



DataScientest • com

PHUOC NHAT DANG
SANA MZALI

07/02/2020

TWITTER SENTIMENT ANALYSIS



- ▶ Problématique
- ▶ Exploration des données
- ▶ Visualisation des données
- ▶ Itération: construction du modèle et évaluation
- ▶ Perspectives
- ▶ Conclusion



- ▶ Problématique
- ▶ Exploration des données
- ▶ Visualisation des données
- ▶ Itération: construction du modèle et évaluation
- ▶ Perspectives
- ▶ Conclusion



- ▶ Problème: Analyse de sentiments exprimés dans les tweets
- ▶ Objectif: Analyser des émotions, sentiments et opinions exprimés dans des messages postés sur Twitter classés positifs ou négatifs
- ▶ Exemples:
 - ▶ “my whole body feels itchy and like its on fire”: négatif
 - ▶ “whoo hoo! congrats congrats! didn’t know you could find out so soon”: positif

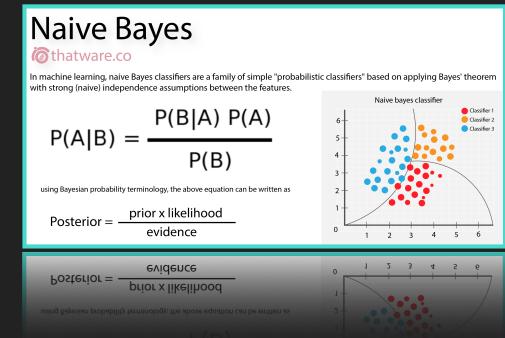
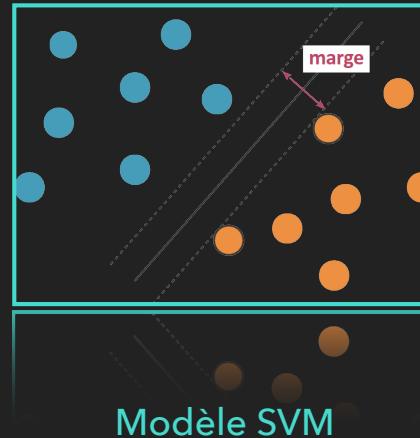


MÉTHODES SCIENTIFIQUES (1/2)



DataScientest • com

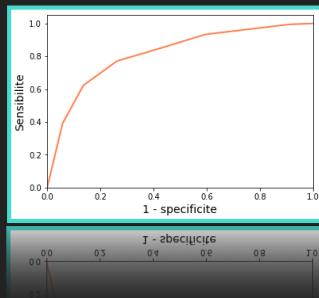
- ▶ Type de problème: Analyse textuelle
- ▶ Tâche de ML: Analyse de sentiment
- ▶ Type de données : non structurées (texte)
- ▶ Type de modèle: classification en apprentissage supervisé
- ▶ Modèles prédictifs:
 - ▶ SVM (Support Vector machines)
 - ▶ Naive Bayes Classifier



Modèle bayésien naïf

		Classe réelle	
		-	+
Classe prédictée	-	True Negatives (vrais négatifs)	False Negatives (faux négatifs)
	+	False Positives (faux positifs)	True Positives (vrais positifs)
réel	+	True Positives (vrais positifs)	False Negatives (faux négatifs)

Exemple de matrice de confusion



Exemple de courbe de ROC

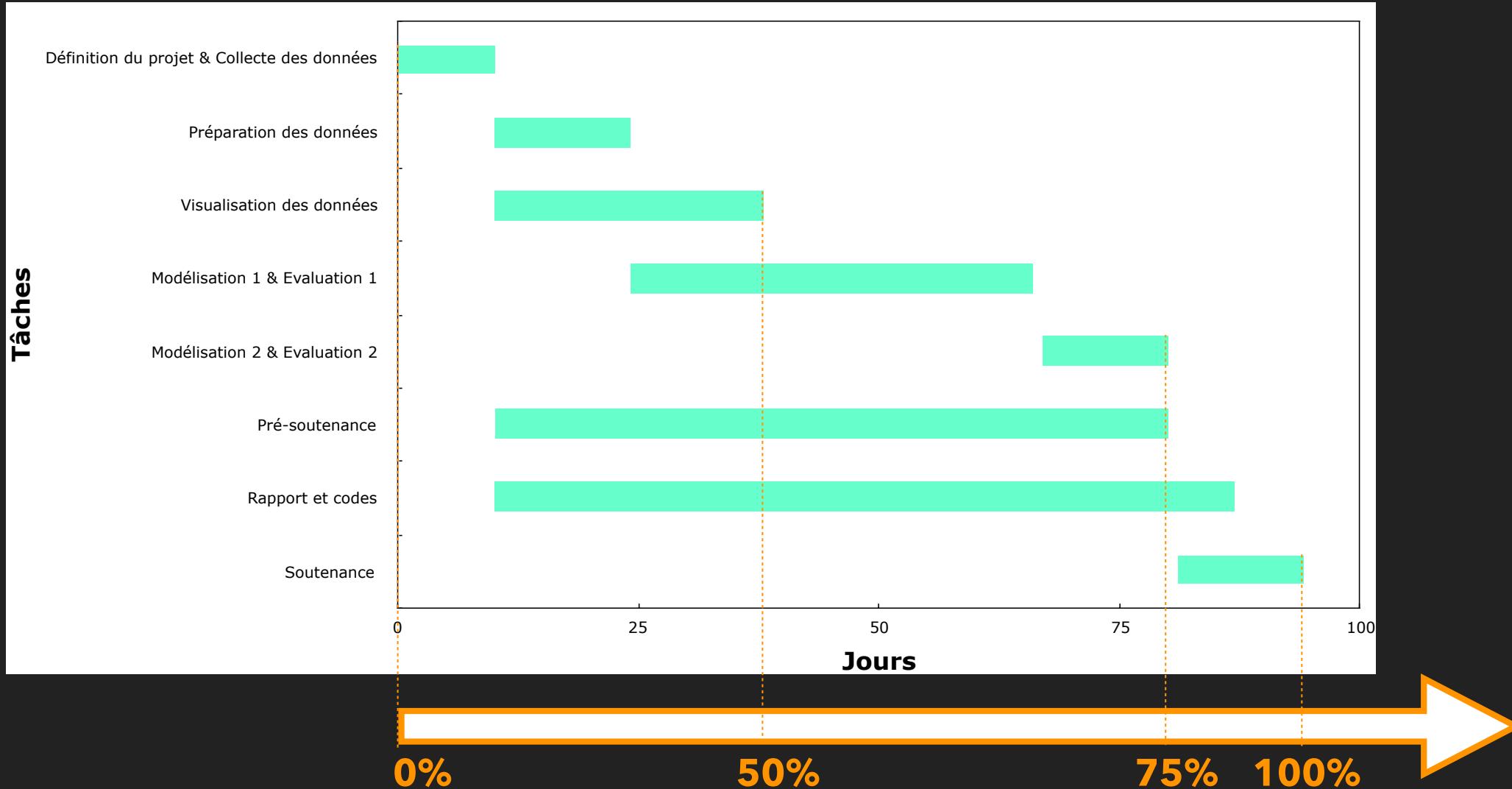
- ▶ Métriques de mesure de performance:
- ▶ Accuracy score
- ▶ Matrice de confusion
- ▶ Rappel, précision, f1-score
- ▶ Courbe de ROC
- ▶ ROC AUC Score

MÉTHODES SCIENTIFIQUES (2/2)



DataScientest • com

- ▶ Etape 1: Pré-traitement des données (Stopwords, Tokenization, Stemming, Pos Tagging)
- ▶ Etape 2: Visualisation des Tweets (Wordcloud, barplot)
- ▶ Etape 3: Transformation du texte en vecteur (CountVectorizer, TF-IDF)
- ▶ Etape 4: Construction du modèle (Modèles de classification: K-NN, Random Forest, SVM,...)
- ▶ Etape 5: Evaluation (F1-score, la courbe precision-recall, la courbe de ROC, la courbe de AUC...)
- ▶ Etape 6: Amélioration (k-fold cross validation, Gridsearch, XGBoost, deep learning, ...)





