

Projet 7 : Détectez les Bad Buzz grâce au Deep Learning

Xiaofan LEI



Ordre du jour

- Contexte
- Jeu de données
- Prétraitement du texte
- Trois approches
 - approche “API sur étagère”
 - approche “Modèle sur mesure simple”
 - approche “Modèle sur mesure avancé”
- Déploiement du modèle
- Démonstration
- Conclusion

Contexte

Demande initiale

- Préparer un prototype fonctionnel du modèle pour détecter les bad buzz sur les réseaux sociaux suite à la demande de “Air Paradis”

Scope du projet

- approche “API sur étagère”
- approche “Modèle sur mesure simple”
- approche “Modèle sur mesure avancé”

Jeu de données

- Open-source Sentiment140:
 - 1,6 million tweets
 - Deux polarités basée sur les emoticons

```
df_tweets.polarity.value_counts()
```

```
0    800000  
4    800000  
Name: polarity, dtype: int64
```

- Jeu de données d'entraînement : 1120 tweets (et 480 pour le test)

```
df_tweet.polarity.value_counts()
```

```
0     811  
4     789  
Name: polarity, dtype: int64
```

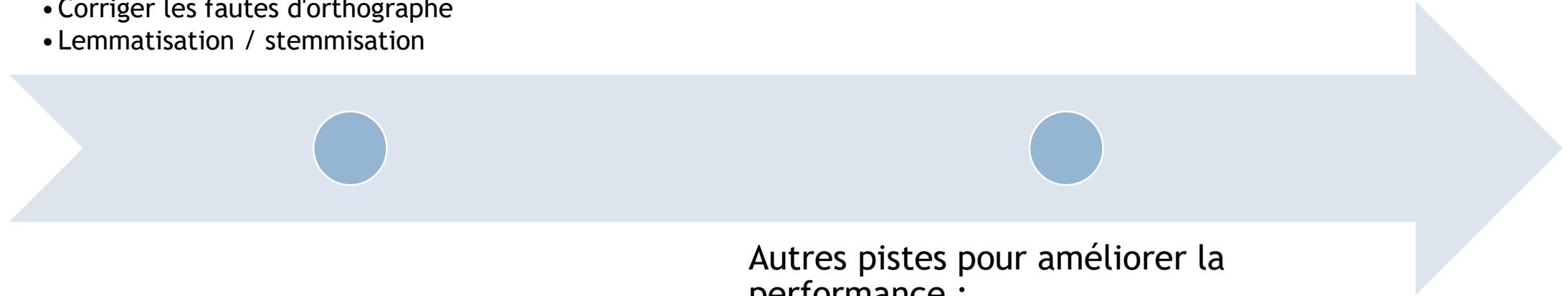
```
training size: 1120
```

```
testing size: 480
```

Pre-traitement du texte

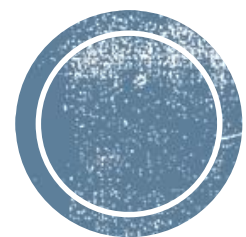
Dans le cadre du projet, chaque tweet est analysé indépendamment en suivant les étapes ci-dessous:

- Suppression des mots non informatifs
 - Suppression des hashtags, liens, mails, chiffres, ponctuations
- Mettre toutes les lettres en minuscule
- Corriger les fautes d'orthographe
- Lemmatisation / stemmisation



Autres pistes pour améliorer la performance :

- La mise en relation avec l'auteur
- L'association avec des événements
- Etiquetage des tweets en « initial » ou « réponse »



API sur étagère



Azure cognitive service

Résultat
de
requête

	sentiment	positive_score	neutral_score	negative_score
0	negative	0.0	0.0	1.0
1	negative	0.0	0.0	1.0
2	neutral	0.04	0.92	0.04
3	neutral	0.08	0.82	0.1
4	neutral	0.02	0.97	0.01
...
1595	negative	0.44	0.02	0.54
1596	positive	1.0	0.0	0.0
1597	negative	0.0	0.01	0.99
1598	positive	0.54	0.36	0.1
1599	negative	0.02	0.0	0.98

