papers and converted each of them into the requested format.

Once we've obtained the converted CSV files we then proceeded to searching inside each paper, all the articles containing words relative to our field of interest, in the specific we used a small dictionary containing words related to the data mining and sentiment analysis field.

3. Text Cleaning + Preprocessing

Once we obtained the filtered papers we then proceeded to clean the text by

- Removing stopwords
- Removing punctuation
- Convert the text into lower case
- Remove tags, special characters, digits, links

After that we applied

- Stemming: in order to remove suffixes from words
- Lemmatization: used to group together words by their inflicted form.

4. Topic Extraction

We adopted different strategies based on the NLP techniques in order to retrieve the possibles and most important topics inside the papers.

- Word Cloud
- One,Two,Three words n-gram
- TF-IDF technique

We chose these techniques in order to retrieve in an easy and readable way for the user the topics of interest.

4. Topic Extraction

World cloud

Returns an image showing the most frequently used words inside the document based on the size of the text itself. The bigger the text the higher is the frequency of a single word.

4. Topic Extraction

One,Two.Three words n-gram

We proceeded to create three different n-grams showing the frequency of n group of words inside the document.

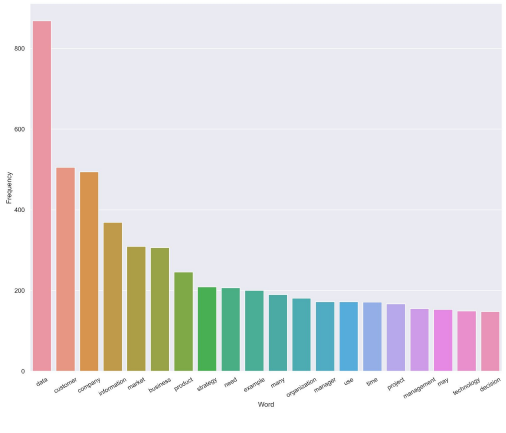
In order to do so we prepared the text in order to be recognized by the machine learning algorithms by **tokenizing** and **vectorizing** it.

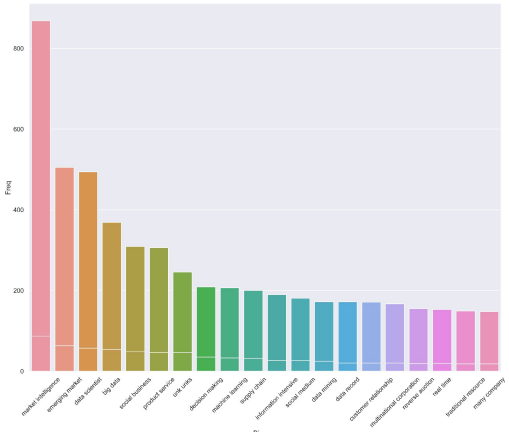
The tokenization transforms the continuous text into a list of single words, the vectorization will then convert the list into a matrix of integers.

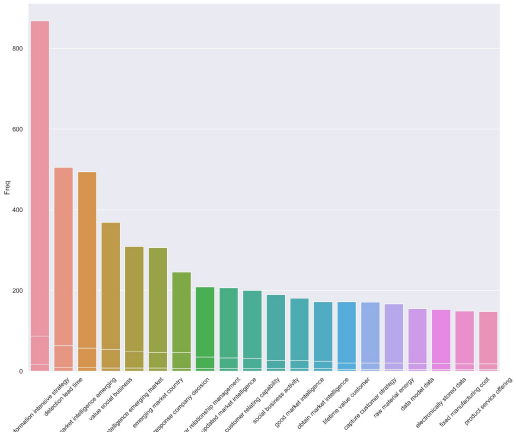
For doing so we used the bag of word model which ignores the sequence of words and only considers their frequency.

4. Topic Extraction

One,Two.Three words n-gram







4. Topic Extraction

Tf-Idf

The last npl process applied to the text is the TF-IDF vectorization. CountVectorizer focuses on the frequency of the words. By doing so it could happen that words with more important meaning but lower frequency may be discarded by the algorithm. By applying the TF-IDF vectorization we penalize in this case the words that appear several times in the document highlighting the ones which are more important to the context.