- ▶ Problématique
- ▶ Exploration des données
- ▶ Visualisation des données
- ▶ Itération: construction du modèle et évaluation
- ▶ Perspectives
- ▶ Conclusion

DONNÉES (1/2)



DataScientest • com

Nom: Le jeu de données **Sentiment140**

Source: Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." *CS224N project report, Stanford 1.12 (2009)*: 2009.

Jeux de donnée (.csv) comprenant:

- ▶ 1,6 million de tweets
- ▶ 6 colonnes:

- ◆ target: la nature/classe du tweet (0 = négatif, 4 = positif).
- ◆ ids: l'identifiant du tweet
- ◆ date: la date du tweet sous la forme: (Sat May 16 23:58:44 PDT 2009)
- ◆ flag: existence ou non d'une requête
- ◆ user: l'identifiant de l'utilisateur qui a posté le tweet
- ◆ text: l'intégralité du texte du tweet

0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zI - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D
0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scothamilton	is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah!
0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattykus	@Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds
0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwidelass no, it's not behaving at all. i'm mad. why am i here? because I can't see you all over there.
0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug
0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rains a bit ,only a bit LOL , I'm fine thanks , how's you ?
0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope they didn't have it
0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	@twittera que me muera ?
0	1467812416	Mon Apr 06 22:20:16 PDT 2009	NO_QUERY	erinx3leannexo	spring break in plain city... it's snowing
0	1467812579	Mon Apr 06 22:20:17 PDT 2009	NO_QUERY	pardonlauren	I just re-pierced my ears
0	1467812723	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	TLeC	@caregiving I couldn't bear to watch it. And I thought the UA loss was embarrassing
0	1467812771	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	robobbierobert	@octolinz16 It it counts, idk why I did either. you never talk to me anymore

DONNÉES (2/2)



DataScientest • com

Nom: Le jeu de données **Sentiment140**

- ◆ **Variable cible : «target» (positif ou négatif)**
- ◆ **Variables explicatives significatives: «text» et «date»**

	ID	Date	Label	Text	Source
0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_ @switchfoot http://twitpic.com/2y1zI - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ,D	
0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah!	
0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattykus @Kenichan I dived many times for the ball. Managed to save 50%. The rest go out of bounds	
0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF my whole body feels itchy and like its on fire	
0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli @nationwideclass no, it's not behaving at all. i'm mad. why am i here? because i can't see you all over there.	
0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf @Kwesidei not the whole crew	
0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch Need a hug	
0	1467811593	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ @LOLTrish hey long time no see! Yes. Rains a bit, only a bit LOL , I'm fine thanks , how's you ?	
0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollywood @Tatiana_K nope they didn't have it	
0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	minimiso @Twittera que me muera ?	
0	1467812416	Mon Apr 06 22:20:16 PDT 2009	NO_QUERY	erinx3leanexo spring break in plain city... it's snowing	
0	1467812579	Mon Apr 06 22:20:17 PDT 2009	NO_QUERY	pardonlauren I just re-pierced my ear	
0	1467812723	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	TLeC @caregiving I couldn't bear to watch it. And I thought the UA loss was embarrassing	
0	1467812771	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	robrobberrobert @octolinz16 if it counts, idk why I did either. you never talk to me anymore	

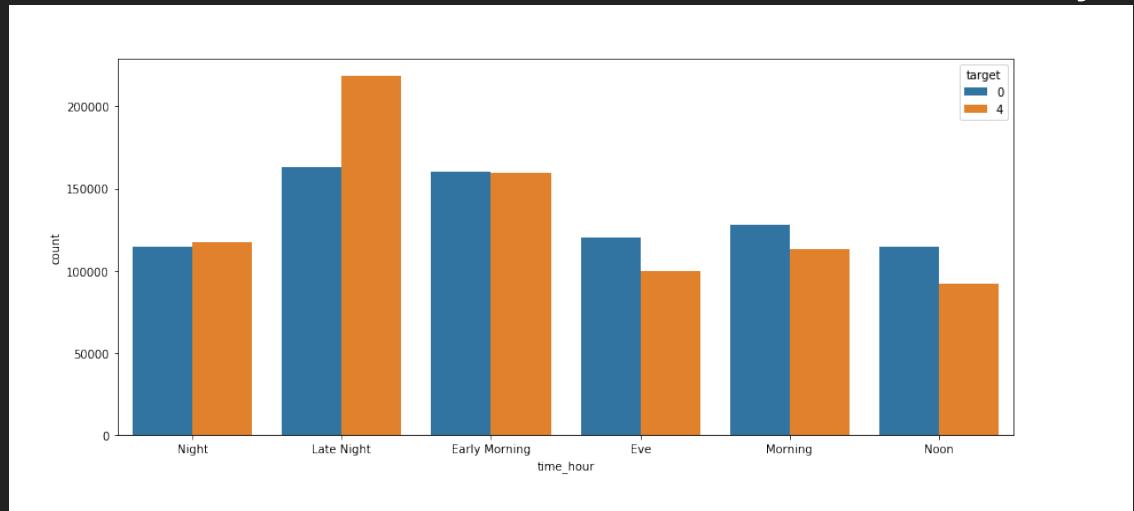
Traitement des données:

- ◆ Absence de valeurs manquantes
- ◆ Absence de doublons et retweets
- ◆ Jeu de données équilibré: 800.000 tweets + et 800.000 tweets -
- ◆ Réduction des données à 200,000 tweets



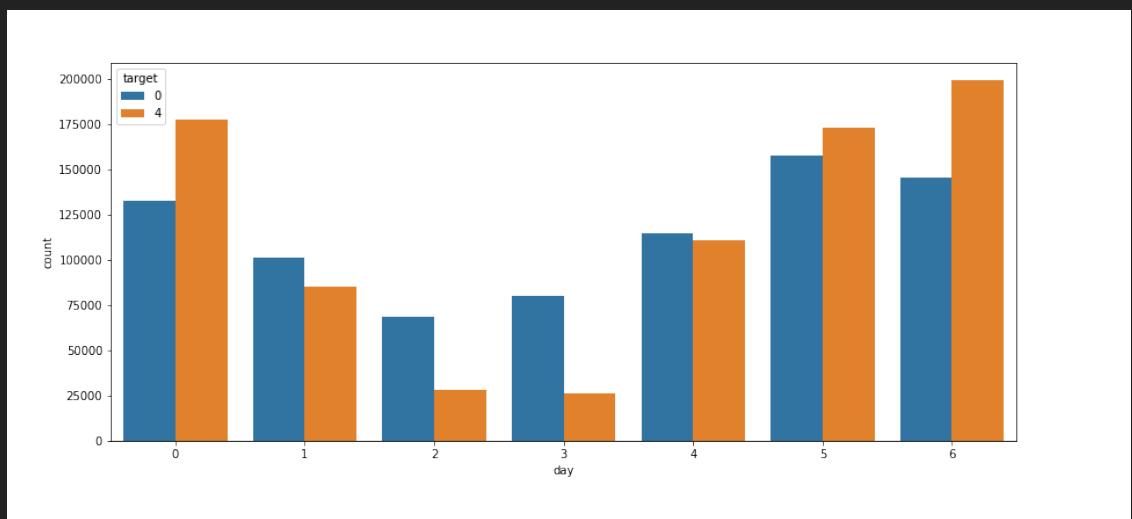
- ▶ Problématique
- ▶ Exploration des données
- ▶ Visualisation des données
- ▶ Itération: construction du modèle et évaluation
- ▶ Perspectives
- ▶ Conclusion

► Distribution du nombre de tweets durant la journée selon la nature du tweet



- ▶ Les tweets négatifs sont prédominants le long de la journée alors que les tweets positifs sont plus présents le soir.