1600 rows × 4 columns



positive	563
negative	553
neutral	326
mixed	158

Tarification



Gratuit pour moins de 5000 d'enregistrement de texte par mois

De 0.0 à 0.5 millions d'enregistrements de texte - **0,8998 €** par
1 000 enregistrements texte

De 0.5 à 2.5 millions d'enregistrements de texte - **0,6749 €** par
1 000 enregistrements texte

De 2.5 à 10.0 millions d'enregistrements de texte - **0,2700 €** par
1 000 enregistrements texte

Enregistrements de texte 10.0M+ - **0,2250 €** par
1 000 enregistrements texte

Un exemple

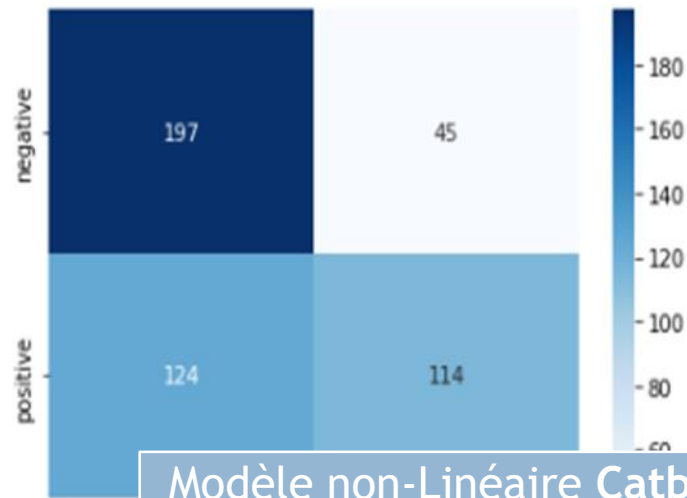
The screenshot shows a web browser at localhost:8501 with a dark theme. The page title is "sentiment analysis". It has a form with a language code input set to "en" and a text input area containing "good luck for your project!". A "get detailed result" button is below the text input. The results section shows "Overall sentiment: positive" followed by "positive_score 0.99 neutral_score 0.01 negative_score 0.0". Below this, it says "break down the analysis by each sentence" and shows "sentence # 1 : good luck for your project!" with "positive_score 0.99 neutral_score 0.01 negative_score 0.0".



