Analyse des Sentiments sur Twitter

Membres du groupe :

SADDEDINE Sabrina

BEN NASR Oumayma

BOULARES Nour

1. Introduction

L'analyse des sentiments est une approche permettant d'extraire et d'interpréter les émotions exprimées dans des textes. Ce projet se concentre sur l'analyse des sentiments des utilisateurs de Twitter en exploitant un ensemble de données dédié. L'objectif est d'identifier les tendances émotionnelles et de développer des modèles de classification pour prédire le sentiment des tweets.

2. Préparation et Nettoyage des Données

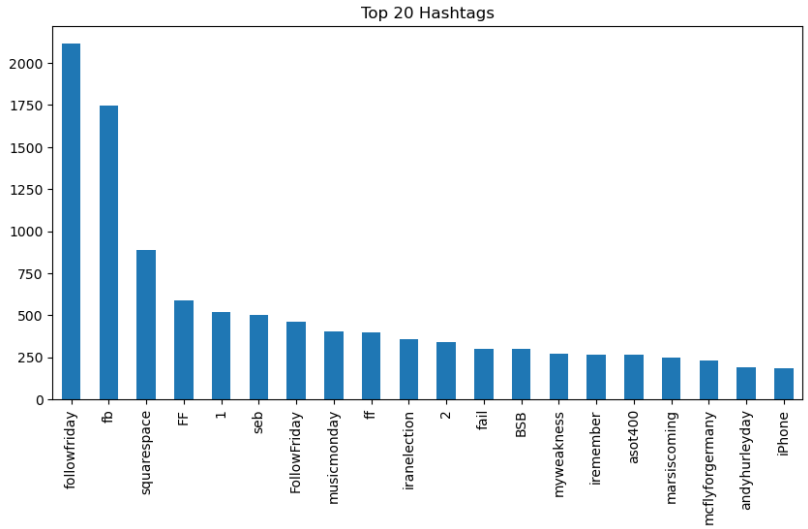
Nous avons procédé aux étapes suivantes pour préparer et nettoyer les données :

- Attribution de noms explicites aux colonnes.
- Suppression des valeurs manquantes (NA).
- Nettoyage des tweets en supprimant les hashtags, mentions (@), URLs et en normalisant la casse.
- Extraction et transformation des dates et heures.
- Création d'une colonne indiquant le moment de la journée (matin, midi, après-midi, soir, nuit).

3. Analyse et Visualisation des Données

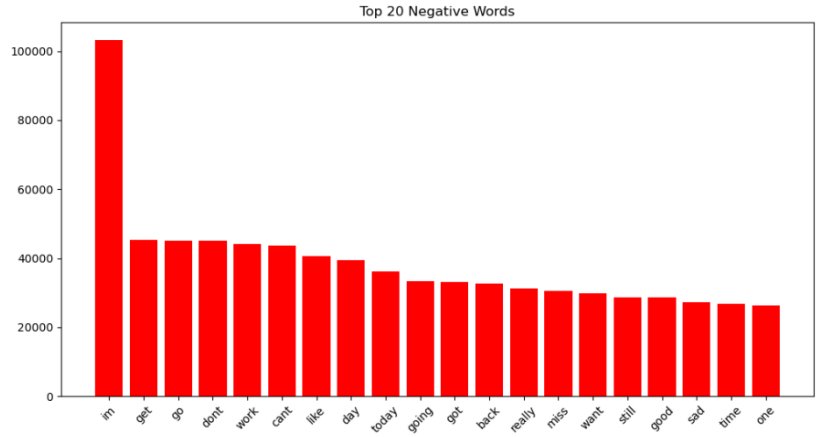
Afin de mieux comprendre la distribution des sentiments, plusieurs analyses graphiques ont été réalisées :

- **Visualisation des hashtags les plus utilisés** :



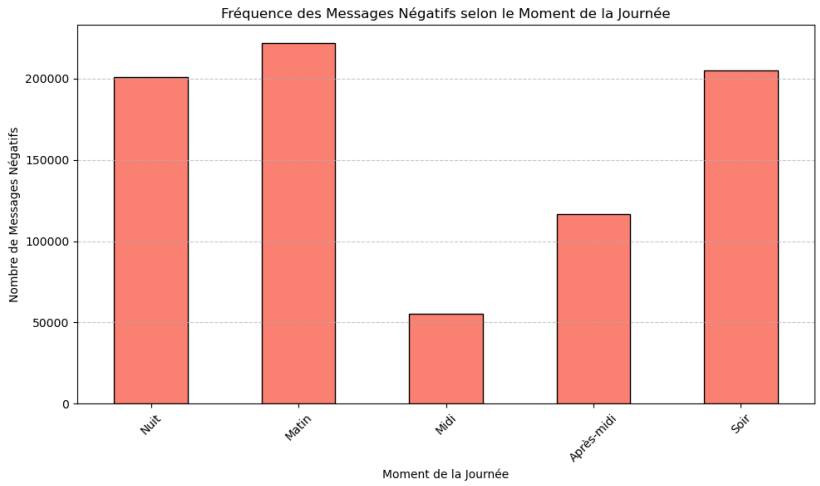
Les 20 hashtags les plus fréquents ont été identifiés, avec en tête #followfriday (>2000 occurrences) et #fb (1750 occurrences).

