### **REPUBLIQUE DU BENIN**

### **\*\*\*\*\*\*\*\*\***

### **UNIVERSITE D’ABOMEY-CALAVI**

### **INSTITUT DE FORMATION ET DE RECHERCHE INFORMATIQUE**

### 

### **Matière: Traitement du langage Naturel**

**GROUPE 4**

**Thème: Détecteur automatique de fake news**

### **Professeur:** Gracieux HOUNNA

### **Membres:**

### **OKE Michel**

### 

### **LAOUROU Jesukristidujuoyinlo**

### 

### **RADJI Anlim Ishola Adéwalé**

### **ADJIBADE Sadiyath**

### **TEHOUENOU Gédéon**

### **Année scolaire: 2024-2025**

### Table des matières

### I – Introduction

### II – Objectifs du projet

### III – Méthodologie et analyse des résultats

### Collecte des données Prétraitement des Données Modèles utilisés Entrainement et évaluation

### Visualisation avec wandb

### IV – Conclusion

### V – Références

### 

### 

### 

### 

### 

### 

### 

### 

### **Introduction**

Avec l’essor des réseaux sociaux et des plateformes numériques, la propagation des fausses informations (**Fake News**) est devenue un défi majeur. Ces fausses nouvelles peuvent avoir des conséquences importantes sur l’opinion publique, la politique et la santé.

Dans ce projet, nous développons un modèle de **Machine Learning et de Traitement du Langage Naturel (NLP)** pour détecter automatiquement les fake news.

**Objectifs du projet**

Les objectifs de ce projet sont les suivants :

* Développer un modèle de machine learning pour classer les articles comme faux ou vrais (fake news ou non).
* Analyser les différentes caractéristiques du texte (comme les mots-clés, le ton, etc.) pour détecter les fausses informations.
* Évaluer la performance du modèle en termes de précision, rappel et F1-score.
* Créer une interface utilisateur simple où les utilisateurs peuvent soumettre un article et obtenir une évaluation du modèle.

**Méthodologie**

## **1. Collecte des données**

### **Sources et jeux de données**

* **Dataset général :** Pour le projet, nous avons utilisé le dataset « Fake and Real News » disponible sur Kaggle. Ce jeu de données comprend environ 40000 articles, répartis entre articles « fake » et « real », et couvre une large période avec des articles issus de divers domaines.
* **Dataset COVID-19 :** Un second jeu de données, axé sur les fake news relatives à la pandémie de COVID-19, a été fourni dans le cadre d’une compétition sur Codalab. Ce dataset permet d’aborder des exemples spécifiques à la crise sanitaire et d’enrichir l’analyse.

### **Justification du choix des données**

Ces jeux de données ont été sélectionnés pour leur diversité et leur représentativité. Le premier offre une vue d’ensemble sur la désinformation, tandis que le second permet de tester le modèle sur des cas récents et thématisés, garantissant ainsi une analyse robuste et pertinente.

## **2. Prétraitement des données**

### **Nettoyage et normalisation**

* **Suppression des éléments non pertinents :** Les URL, mentions, hashtags et caractères spéciaux sont retirés afin de ne conserver que le contenu textuel essentiel.
* **Gestion des valeurs manquantes et des doublons :** Une vérification approfondie a permis de confirmer l’absence de doublons et de valeurs manquantes, garantissant ainsi la qualité du dataset.

### **Transformation du texte**

* **Conversion en minuscules :** Tous les textes sont convertis en minuscules pour uniformiser l’ensemble des données.
* **Tokenization :** Le texte est découpé en tokens (mots ou sous-mots) à l’aide de NLTK afin de faciliter l’analyse.
* **Suppression des stopwords :** Les mots courants (ex. « le », « la », « et », etc.) qui n’apportent pas de valeur discriminante sont supprimés.
* **Lemmatisation :** Chaque token est ramené à sa forme de base pour améliorer la cohérence et la performance lors de l’analyse textuelle.