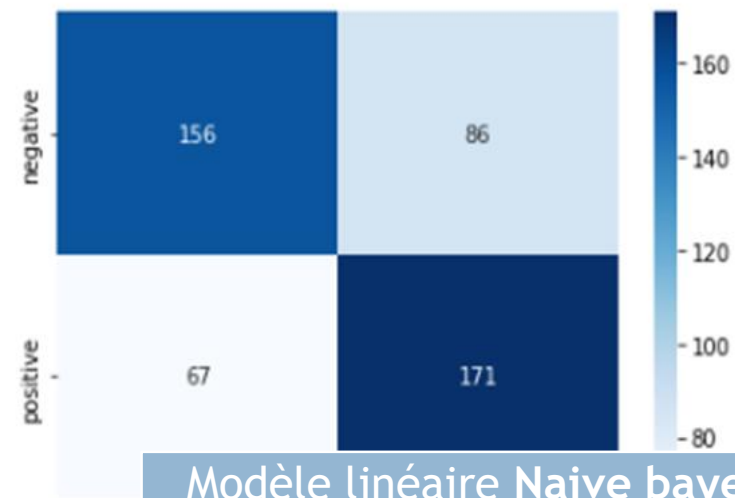
Modèles sur mesure simple



Avec Jupyter notebook

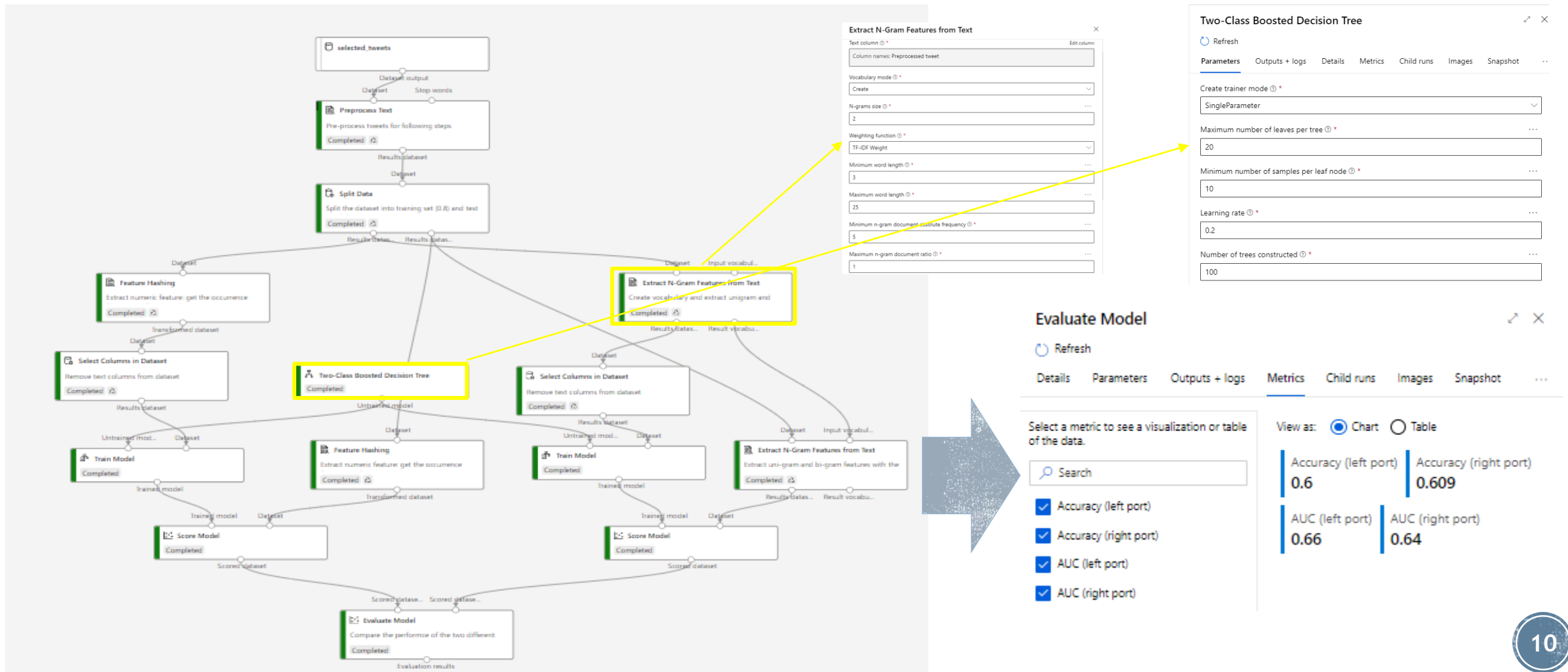


Modèle non-Linéaire Catboost
classifieur (Accuracy de test :
0,65)



Modèle linéaire Naive bayes
classifieur (Accuracy de test :
0,68)

Avec Azure Designer



Avec Azure automated ML

Algorithm name	Explained	Accuracy ↓	Sampling	Submitted time	Duration	Hyperparameter
MaxAbsScaler, ExtremeRandomTrees	View explanation	0.70313	100.00 %	4 mars 2022 12:38	23s	bootstrap : true class_weight : be ...
StandardScalerWrapper, XGBoostClassifier		0.68750	100.00 %	4 mars 2022 12:46	1m 5s	booster : gbtrees colsample_bytre ...
SparseNormalizer, LightGBM		0.67969	100.00 %	4 mars 2022 12:38	23s	boosting_type : gbdt colsample ...
MaxAbsScaler, LogisticRegression		0.67188	100.00 %	4 mars 2022 12:45	1m 1s	C : 0.2682695795279725 class_w ...
MaxAbsScaler, LogisticRegression		0.66406	100.00 %	4 mars 2022 12:47	1m 16s	C : 109.85411419875572 class_w ...
MaxAbsScaler, LogisticRegression						
MaxAbsScaler, LogisticRegression						
MaxAbsScaler, LogisticRegression						
MaxAbsScaler, XGBoostClassifier						