TF determines the Term Frequency of a single term.

IDF determines the inverse document frequency.

$$TF = \frac{Frequency\ of\ a\ single\ term}{Total\ number\ of\ terms\ in\ the\ document}$$

$$IDF = \frac{log(Total\ documents)}{Number\ of\ documents\ with\ the\ term}$$

03

Result Discussion

TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS

We now proceed to show a small relation about the results gained from the analysis of the topics retrieved by single documents and then by all the documents joined together.

The analysis of the single documents gave us incredible results that showed how companies evolved along the technologic progresses during the last 20 years.

TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS

1998-1999

15 October 1998: “Data as a Resource”

15 July 1999: “Winning in Smart Markets”

Just by taking a look at the graphs obtained through the use of the npl process we can observe that in the year 1998 companies already started to interact with electronic data gained from the internet by considering the uses of the user’s information and the problems of storage and maintenance of the data itself comparing them to the traditional ways of data retrieval.

TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS

1998-1999

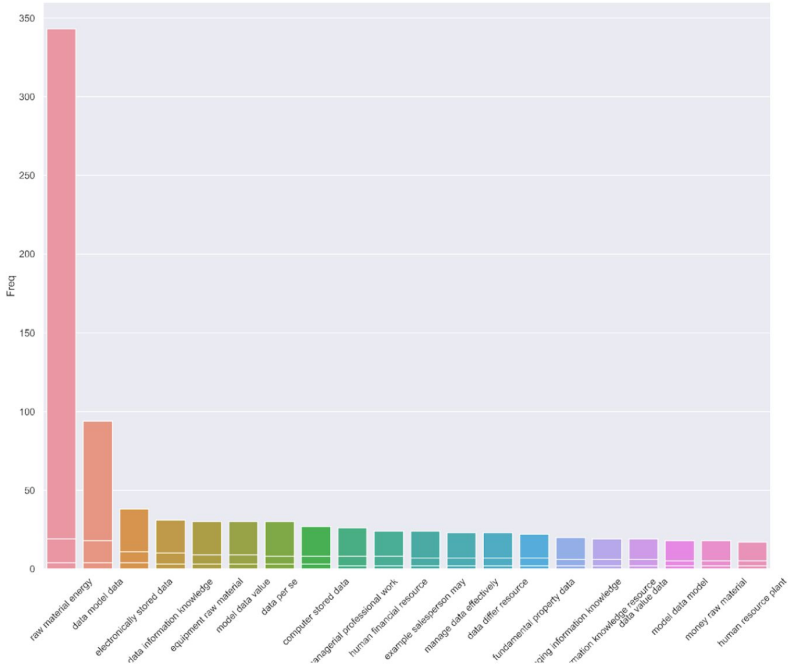
15 October 1998: “Data as a Resource”

15 July 1999: “Winning in Smart Markets”

After a year from the first paper we can see that the market companies already started to take advantage of the data analysis seeing in it a great opportunity in order to gain information about the customers preferences and behaviors and use them at they own advantage against the ones who didn’t.

“capture customers strategy”, “data information knowledge”, ”help marketing choice”, and “model data value” says a lot about how companies took interest in the digital data field.

