Reconnaissance des locuteurs et analyse de sentiments des reviews

March 19, 2023

- **YOUCEF KHODJA Amine 21113585**
- **KEMICHE Koceila 21114731**

Recherche d'Information et Traitement Automatique du Langage

RITAL

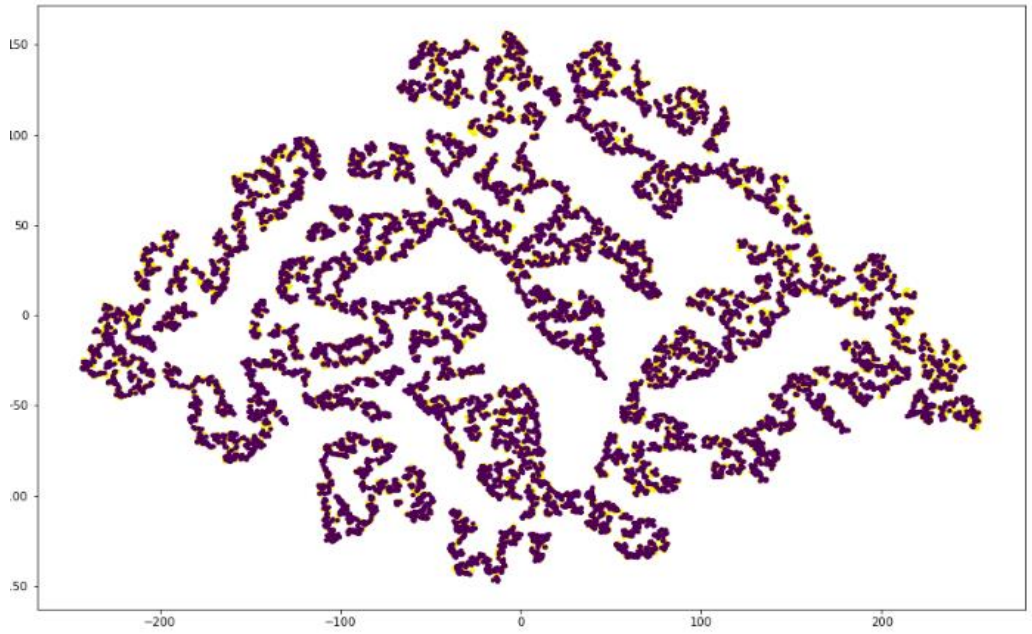
1 Introduction

Le présent rapport expose une série d'expériences menées dans le but de résoudre deux problématiques de classification de documents textuels en se servant des techniques de Machine Learning et de Traitement Automatique du Langage Naturel. La première problématique consiste en la reconnaissance des locuteurs dans des discussions présidentielles entre Chirac et Mitterrand, tandis que la seconde se concentre sur la prédiction des sentiments positifs ou négatifs exprimés dans les critiques de films. Pour atteindre ces objectifs, nous avons mis en place une chaîne de traitement souple permettant d'entraîner des modèles d'apprentissage avec des paramètres optimaux.

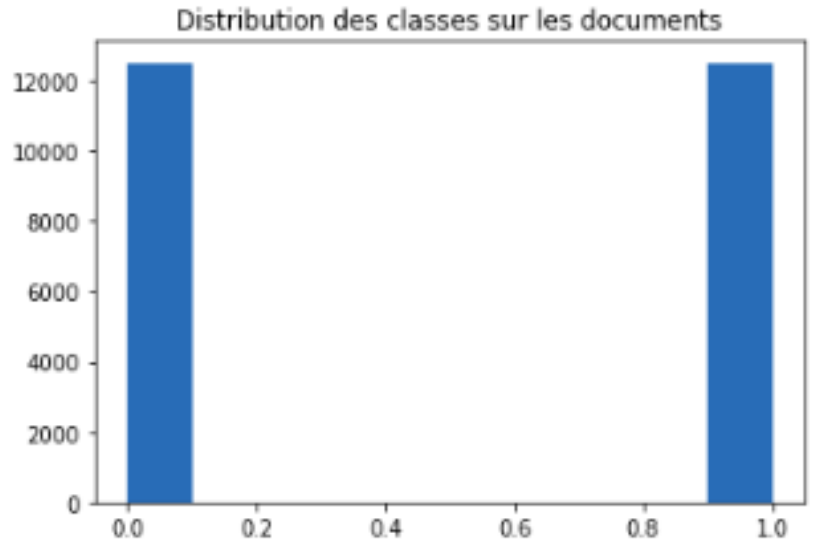
1.1 Visualisation des données :

Nous commençons d'abord par visualiser un nuage de points de nos don-nées en faisant une TSNE pour avoir une information sur la distributiondes classes .