Model performance

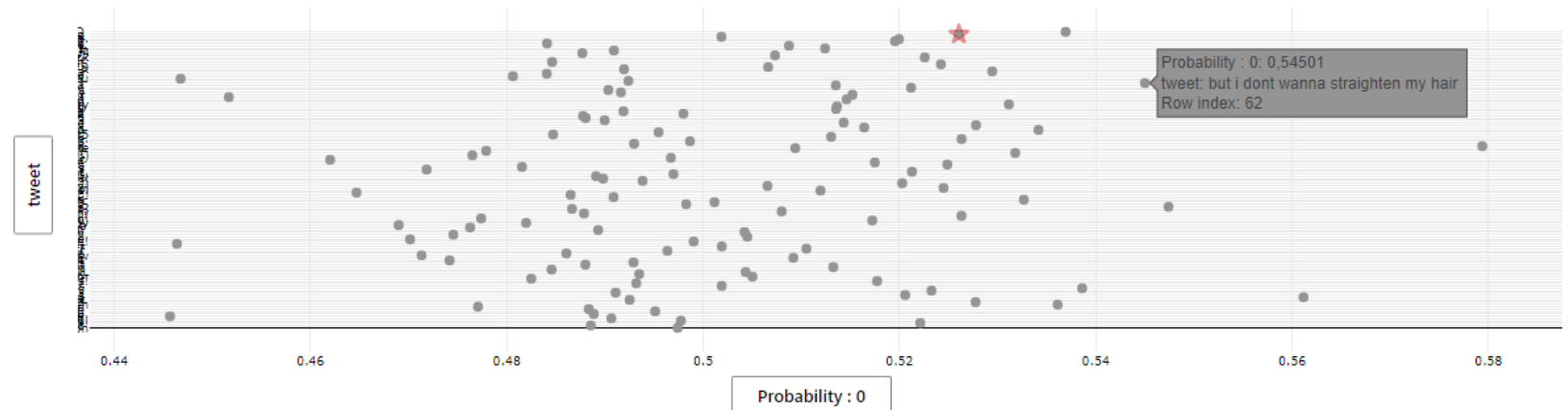
Dataset explorer

Aggregate feature importance

Individual feature importance

Select an individual datapoint by clicking on a datapoint in the scatterplot to view its local feature importance values below and feature values in the panel on the right.

Select a dataset cohort to explore





Modèles sur mesure avancés



Word embedding



le contexte
des mots
n'est pas
maintenu

Incorporation à
chaud

Incorporation
basée sur la
fréquence

Countvectorizer

Tfidf

Incorporation à
l'intégration

Word2Vec

GLOVE

FastText

Couche
d'intégration

entraîne lui-
même

entraîne lui-
même

entraîne lui-
même

Pré-entraînée

Pré-entraînée

Pré-entraînée



comprend la
signification
contextuelle

Des représentations numériques
similaires pour des mots similaires

génère de meilleures incorporations de mots
pour les mots rares, ou même des mots non vus

- word2vec :
 - crée un certain nombre de paires de mots (une variable indépendante et dépendante) en fonction de la taille de la fenêtre
 - minimise la perte de prédiction des mots cibles étant donné les mots de contexte.
- GLOVE :
 - construit une matrice de cooccurrence
 - factorise cette matrice pour obtenir une représentation de dimension inférieure
- FastText : utilise des caractères n-grammes comme la plus petite unité

Réseaux de neurones (avant 2018)

Apprentissage

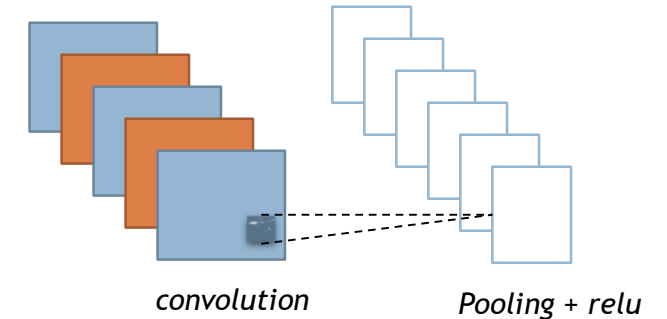
le Réseau neuronal convolutif (CNN)

particulièrement utilisé afin de classifier des images.

