



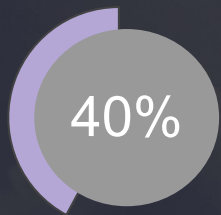
**Air Paradis**

# Projet 7

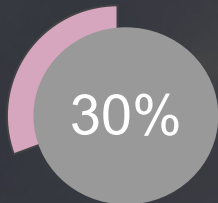
Détectez les Bad Buzz grâce au Deep Learning

**OPENCLASSROOMS**

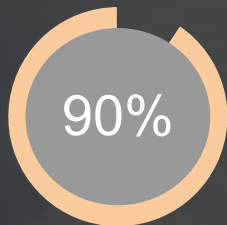
Yann Héreng



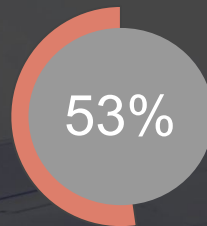
e-réputation = risque numéro 1



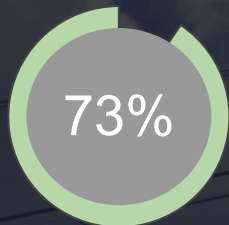
Bad buzz relayés mondialement dans l'heure qui suit la première annonce.



crises < 72 heures



cours de bourse inférieur 1 an plus tard



internauts ont une image plus positive de l'entreprise lorsque celle ci engage sur les réseaux sociaux

Sources :

(1) Selon l'éditeur Digimind ([www.digimind.fr](http://www.digimind.fr)) (2) [http://www.deloitte.com/assets/Dcom-France/Local%20Assets/Documents/publications/1311%20Exploring%20Strategic%20Risk/Exploring\\_strategic\\_risk\\_nov2013.pdf](http://www.deloitte.com/assets/Dcom-France/Local%20Assets/Documents/publications/1311%20Exploring%20Strategic%20Risk/Exploring_strategic_risk_nov2013.pdf) (3) <http://www.cadic-services.com/gerer-son-e-reputation/#.Uu9Gz3d5PxZ> (4) selon l'agence américaine de communication Digital Firefly.

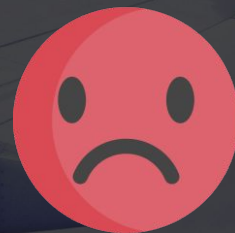
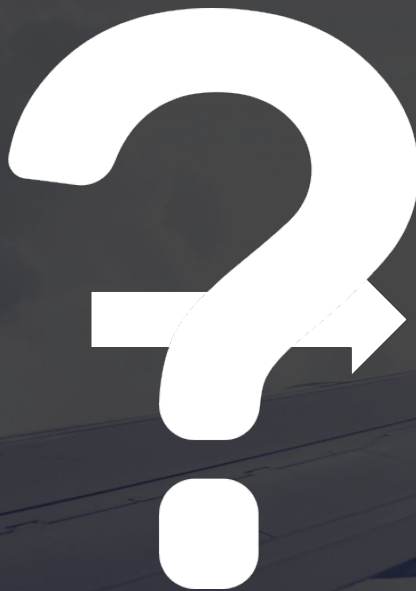
'Il faut 20 ans pour bâtir une réputation et  
cinq minutes pour l'anéantir'