► Distribution du nombre de tweets selon le jour de la semaine et la nature du tweet

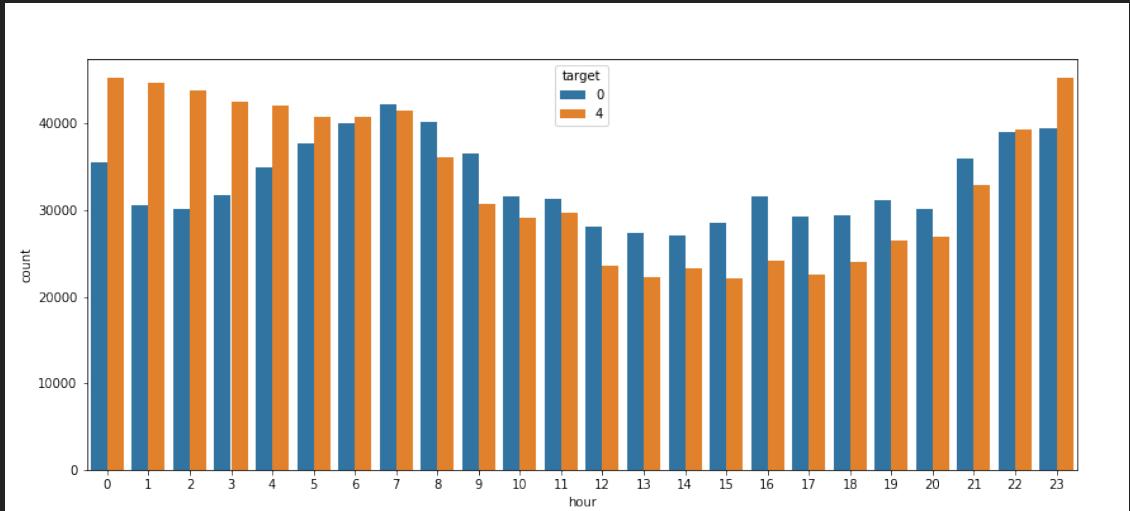


- ▶ Les tweets sont plus de nature positive le weekend et en début de semaine (le lundi) alors qu'ils sont plus négatifs le reste de la semaine.

DATA'VIZ - << DATE >> (2/2)

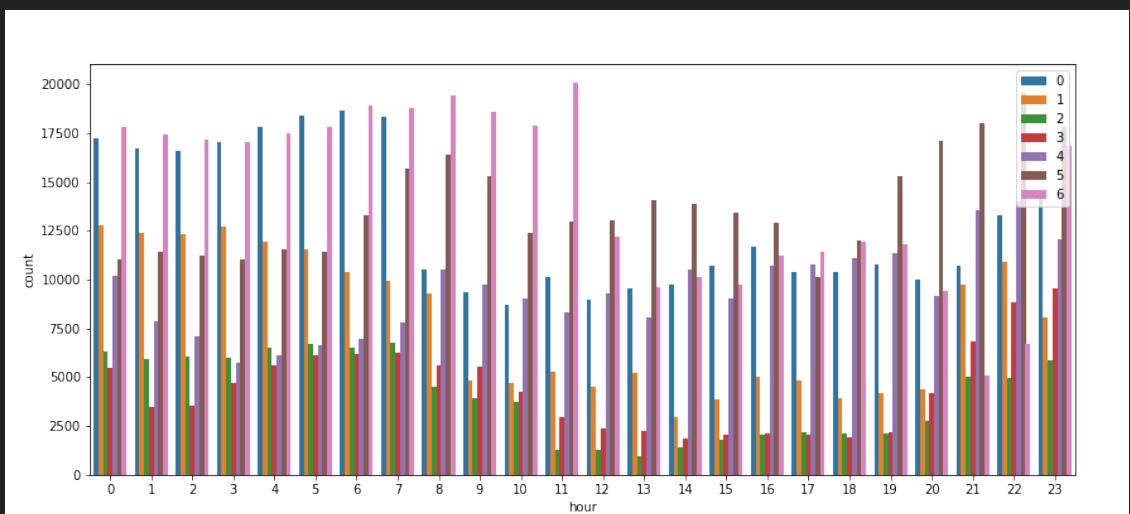


▶ Distribution du nombre de tweets selon l'heure du tweet et la nature du tweet



- ▶ Les tweets sont plus de nature positive le soir à partir de 23h et la nuit jusqu'à 6h du matin alors qu'ils sont plus négatifs durant la journée.

▶ Distribution du nombre de tweets selon le jour et l'heure du tweet

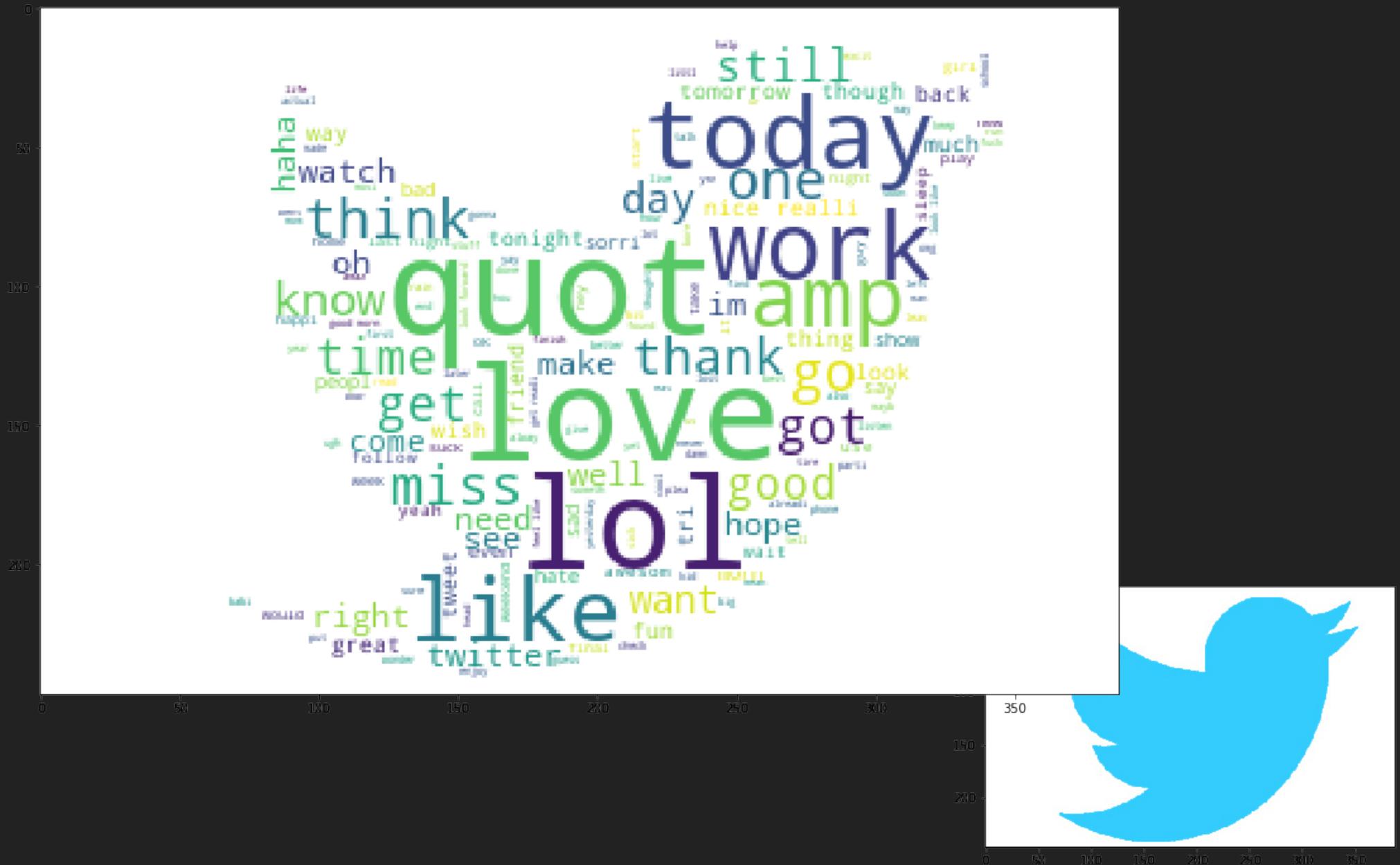


- ▶ Les tweets postés la nuit sont plus ceux du week-end (la nuit du samedi au dimanche et surtout la nuit du dimanche au lundi); ceux de la matinée sont ceux du dimanche suivi par ceux du lundi et enfin les tweets de l'après-midi sont plus postés le samedi.

DATA'VIZ - « WORDCLOUD » (1/2)



- ▶ WordCloud des mots regroupant les tweets positifs et négatifs



DATA'VIZ - << WORDCLOUD >> (2/2)



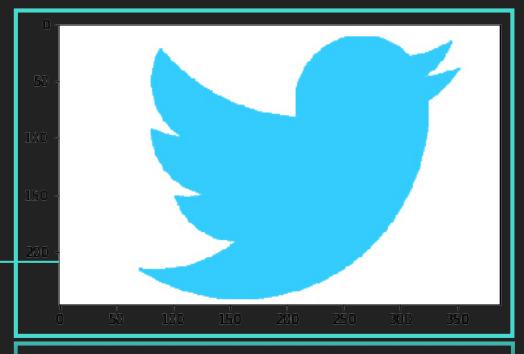
▶ WordCloud des mots issus des tweets positifs



▶ WordCloud des mots issus des tweets négatifs



- ▶ Sur le WordCloud des tweets positifs, on remarque la prédominance de certains mots à connotation positive comme: "love", "thank", "like", "lol", "good" etc.



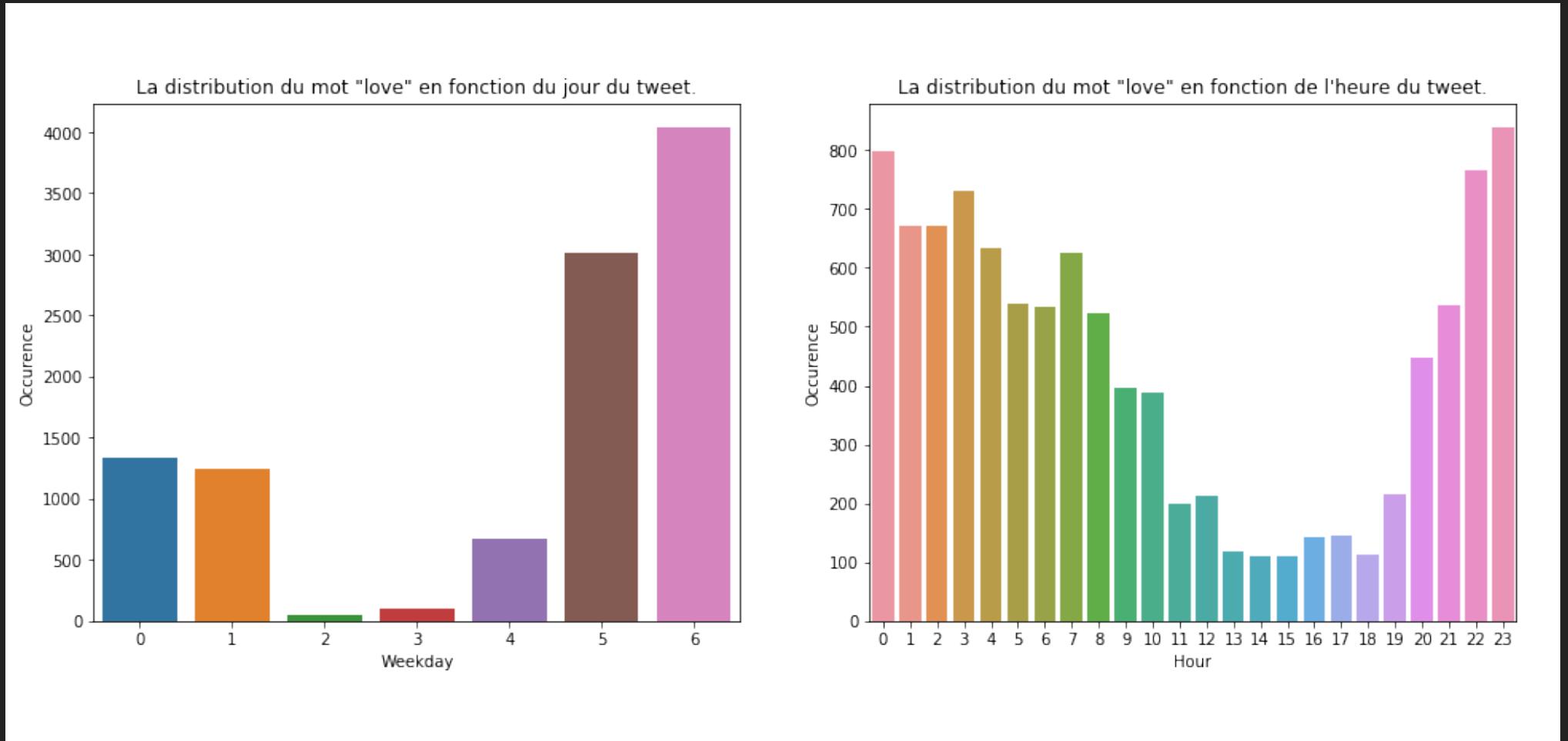
- ▶ Sur le WordCloud des tweets négatifs, on remarque la présence de certains mots à connotation négative comme: "sad", "hate", "bad" mais également d'autres à connotation positive comme "like" et "good" qui peuvent soit être précédés d'une négation soit utilisés dans un contexte ironique.

DATA'VIZ - « MOT » (1/2)



DataScientest • com

- ▶ Distribution d'un mot spécifique selon l'heure et le jour de la semaine



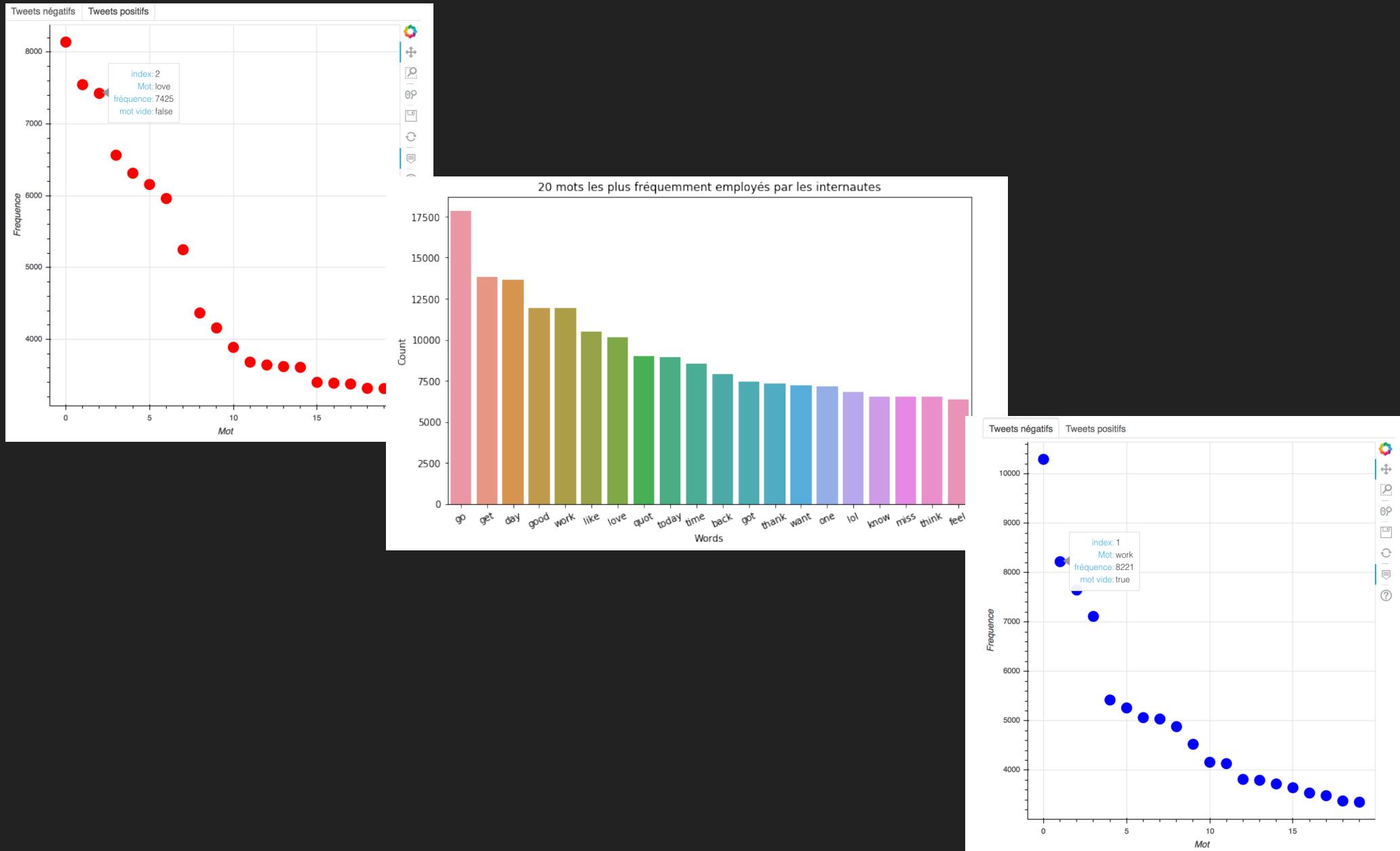
- ▶ Par exemple, si on prend le mot "love", on voit qu'il apparaît principalement dans les tweets du week-end et pendant la soirée ou la nuit.

DATA'VIZ - « MOT » (2/2)



DataScientest • com

► Histogramme des mots les plus fréquemment utilisés





- ▶ Problématique
- ▶ Exploration des données
- ▶ Visualisation des données
- ▶ Itération: construction du modèle et évaluation
- ▶ Perspectives
- ▶ Conclusion

TRAITEMENT DES DONNÉES



DataScientest • com

► Types de traitement de données

- ▶ A: Sans traitement des données
- ▶ B: Avec traitement des données (Stopwords, Tokenisation, La racinisation: stemming, etc.)

```
def create_corpus(df):  
    text = []  
    for i in range(0,df.shape[0]):  
        #Removing all except a-z, A-Z letters and replace by space:  
        review = re.sub('(^0-9A-Za-z |\t)|(\w+:\//\S+)', ' ', df.iloc[i])  
        #Making all letters to lower case:  
        review = review.lower()  
        #Split the review to the list of words  
        review = review.split()  
        #Remove Stopwords (the, a, an, on, and, in...)  
        #and Stemming (derive the root of the word):  
        ps = PorterStemmer()  
        review = [ps.stem(word) for word in review if not word in set(stopwords.words('english'))]  
        #Joining the word list to a sentence (separate by space):  
        review = ' '.join(review)  
        text.append(review)  
    return text
```

- ▶ C: Avec traitement des données (Stopwords, Tokenisation, La racinisation: stemming, etc.)
+ Pos Tagging

```
def create_corpus(df):  
    text = []  
    for i in range(0,df.shape[0]):  
        #Removing all except a-z, A-Z letters and replace by space:  
        review = re.sub('(^0-9A-Za-z |\t)|(\w+:\//\S+)', ' ', df.iloc[i])  
        #Making all letters to lower case:  
        review = review.lower()  
        #Split the review to the list of words  
        review = review.split()  
        #Remove Stopwords (the, a, an, on, and, in...)  
        #and Stemming (derive the root of the word):  
        ps = PorterStemmer()  
        review = [ps.stem(word) for word in review if not word in set(stopwords.words('english'))]  
        #Joining the word list to a sentence (separate by space):  
        review = ' '.join(review)  
        text.append(review)  
    return text
```

```
# POS Tagging  
def POSTagging(Corpus):  
    text = []  
    for i in range(0,len(Corpus)):  
        tokenized = sent_tokenize(Corpus[i])  
        for j in tokenized:  
            wordsList = nltk.word_tokenize(j)  
            tagged = nltk.pos_tag(wordsList)  
            text.append(tagged)  
    return text
```

```
reflex =  
    reflex = sdbbeng(fsddeq)  
    fsddeq = nJFK·sof-fsd(mordarif)  
    MOTCPPTBZ = nTCK·MOTD·soCQNTSC()
```

► Transformation du texte en vecteur:

- ▶ CountVectorizer
- ▶ TF-IDF

CONSTRUCTION DES MODÈLES



DataScientest • com

- ▶ 1. Régression logistique
- ▶ 2. Naif bayésien BernoulliNB
- ▶ 3. Naif bayésien MultinomialNB
- ▶ 4. K plus proches voisins
- ▶ 5. Arbre de décision
- ▶ 6. Forêt aléatoire
- ▶ 7. XGBOOST
- ▶ 8. Gradient Boosting
- ▶ 9. Adaboost

EVALUATION DES MODÈLES (1/3)



DataScientest • com

► Exemple d'évaluation détaillée: Régression logistique

Evaluation détaillée de la Classification par régression logistique :				
	precision	recall	f1-score	support
0	0.80	0.78	0.79	19887
4	0.78	0.81	0.80	20102
accuracy			0.79	39989
macro avg	0.79	0.79	0.79	39989
weighted avg	0.79	0.79	0.79	39989

La matrice de confusion pour la Classification par régression logistique :				
Classe prédictive	0	4		
Classe réelle	0	15437	4450	
0	15437	4450		
4	3894	16208		

Le score pour la Classification par régression logistique : 0.7913426192202856

► Comparaison des scores obtenus pour les différents modèles entraînés

Tableau 1-1

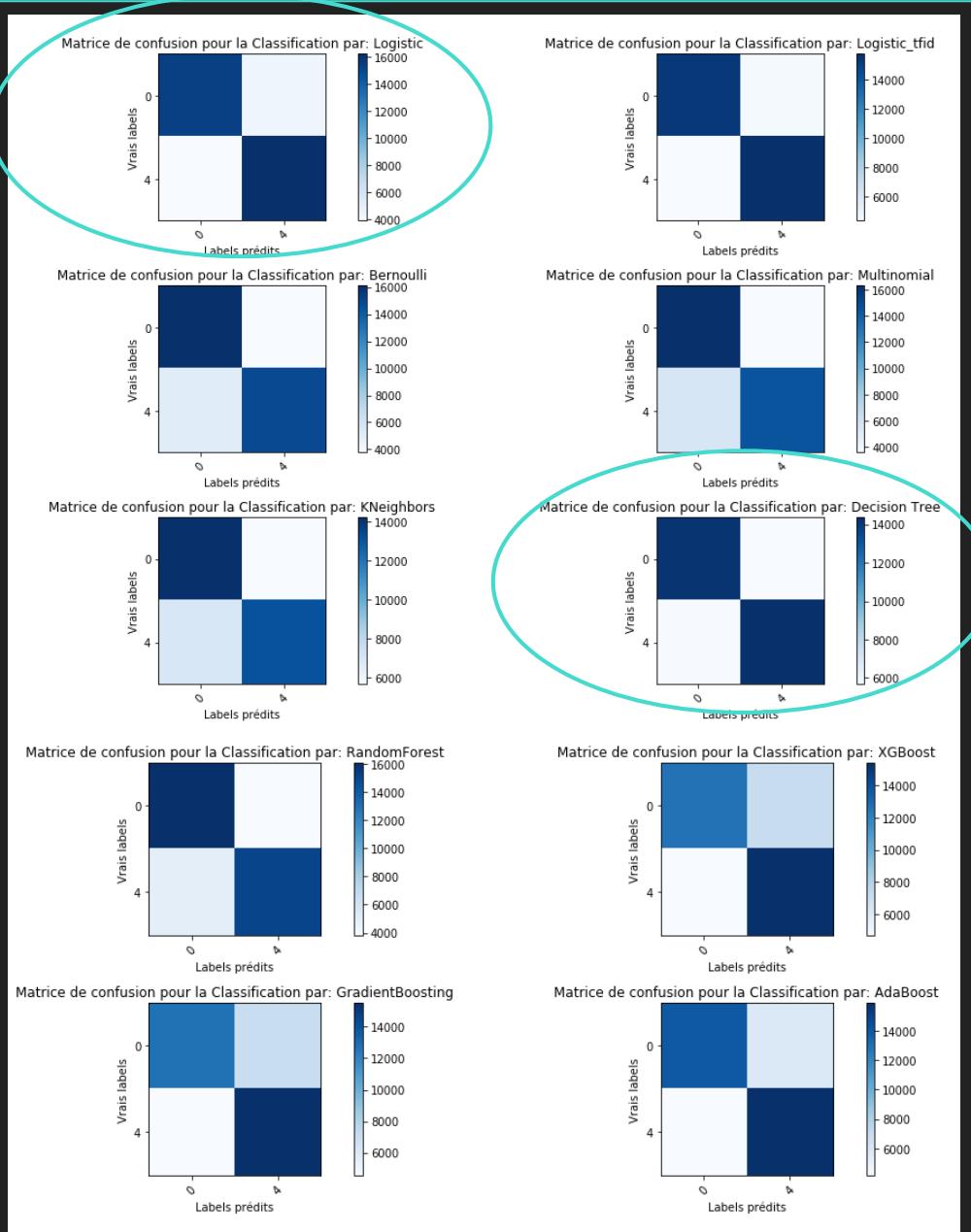
Modèle	Régression logistique	Régression logistique/TF-IDF	Naïf bayésien BernoulliNB	Naïf bayésien MultinomialNB	KNN	Arbre de décision	Forêt aléatoire	XGBOOST	Gradient Boosting	Adaboost
Score (200,000 tweets)	0.79	0.78	0.77	0.77	0.68	0.71	0.78	0.70	0.70	0.75

- Le modèle de régression logistique semble être le plus performant avec un score de 79% et une aire sous la courbe (AUC) de 0.79.
- Le Pos Tagging tout comme le traitement du texte (stopwords, minuscule, ponctuation, etc.) ont eu peu d'influence sur les résultats des différents modèles.
- Amélioration: un modèle MLP a donné un résultat similaire avec un score de 80%.

EVALUATION DES MODÈLES (2/3)