La *convolution* : applique un filtre pour extraire les *features*

Le *pooling* : prend la valeur maximale ou moyenne de chaque « morceau »

La fonction d'activation de type ReLU : conserve les valeurs positives

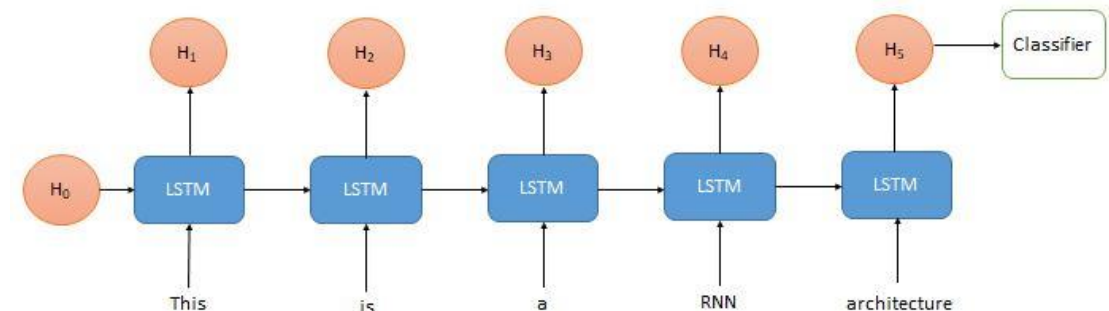


Les réseaux de neurones récurrents (RNN)

Prend en compte l'ordre des mots,
l'information se perd au fur à mesure (vanishing gradients)
ne peut pas paralléliser les calculs.

LSTM (Long Short-Term Memory) : crée une mémoire via un système de portes (gates) et d'états.

GRU (Gated Recurrent Unit) : une variante plus simple de LSTM



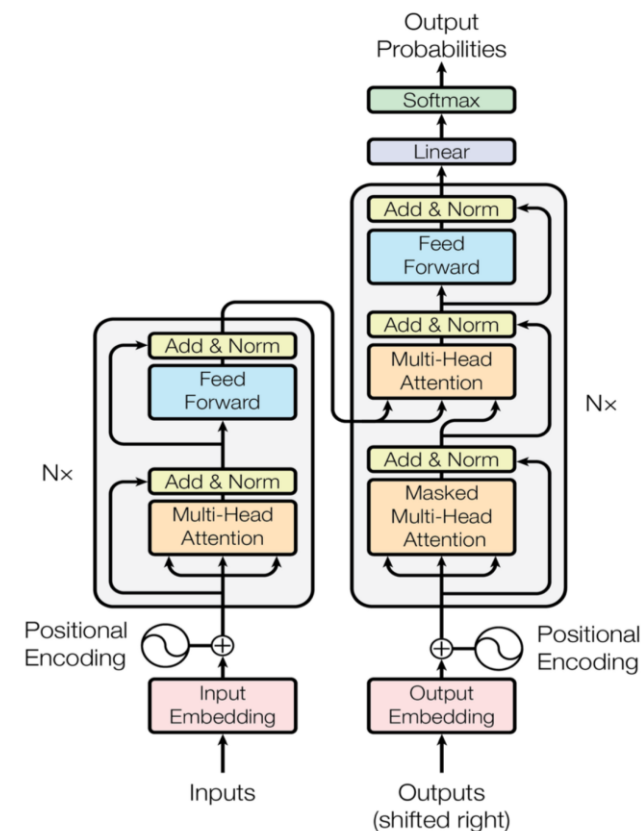
Transformers (à partir de 2018)

Transformers : modèles séquence à séquence (seq2seq)

- **Attention** : mesure à quel point deux éléments de deux séquences sont liés
- **Multi-head Attention**: a pour but d'avoir plusieurs « sous-espaces de représentation » qui empêchent que la représentation soit totalement biaisée si une couche d'*attention* l'est.
- **Position représentation** : stock l'importance de l'ordre des mots dans les vecteurs

BERT : Bidirectional Encoder Representations from Transformers

- Proposé par Google AI fin 2018,
- Plus performant, plus rapide dans l'apprentissage.
- Une fois pré-entraîné, peut être entraîné en mode incrémental pour spécialiser le modèle rapidement et avec peu de données
- Multi-tâches



From "Attention Is All You Need" by Vaswani et al.