- Warren Buffett -



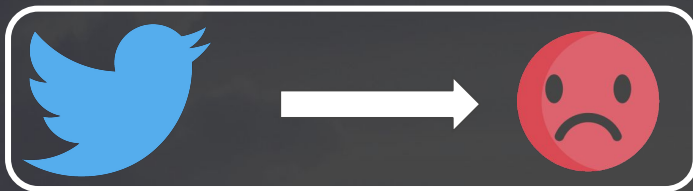


Air Paradis





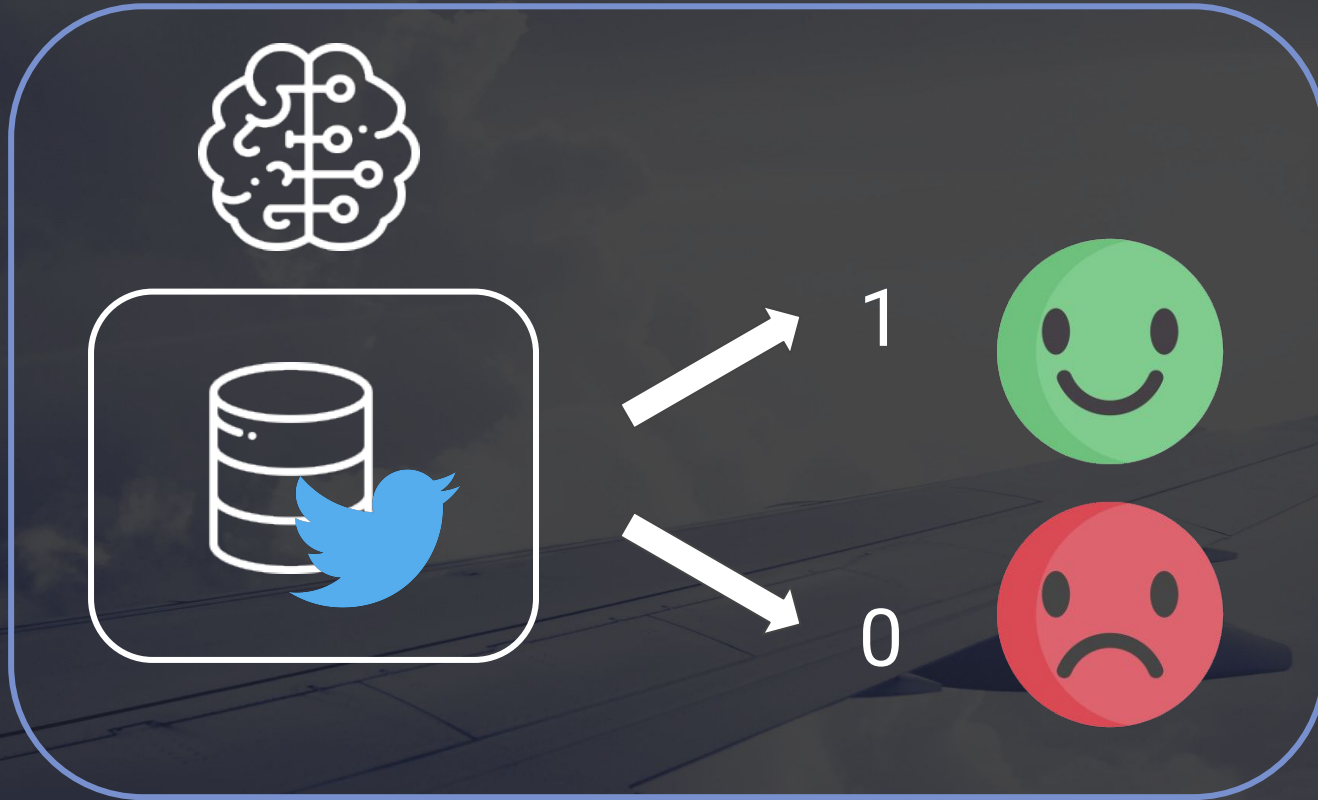
Notre solution



Valeur ajoutée



# Classification supervisée





# Méthodologie



Data



LAB



Cloud





**Data**



**LAB**



**Cloud**

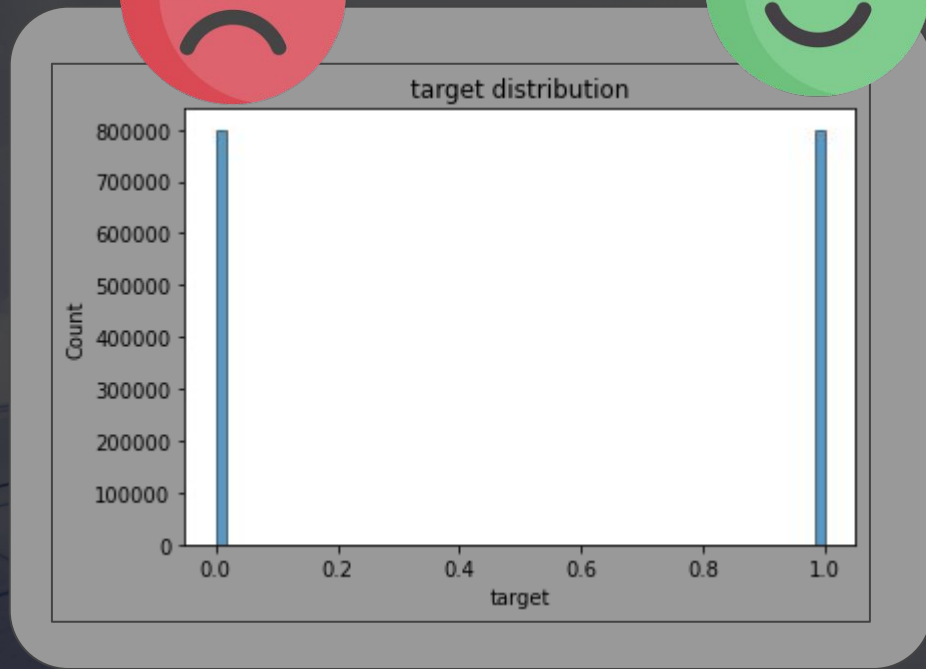




# Data



1'600'000  
tweets



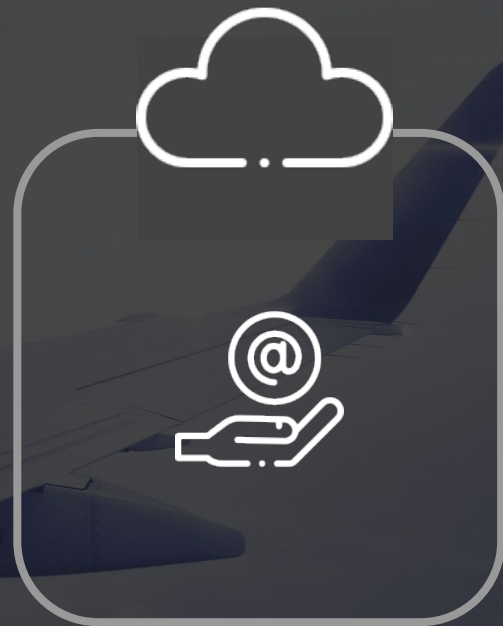
Data



LAB



Cloud





# LAB



Pre-processing



Modèles

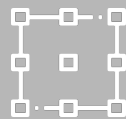


Optimisations





# LAB



## Pre-processing

### format



Majuscules  
Chiffres  
Symboles

### pré-traitement



Stemming  
Lemmatisation

### plongement



FastText  
GloVe  
Word2Vec





LAB



Modèle

RAPIDITE



EXACTITUDE

COUT

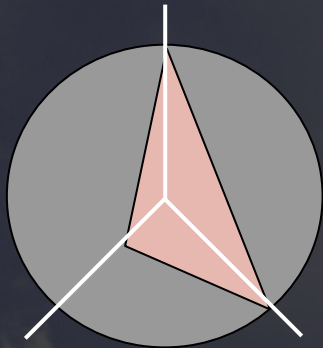


# LAB



## Modèle

RAPIDITE



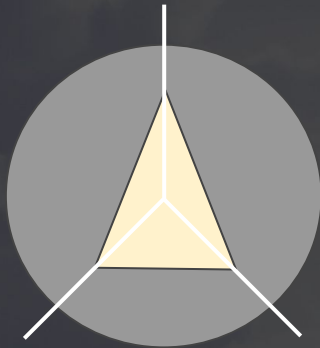
EXACTITUDE

COUT

Machine Learning  
Regression Logistic  
SVM

statistiques