► Matrices de confusion colorées

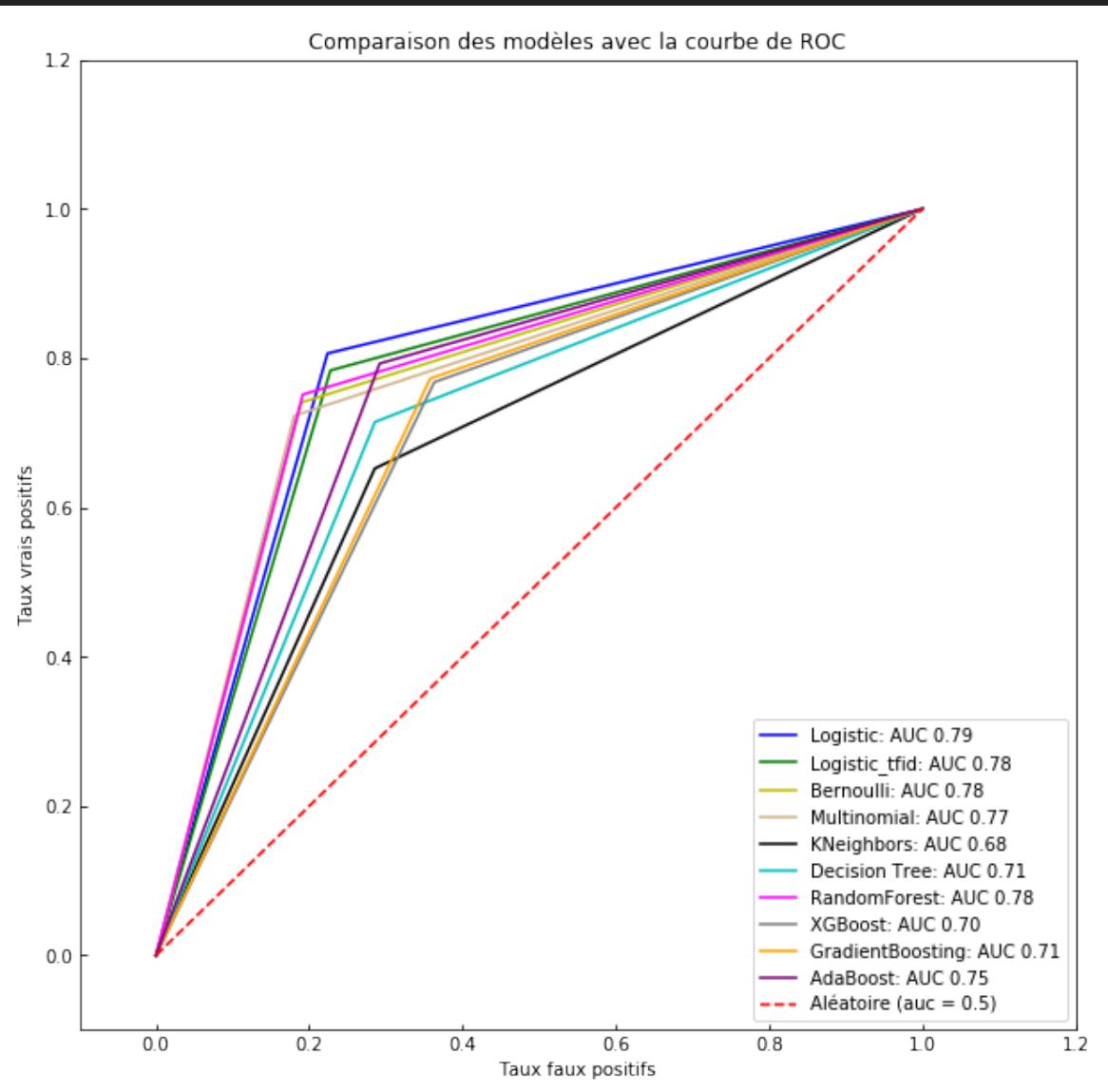


► Les matrices de confusion colorées reflètent bien les résultats des différents modèles avec une intensité de couleur plus prononcée sur la diagonale pour les méthodes de régression logistique et l'arbre de décision.

EVALUATION DES MODÈLES (3/3)



► Courbes de ROC



► Ce graphique confirme les résultats antérieurs avec des modèles bien plus performants que d'autres et notamment: la régression logistique, le naïf bayésien et le random forest.

COMPARAISON A L'ÉTAT DE L'ART



DataScientest • com

► Etat de l'art

	Source	Taille des données	Label	Modèle	Amélioration	Métrique
1	Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." <i>Proceedings of the Workshop on Language in Social Media (LSM 2011)</i> .	11 875 tweets	Positive Negative Neutral	Support Vector Machines	5-fold cross-validation	Average accuracy = 75.39%
2	Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." <i>LREc</i> . Vol. 10. No. 2010. 2010.	300 000 tweets	Positive Negative Neutral	Naïve Bayes classifier	Discriminating common n-grams,	F-0.5=0.65
3	Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." <i>CS224N Project Report, Stanford</i> 1.12 (2009): 2009.	1 600 359 tweets	Positive Negative	Naïve Bayes Maximum Entropy SVM		Accuracy = 81% Accuracy = 80% Accuracy = 82%
4	Sailunaz, Kashfia, and Reda Alhajj. "Emotion and sentiment analysis from Twitter text." <i>Journal of Computational Science</i> 36 (2019): 101003.	7 246 tweets	Positive Negative Neutral	Naïve Bayes SVM Random Forest	3-fold, 5-fold, 10-fold cross validation	Accuracy = 66% Accuracy = 23% Accuracy = 55%
5	Kharde, Vishal, and Prof Sonawane. "Sentiment analysis of twitter data: a survey of techniques." <i>arXiv preprint arXiv:1601.06971</i> (2016).	90 000 tweets	Positive Negative	Naïve Bayes Maximum Entropy SVM		Accuracy = 74% Accuracy = 75% Accuracy = 76%
6	Arias, Marta, Argimiro Arratia, and Ramon Xuriguera. 2013. "Forecasting with Twitter Data." <i>ACM Transactions on Intelligent Systems and Technology</i> 5(1):1–24.	20 000 tweets	Positive Negative	SVM, Neural Networks		Accuracy = 45,98%, Accuracy = 74,07%
7	Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. "Twitter Mood Predicts the Stock Market." <i>Journal of Computational Science</i> 2(1):1–8	9 853 498 tweets	Positive Negative	A Granger causality analysis, Self-Organizing Fuzzy Neural Network		Accuracy = 87.6%
8	Zhou, Xujuan, Xiaohui Tao, Jianming Yong, and Zhenyu Yang. 2013. "Sentiment Analysis on Tweets for Social Events." <i>Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2013</i> (April 2010):557–62.	57 000 tweets	Positive Negative	Lexicon-based sentiment Analysis intelligent system		Accuracy
9	Stieglitz, Stefan and Linh Dang-Xuan. 2012. "Political Communication and Influence through Microblogging - An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior." <i>Proceedings of the Annual Hawaii International Conference on System Sciences</i> 3500–3509	64,431 tweets	Positive Negative	Linguistic inquiry and word count (LIWC) software, Poisson regression		The number of times the tweet has been retweeted, LIWC categories.

► Notre modèle présente des résultats proches de ceux publiés.



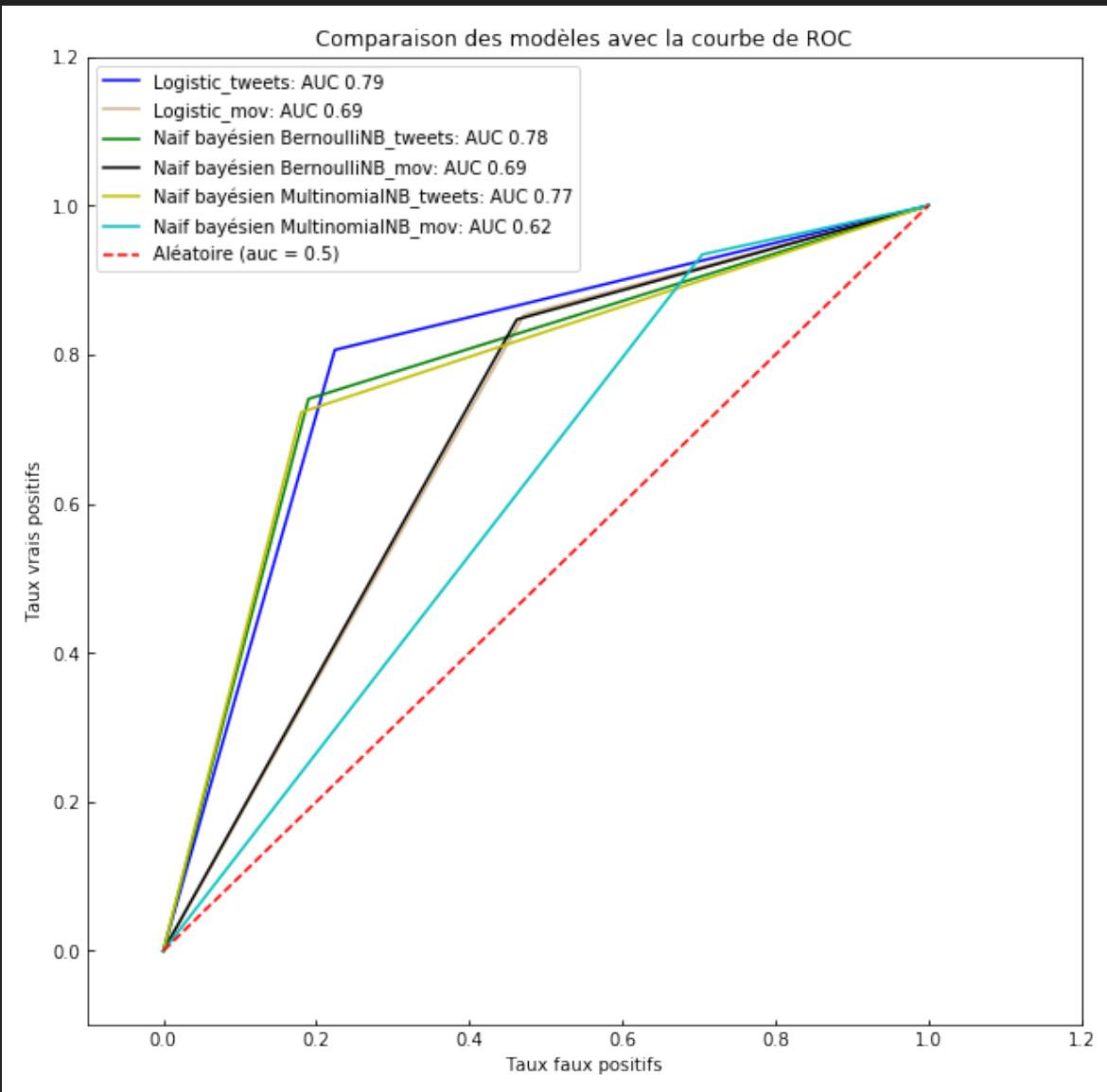
- ▶ Problématique
- ▶ Exploration des données
- ▶ Visualisation des données
- ▶ Itération: construction du modèle et évaluation
- ▶ Perspectives
- ▶ Conclusion