### **Vectorisation**

Pour le modèle XGBoost, le texte prétraité est transformé en vecteurs numériques à l’aide d’un TF-IDF Vectorizer. Cette étape permet de quantifier l’importance des termes dans chaque document en fonction de leur fréquence dans le corpus.

## **3. Modèles utilisés**

### **XGBoost**

* **Principe :** XGBoost (eXtreme Gradient Boosting) est un algorithme de boosting qui combine plusieurs arbres de décision pour optimiser la performance prédictive.
* **Pipeline :** Les vecteurs issus du TF-IDF servent d’entrée à l’algorithme. L’entraînement se fait avec un ajustement fin des hyperparamètres (tels que learning\_rate, max\_depth, et n\_estimators) à l’aide d’un GridSearchCV.
* **Avantages :** Rapidité et simplicité d’implémentation, ce qui en fait une solution efficace dans des environnements contraints en ressources.

### **BERT**

* **Architecture :** BERT (Bidirectional Encoder Representations from Transformers) est un modèle de langage pré-entraîné qui utilise l’architecture Transformer et un mécanisme d’attention pour capturer les relations contextuelles entre les mots.
* **Adaptation à la classification :** Pour la détection de fake news, une couche de classification (linear layer) est ajoutée sur le token [CLS] pour effectuer une classification binaire (fake vs. real).
* **Détails du fine-tuning :** La version « bert-base-uncased » est utilisée et fine-tunée sur le jeu de données avec :
  + Nombre d’époques : 3
  + Taille du batch pour l’entrainement : 8
  + Taille du batch pour l’évaluation : 16

## **4. Entraînement et évaluation**

### **Séparation des données**

Les données sont réparties en deux ensembles afin d’éviter le surapprentissage :

* **Entraînement :** 80%
* **Test :** 20%

Toutefois, pour le modèle avec BERT, les données initiales ont été séparées en deux lots. Seul le premier lot est utilisé (entraînement, validation) pour limiter la quantité de données étant donné que BERT est un modèle pré-entrainé. Une partie du second lot a quand même été utilisée par la suite pour tester le modèle.

### **Processus d’entraînement**

* **Pour XGBoost :** Le modèle est entraîné à partir des vecteurs TF-IDF. Un GridSearchCV est utilisé pour optimiser les hyperparamètres et assurer la robustesse des résultats.
* **Pour BERT :** Le fine-tuning se fait en surveillant l’évolution des valeurs des pertes d’apprentissage et de validation au fil des époques.

### **Métriques d’évaluation**

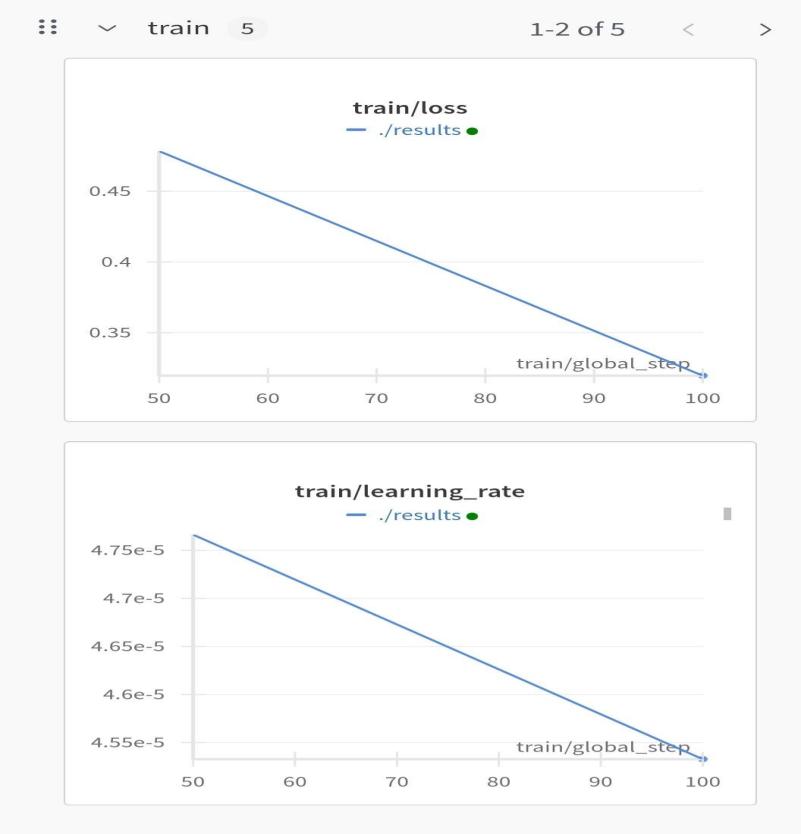
Les performances des modèles sont mesurées à l’aide des indicateurs suivants :