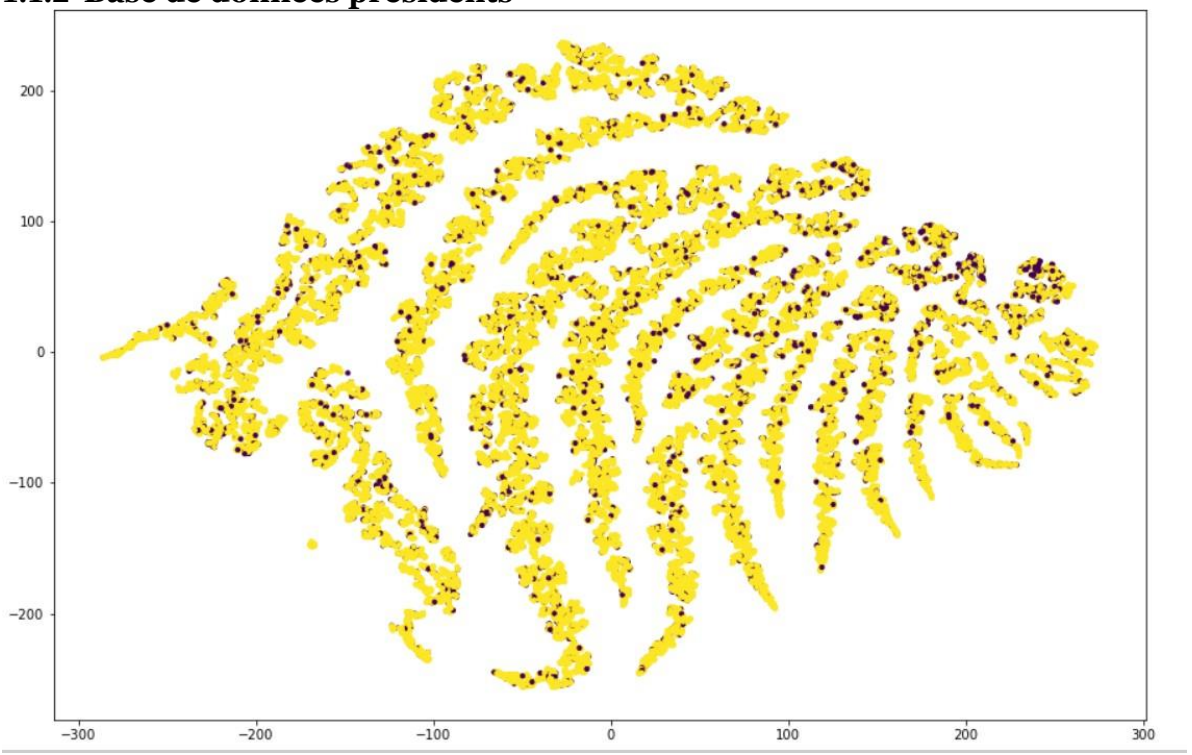
1.1.1 Base de données movies



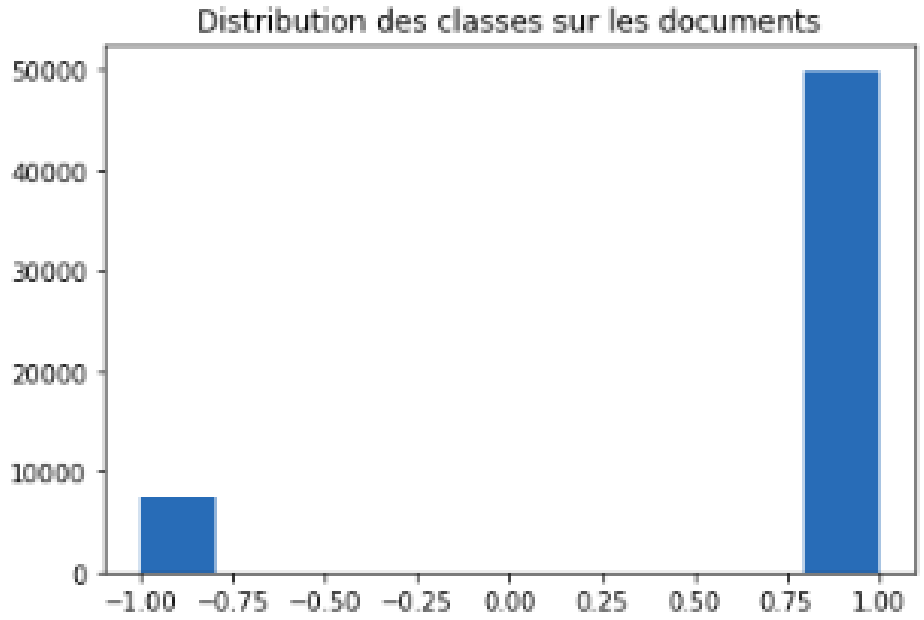
```
Nombre de documents classés positifs  12500
Nombre de documents classés négatifs  12500
```



1.1.2 Base de données présidents



```
Nombre de documents classés positifs  7523
Nombre de documents classés négatifs  49890
```



1.2 Prétraitements

Dans cette partie du rapport, plusieurs stratégies de prétraitement des données ont été adoptées, notamment la suppression des chiffres et des caractères spéciaux, la mise en minuscules de tous les mots, la trans- formation des mots entièrement en majuscules en marqueur spécifique, et la conservation d'une partie du texte seulement. Des techniques de stemming ont également été utilisées. De plus, des vectoriseurs tels que Count Vectorizer et TfIDF Vectorizer ont été créés avec différents paramètres tels que les bigrammes et les trigrammes, qui ont été utilisés dans la partie apprentissage.

- **Preprocess 0 : aucune modification sur la base de données.**
- **Preprocess 1 : suppression des chiffres et des caractères spéciaux et mise en minuscules de tous les mots.**
- **Preprocess 2 : transformation des mots entièrement en majuscule en marqueur spécifique en plus du prétraitement de la stratégie 1 .**
- **Preprocess 3 : transformation des mots entièrement en majuscule en marqueur spécifique en plus du prétraitement de la stratégie 1 sans suppression des chiffres.**
- **Preprocess 4 : suppression des chiffres et de la ponctuation en plus du stemming.**
- **Preprocess 5 : conservation d'une partie du texte seulement (seule- ment les trois premières lignes et les trois dernières lignes) en plus dustemming.**
- **Preprocess 6 : conservation d'une partie du texte seulement (seulement les trois premières lignes et les trois dernières lignes) , stemming et transformation des mots entièrement en majuscule en marqueur .**

1.3 Extraction du vocabulaire

1.3.1 Exploration préliminaire des jeux de données

Base de données movies

La taille du vocabulaire en gardant les 100 mots les plus fréquents est : 62851

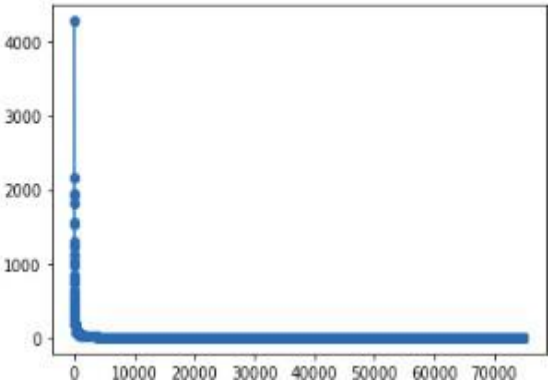
Out[34]: (-0.5, 399.5, 199.5, -0.5)

Les 100 mots les plus fréquents en terme de fréquence documentaire

Out[35]: (-0.5, 399.5, 199.5, -0.5)

On remarque que la distribution suit la loi Zipf .

Out[36]: [<matplotlib.lines.Line2D at 0x193d750b160>]



Les 100 mots les plus fréquents en terme de odds ratio

Les 100 mots les plus fréquents dans un Bag of Word Binaire

Out[39]: (-0.5, 399.5, 199.5, -0.5)

Nous avons également étudié les 100 bigrammes les plus fréquents dans les 100 trigrammes .