SCALABILITÉ DU MODÈLE



DataScientest • com

▶ Evaluation de la « scalabilité » du modèle sur le jeu de données: Movie Reviews



	Accuracy
1. Régression logistique avec Vectorizer	0.69
2. Naïf bayésien BernoulliNB	0.69
3. Naïf bayésien MultinomialNB	0.62
4. K plus proches voisins	0.54
5. Arbre de décision	0.52
6. Forêt aléatoire	0.58
7. XGBOOST	0.6
8. Gradient Boosting	0.61
9. Adaboost	0.61

- ▶ les courbes de ROC donnent un aperçu graphique des résultats obtenus sur les scores avec des aires sous la courbe d'environ 0.7 sur le jeu de données des Movie Reviews.
- ▶ Bien que ces résultats sont inférieurs à ceux obtenus avec le jeu de données des tweets et à l'état de l'art (aux alentours des 80%), ils sont néanmoins assez satisfaisants en général vu que le modèle n'a pas été entraîné sur le même dataset.

SCALABILITÉ DU MODÈLE

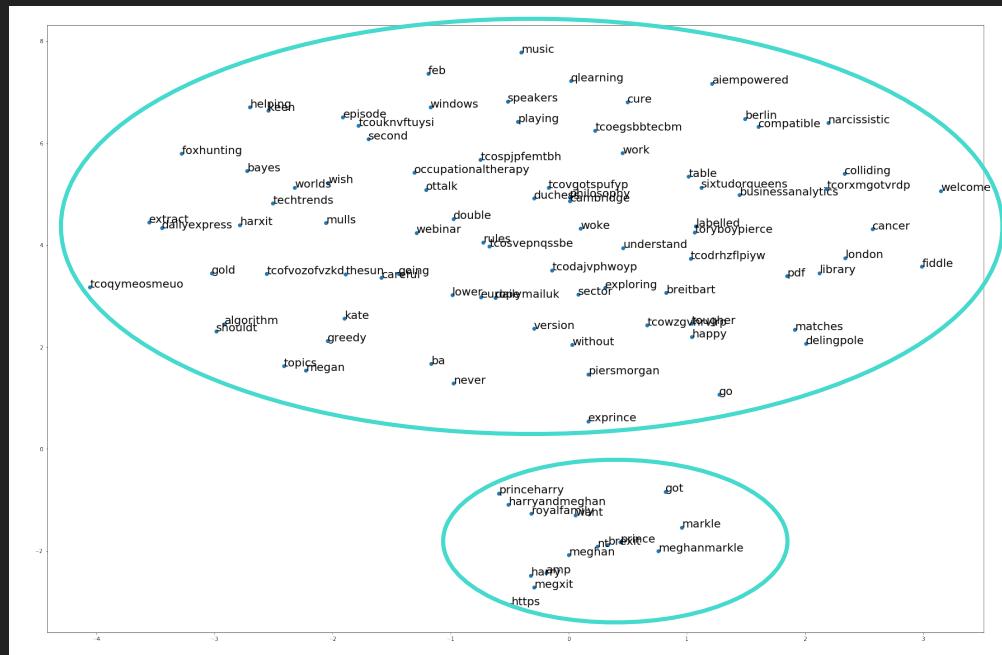


DataScientest • com

- ▶ Evaluation de la « scalabilité » du modèle sur le jeu de données: Trump tweets
- ▶ Avant élection 2015-2016 (pendant la campagne présidentielle)
 - ▶ Sur 7375 tweets de Donald Trump avant son élection à la maison blanche en novembre 2016, il y a 5202 tweets positifs et 2173 tweets négatifs. 70.53% des tweets sont positifs selon les prédictions de notre modèle.
- ▶ Après élection 2017-2018
 - ▶ Sur 3855 tweets de Donald Trump après son élection à la maison blanche, il y a 2715 tweets positifs et 1140 tweets négatifs. 70.42% des tweets sont positifs selon les prédictions de notre modèle.
- ▶ Après élection 2018-2019
 - ▶ Sur 7031 tweets de Donald Trump après son élection à la maison blanche, il y a 4842 tweets positifs et 2189 tweets négatifs. 68.86% des tweets sont positifs selon les prédictions de notre modèle.
- ▶ Après élection 2016-2020 (depuis son élection à aujourd’hui)
 - ▶ Sur 12420 tweets de Donald Trump après son élection à la maison blanche, il y a 8504 tweets positifs et 3916 tweets négatifs. 68.47% des tweets sont positifs selon les prédictions de notre modèle.
- ▶ Pour conclure, on peut dire que finalement l'humeur de Trump n'a pas beaucoup changé depuis son élection. Il reste donc fidèle à lui-même!

WEBSRAPING

- ▶ Scraping + Word2Vec des tweets comprenant: #Megxit et #DataScience



- ▶ Ce graphe fait apparaître deux clusters dont: le cluster en bas qui représente des mots relatifs au **Megxit** (`princeharry`, `harryandmeghan`, `royalfamily`, `meghan`, `megxit` etc).

- ## ► Prédiction de la nature des tweets

- ▶ Sur 400 tweets de #Megxit et #DataScience, il y a 280 tweets positifs et 120 tweets négatifs; 70.0% tweets sont **positifs** (selon les prédictions de notre modèle).
 - ▶ Le tweet: «This coastal Canadian Town could be Harry and Meghan Markle's new home.... #Megxit » est **positif**.
 - ▶ Le tweet: «How can Hollywood or netflix welcome them,no one wants 2 make a movie abt an American grifter,who didn't invite her dad 2 her publicized wedding #Megxit » est **négatif**.



- ▶ Problématique
- ▶ Exploration des données
- ▶ Visualisation des données
- ▶ Itération: construction du modèle et évaluation
- ▶ Perspectives
- ▶ Conclusion

CONCLUSION



DataScientest • com

- ▶ Comprendre la **problématique** posée par le sujet
- ▶ Construire un **modèle** satisfaisant avec des résultats proches de ceux de l'état de l'art
- ▶ Evaluer le modèle retenu avec plusieurs **métriques** de mesure de performance
- ▶ Evaluer le modèle en testant sa « **scalabilité** » sur d'autres jeux de données
- ▶ Aller plus loin avec le **webscraping** et le **word2vec**
- ▶ Améliorer les performances obtenues avec du **Deep Learning**, **SVM** etc.

A black and white photograph showing two wind turbines against a dramatic, cloudy sky. The turbines are white and have three blades each. The one on the left is larger and more prominent, while the one on the right is smaller and partially visible. The clouds are thick and varied in texture.

MERCI DE VOTRE
ATTENTION