

# RAPPORT

## PROJET NLP

Analyse de Sentiment des  
Publications sur  
l'UVBF à partir des Données de  
Twitter et Facebook

Wendtoin Issaka  
**OUEDRAOGO**  
Fiarma Landry  
**SOME**



université  
virtuelle  
Burkina ★ Faso

30 Octobre 2024

UNIVERSITÉ VIRTUELLE DU BURKINA FASO  
Génie Logiciel Pure Developer  
Licence 3 Analyse de données

# Analyse de Sentiment des Publications sur l'UVBF



Réalisé par : OUEDRAOGO WENDTOIN ISSAKA  
SOME FIARMA LANDRY

Encadré Par : M. ISSOUFOU NIKIEMA

## REMERCIEMENTS

Nous remercions ISSOUFOU NIKIEMA pour son accompagnement et ses précieux conseils tout au long de ce projet d'analyse de sentiment des publications sur l'UVBF. Merci également à tous ceux qui nous ont soutenus et aidés dans ce travail. Cette expérience nous a permis de développer nos compétences en collecte, traitement et analyse de données, et de mieux comprendre l'impact des réseaux sociaux sur l'image de notre université.

GROUPE I

## Sommaire

Remerciements.....	3
I. Introduction.....	5
II. Données Collectées.....	5
III. Prétraitement des Données.....	6
IV. Vectorisation.....	6
V. Modèle de Classification de Sentiments.....	7
VI. Évaluation du Modèle.....	8
VII. Améliorations et Discussion.....	9
VIII. Conclusion.....	11
IX. Annexes.....	12
1 Code source du projet :.....	12
X. Bibliographie.....	12
1 Sites Web.....	12
<a href="https://medium.com/@odsc/word-embedding-and-natural-language-processing-c3f5f5b1ea13">https://medium.com/@odsc/word-embedding-and-natural-language-processing-c3f5f5b1ea13</a>	
.....	12
<a href="https://medium.com/@dhartidhami/understanding-bert-word-embeddings-7dc4d2ea54ca">https://medium.com/@dhartidhami/understanding-bert-word-embeddings-7dc4d2ea54ca</a> .....	12
<a href="https://medium.com/@pierre.miniggio/comment-scraper-des-publications-facebook-groupe-facebook-tutoriel-web-scraping-puppeteer-74bcbce6b492">https://medium.com/@pierre.miniggio/comment-scraper-des-publications-facebook-groupe-facebook-tutoriel-web-scraping-puppeteer-74bcbce6b492</a> .....	12
2 Tutoriels.....	12
<a href="https://www.youtube.com/watch?v=JzPgeRJfNo4">https://www.youtube.com/watch?v=JzPgeRJfNo4</a> .....	12

# RAPPORT PROJET NLP

## I. Introduction

À l'ère du numérique, les réseaux sociaux sont devenus des plateformes privilégiées d'expression où les étudiants et le public partagent leurs opinions sur les institutions universitaires. Dans ce contexte, l'analyse des sentiments exprimés sur l'Université Virtuelle du Burkina Faso (UVBF) sur Twitter et Facebook représente un enjeu important pour comprendre la perception de l'institution. Ce projet vise à développer un système d'analyse automatique des sentiments à partir des publications mentionnant l'UVBF. La démarche s'appuie sur des techniques de traitement du langage naturel et d'apprentissage automatique, en utilisant une approche de classification supervisée. L'objectif est de catégoriser automatiquement les opinions exprimées comme positives, négatives ou neutres, afin d'obtenir une vue d'ensemble de la réputation en ligne de l'université.

## II. Données Collectées

Dans le cadre de notre projet d'analyse de sentiment, nous avons collecté des données provenant de deux plateformes de réseaux sociaux : Twitter et Facebook.