* **Accuracy :** Mesure globale du taux de classification correcte.
* **Précision, Rappel et F1-score :** Évaluent la performance sur chaque classe et permettent de gérer d’éventuels déséquilibres dans le dataset.

## **5. Visualisation et analyse des résultats**

### **Suivi de l’entraînement avec Weights & Biases (wandb)**

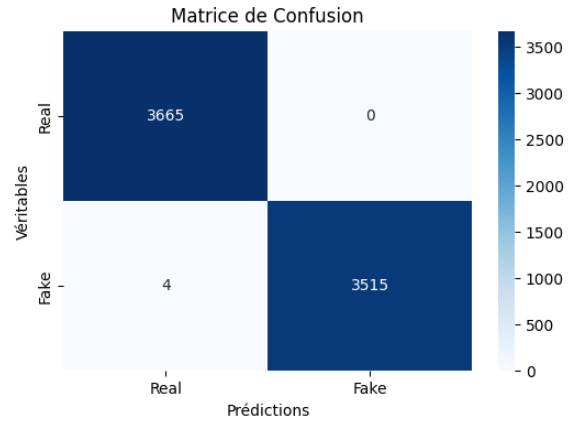
Nous avons utilisé l’outil wandb qui permet de suivre en temps réel l’évolution de l’entraînement :



* **Courbe de la loss :** Visualisation de la diminution de la perte au fil des époques, ce qui est un bon signe. Cela signifie que le modèle s’améliore au fil des itérations.

### **Présentation des résultats**

* **Matrice de confusion :** Elle permet d’identifier les faux positifs et faux négatifs, afin de déterminer les axes d’amélioration.



### Cette matrice issue de l’évaluation du modèle BERT sur l’ensemble de test prédit une infirmier portion de fake news comme vraies.

### **Analyse comparative**

* **BERT :** Le modèle BERT montre une capacité supérieure à comprendre le contexte, atteignant une précision d’environ 99% sur l’ensemble de test. Il excelle particulièrement dans la réduction des faux positifs.
* **XGBoost :** Même si XGBoost obtient une précision d’environ 91% sur le dataset COVID-19, sa rapidité et sa simplicité en font une alternative intéressante, notamment lorsque les ressources de calcul sont limitées.

### **Limites et pistes d’amélioration**

* **Limitations :**
  + Temps de calcul important pour le fine-tuning de BERT.
* **Améliorations futures :**
  + Exploration de modèles basés sur des Transformers alternatifs (ex. DistilBERT, RoBERTa).
  + Utilisation de méthodes d’optimisation plus avancées pour le tuning des hyperparamètres.

## 

## 

## 

## 

## 

## 

## **Conclusion**

Cette méthodologie rigoureuse et l’analyse comparative des modèles démontrent l’efficacité des approches basées sur les Transformers pour la détection automatique de fake news. Alors que XGBoost offre une solution rapide et opérationnelle, BERT se révèle supérieur en termes de compréhension contextuelle et de précision. Ce modèle peut être donc déployé dans des systèmes d'analyse de contenu sur les plateformes de médias sociaux, pour détecter et signaler les fake news. Dans le cadre de notre projet, ces modèles seront intégrés à une plateforme web à l’aide de React , fastapi, deepl.

Ces résultats ouvrent la voie à de nouvelles perspectives pour améliorer la détection et la classification des fake news dans des environnements réels.

**Références**

Source: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset