Analyser les retours clients sur Twitter

Air Paradis

Contexte & objectif du projet

Air Paradis

Air Paradis est une compagnie aérienne qui souhaite surveiller sa réputation sur les réseaux sociaux.

Pour ce faire, il nous a été demandé de créer un prototype de produit IA permettant de prédire le sentiment associé à un tweet.

Objectif du projet

Air Paradis

Pour mettre en place ce prototype permettant de prédire le sentiment associé à un tweet, nous avons essayé plusieurs types de modèles:

- des **modèles simples** pour établir une référence
- des **modèles RNN** pour faire notre prototype
- des **modèles BERT** pour ouvrir de nouveaux horizons

Analyse de sentiments

Air Paradis

	Modèles classiques	Modèles réseaux neuronaux
COLLECTE		
PRE-TRAITEMENT	Tokenization, Stemming, Part Of Speech, Lemmatization, Stop-words, Named Entity Extraction, ...	
EXTRACTION DE FEATURES	Count, TF-IDF, One Hot Encoding, ...	
EMBEDDING		Word2Vec, FastText, GloVe, ...
MODELISATION	Naive bayes, Decision Tree, Logistic Regression, ...	
MODELISATION		Various configurations with: SimpleRNN layers, LSTM layers, GRU layers, ...
INFERENCE		

Analyse exploratoire des données

Air Paradis

Présentation du jeu de données

Pour ce problème, nous avons travaillé avec un jeu de données open-source contenant des tweets rédigées en 2009 accompagnés d'un label de polarité (positif / négatif).

- **1.600.000 tweets décrits par 6 variables (target, ids, date, flag, user, text)**

Analyse exploratoire des données

Air Paradis

Vérifications de base

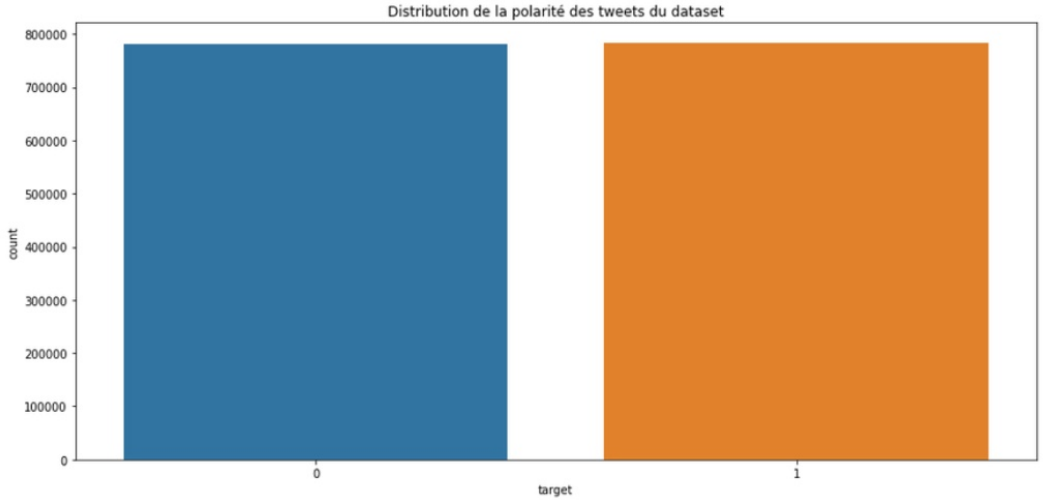
Les vérifications des erreurs les plus fréquentes
(valeurs manquantes, doublons, outliers, erreurs de format, erreurs lexicales, contenus multiples)
puis les analyses univariées et multivariées
ont permis de déceler quelques problèmes qui
ont été corrigés avec les actions suivantes:

- Suppression des colonnes inutiles au projet
- Remplacement des balises HTML
- Suppression des encodage exotiques
- Traitement des outliers (plus de 140 caractères)

Analyse exploratoire des données

Air Paradis

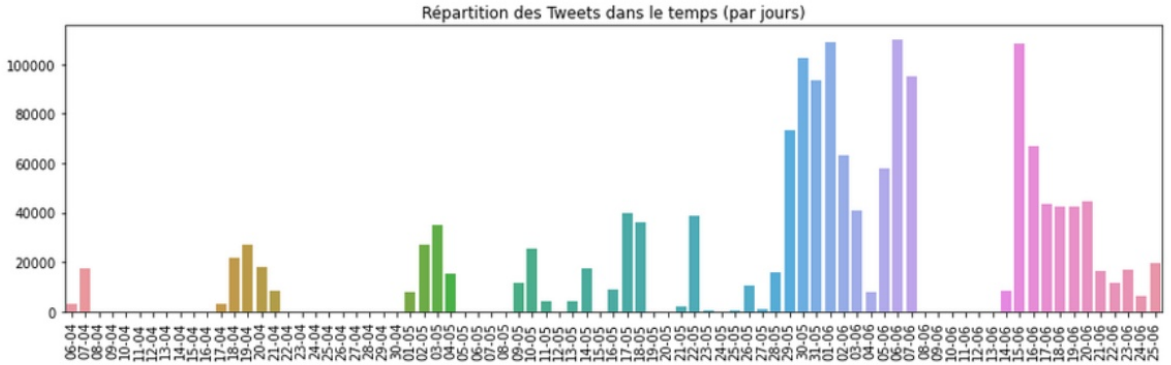
Répartition des polarités > binarisation



Analyse exploratoire des données

Air Paradis

Temporalité des tweets



Tous les tweets ont été postés entre Mai et Juin 2009

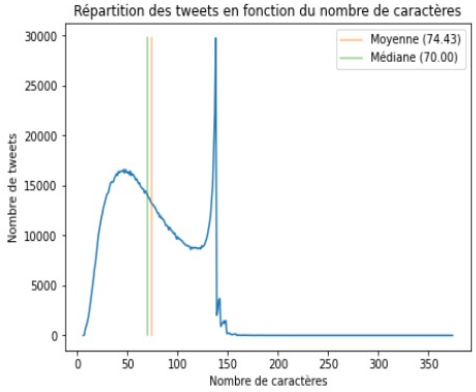
- En 2009 les tweets étaient limités à 140 caractères

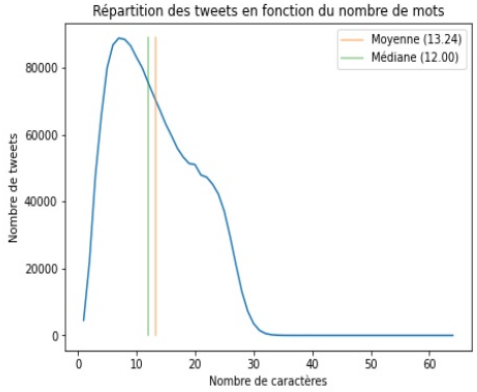
Analyse exploratoire des données

Air Paradis

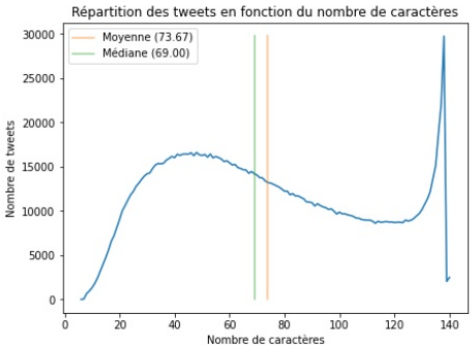
Nombre de caractères & mots / tweet

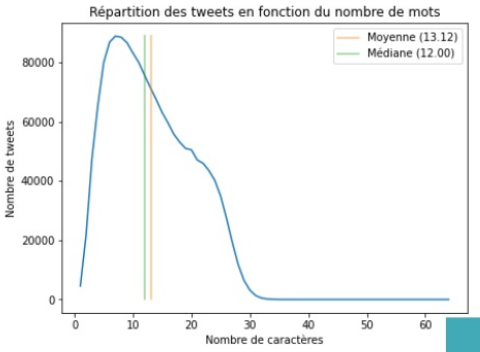
AVANT





APRÈS





Pré-traitement des textes

Air Paradis

ORIGINAL

- Aucun pré-traitement

Pré-traitement des textes

Air Paradis

RAW

- Suppression des balises HTML

Prezi

Pré-traitement des textes

Air Paradis

PREPROCESS01

- Nettoyage avec Twitter-preprocessor

Prezi

Pré-traitement des textes

Air Paradis

PREPROCESS02

- Nettoyage avec Twitter-preprocessor
- Tokenization

Prezi

Pré-traitement des textes

Air Paradis

PREPROCESS03

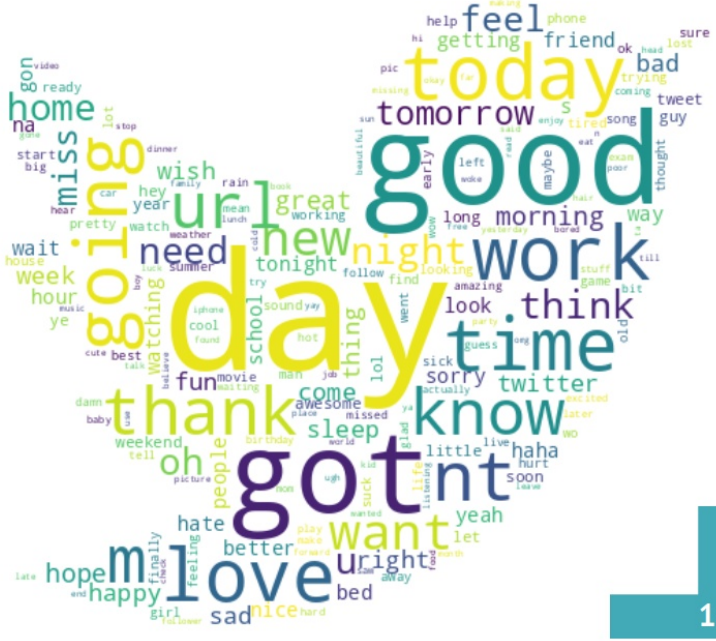
- Nettoyage avec Twitter-preprocessor
- Tokenization
- Filtrage avancé

Pré-traitement des textes

Air Paradis

PREPROCESS03_simple

- Nettoyage avec Twitter-preprocessor
- Tokenization
- Filtrage simple



Prezi

Pré-traitement des textes

Air Paradis

PREPROCESS04

- Nettoyage avec Twitter-preprocessor
- Tokenization
- Filtrage avancé
- Lemmatization

Prezi

Pré-traitement des textes

Air Paradis

PREPROCESS04_simple

- Nettoyage avec Twitter-preprocessor
- Tokenization
- Filtrage simple
- Lemmatization

Pré-traitement des textes

Air Paradis

PREPROCESS04_nofilter

- Nettoyage avec Twitter-preprocessor
- Tokenization
- AUCUN Filtrage
- Lemmatization

Recherche du modèle le plus adapté

Air Paradis

Les métriques

Pour un problème de classification,
les mesures de référence sont:

- **Accuracy** : taux de bonnes prédictions au total
- **Precision** : taux de vrai-positifs parmi les positifs
- **Recall** : capacité à détecter les vrai-positifs
- **Specificity** : capacité à éviter les faux-positifs

- **F1-Measure** : moyenne harmonique
  entre Precision et Recall -> 2*(P*R)/(P+R)

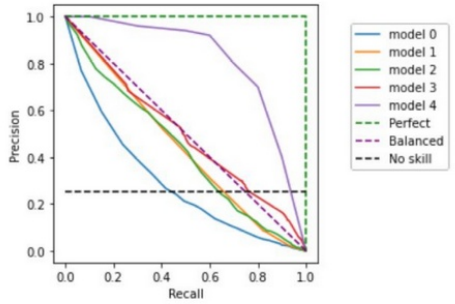
Recherche du modèle le plus adapté

Air Paradis

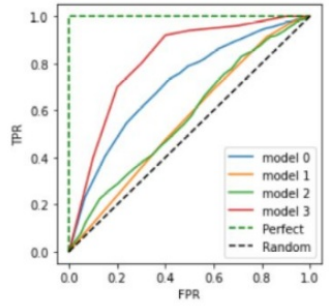
Les métriques

Mais les métriques que nous venons de voir sont calculées sur les labels prédits et ne prennent donc pas en compte un éventuel seuillage.

- **PR AUC : qui supporte bien le déséquilibrage, mais va favoriser la classe positive**



- **ROC AUC : qui supporte les déséquilibrage jusqu'à un certain point, mais va donner autant de poids aux deux classes**



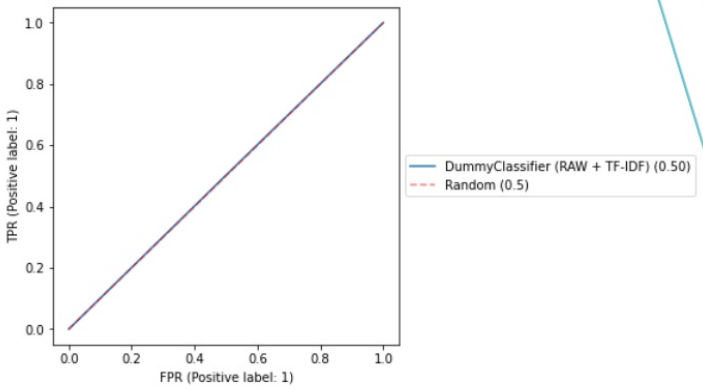
Recherche du modèle le plus adapté

Air Paradis

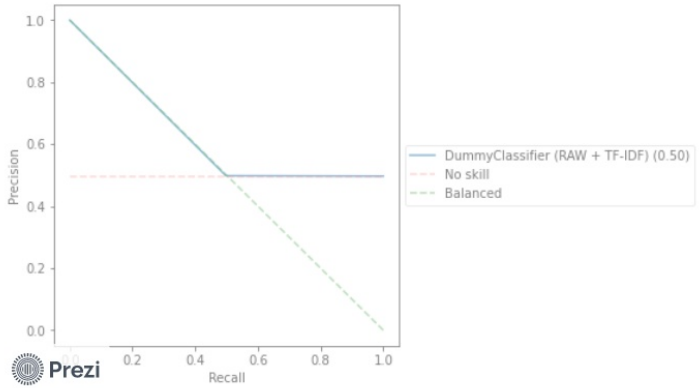
Modèle naïf

Nous avons utilisé un DummClassifier avec tous les samples du dataset pré-processé RAW sur lesquels nous avons appliqué un TF-IDF.

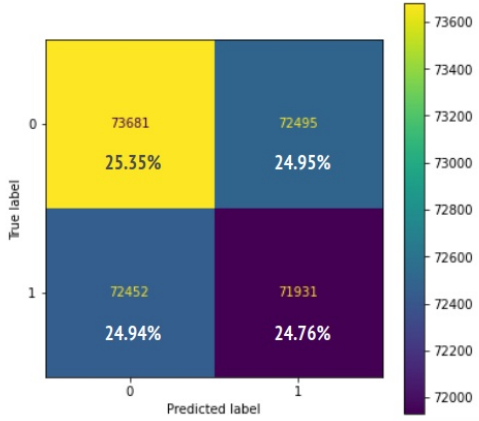
--- ROC AUC ----------



--- PRECISION RECALL AUC ----------

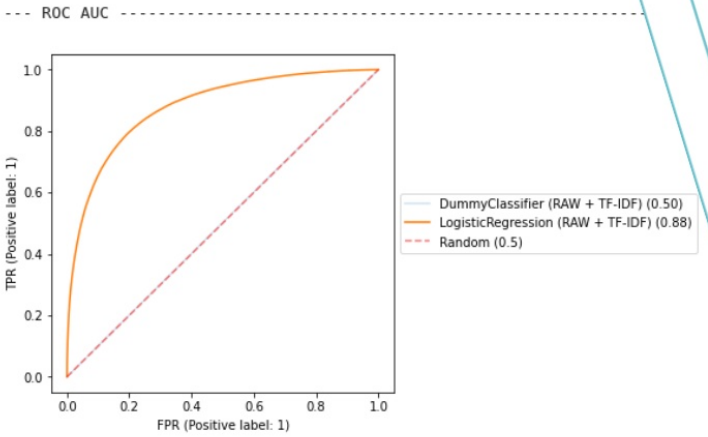


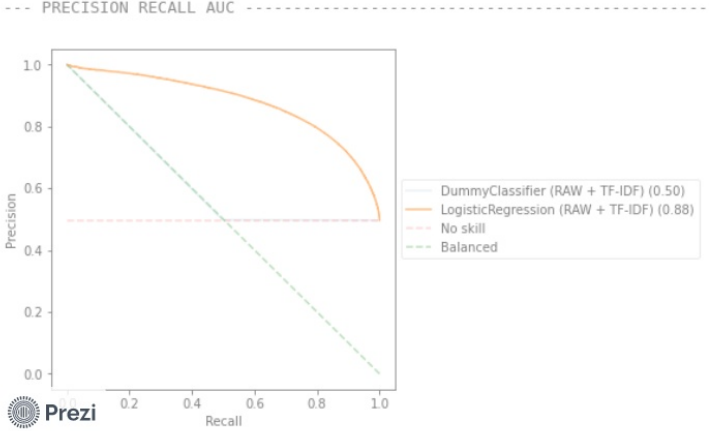

```
--- DummyClassifier (RAW + TF-IDF) ---

              ROC_AUC : 0.5011
                   F1 : 0.4981
             ACCURACY : 0.5011
            PRECISION : 0.4980
               RECALL : 0.4982
    AVERAGE_PRECISION : 0.4975

       TRAINING-TIME : 2.1800
       INFERENCE-TIME : 0.0269
```