Résultats de Deep learning

Accuracy global / IoU sur les sentiments négatifs

	Keras embedding		Fasttest embedding pre-entraîné	
	accuracy	IoU (0)	accuracy	IoU (0)
CNN	65%	52,91%		
LSTM			71%	52.91%

*iou = true_positives / (true_positives + false_positives + false_negatives)

Comparaison avec BERT


	Accuracy	IoU	Temps d'apprentissage
fasttext + LSTM	73,33%	52,91%	3min 7s
BERT	77.81%	66.82%	58min 21s

Fine-tuning

Accuracy sur la base de validation	Couche d'embedding pre-entraînée	Taux d'apprentissage	Taux de dropout	Cellules de LSTM	Nombre d'époque	Taille de batch
59,8%	fasttext	0,003	0,3	14	12	19
58,9%	fasttext	0,001	0,4	14	18	12
58,0%	fasttext	0,003	0,3	15	13	17
57,1%	fasttext	0,003	0,3	17	12	16
56,7%	fasttext	0,001	0,5	6	9	8
56,7%	fasttext	0,003	0,3	14	11	20
54,0%	glove	0,003	0,3	7	14	30
53,1%	fasttext	0,0003	0,3	6	17	18
53,1%	fasttext	0,001	0,5	23	11	32
52,7%	fasttext	0,003	0,3	14	12	19
52,2%	glove	0,003	0,5	23	9	17
51,8%	glove	0,0003	0,5	16	15	30
50,9%	fasttext	0,003	0,3	18	10	17
50,4%	fasttext	0,00003	0,4	23	10	26
50,4%	fasttext	0,003	0,3	12	14	19
49,1%	fasttext	0,003	0,4	14	14	15
48,7%	fasttext	0,0003	0,3	7	10	32
48,7%	fasttext	0,001	0,4	15	19	8
46,4%	glove	0,001	0,5	11	16	15
44,6%	fasttext	0,0001	0,5	31	18	23

Entraînement du modèle BERT sur Azure (Optionnel)

- Connecter au Workspace
- Créer un calcul

	Standard_DS12_v2 4 cores, 28GB RAM, 56GB storage	Memory optimized	Data manipulation and training on medium-sized datasets (1-10GB)	6 cores	\$0.47/hr
-----------------------------------------------------------------------------------	-----------------------------------------------------	------------------	---------------------------------------------------------------------	---------	-----------

- Créer une expérimentation
- Récupérer les données d'entraînement
- Configurer un environnement avec les bibliothèques nécessaires
- Entraîner le modèle
 - Créer un répertoire d'entraînement
 - Créer un script pour exécuter l'entraînement
 - Lancer l'entraînement sur un cluster
 - Enregistrer le modèle après l'entraînement
- Enregistrer le modèle final dans le workspace

Mise en production

- **Etape 1 : Déployer le modèle en tant que web service**
 - Créer le script de scoring : script pour interroger le modèle
 - Créer le fichier de configuration, surtout l'environnement dans lequel le script doit être exécuté
 - Déploiement dans ACI (Azure Container Instances)
 - Récupérer le **endpoint** http
- **Etape 2 : Créer un site web avec Streamlit**
 - Créer le script pour interagir avec le endpoint
 - Exécuter la commande pour lancer streamlit



localhost:8501

sentiment analysis

english only

Enter your text here:

@ColdHearted19 ooo will watch out for itt, final auditions next week

submit

sentiment detected: positive

Conclusion

- Approche “API sur étagère”
 - Facile à mettre en place
 - Peu coûteux
 - Ne maîtrise pas la technologie dernière
- Approche “Modèle sur mesure simple”
 - Relativement facile à mettre en place
 - La performance n’est pas au rendez-vous
- Approche “Modèle sur mesure avancé”
 - Performant
 - BERT peut répondre à nos besoins mais exigeant vis-à-vis de l’architecture matérielle, même pour le fine-tuning et le déploiement

Merci pour votre attention

- Pour plus de détail, veuillez consulter mon blog :
- <https://medium.com/@lei.xiaofan/quick-start-building-sentiment-analysis-models-8c1e78c30b2c>