Base de données présidents

La taille du vocabulaire en gardant les 100 mots les plus fréquents est : 28448

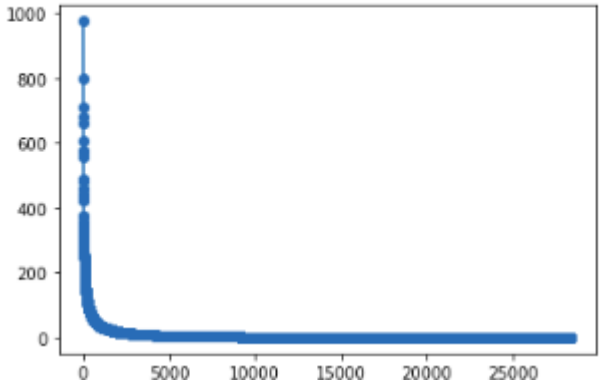
Out[10]: (-0.5, 399.5, 199.5, -0.5)

Les 100 mots les plus fréquents en terme de fréquence documentaire

Out[11]: (-0.5, 399.5, 199.5, -0.5)

On remarque que la distribution suit la loi Zipf .

Out[13]: [<matplotlib.lines.Line2D at 0x24ba87cbf70>]



sans les stop words

Out[12]: (-0.5, 399.5, 199.5, -0.5)

Les 100 mots les plus fréquents en terme de odds ratio

Les 100 mots les plus fréquents dans un Bag of Word Binaire

Out[15]: (-0.5, 399.5, 199.5, -0.5)

Nous avons également étudié les 100 bigrammes les plus fréquents dans les 100 trigrammes .

1.3.2 variantes de BoW

Avantages et inconvénients de ces variantes

- Pour le modèle binaire : bonne performance lorsque la présence d'un mot ou non est plus discriminante que la fréquence de ce mot, son principale inconvénient est qu'elle ne prend pasen copte la fréquence du mot, par exemple un document qui contient un mot 50 fois va être représenté de la même manière qu'un document qui en a qu'un seul, alors qu'ils sont très différents

- Bi-gramme, tri-gramme : capture la relation sémantique entre les mots, mais l'inconvénient est que si on avait beaucoup de dimension au début, cette dernière accroît d'une manière très grande ce qui rends le model complex.

- Réduire la taille du vocabulaire : l'avantage principale est qu'il permet de réduire la dimension du vocabulaire ce qui rends le mode moins complex, cependant, on peut éliminer desmots discriminant pour un document, donc une perte d'information très importante

- TF-IDF : le plus grand avantage c'est qu'il détecte les mots les plus discriminant dans chaquedocument par rapport au corpus, cependant, il peut attribuer des poids importants pour des mots rares qui n'ont pas de signification précise pour un genre de document

1.4 Optimisation et apprentissage

Le processus d'entraînement de nos modèles commence par la création de CountVectorizer et TfidfVectorizer avec différentes techniques de prétraitement de texte. Ces vecteurs sont ensuite utilisés pour entraîner plusieurs modèles de machine learning, avec différentes valeurs de paramètres et de pénalités. Pour chaque modèle, une validation croisée est effectuée en utilisant une métrique globale qui pondère les métriques Accuracy, Auc et F1_score selon leur importance dans le contexte du problème à résoudre. Cette étape permet de sélectionner les meilleurs hyperparamètres pour chaque modèle. Une fois les meilleurs hyperparamètres sélectionnés, le modèle correspondant est entraîné sur les données d'entraînement et utilisé pour faire des prédictions sur les données de test.