Recherche du modèle le plus adapté

Air Paradis

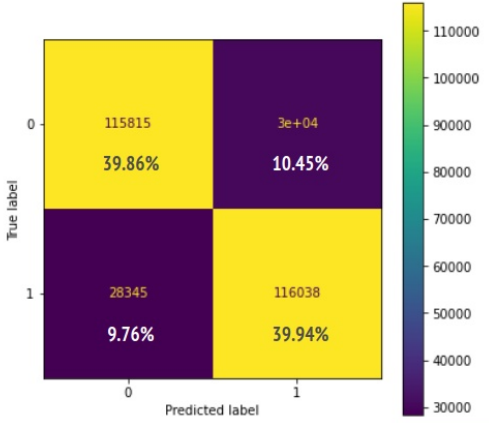




Modèle sur mesure simple

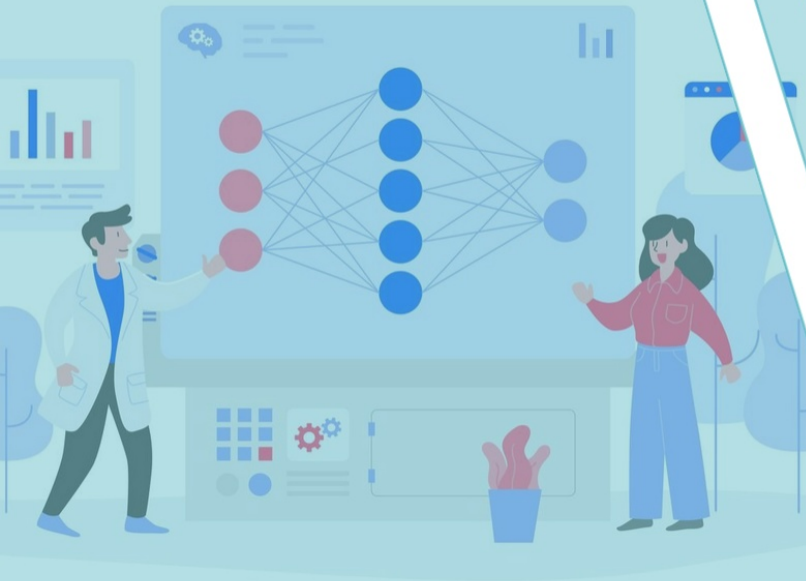
Nous avons utilisé une LogisticRegression avec GridSearch sur tous les samples du dataset pré-processé RAW préparés avec un TF-IDF.

```
--- LogisticRegression (RAW + TF-IDF) ---

          ROC_AUC : 0.8783
               F1 : 0.7981
         ACCURACY : 0.7980
        PRECISION : 0.7926
           RECALL : 0.8037
AVERAGE_PRECISION : 0.8786

    TRAINING-TIME : 1104.5527
    INFERENCE-TIME : 0.0711
```



Recherche du modèle le plus adapté

Air Paradis



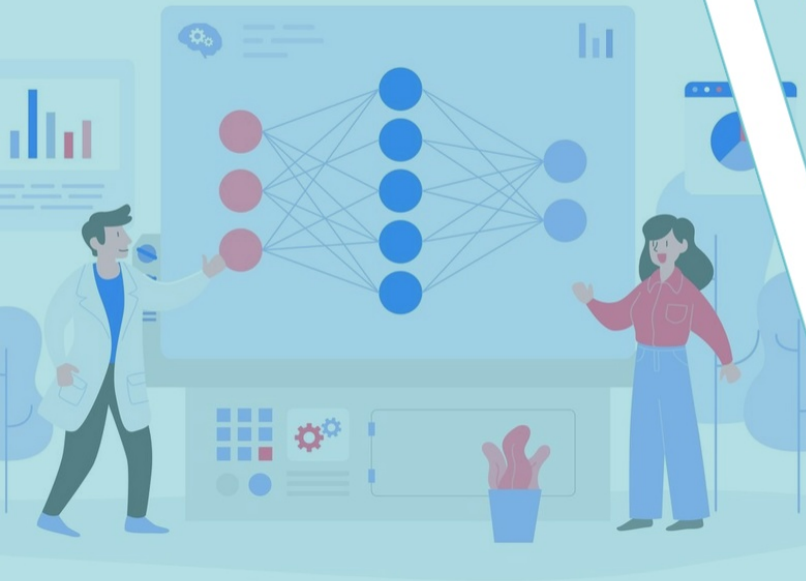
Modèle sur mesure avancé
Recherche du pré-processing

Chacun des 8 datasets pré-processés a été préparé avec un Tokenizer keras + padding, puis ils ont été évalués avec un modèle RNN commun.

Method	ROC AUC	F1 score	Accuracy	Precision	Recall	Training time	Inference time
PREPROCESS04_nofilter + Tokenizer4500	0.896341	0.814513	0.813993	0.807288	0.821870	1400.868720	24.291804
RAW + Tokenizer4500	0.895449	0.810211	0.812912	0.816891	0.803640	2231.321571	23.848580
PREPROCESS01 + Tokenizer4500	0.894032	0.808555	0.811642	0.816821	0.800454	1195.012966	23.674459
PREPROCESS02 + Tokenizer4500	0.893776	0.808913	0.811426	0.814675	0.803232	1195.012966	24.195830
PREPROCESS04_simple + Tokenizer4500	0.846138	0.763402	0.764069	0.760844	0.765977	960.968502	24.119112
PREPROCESS03 + Tokenizer4500	0.845024	0.764767	0.764568	0.759442	0.770167	745.887899	23.932140
PREPROCESS03_simple + Tokenizer4500	0.844488	0.763473	0.764685	0.762669	0.764280	702.937229	23.920813
PREPROCESS04 + Tokenizer4500	0.840828	0.753618	0.759687	0.768160	0.739616	675.146728	23.966787