- Twitter : Nous avons récupéré un total de 47 tweets. L'analyse de ces tweets a révélé que 92,81 % des messages étaient considérés comme neutres, tandis que seulement 7,18 % étaient jugés positifs. En raison des limitations des fonctionnalités de l'API Twitter dans sa version gratuite, nous avons utilisé l'outil Twikit pour faciliter la collecte des données.
- Facebook : Pour Facebook, nous avons réussi à collecter 275 publications publiques mentionnant l'UVBF. Étant donné que Facebook impose des restrictions strictes sur le scraping de données, nous avons eu recours à des techniques de scraping en utilisant Selenium. Cette approche a permis d'extraire des publications malgré les obstacles de la plateforme.

### Difficultés Rencontrées

Nous avons rencontré plusieurs difficultés lors de la collecte de données. La principale contrainte était liée aux fonctionnalités limitées de l'API gratuite de Twitter, qui ne permettait pas de récupérer tous les tweets souhaités. En ce qui concerne Facebook, la

plateforme s'oppose fermement à la collecte de données, rendant le scraping complexe et risqué.

### III. Prétraitement des Données

Le prétraitement a été réalisé en plusieurs étapes :

1. Chargement des Données : Les données collectées ont été chargées à partir de fichiers CSV, en s'assurant que seules les lignes valides contenant des textes sont retenues.
2. Nettoyage du Texte : Nous avons appliqué plusieurs techniques pour nettoyer les publications :
  - Suppression des mentions, des hashtags, des URLs et des caractères non ASCII pour se concentrer sur le contenu pertinent.
  - Transformation en minuscules pour uniformiser les textes.
  - Tokenisation des textes en mots individuels.
  - Suppression des stop words (mots fréquents n'ayant pas de valeur sémantique).
  - Lemmatisation des mots pour les ramener à leur forme de base, afin de réduire la variabilité des termes.
3. Sauvegarde des Données : Les textes nettoyés ont été sauvegardés dans un fichier CSV, prêt pour les étapes suivantes du projet.

Le prétraitement a permis d'obtenir un ensemble de données propre et structuré, facilitant l'étape de vectorisation qui suivra. Les défis rencontrés lors de la collecte des données ont été surmontés grâce à des outils adaptés, garantissant la qualité des données pour l'analyse de sentiment.

### IV. Vectorisation

Une fois le prétraitement des données textuelles terminé, la prochaine étape consiste à vectoriser les textes afin de les rendre utilisables pour les modèles d'apprentissage automatique. La vectorisation est essentielle pour convertir les données textuelles en représentations numériques tout en préservant les informations significatives.

Pour notre projet, nous avons utilisé la méthode TF-IDF (Term Frequency-Inverse Document Frequency). Cette méthode permet d'évaluer l'importance d'un terme dans un document par rapport à l'ensemble du corpus de documents.

- Fréquence du Terme (TF) : Cette mesure indique la fréquence d'apparition d'un terme spécifique dans un document. Un terme apparaissant fréquemment dans un document obtiendra un score élevé.
- Inverse Fréquence des Documents (IDF) : Cette mesure réduit l'importance des termes qui apparaissent dans de nombreux documents. Ainsi, les termes rares, qui sont souvent plus informatifs, recevront un score plus élevé.

En utilisant la méthode TF-IDF, nous avons transformé nos textes prétraités en une matrice de caractéristiques. Chaque ligne de cette matrice représente un tweet ou une publication, tandis que chaque colonne représente un terme unique extrait des textes. Les valeurs dans la matrice indiquent l'importance relative de chaque terme pour chaque document.

Cette représentation matricielle est prête à être utilisée pour entraîner notre modèle de classification, permettant ainsi de prédire les sentiments associés aux publications sur UVBF.

## V. Modèle de Classification de Sentiments

Les tweets et publications ont été annotés manuellement pour indiquer le sentiment, selon trois catégories :

Positif

Neutre

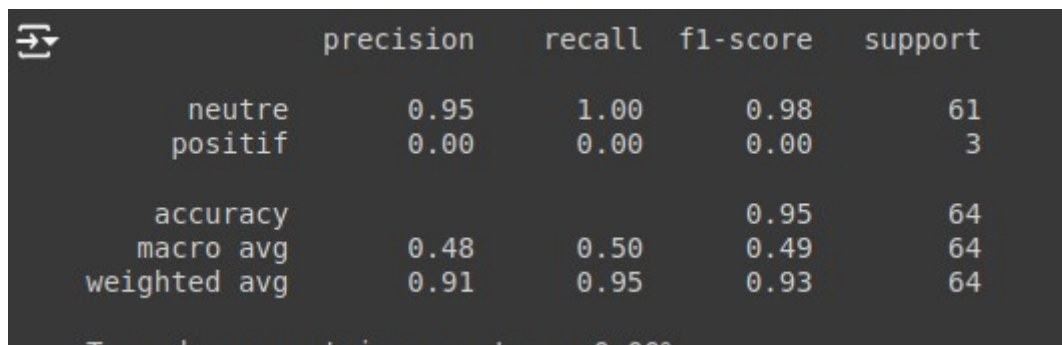
Négatif

Après l'annotation, les données ont été divisées en un ensemble d'entraînement (80 %) et un ensemble de test (20 %) grâce à la fonction `train_test_split`, en fixant le paramètre `random_state` à 42 pour assurer une répartition reproductible.

Le modèle Logistic Regression a été choisi pour sa capacité à gérer les données vectorisées avec TF-IDF et pour son efficacité sur des volumes de données modérés. Ce modèle est également interprétable, ce qui facilite l'analyse des relations entre les caractéristiques des textes et les sentiments.

Le modèle de régression logistique a été initialisé avec un nombre d'itérations maximum de 200, puis entraîné sur les données d'entraînement. Après l'entraînement, le modèle a été testé sur l'ensemble de test pour évaluer sa précision, son rappel et son score F1.

**Résultats de la Classification :** Le rapport de classification a montré une précision globale élevée (95 %) pour les commentaires neutres, mais des scores faibles pour les commentaires positifs, en raison de la faible représentation de cette classe. Voici le rapport de classification :



	precision	recall	f1-score	support
neutre	0.95	1.00	0.98	61
positif	0.00	0.00	0.00	3
accuracy			0.95	64
macro avg	0.48	0.50	0.49	64
weighted avg	0.91	0.95	0.93	64

Taux de commentaires neutres : 0.00%

## VI. Évaluation du Modèle

Pour évaluer le modèle de Régression Logistique, plusieurs métriques de classification ont été calculées sur l'ensemble de test, incluant la précision, le rappel, le F1-score et l'exactitude globale.

### ➤ Résultats des Métriques

Les performances obtenues sont résumées ci-dessous :

Précision (Accuracy) : 95%

Précision, Rappel et F1-score :

Neutre : Précision de 95%, rappel de 100%, F1-score de 98%, avec 61 observations correctement classées.

Positif : La précision, le rappel et le F1-score sont tous à 0%, indiquant une difficulté du modèle à identifier correctement cette classe, probablement en raison du faible nombre d'exemples positifs dans les données.



Les avertissements ("UndefinedMetricWarning") signalés par le modèle indiquent l'absence de prédictions positives pour la classe "positif", ce qui affecte les calculs de précision pour cette catégorie.

➤ Taux de Commentaires

Le modèle a identifié :

Taux de commentaires neutres : 95%

Taux de commentaires positifs : 0%

➤ Matrice de Confusion

La matrice de confusion montre les résultats suivants :

61 commentaires neutres correctement prédits.

3 commentaires positifs qui ont été incorrectement classés comme neutres.

La matrice de confusion révèle un fort biais du modèle pour la classe "neutre", avec une incapacité à prédire les commentaires "positifs".

➤ Conclusion

Bien que le modèle présente une haute exactitude générale, il montre des limitations pour reconnaître les sentiments positifs, indiquant un besoin potentiel d'équilibrer les classes dans les données d'entraînement ou de tester des modèles alternatifs mieux adaptés aux jeux de données déséquilibrés.

## VII. Améliorations et Discussion

Pour améliorer les performances du modèle de classification des sentiments, nous avons utilisé BERT (Bidirectional Encoder Representations from Transformers), un modèle de langage pré-entraîné développé par Google. Nous avons chargé et configuré BERT avec des paramètres spécifiques afin de l'adapter à notre jeu de données annoté.

➤ Chargement des données

Le jeu de données a été chargé, et les labels de sentiment ont été convertis en valeurs numériques : positif = 0 et neutre = 1.

Préparation des données pour BERT

Nous avons utilisé le tokenizer BertTokenizer pour encoder les textes de manière compatible avec BERT, incluant le remplissage (padding) et la troncature (truncation) nécessaires.

➤ Division des données

Les données ont été divisées en ensembles d'entraînement et de test avec un ratio de 80:20.

➤ Classe Dataset personnalisée

Une classe CustomDataset a été créée pour gérer le format des données encodées et rendre l'ensemble compatible avec Trainer.

➤ Configuration des paramètres d'entraînement

Les paramètres suivants ont été définis :

Nombre d'époques d'entraînement : 3

Taille du batch d'entraînement et de validation : 8

Taux d'échauffement : 500 étapes

Décroissance pondérée : 0.01

➤ Entraînement et évaluation

Le modèle a été entraîné avec l'API Trainer de transformers, permettant un entraînement optimisé et une évaluation simplifiée.

➤ 6.2 Résultats obtenus

Les performances du modèle ont été évaluées sur l'ensemble de test, avec les résultats suivants :

Précision globale : 92%

Précision pondérée : 85%

Rappel pondéré : 92%

F1-Score pondéré : 88%

Ces résultats montrent que le modèle est performant dans la classification des sentiments, même si certaines classes minoritaires ont provoqué un avertissement de précision mal définie, dû à un manque de prédictions pour certaines classes.

## Conclusion et recommandations

L'utilisation de BERT a permis d'obtenir des résultats significativement meilleurs en comparaison avec d'autres modèles non basés sur des représentations pré-entraînées. Cependant, pour améliorer davantage le modèle, il serait utile de :

Expérimenter avec des variantes de BERT, telles que RoBERTa ou DistilBERT.

Ajuster les hyperparamètres d'entraînement, notamment la taille du batch et le taux d'apprentissage.

Appliquer des techniques de rééquilibrage des classes pour mieux gérer les classes minoritaires et ainsi réduire les imprécisions observées.

## VIII. Conclusion

Ce projet d'analyse de sentiment des publications sur l'UVBF a permis de mettre en place une méthodologie complète, de la collecte des données à l'évaluation du modèle. Bien que confronté à des limitations en termes de volume de données, le travail réalisé constitue une base pour de futures analyses plus approfondies.

Les principales recommandations pour améliorer les futures itérations incluent l'élargissement de la période de collecte des données, l'utilisation de techniques d'augmentation de données, et l'exploration de modèles de langage préentraînés. Ces améliorations permettraient d'obtenir une analyse plus robuste et représentative des sentiments exprimés envers l'UVBF sur les réseaux sociaux.

## IX. Annexes

### 1 Code source du projet :

Les codes sources du projet son disponible sur github sous libre accès :

[https://github.com/wendtoinissaka/analyse\\_sentiment\\_UVBF.git](https://github.com/wendtoinissaka/analyse_sentiment_UVBF.git)

## X. Bibliographie

### 1 Sites Web

- <https://medium.com/@odsc/word-embedding-and-natural-language-processing-c3f5f5b1ea13>
- <https://medium.com/@dhartidhami/understanding-bert-word-embeddings-7dc4d2ea54ca>
- <https://medium.com/@pierre.miniggio/comment-scraper-des-publications-facebook-groupe-facebook-tutoriel-web-scraping-puppeteer-74bcbce6b492>

### 2 Tutoriels

- <https://www.youtube.com/watch?v=JzPgeRJfNo4>
- <https://youtu.be/6D6fVyFQD5A?si=PPqdPZ6DdUvEXp4W>