RAPIDITE



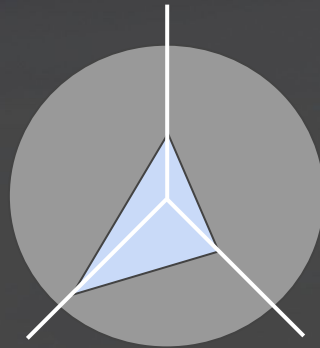
EXACTITUDE

COUT

Réseau de neurones  
Perceptrons  
Multicouches

Pattern  
statistique

RAPIDITE



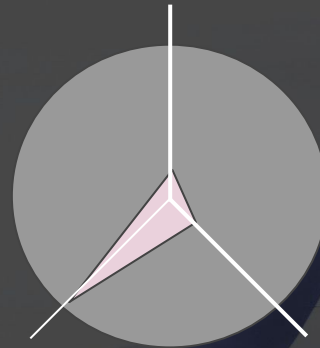
EXACTITUDE

COUT

Réseaux récurrents  
LSTM

Pattern  
temporel

RAPIDITE



EXACTITUDE

COUT

Réseaux récurrents  
bi-directionnel  
biLSTM

Pattern temporel  
bi-directionnel

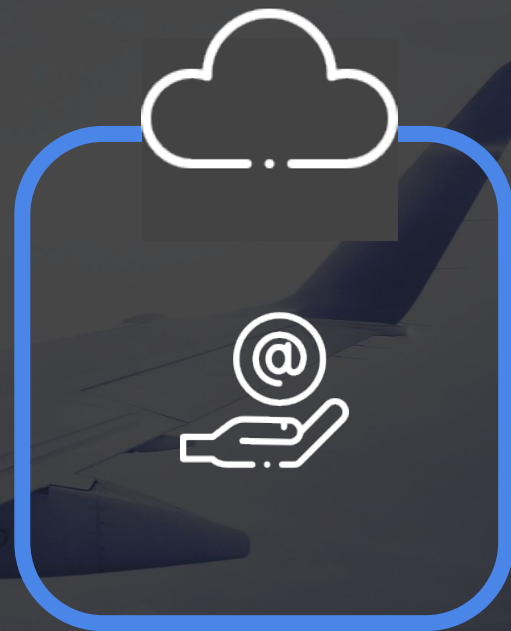
Data



LAB



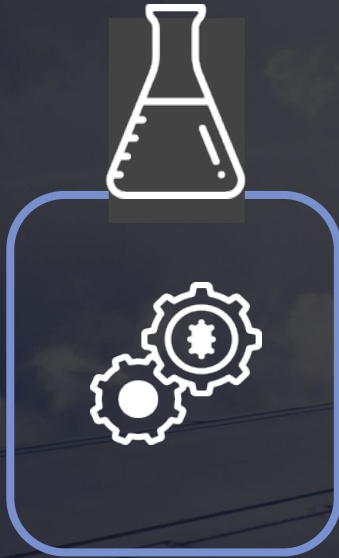
Cloud




 Cloud

 Compute

LAB



 Azure



GPU

Cloud

Deploy





MAINTENANCE



SCALABILITE



DISPONIBILITE



TCO



# Résultats



DEMO



Résultats

Séparation des jeux de données



1'600'000  
tweets



160'000



1'440'000

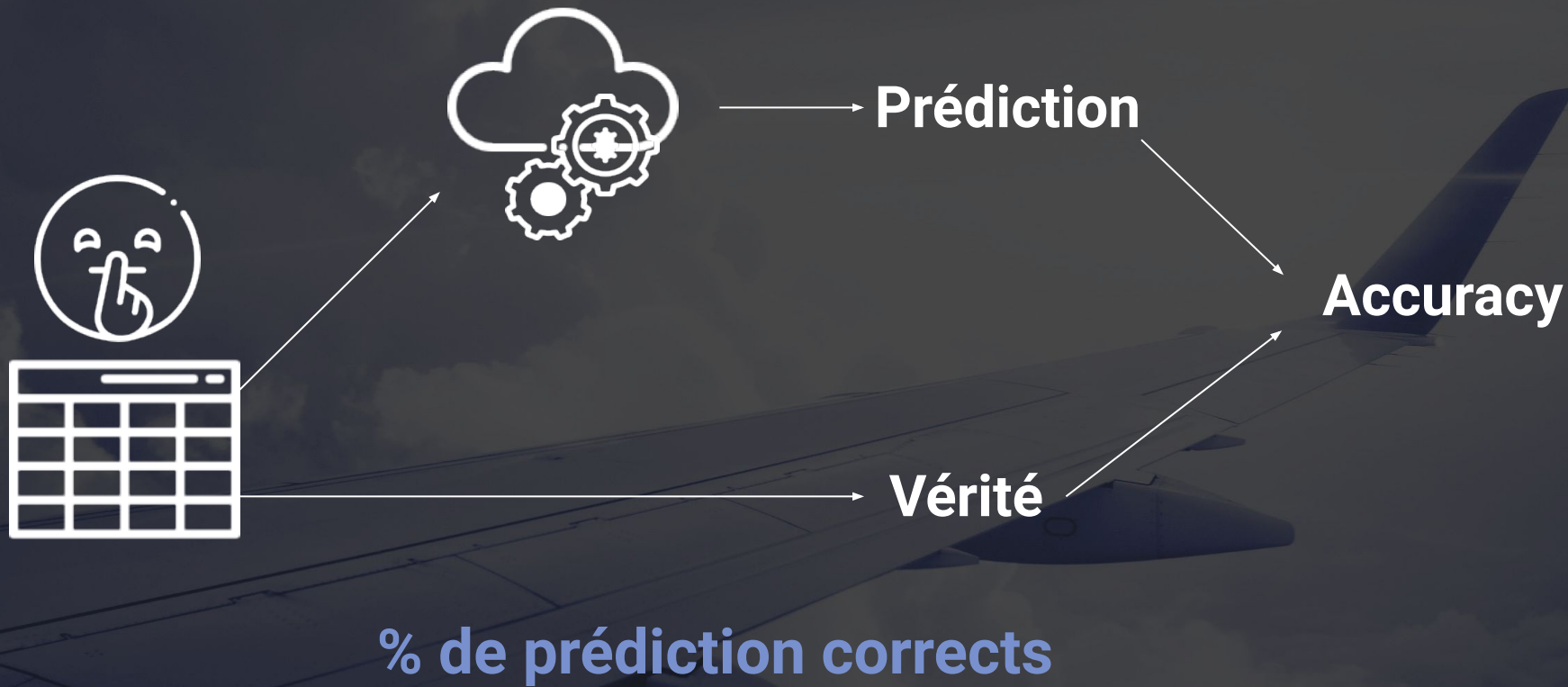


LAB



Résultats

Métrique





Test = 1600 tweets

Accuracy = 80%

|               |   | Classe Prédite |           |
|---------------|---|----------------|-----------|
|               |   | 0              | 1         |
| Classe Réelle | 0 | 637<br>TN      | 143<br>FP |
|               | 1 | 172<br>FN      | 577<br>TP |



@bubbleMAMI LOL peedi crack is cute I just saw his video from 2yrs ago..don't know where he is now  
SKO Almost the end. Nicola smothers us with affection. &lt

@rxtheride thank you gelli. i was just really frustrated with things last night. thanks you tho! &lt

@bendaubney As soon as I find a small component of my mp3 which makes it possible...  
Spent the day designing a logfiles online course and the first 2 learning objects...only 5 more to go

@TheRawBee yeh its fucked yeh going again for my bro's confo. waat u doin?  
is doing a oh so interesting scale drawing at college

Woo hoo .. juz found the photo that was taken in 1999 - together with my mom !! i look so young thinking if i  
should upload here or fb

@dheaasa @feliciasfelicia yes,she has a javelin! ohh,im so jealous of her.. i mean its expensive! she  
bought it ysterday..

A photograph taken from an airplane window, showing the wing and tail of the aircraft against a sunset sky. The sky is a mix of orange, yellow, and blue, with a layer of white clouds visible below the plane. The word "Conclusion" is written in white text across the center of the image.

Conclusion

# Autres options



Accuracy = 71%



Azure ML Studio

Accuracy = 75%

# Suggestions



Benchmark architectures / coût / performances



Développer les use cases avec des métriques



Maintenance du modèle, ré-entraînement

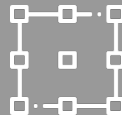


A photograph taken from an airplane window, showing the wing and tail of the aircraft against a sunset sky. The sky is a mix of orange, yellow, and blue, with a layer of white clouds visible below the plane. The word "Merci" is written in white text in the center of the image.

Merci



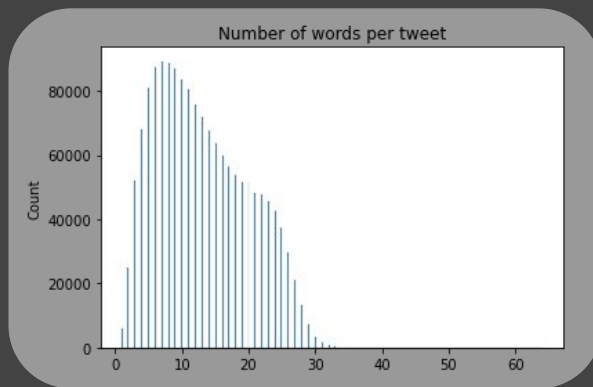
LAB



Pre-processing

Vocabulaire 60'000 mots → EDA montre que nb\_occure>10 concerne 65637 mots

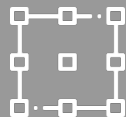
Tweet limité à 30 mots



Tokenization: minuscules, nombre, ponctuation, caractères spéciaux



# LAB



Pre-processing

|                          | training_time (s) | Accuracy |
|--------------------------|-------------------|----------|
| Fasttext - No preprocess | 2.917658          | 0.510    |
| Fasttext - Lemmatization | 8.671020          | 0.540    |
| Fasttext - stemming      | 7.859769          | 0.515    |
| Glove - No preprocess    | 2.781776          | 0.740    |
| Glove - lemmatization    | 6.510423          | 0.715    |
| Glove - stemming         | 1.543808          | 0.695    |
| Doc2vec - No preprocess  | 3.841180          | 0.595    |
| Doc2vec - lemmatization  | 11.377107         | 0.610    |
| Doc2vec - stemming       | 4.476753          | 0.595    |

Training 50k tweets



LAB



Modèle

|                                 | AUC      | Accuracy | Epoch max reach | total training time | training time to opt |
|---------------------------------|----------|----------|-----------------|---------------------|----------------------|
| simple NN - raw text            | 0.835903 | 0.7559   | 2.0             | 68.423138           | 9.123085             |
| simple NN - Basic preprocessing | 0.834373 | 0.7517   | 2.0             | 59.753760           | 7.967168             |
| simple NN - Stemming            | 0.836105 | 0.7571   | 3.0             | 53.780293           | 10.756059            |
| simple NN - GLoVe embedding     | 0.809682 | 0.7375   | 7.0             | 17.404001           | 4.060933             |
| LSTM - Own embedding            | 0.844849 | 0.7638   | 1.0             | 399.488908          | 13.316297            |
| LSTM - GLoVe embedding          | 0.871638 | 0.7877   | 28.0            | 288.322934          | 269.101405           |
| biLSTM - GLoVe embedding        | 0.873431 | 0.7884   | 27.0            | 566.732794          | 510.059515           |

Training 50k tweets



# LAB



## Optimisation

|                        | AUC  | Accuracy_val | Accuracy_train | Epoch max reach | dropout | learning_rate | training time |
|------------------------|------|--------------|----------------|-----------------|---------|---------------|---------------|
| LSTM - GLoVe embedding | 0,88 | 0,79         | 0,81           | 26              | 0,2     | 0,001         | 610,22        |
| LSTM - GLoVe embedding | 0,88 | 0,80         | 0,86           | 8               | 0,2     | 0,01          | 179,96        |
| LSTM - GLoVe embedding | 0,85 | 0,77         | 0,77           | 9               | 0,2     | 0,1           | 236,49        |
| LSTM - GLoVe embedding | 0,88 | 0,80         | 0,98           | 6               | 0       | 0,01          | 131,28        |
| LSTM - GLoVe embedding | 0,88 | 0,80         | 0,90           | 9               | 0,1     | 0,01          | 207,75        |
| LSTM - GLoVe embedding | 0,88 | 0,81         | 0,81           | 28              | 0,4     | 0,01          | 630,54        |
| LSTM - GLoVe embedding | 0,85 | 0,77         | 0,71           | 28              | 0,8     | 0,01          | 739,56        |

Training 50k tweets