Recherche du modèle le plus adapté

Air Paradis



Modèle sur mesure avancé
Recherche de l'embedding

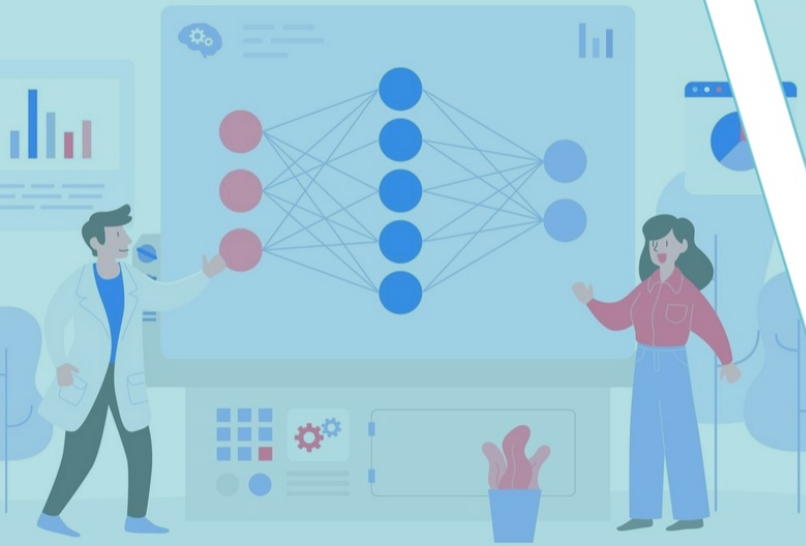
C'est une représentation vectorielle qui encode la signification contextuelle des mots, de sorte que les mots qui sont proches dans l'espace vectoriel doivent avoir une signification similaire.

Roi - Homme + Femme = Reine

J'ai le droit != Tiens-toi droit

Recherche du modèle le plus adapté

Air Paradis



Modèle sur mesure avancé
Recherche de l'embedding

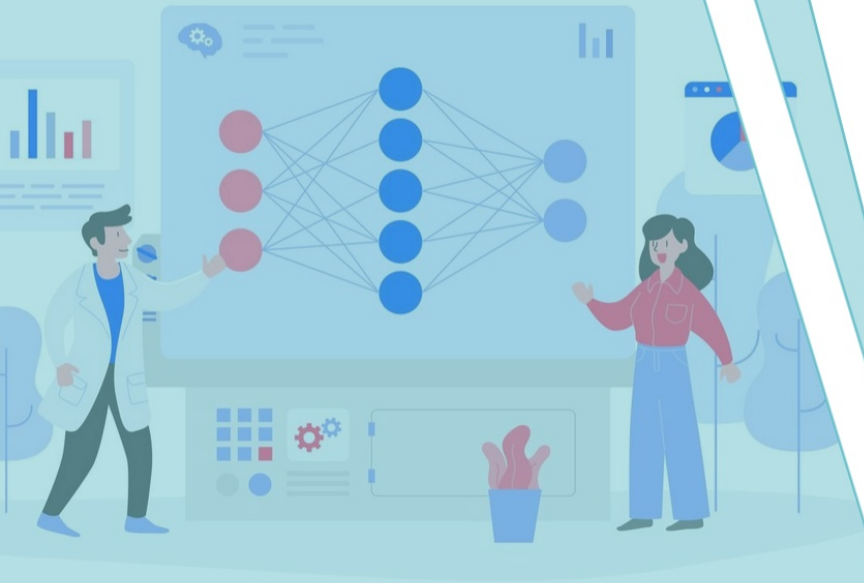
Nous avons donc essayé des modèles séquentiels avec des embeddings (pré-entrainés ou non) :

- Word2Vec
- FastText
- GloVe
- GloVe-Twitter
- Keras

Que nous avons comparé avec des modèles bag-of-words / bag-of-ngrams
(donc sans embedding mais avec feature extraction)