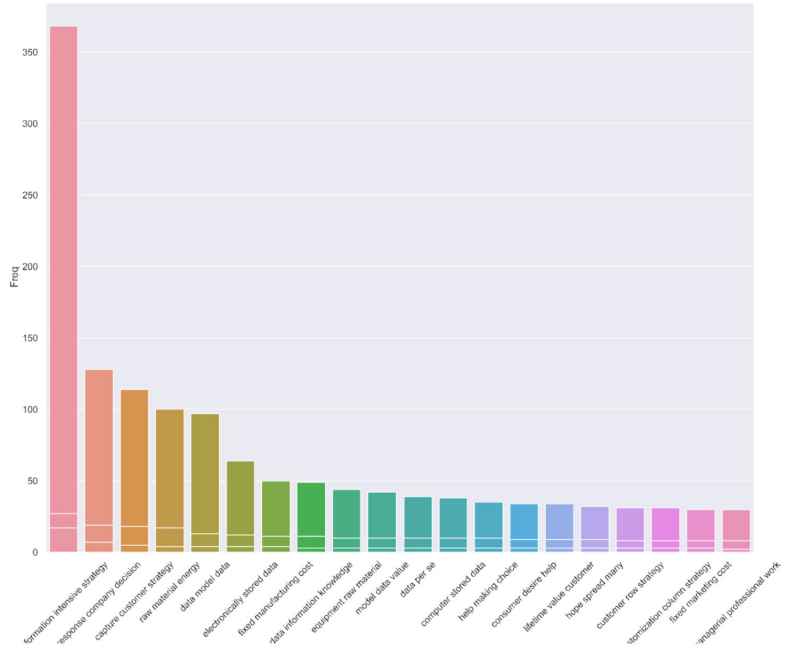
TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS



1998-1999

15 October 1998: “Data as a Resource”

TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS



1998-1999

15 July 1999: “Winning in Smart Markets”

TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS

2003-2008

15 April 2003: “Creating a Superior Customer-Relating Capability”

15 January 2005: “E-procurement”

01 October 2005: “Evolving from information to insight”

After almost 4 years since the last paper we analyzed, we can see how companies completely integrated the digital data analysis into their marketing strategy.

TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS

2003-2008

In this papers are mentioned topics as Smart markets, different types of market strategies related to the customer's data usage, responsive companies decisions, lifetime value customer (which it's started to be also correlated to the feedbacks and the behaviours of the subject not also by analyzing his purchases but also by the digital data gained from a larger view) and mass customization.

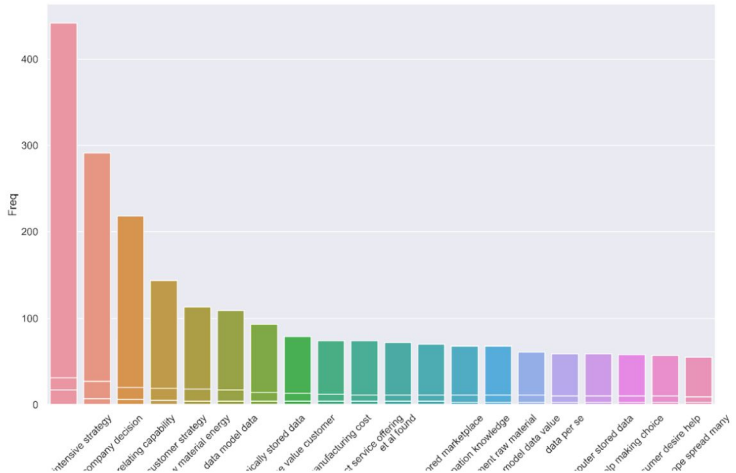
Companies started to find new ways and strategy in order to use the new huge amount of data at their disposal in order to achieve the best results.

The market started to adapt very rapidly basing his production decision on the customer's tastes, creating new and rapid strategies in order to prevail on other companies.

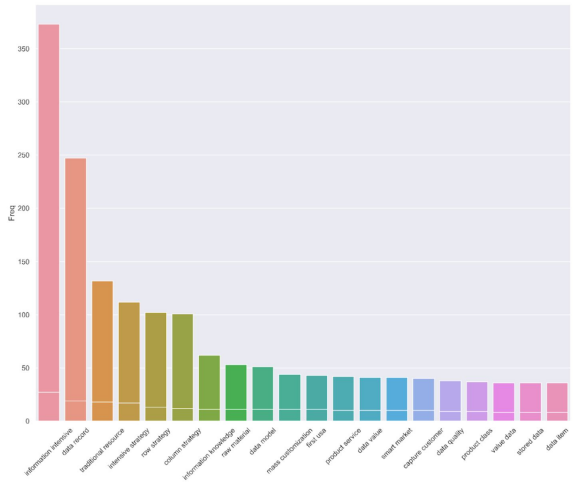
TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS

2003-2008

15 April 2003: “Creating a Superior Customer-Relating Capability”



01 October 2005: “Evolving from information to insight”



TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS

2010-2019

26 June 2010 “The IT Audit That Boosts Innovation”.

19 June 2012 “What Managers Really Think About Social Business”.

18 June 2013 “How to Drive Customer Satisfaction”

16 September 2014 “Getting Value From Your Data Scientists”.

15 September 2015 “Preparing for Disruptions Through Early Detection”

16 March 2015 “Reducing Unwelcome Surprises in Project Management”

14 June 2016 “Using Artificial Intelligence to Set Information Free”

14 November 2016 “Why Big Data Isn’t Enough”

5 Dicember 2016 “Mastering the Market Intelligence Challenge”

13 March 2017 “Building a More Intelligent Enterprise”

16 November 2018 “New Ways to Gauge Talent and Potential”

11 June 2019 “Strategy For and With AI”

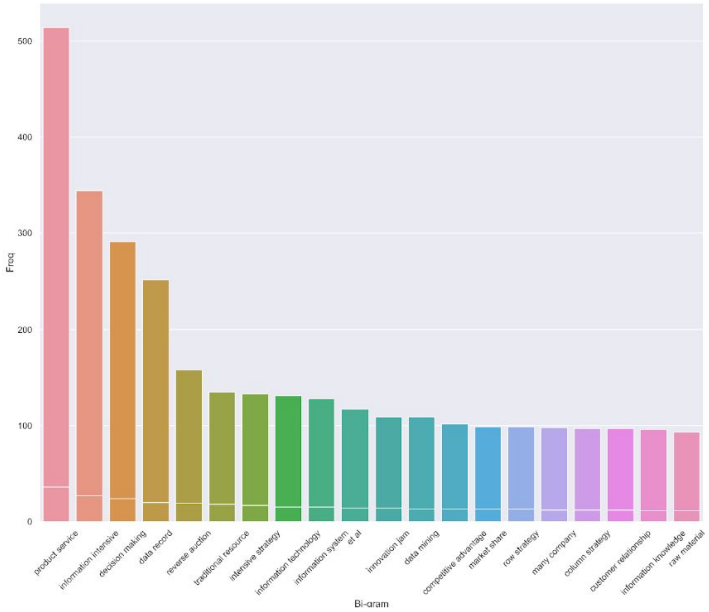
TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS

2010-2019

This are the years in which the main and consistent topics and evolutions of the market business are in and in which we'll focus our attention from now on.

During these years we have the advent of a new crucial figure in the data analysis, the Data Scientist, the advent of the social webs (in a total and preponderant way respect to the past)

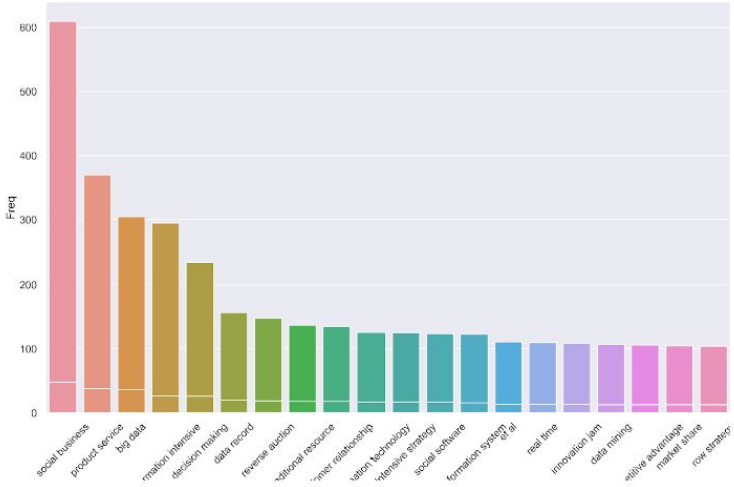
TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS



26 June 2010 “The IT Audit That Boosts Innovation”.

Innovation Jam is the keyword to the new digital world that was about to take life.
In 2010 the social websites were starting to take place and have a huge impact on people’s life and also opening new paths for the market to take along with new ways to acquire people and customer’s datas.

TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS

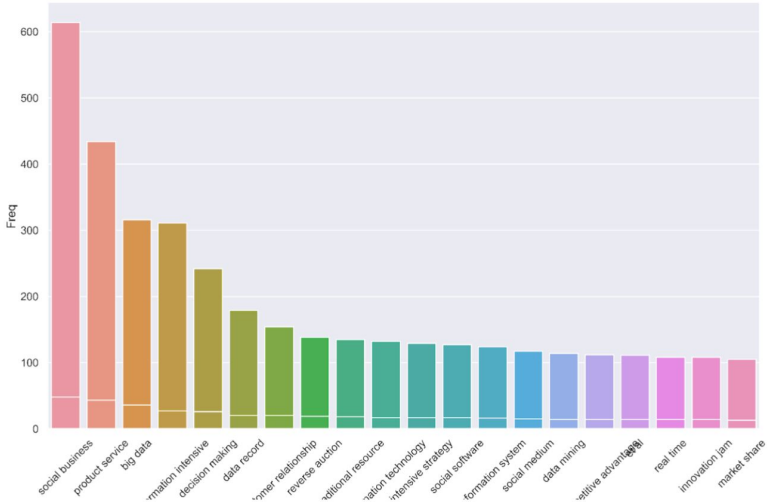


19 June 2012 “What Managers Really Think About Social Business”.

If until now the main topics found were about strategies,data treatment and acquisition here the main one becomes the social business along with intensive strategies.

As we said social networks were starting to have a huge impact on people’s life and the business field took interest in the phenomena starting to formulate new marketing actions in order to grow along the new web event that was spreading.

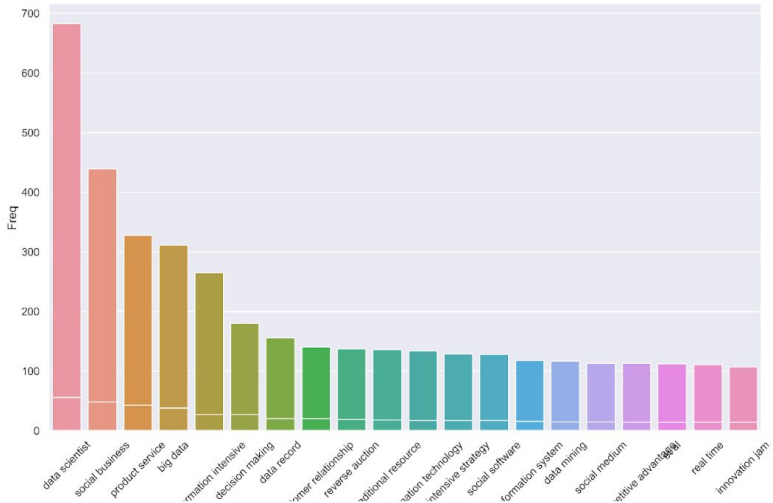
TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS



18 June 2013 “How to Drive Customer Satisfaction”

Social web component has already become a fundamental part in the marketing business field. Companies not only started to use the social websites like facebook and instagram in order to analyze trends and customers tastes but also to redirect customers own tastes and choices throw their products by precise and direct advertisements strategies.

TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS

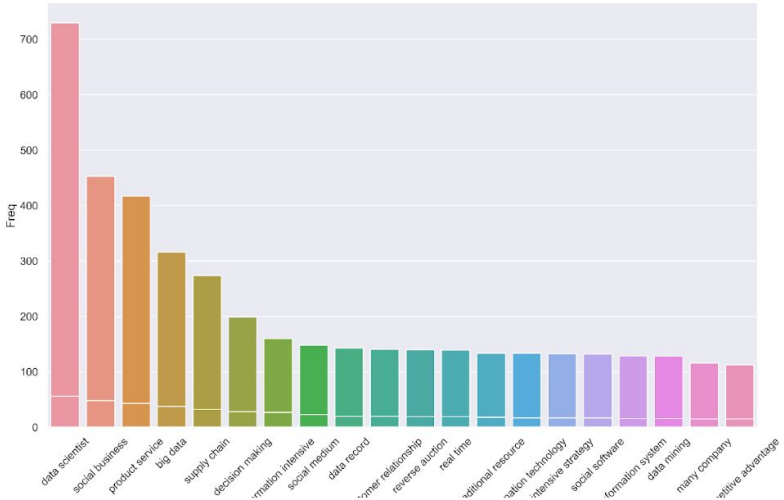


16 September 2014 “Getting Value From Your Data Scientists”

In 2014 and the data available for companies has become immensely larger than before so companies started to search for new figures that were capable to interact and operate with this huge amount of data rapidly and efficiently.

He’s capable of working with huge amounts of data by using machine learning algorithms and techniques in order to always retrieve the best quality information from them.

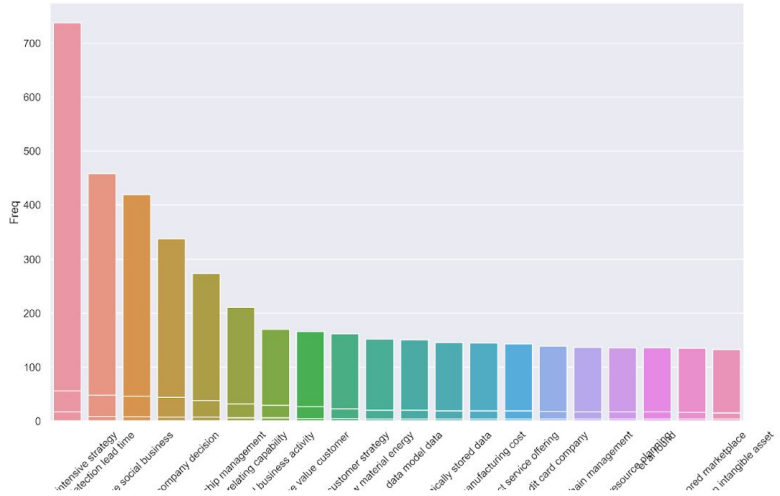
TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS



15 September 2015 “Preparing for Disruptions Through Early Detection”

I after a year the data scientist was a consolidated figure in the market

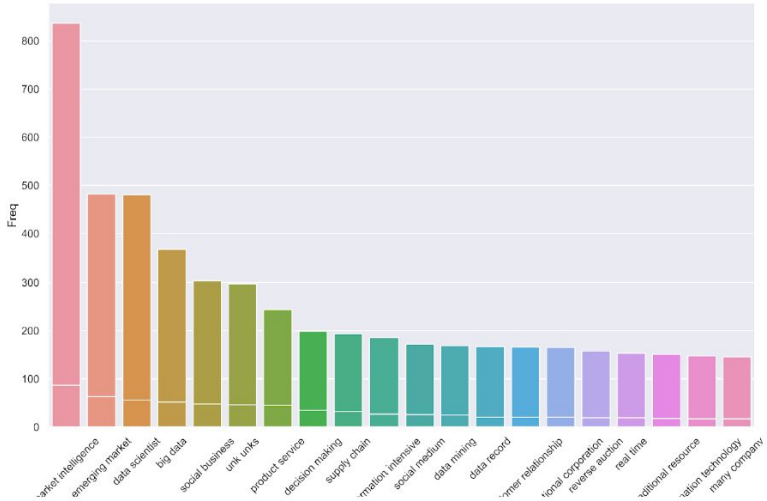
TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS



16 March 2015 “Reducing Unwelcome Surprises in Project Management”

Even working with huge amount of data, it is imperative in this field to be as fast as possible when working with data and this not only confirms the importance that the new figure of the data scientist covered but also how being fast in the data analysis could also save companies from taking or avoiding bad decisions also for prevailing on the concurrence and correct bad management mistakes in a rapid way.

TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS

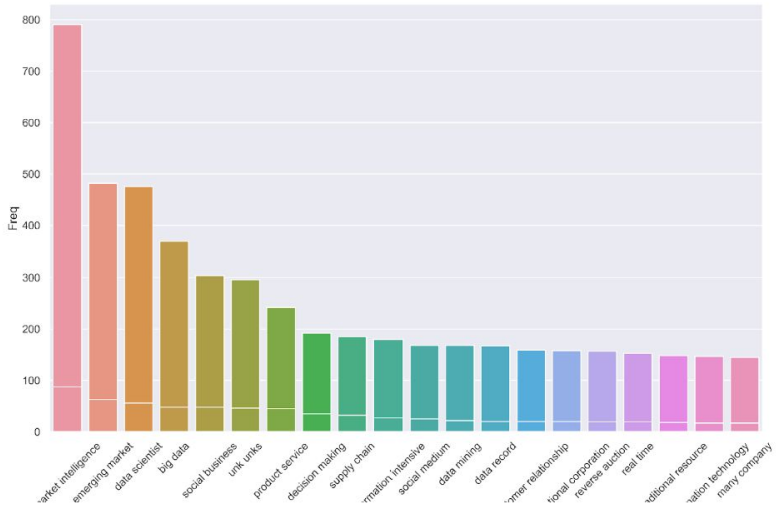


14 November 2016 “Why Big Data Isn’t Enough”

A.I. become the main argument for the marketing field. Companies started to adapt their strategies to the growing technologies rebuilding their way of action.The three main topics let us understand how the whole market nowadays is based on three key figures, the artificial intelligence, data scientist and big data analysis.

The growth of data on the web increases so rapidly that, in order to product a rapid and efficient analysis, always new techniques are required.

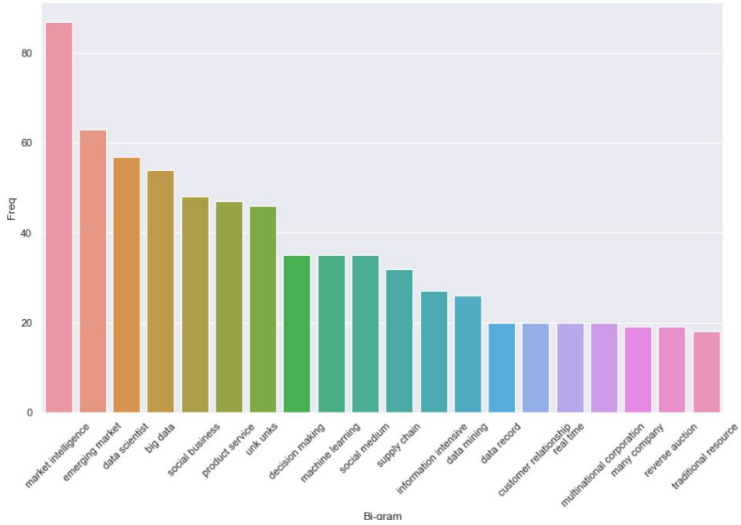
TOPICS FOR SINGLE PAPERS
THE EVOLUTION OF THE BUSINESS DATA ANALYSIS



5 December 2016 “Mastering the Market Intelligence Challenge”

Market Intelligence was the name given by the process with which a company take and analyze big quantity of data retrieved on many fields and use them in order to not only understand people’s feelings,tastes and moving trends but also to use the information retrieved in order to formulate marketing strategies and make predictions about market directions and customer choices.

TOPICS FOR THE WHOLE DATASET



We can see that the core of the market strategies is the usage of the AI and machine learning techniques. The data scientist became a fundamental role for business decisions and strategy adoption along with the big data analysis. Social networks are considered the main field on which a company can interact with customers and create new ones. The internet is now the principal field on which companies uses advertisement sell products and interact with people.

CONCLUSION

Topic analysis let us understand how and how much the business and market field transformed and adapted itself during those years.

We can understand by this study how much this two subjects are strictly related with the technology evolution and how is important for a company to always keep up with the world’s digital evolution both from a technological point of view and from the customer’s perspective.

Sentiment Analysis plays in this case one of the key roles in the social data analysis. Throw this technique a company will always be able to know the actual perception of people about a product or the company itself. Based on these feedback firms can improve their image along with their products in order to gain people's preference.