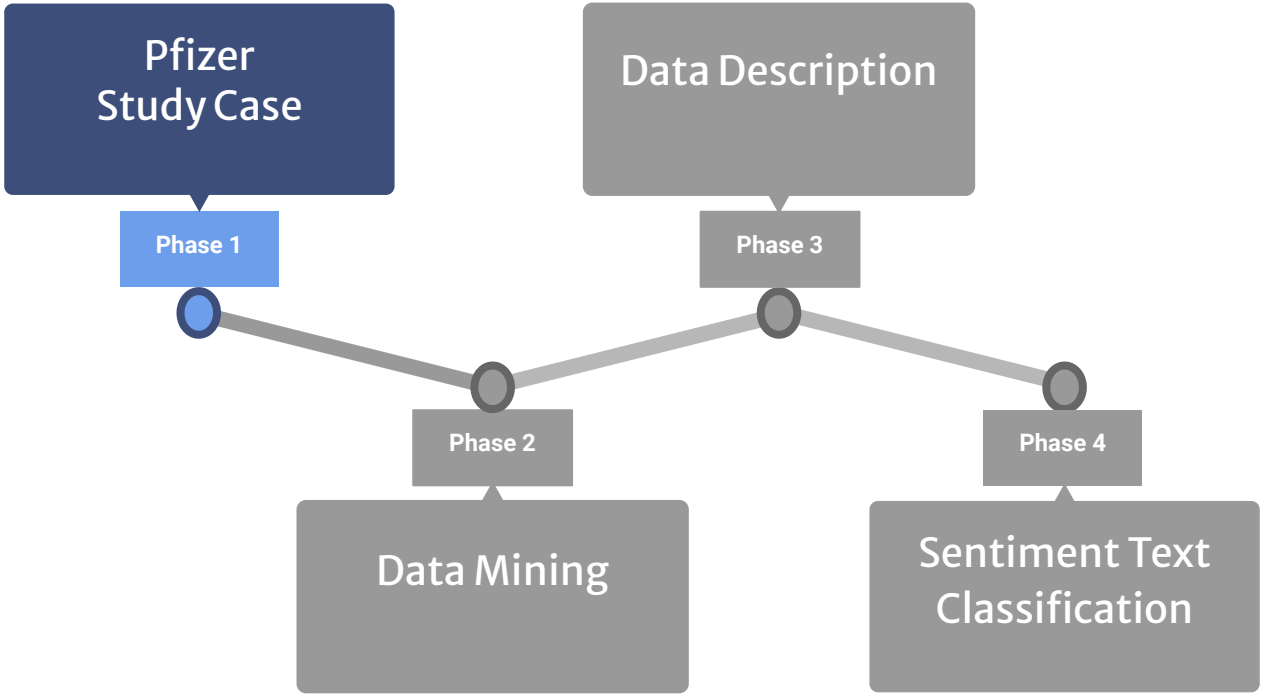
Natural Language Processing

Text Classification on Pfizer' tweets for Sentiment Analysis

Student:
Pietro Morichetti

RoadMap


Pfizer
Study Case
Data Description
Phase 1
Phase 3
Phase 2
Phase 4
Data Mining
Sentiment Text
Classification

Pfizer & Covid-19 case study

31/12/19

China Authorities

Wuhan (China) some cases of a suspect pneumonia have been observed among the open market for animals

31/01/20

First cases of this diseases in Italy

Prime Minister declares the sanitary emergency after two chinese tourists had the pneumonia

11/04/20

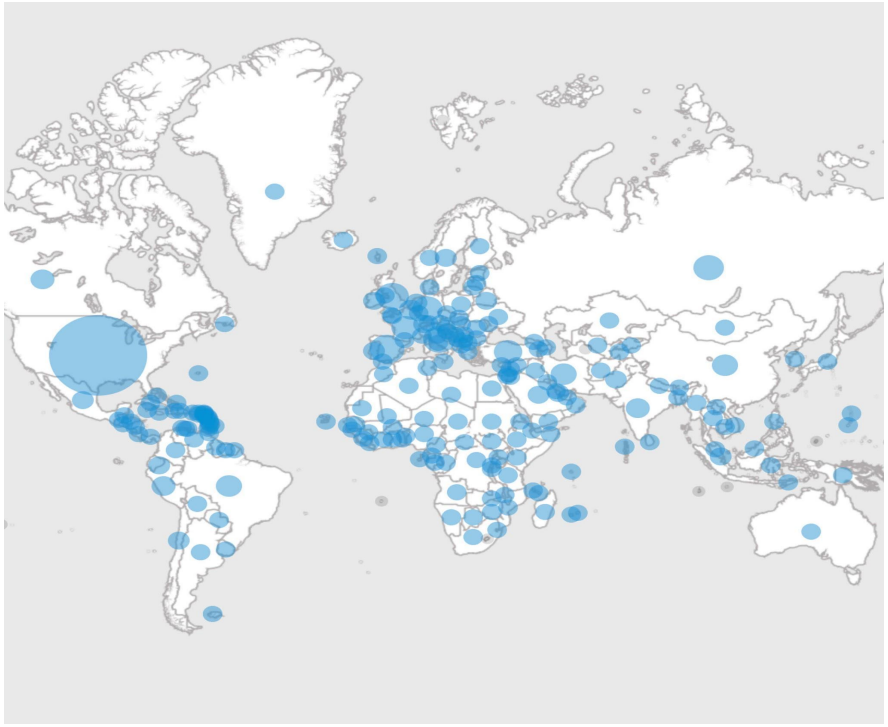
Global Pandemic

WHO Director-General called the spread of Covid-19 no longer an epidemic confined to certain geographical areas, but a pandemic spread across the planet.

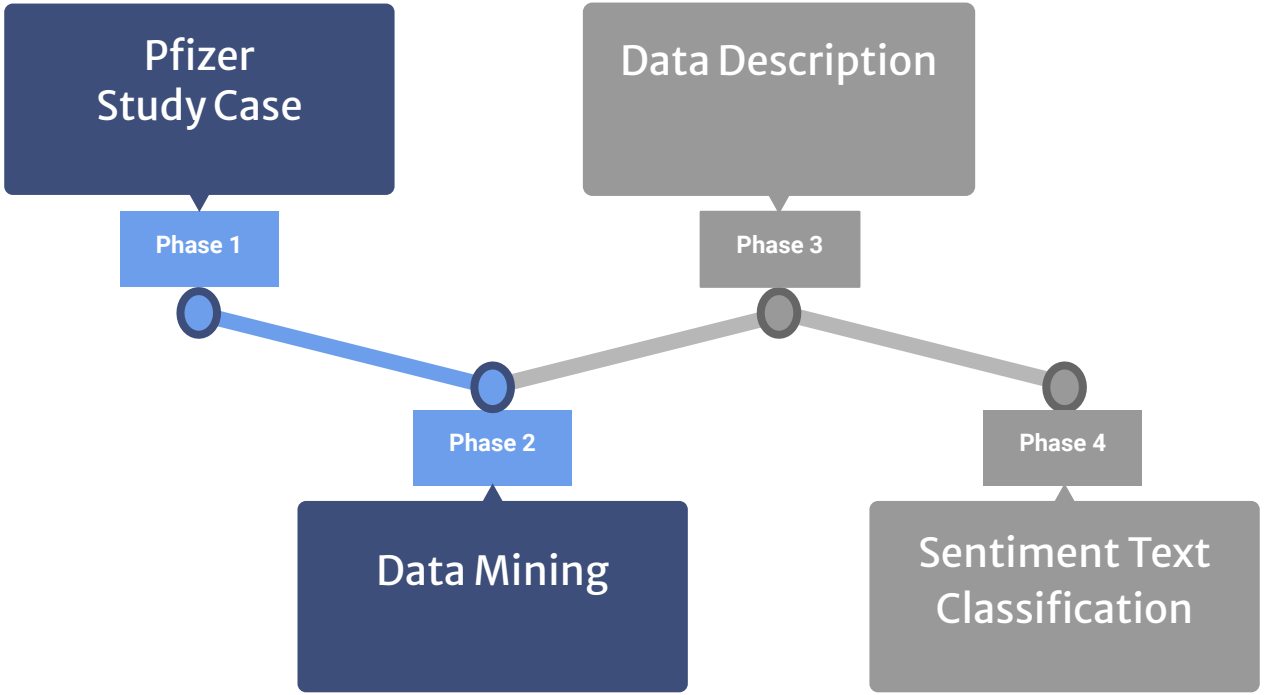
08/12/20

Vaccin Day

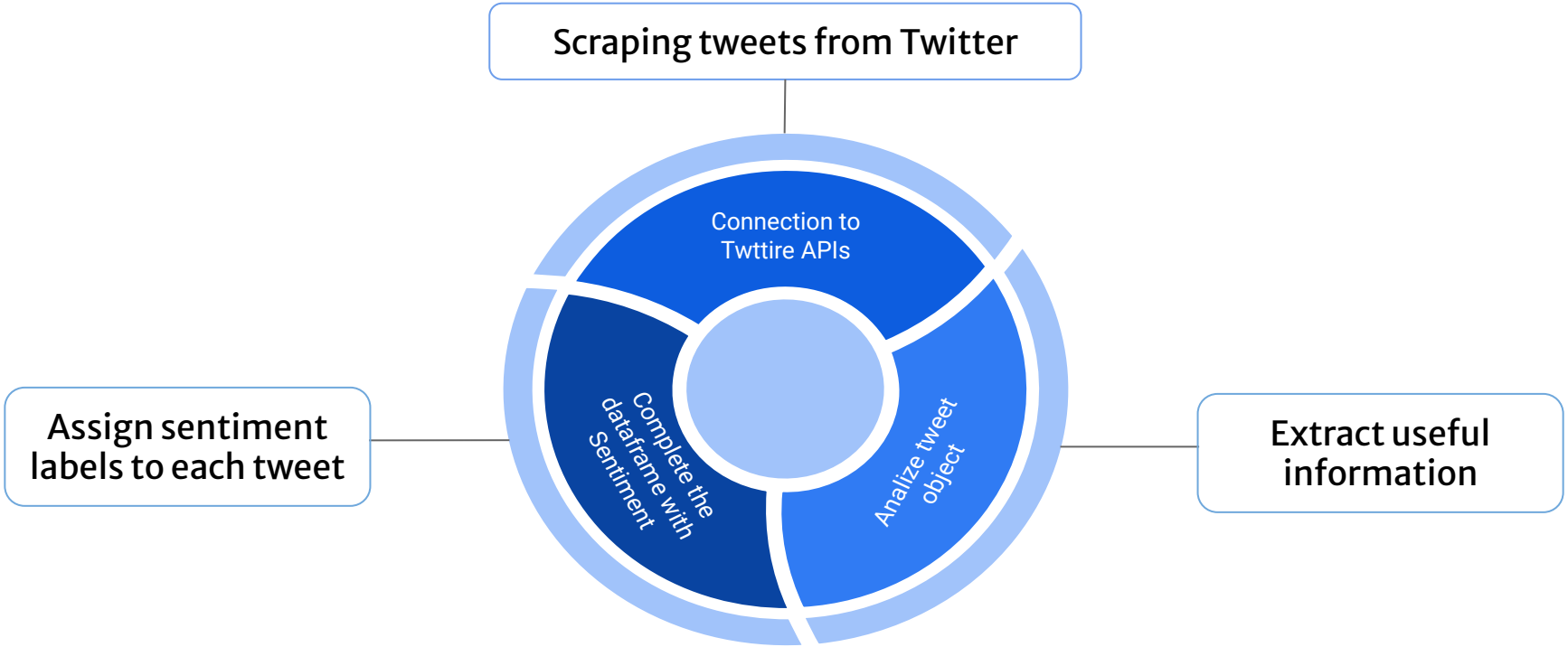
England the first country in the world to administer the first dose of Covid-19 vaccine: Pfizer BioNTEch vaccine



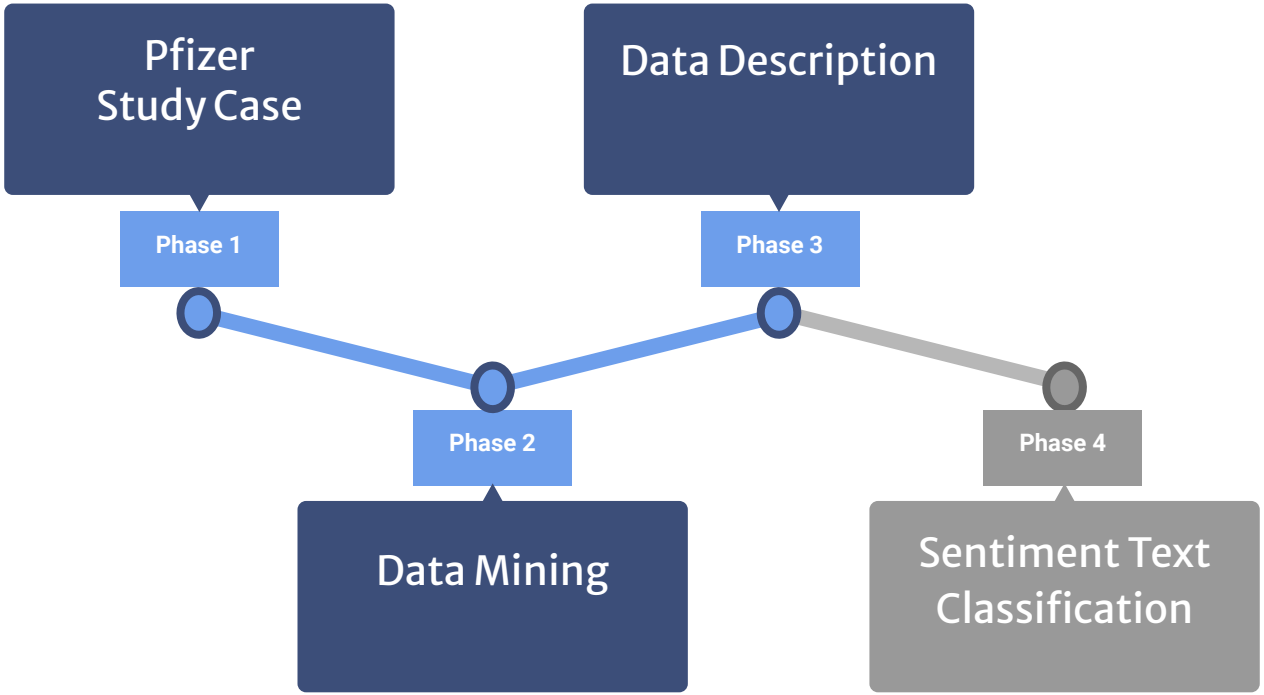
RoadMap


Pfizer
Study Case
Data Description
Phase 1
Phase 3
Phase 2
Phase 4
Data Mining
Sentiment Text
Classification

Data Mining: get access to Twitter Uninverse



RoadMap


Pfizer
Study Case
Data Description
Phase 1
Phase 3
Phase 2
Phase 4
Data Mining
Sentiment Text
Classification

Data Description: NLP Analysis

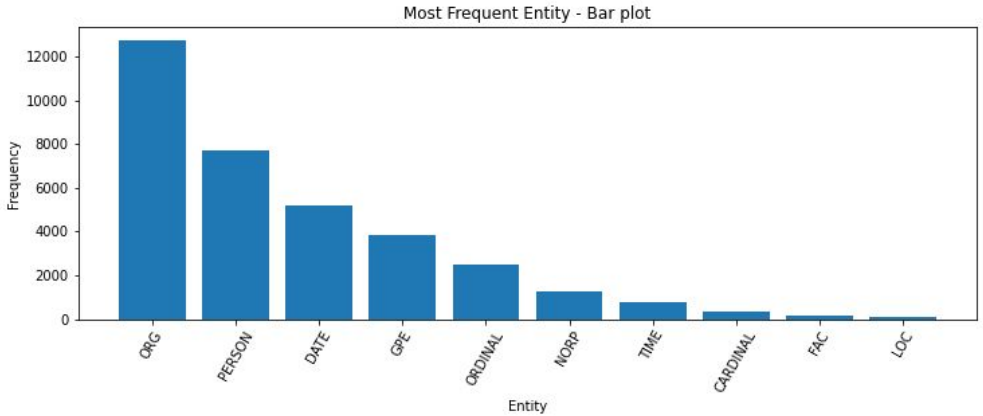
Topic Modelling

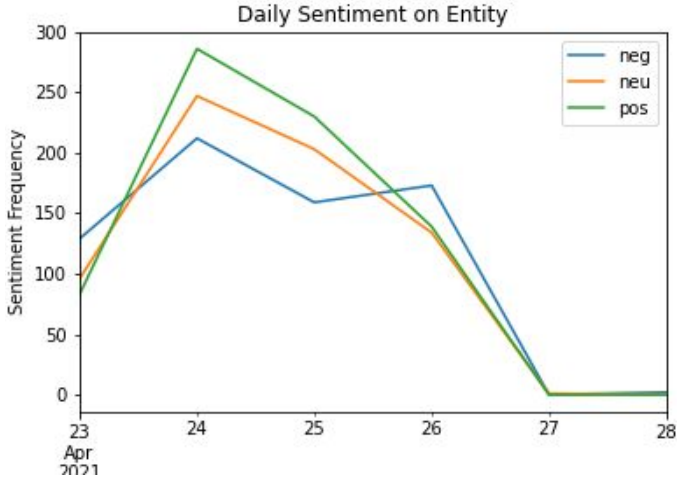
Topics	Related Words									
T1	jab	wait	fulli	think	go	work	look	better	tomorrow	week
T2	dose	got	shot	second	effect	today	get	feel	arm	day
T3	covid	peopl	receiv	israel	heart	case	india	inflamm	world	examin
T4	moderna	deal	time	know	said	booster	jampj	canada	come	problem
T5	avail	appoint	covid	new	need	moderna	pfizerbiontech	amp	astrazeneca	read

- Why Topic Modelling? Be sure most of the tweets are related to Pfizer (and Covid-19, vaccines, ...)
- A preliminary phase is prefered to reduce the variance on tweet text
- Topic Modelling is a stochastic procedure: new iteration... new terms!
- Improvement by tuning the hyper-parameters

Data Description: NLP Analysis

Entity Analysis





- Useful to better understand what people are more focused on
- Pfizer, Moderna, Vaccin, India, Washington, Jeff Bezos,...
- Trend Sentiments plot for **pfizer** entity: positive sentiment is mostly dominant!

Data Description: NLP Analysis

Table shows the collections set of the most likely terms to be seen together

In this PMI there is metrics on the score

Seems people are quite worried about the negative effects of the Covid-19 and vaccines.

PMI Analysis

Terms	Score
heart_inflamm	331.863
raw_materi	238.647
israel_examin	198.439
examin_heart	182.204
inflamm_case	165.408
pfizer_vaccin	142.228
neurolog_degen	137.768
covid_vaccin	116.6277
caus_neurodegen	112.576
al_neurolog	109.096

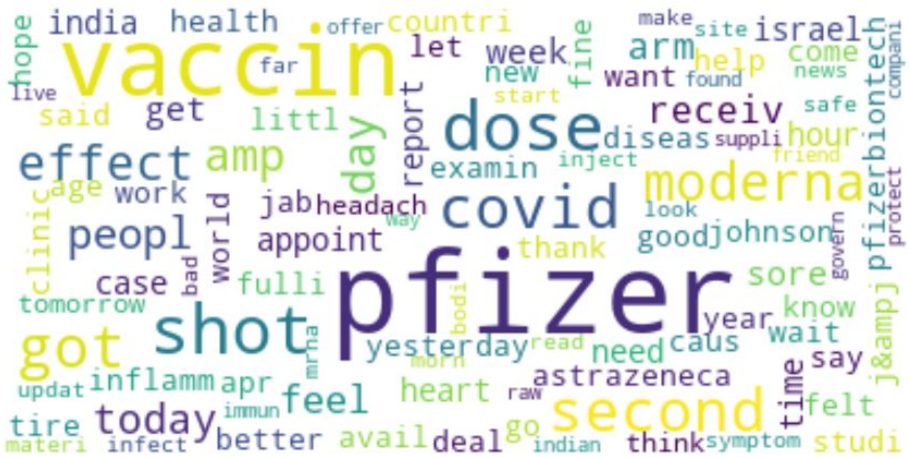
Terms	Score
alzheim_al	106.588
sore_arm	104.236
second_dose	100.465
trigger_alzheim	97.504
bee_gee	85.0
pfizer_shot	84.327
neurodegen_diseas	82.586
gee_singer	76.246
barri_gibb	73.751
singer_barri	71.579

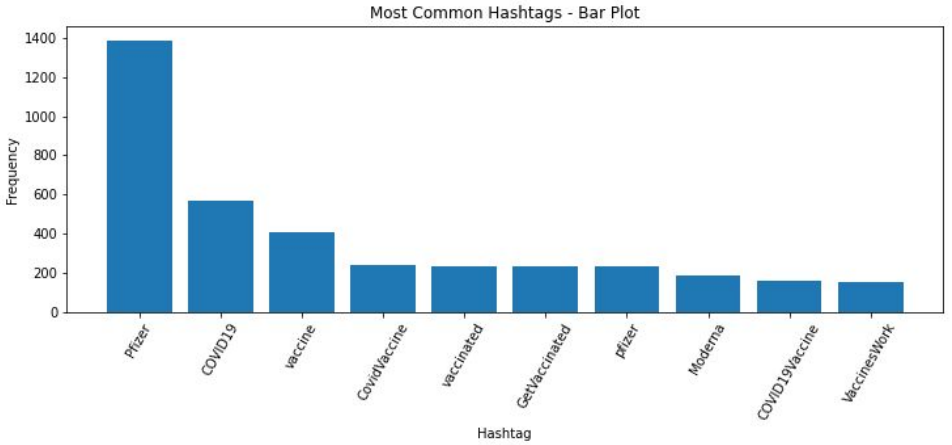
Data Description: Descriptive Analysis

Pre-processing applied, most frequent terms are related with the pandemic situation

Also from the point of view of the most used hashtags...

... having the *potential* to move the conscious of a lot of people



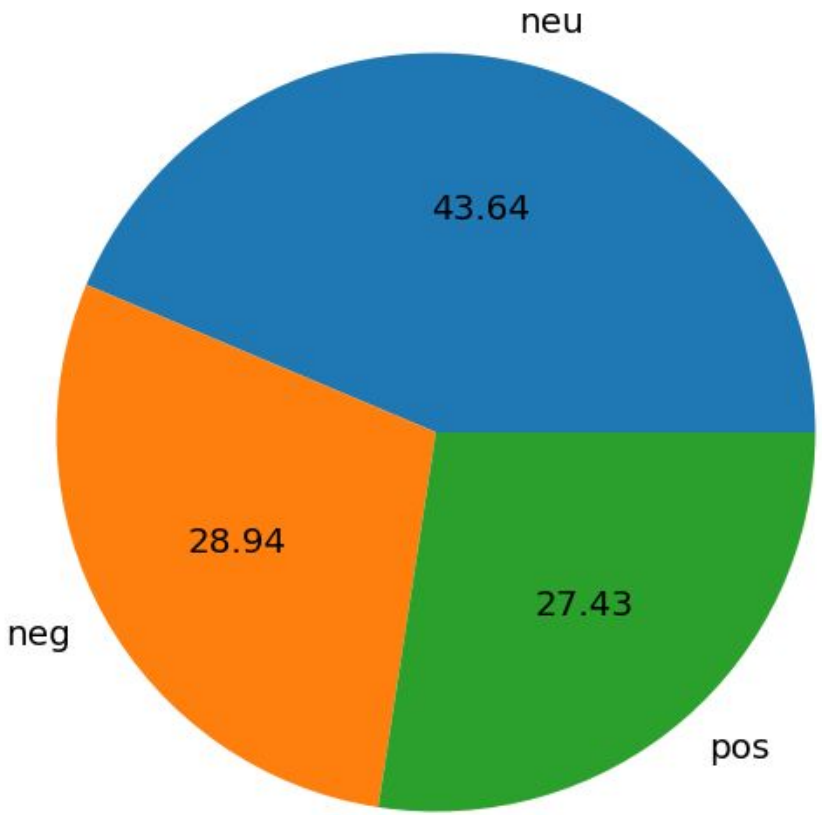


Data Description: Tweet Sentiment

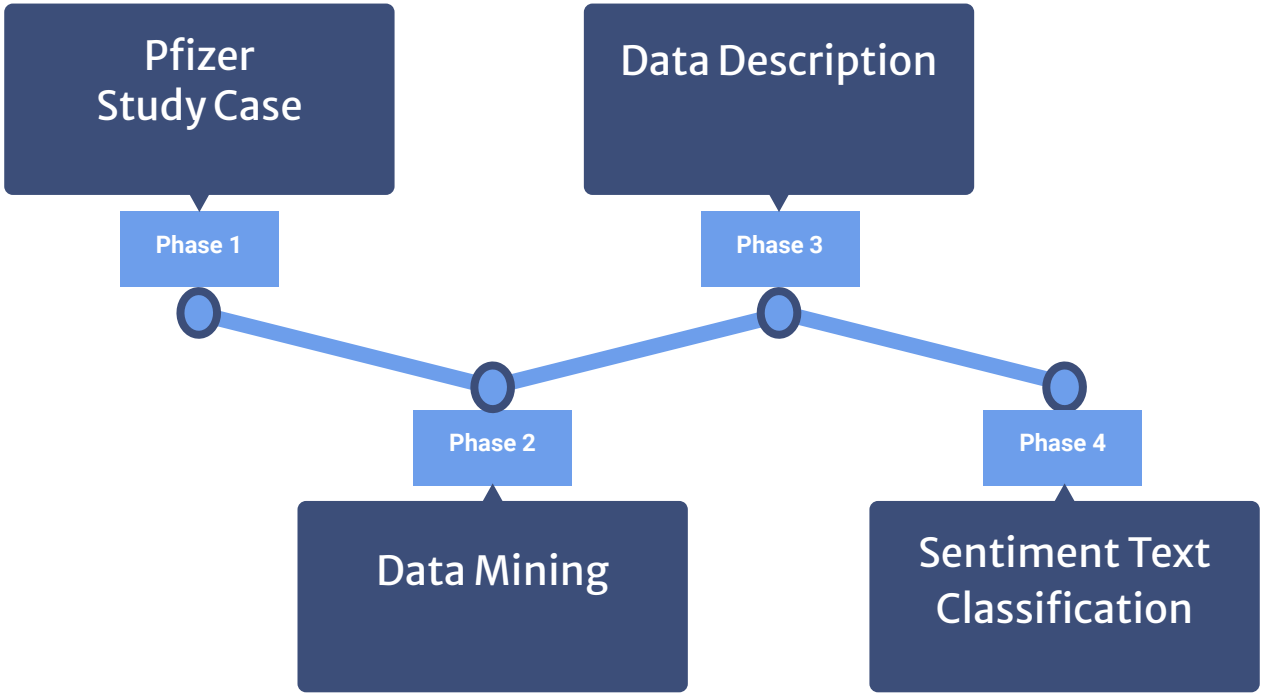
Pie chart shows the percentage of sentiment among: negative, neutral and positive

Most of the tweets have a neutral sentiment

The dataset is balanced enough to avoid oversampling or undersampling

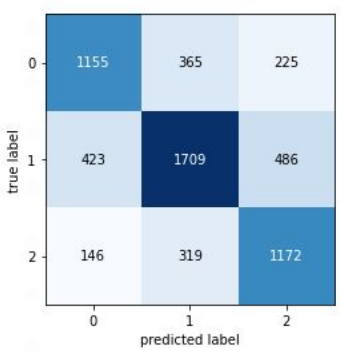


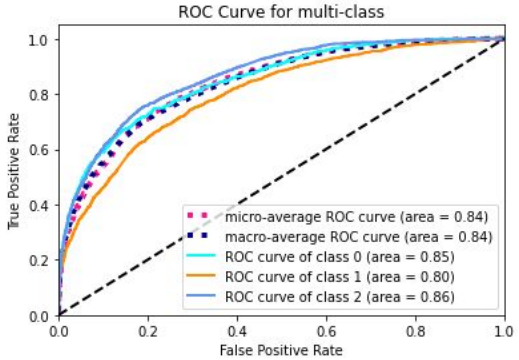
RoadMap

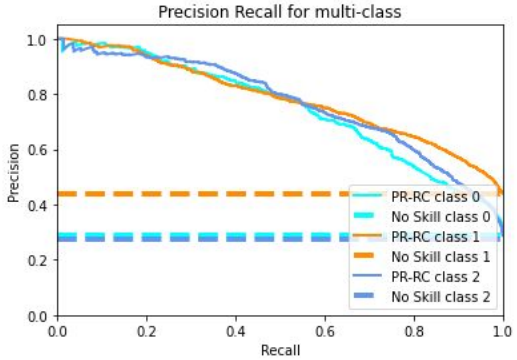

Pfizer
Study Case
Data Description
Phase 1
Phase 3
Phase 2
Phase 4
Data Mining
Sentiment Text
Classification