Recherche du modèle le plus adapté

Air Paradis

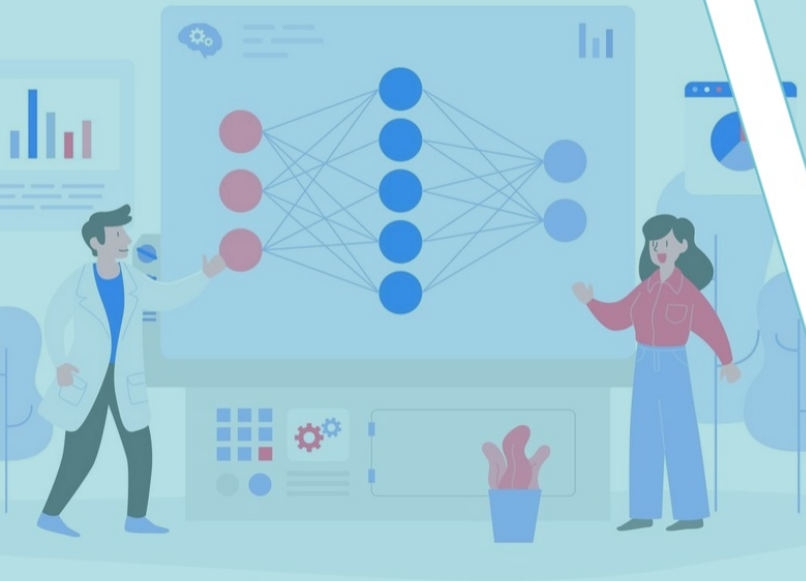


Modèle sur mesure avancé
Recherche de l'embedding

Method	ROC AUC	F1 score	Accuracy	Precision	Recall	Training time	Inference time
RNN | 1gram | int | GloveTwitter 100d + MASK + NT	0.873813	0.786117	0.79110	0.805328	0.7678	426.015586	8.673374
RNN | 1gram | int | GloveTwitter 200d + MASK + NT	0.873468	0.785758	0.79180	0.809241	0.7636	417.524911	9.157506
RNN | 2gram | int | Keras embedding + MASK	0.867917	0.788273	0.79165	0.801260	0.7757	219.931416	8.449884
NN | 2gram | multi_hot	0.867186	0.793385	0.79385	0.795178	0.7916	75.339019	2.850952
RNN | 1gram | int | Word2Vec300_GoogleNews + M...	0.866871	0.782966	0.78620	0.794991	0.7713	681.599748	9.969062
NN | 2gram | count	0.866717	0.792188	0.79410	0.799613	0.7849	101.619375	2.707457
RNN | 2gram | int | Keras embedding	0.861687	0.786254	0.79195	0.808387	0.7653	168.930055	6.789818
RNN | 1gram | int | FastText300 + MASK + NT	0.860322	0.774277	0.78115	0.799382	0.7507	810.605857	10.895422
NN | 2gram | tf-idf	0.859499	0.780317	0.78615	0.802197	0.7596	58.853025	2.635296
NN | 1gram | multi_hot	0.855491	0.781014	0.78085	0.780429	0.7816	110.710172	2.988151
RNN | 1gram | int | Glove 100d + MASK + NT	0.854755	0.783213	0.77530	0.756570	0.8118	459.133742	8.510441
RNN | 2gram | int | Glove 100d + MASK + NT	0.853224	0.774258	0.77340	0.771338	0.7772	415.313036	8.654036
LogisticRegression | TfidVectorize	0.850403	0.768425	0.76795	0.766856	0.7700	14.999396	0.003437
RNN | 1gram | int | GloveTwitter 25d + MASK + NT	0.834145	0.759330	0.75430	0.744097	0.7752	361.951274	7.954673
NN | 2gram | int	0.514969	0.539147	0.51380	0.512432	0.5688	10.920612	0.[illegible]7

Recherche du modèle le plus adapté

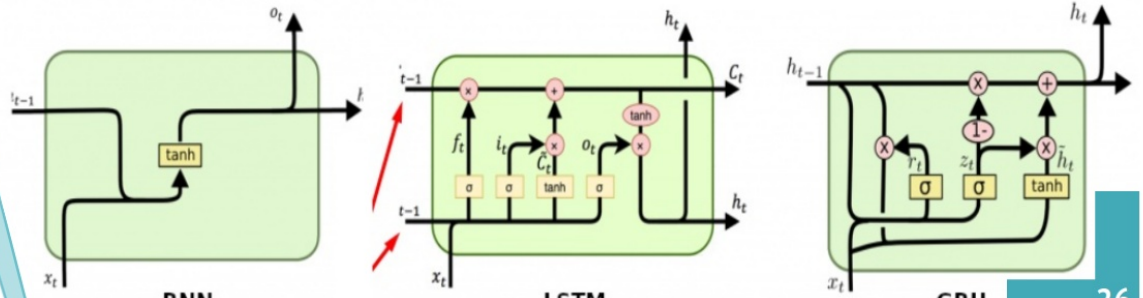
Air Paradis



Modèle sur mesure avancé
Recherche de l'architecture

Nous avons essayé des architectures de réseaux neuronaux avec ou sans couches récurrentes.

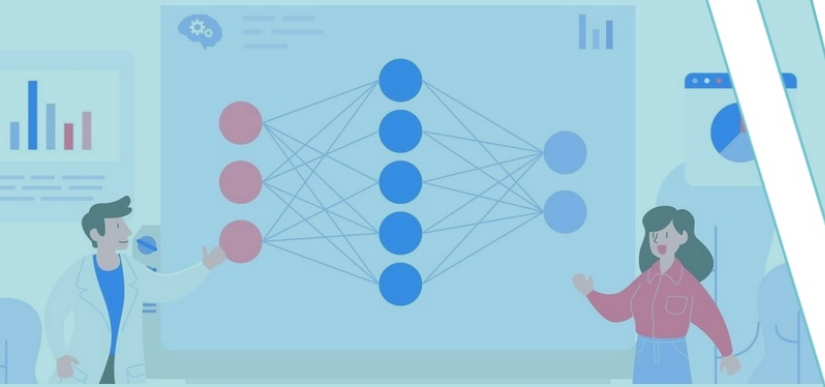
Les RNN sont des réseaux de neurones dont au moins un neurone boucle vers sa propre couche ou une couche précédente.



Recherche du modèle le plus adapté

Air Paradis

```
inputs = keras.Input(shape=(None,), dtype="int64")

x = embedding(inputs)
x = layers.Bidirectional(layers.LSTM(64))(x)
x = layers.Dropout(0.5)(x)

predictions = layers.Dense(1, activation='sigmoid', name=
```



```
inputs = keras.Input(shape=(None,), dtype="int64")

x = embedding(inputs)
x = layers.Bidirectional(layers.LSTM(64, return_sequences=True, input_
x = layers.Dropout(0.2)(x)
x = layers.Bidirectional(layers.LSTM(32))(x)
x = layers.Dropout(0.2)(x)
x = layers.Dense(64, activation='relu')(x)
x = layers.Dropout(0.1)(x)

predictions = layers.Dense(1, activation='sigmoid', name='predictions')(x)
```

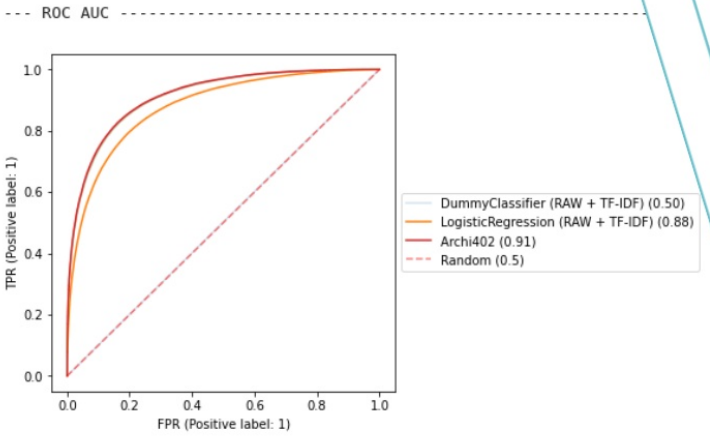
Modèle sur mesure avancé
Recherche de l'architecture