- **Analyse des mots les plus fréquents dans les tweets négatifs** :



Après suppression des mots vides (stopwords), il apparaît que le mot "im" est le plus récurrent, ce qui s'explique par l'expression personnelle fréquente sur Twitter.

- **Répartition des tweets négatifs selon les moments de la journée** :



- Une forte présence de tweets négatifs est observée entre le soir et le matin, atteignant près de 200 000 messages, avec une baisse significative durant la journée.

4. Modélisation

Nous avons testé deux modèles de classification pour prédire les sentiments des tweets :

- **Régression Logistique**
- **Modèle Naïf Bayésien (MultinomialNB)**

4.1 Prétraitement des Données

Les tweets ont été convertis en vecteurs numériques via **TfidfVectorizer**, puis divisés en ensembles d'entraînement (80%) et de test (20%).

4.2 Entraînement et Évaluation des Modèles

Les modèles ont été évalués selon plusieurs métriques, notamment l'exactitude, la précision, le rappel et le score F1.

Résultats du modèle Naïf Bayésien

```
Accuracy: 0.757675
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.77      0.76    159494
           1       0.76      0.75      0.76    160506

    accuracy                           0.76    320000
   macro avg       0.76      0.76      0.76    320000
weighted avg       0.76      0.76      0.76    320000
```

Interprétation des résultats

Le modèle atteint une exactitude de **75,77 %**, ce qui signifie qu'il classe correctement environ **76 % des instances** dans la bonne catégorie.

Précision (Précision) : La précision pour la classe 0 est de 75 %, ce qui signifie que parmi toutes les prédictions faites pour cette classe, 75 % étaient correctes. Pour la classe 1, la précision est de 76 %.

Rappel (Recall) : Le rappel pour la classe 0 est de 77 %, ce qui signifie que 77 % des instances réellement appartenant à cette classe ont été correctement identifiées par le modèle. Pour la classe 1, ce taux est de 75 %.

F1-score : Le score F1, qui est la moyenne harmonique entre la précision et le rappel, est équilibré à 76 % pour les deux classes, ce qui montre une bonne cohérence du modèle.

Le modèle Naïve Bayes présente une performance satisfaisante Il démontre un bon équilibre entre précision et rappel pour les deux classes

Résultats du modèle de Régression Logistique

```
Accuracy: 0.772803125
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.75      0.77    159494
           1       0.76      0.80      0.78    160506

    accuracy                           0.77    320000
   macro avg       0.77      0.77      0.77    320000
weighted avg       0.77      0.77      0.77    320000
```

Interprétation des résultats

Le modèle atteint une exactitude de 77,28 %, ce qui signifie qu'il classe correctement environ 77 % des instances.

Précision (Précision) : La précision pour la classe 0 est de 79 %, ce qui signifie que parmi toutes les prédictions faites pour cette classe, 79 % étaient correctes. Pour la classe 1, la précision est de 76 %.

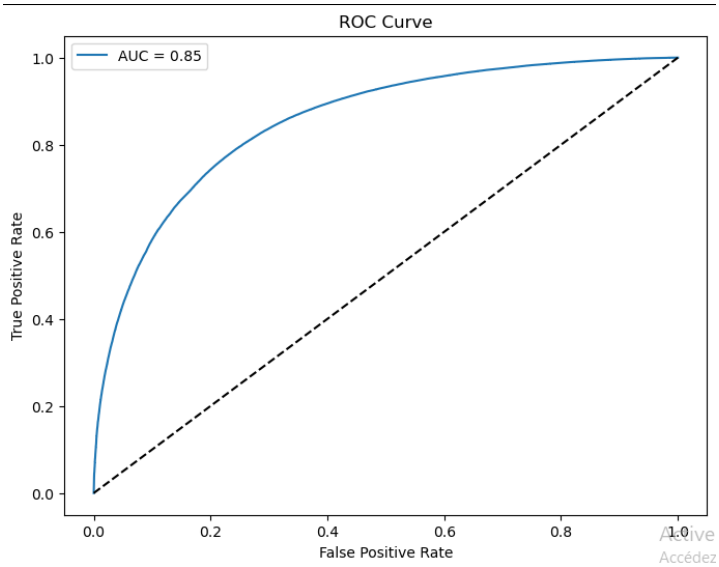
Rappel (Recall) : Le rappel pour la classe 0 est de 75 %, ce qui signifie que 75 % des instances réellement appartenant à cette classe ont été correctement identifiées par le modèle. Pour la classe 1, ce taux est de 80 %, indiquant que le modèle a bien capturé cette classe.

F1-score : Le score F1, qui combine précision et rappel, est de 77 % pour la classe 0 et de 78 % pour la classe 1, ce qui indique un bon équilibre global entre ces deux métriques.

Le modèle **de régression logistique** offre des performances solides avec une **exactitude de 77,28 %**. Comparé au modèle **Naïve Bayes**, il présente une légère amélioration, notamment dans la détection de la classe **1** avec un meilleur rappel (**80 %** contre **75 %** pour Naïve Bayes).

Ces résultats suggèrent que la régression logistique pourrait être plus adaptée à ce problème de classification.nous avons donc retenu LA RÉGRESSION

4.3 Analyse de la Courbe ROC



Le modèle de régression logistique a été évalué via la courbe ROC, qui mesure le compromis entre le taux de vrais positifs (TPR) et le taux de faux positifs (FPR). L'aire sous la courbe (AUC) atteint **0,85**, indiquant une bonne capacité de distinction entre les classes. Cela signifie que 85 % du temps, le modèle attribue une probabilité plus élevée aux échantillons positifs (classe 1) qu'aux négatifs (classe 0).

5. Interprétation des Résultats

5.1 Analyse des Coefficients de la Régression Logistique

L'étude des coefficients et des odds ratios permet d'identifier les termes les plus influents dans la classification des sentiments.

Termes fortement associés aux sentiments positifs

- "smile", "thank", "congratulations", "proud", "welcome", "thanks", "thx" : ces mots augmentent significativement la probabilité qu'un tweet soit classé comme positif.
- "wait" peut également apparaître dans des contextes positifs, comme "Can't wait to see you!".

Top 10 des mots associés à un sentiment positif:

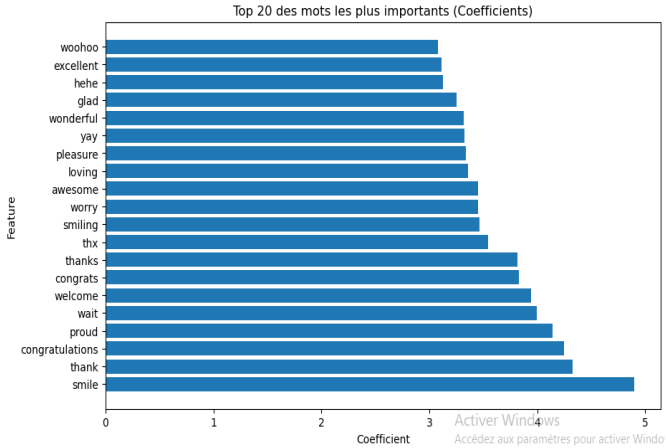
	Feature	Coefficient	Odds Ratio
3921	smile	4.899035	134.160309
4327	thank	4.327494	75.754216
889	congratulations	4.245582	69.796400
3369	proud	4.145059	63.121336
4708	wait	3.995633	54.360265
4777	welcome	3.945176	51.685430
888	congrats	3.827130	45.930537
4330	thanks	3.815382	45.394079
4387	thx	3.546313	34.685205
3924	smiling	3.463523	31.929268

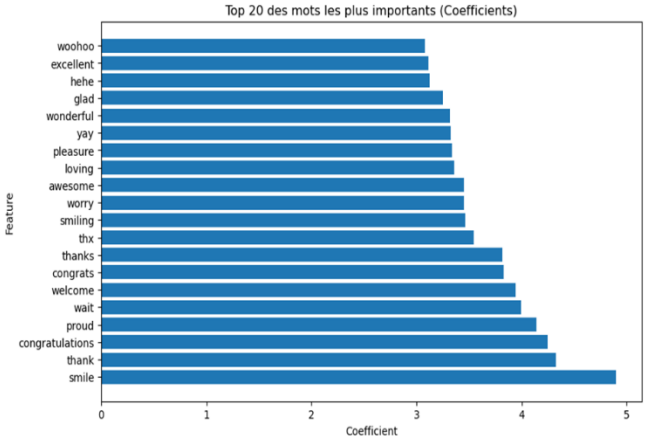
Top 10 des mots associés à un sentiment négatif:

	Feature	Coefficient	Odds Ratio
1172	disappointed	-5.728187	0.003253
598	bummed	-5.758645	0.003155
2756	miss	-5.809430	0.002999
3587	rip	-5.956921	0.002588
1152	died	-6.059894	0.002335
2109	hurts	-6.185727	0.002059
3265	poor	-6.252141	0.001926
4605	unfortunately	-6.644256	0.001301
3652	sadly	-7.064816	0.000855
3650	sad	-9.567591	0.000070

Termes fortement associés aux sentiments négatifs

- "sad" est le mot le plus fortement corrélé à un sentiment négatif.
- "sadly", "unfortunately", "poor" traduisent des regrets ou des déceptions.
- "died", "rip", "miss", "hurts" indiquent des émotions négatives liées à la perte et à la souffrance.
- "bummed" et "disappointed" reflètent des sentiments de frustration ou de tristesse.





Conclusion

L'analyse des sentiments sur Twitter a permis d'identifier des tendances claires et de développer des modèles de classification performants. La régression logistique s’est révélée être le modèle le plus performant avec une AUC de 0,85. Ces résultats ouvrent la voie à des applications avancées, notamment la détection en temps réel des tendances émotionnelles sur les réseaux sociaux.