Sentiment Text Classification

Logistic Regression





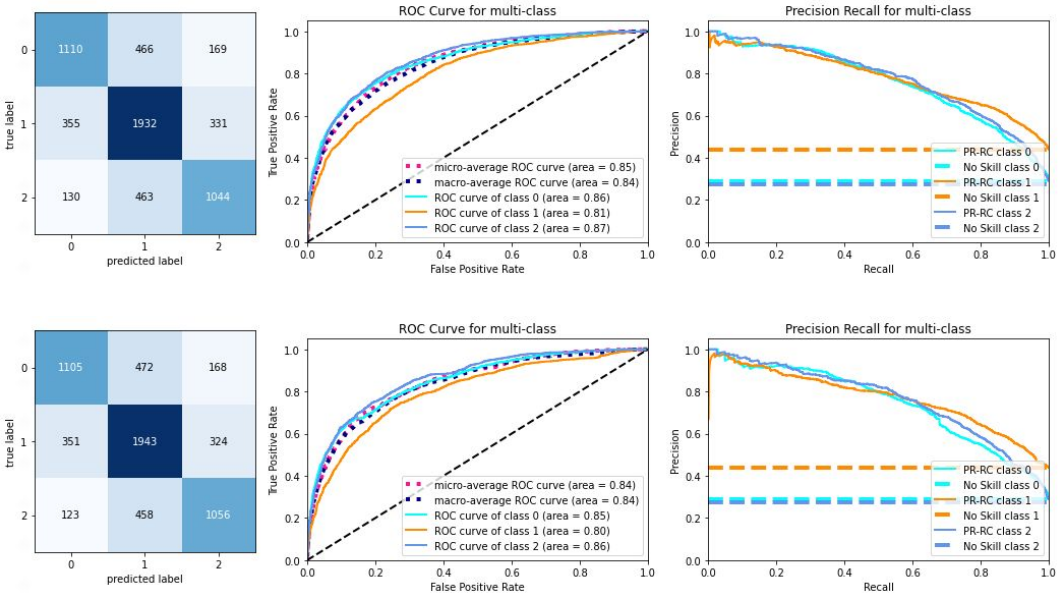


- All classes have mostly the same scores
- Baseline has score 0.43
- Class 2 is the class better predicted, with AUC of 0.86

Sentiment Text Classification

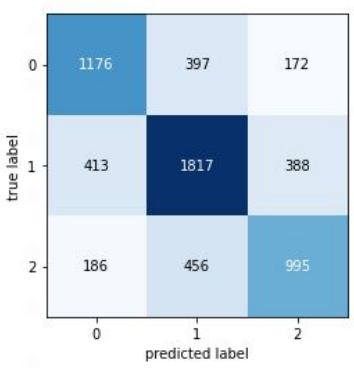
OnevsAll & OnevsOne

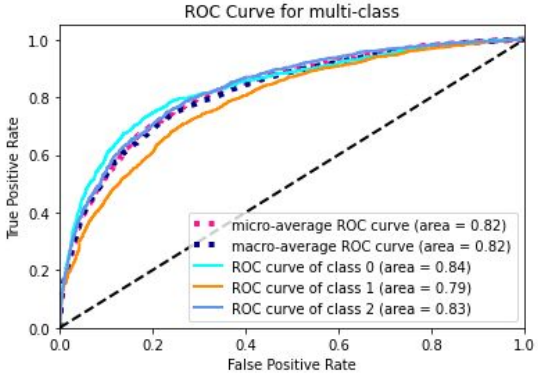
- All classes have mostly the same scores
- Baseline has score 0.43
- Class 2 is the class better predicted, with AUC of 0.87 and 0.88 for OnevsOne and OnevsAll

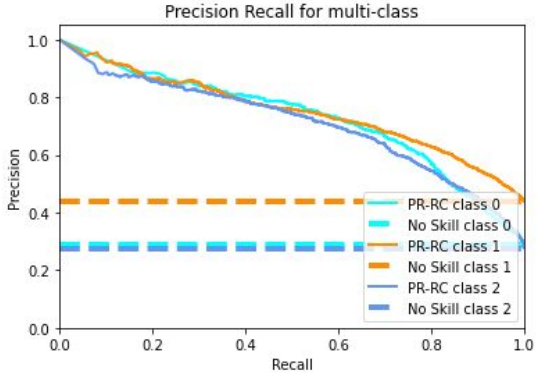


Sentiment Text Classification

Convolutional NN







- All classes have mostly the same scores
- Baseline has score 0.43
- Class 2 is the class better predicted, with AUC of 0.86

Sentiment Text Classification

	Regularized + Feature Selection			CNN		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Class 0	0.69	0.67	0.68	0.66	0.70	0.68
Class 1	0.72	0.68	0.70	0.71	0.65	0.68
Class 2	0.66	0.74	0.70	0.64	0.69	0.66
Accuracy	-	-	0.69	-	-	0.67
Macro Avg	0.69	0.70	0.69	0.67	0.68	0.67
Micro Avg	0.69	0.69	0.69	0.68	0.67	0.67

	One vs All			One vs One		
	Preicsion	Recall	F1 Score	Precision	Recall	F1 Score
Class 0	0.71	0.64	0.67	0.71	0.64	0.67
Class 1	0.69	0.75	0.72	0.69	0.75	0.72
Class 2	0.70	0.67	0.69	0.70	0.67	0.68
Accuracy	-	-	0.70	-	-	0.70
Macro Avg	0.70	0.69	0.69	0.70	0.69	0.69
Micro Avg	0.70	0.70	0.70	0.70	0.70	0.69

Natural Language Processing

Text Classification on Pfizer' tweets for Sentiment Analysis

Student:
Pietro Morichetti

Thanks for your Attention