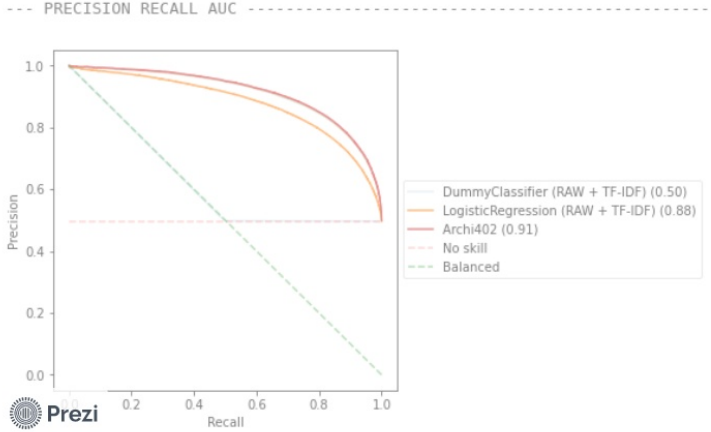
Nous avons donc préparé 11 architectures basées sur des couches SimpleRNN, LSTM ou GRU avec diverses variantes (bidirectionnel, 32 ou 64 units...)

Method	ROC AUC	F1 score	Accuracy	Precision	Recall	Training time	Inference time
Archi402	0.912817	0.825526	0.831119	0.847259	0.804881	829.031158	35.947562
Archi201	0.911483	0.826293	0.830486	0.840865	0.812217	463.956690	21.112510
Archi202	0.910601	0.828341	0.829832	0.829562	0.827123	429.748528	22.041522
Archi302	0.910368	0.830522	0.828923	0.817039	0.844457	405.142050	21.921838
Archi301	0.910134	0.827947	0.829481	0.829359	0.826541	452.679756	19.696889
ArchiREF	0.907486	0.825588	0.826445	0.823674	0.827512	444.809509	42.519483
Archi401	0.905178	0.818621	0.822832	0.832244	0.805436	300.573066	20.122485
Archi200	0.901684	0.816470	0.820705	0.829931	0.803439	213.982050	13.444628
Archi300	0.899944	0.816346	0.817821	0.817009	0.815683	198.676511	12.898922
Archi100	0.875864	0.790294	0.796317	0.808181	0.773182	372.906772	33.734413
Archi000	0.516288	0.484155	0.516850	0.515034	0.456770	18.046688	10.280299

Recherche du modèle le plus adapté

Air Paradis

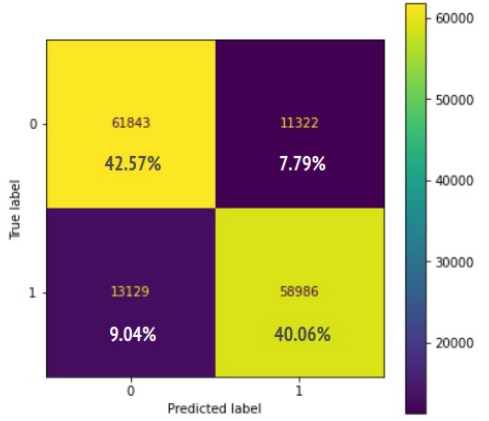




Modèle sur mesure avancé

Nous avons utilisé un RNN avec tous les samples du dataset PREPROCESS04_nofilter que nous avons préparé avec un Embedding GloVe-Twitter 200d.

```
--- Archi402 [w. threshold] ----

          ROC_AUC : 0.9128
               F1 : 0.8283
         ACCURACY : 0.8317
        PRECISION : 0.8390
           RECALL : 0.8179
AVERAGE_PRECISION : 0.9135

    TRAINING-TIME : 829.0312
    INFERENCE-TIME : 40.9845
```

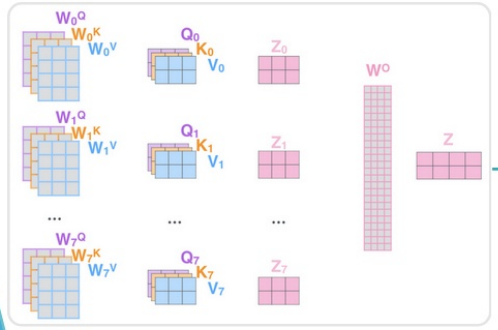


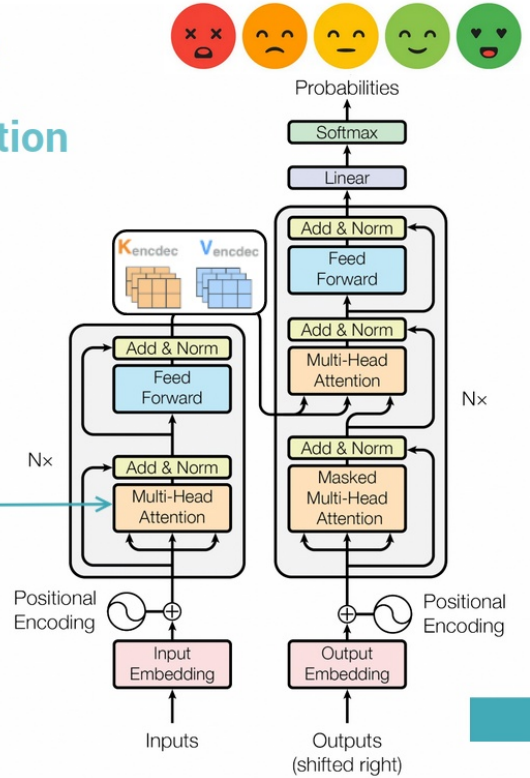
Recherche du modèle le plus adapté

Air Paradis

Modèles BERT
Introduction aux transformers

C'est un réseau neuronal de type seq2seq qui a la particularité d'utiliser un mécanisme d'attention au lieu de RNN ou CNN pour conserver l'interdépendance des mots d'une séquence.





Recherche du modèle le plus adapté

Air Paradis

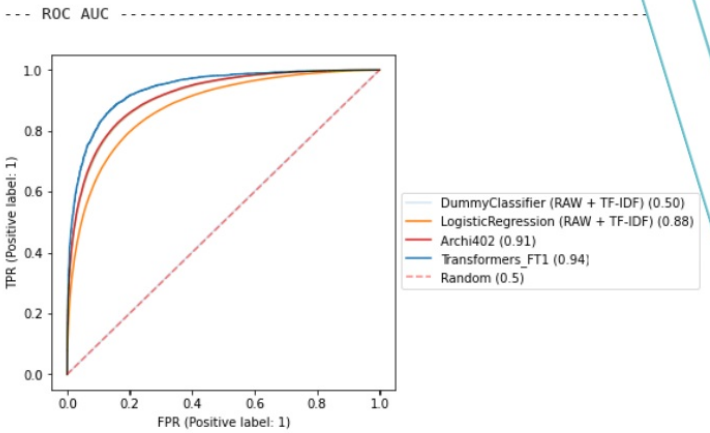
Modèles BERT sans fine-tuning

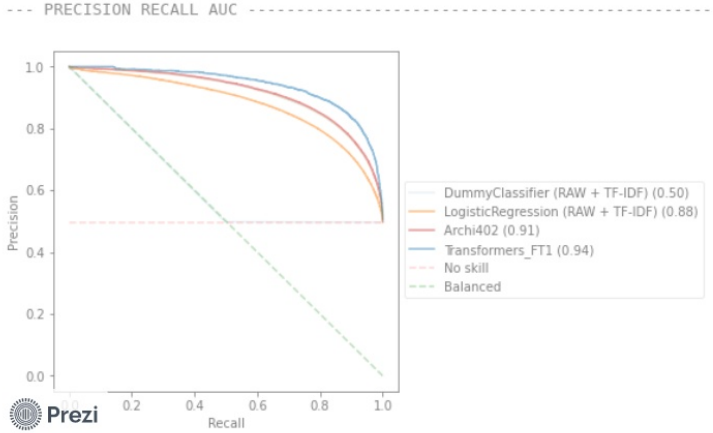
Nous avons utilisé des variantes BERT nommées cardiffnlp/twitter-roberta-base-sentiment et distilbert-base-uncased-finetuned-sst-2-english

Method	ROC AUC	F1 score	Accuracy	Precision	Recall	Training time	Inference time
Archi402	0.912817	0.825526	0.831119	0.847259	0.804881	829.031158	35.947562
cardiffnlp/twitter-roberta-base-sentiment	0.827015	0.684583	0.72360	0.801256	0.597569	0	76.847669
distilbert-base-uncased-finetuned-sst-2-english	0.795734	0.686351	0.71585	0.769554	0.619384	0	41.807902

Recherche du modèle le plus adapté

Air Paradis

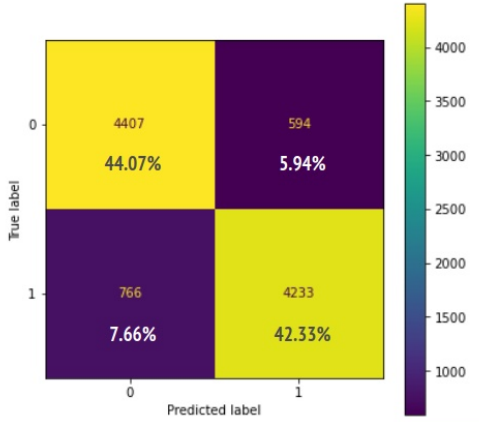




Modèle sur mesure BERT

Nous avons utilisé une variante BERT nommée roberta-base avec 100.000 samples du dataset RAW que nous avons préparé avec l'AutoTokenizer.

```
--- Transformers_FT1 ------------

          ROC_AUC : 0.9380
               F1 : 0.8616
         ACCURACY : 0.8640
        PRECISION : 0.8769
           RECALL : 0.8468
AVERAGE_PRECISION : 0.9389

    TRAINING-TIME : 7363.0000
    INFERENCE-TIME : 40.7827
```



Recherche du modèle le plus adapté

Air Paradis

Modèles BERT > conclusions

- Ne nécessite pas de pré-traitement
- Beaucoup de modèles disponibles
- Peu de données requises pour le fine-tuning
- Bons résultats sans fine-tuning
- Excellents résultats avec fine-tuning

Recherche du modèle le plus adapté

Air Paradis

Logging des scores & comparaisons

GitHub + DagsHub

Analyse de sentiments

Air Paradis

Déploiement du modèle > démo

Flask API
+
Postman

Flask
web development,
one drop at a time

docker

heroku

Axes d'amélioration

Air Paradis

- **Essayer d'autres pré-processing**
  (comme par exemple le **Stemming** qui a honteusement été oublié...)
- **Collecter des données plus récentes**
  (les expressions évoluent et Twitter est passé à 280 caractères)
- **Essayer avec plus de données**
  (pour voir si l'on peut améliorer le modèle de classification)
- **Essayer d'autres algorithmes / architectures**
  (d'autres archi. RNN ou BERT mais aussi **Universal Sentence Encoder**)
- **Surveiller les données collectées**
  (pour éviter les problèmes de data-drift ou de concept-drift)

Merci de m'avoir écouté, évalué et conseillé.

Air Paradis