Base de données movies

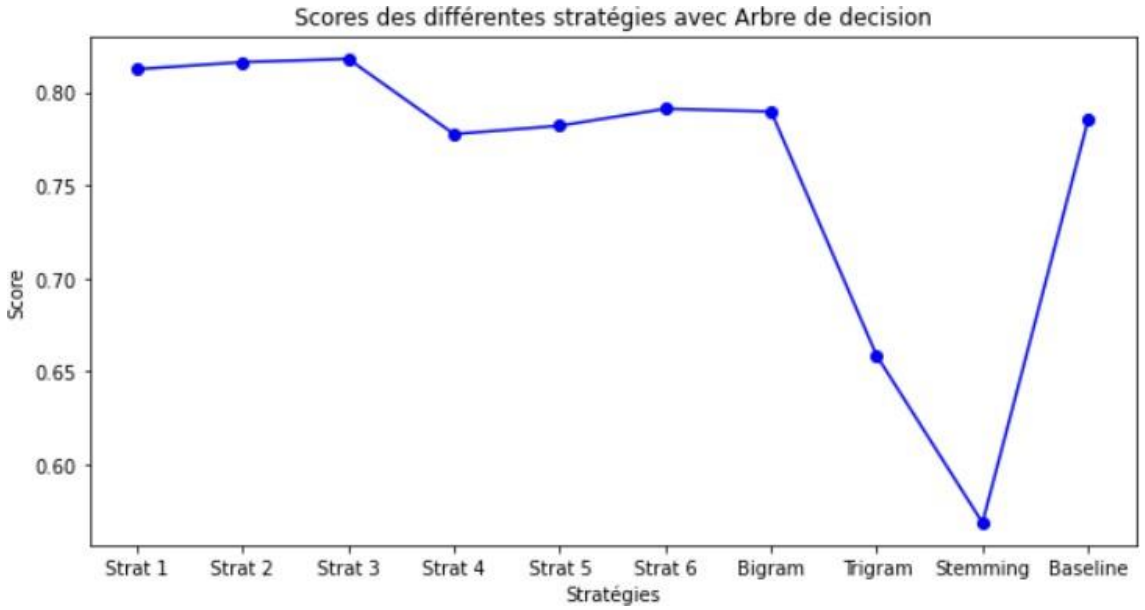
1.4.1 Régression logistique

Les résultats obtenus avec la régression logistique ont montré une bonne performance pour la prédiction des sentiments des critiques de films. Cependant, nous avons également pris en compte la possibilité de sur-apprentissage lors de l'optimisation des hyperparamètres. Pour éviter cela, nous avons effectué une validation croisée pour évaluer les performances des modèles sur des données d'entraînement et de validation distinctes, et avons choisi les hyperparamètres qui ont donné les meilleurs résultats de manière cohérente sur toutes les données. Cette approche nous a permis d'obtenir des modèles de régression logistique avec

des taux d'exactitude élevés sur les données de test, tout en évitant le sur-apprentissage.

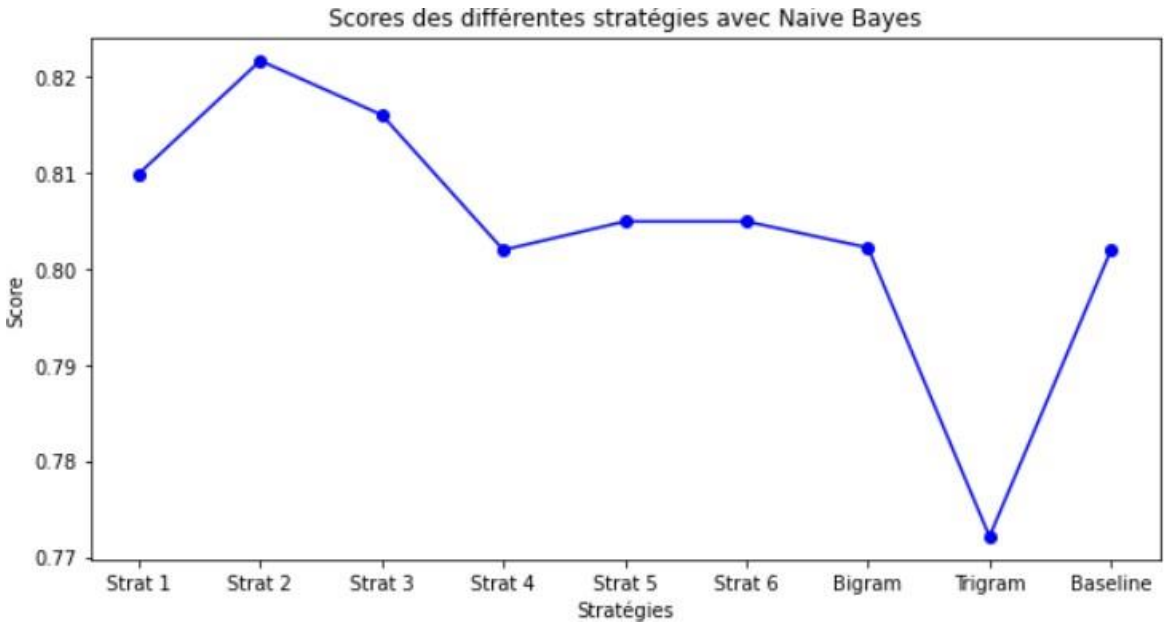
1.4.2 Arbre de décision

Ces résultats représentent les performances de différentes instances du modèle RandomForestClassifier avec une profondeur maximale de 5, évaluées avec la métrique d'accuracy sur un ensemble de données de test distinct. Les résultats varient entre 0.56868 et 0.81804, ce qui montre que la performance du modèle dépend fortement de l'ensemble de données de test utilisé. Cependant, en moyenne, le modèle semble atteindre une précision d'environ 0.78, ce qui est raisonnablement bon pour la classification de texte



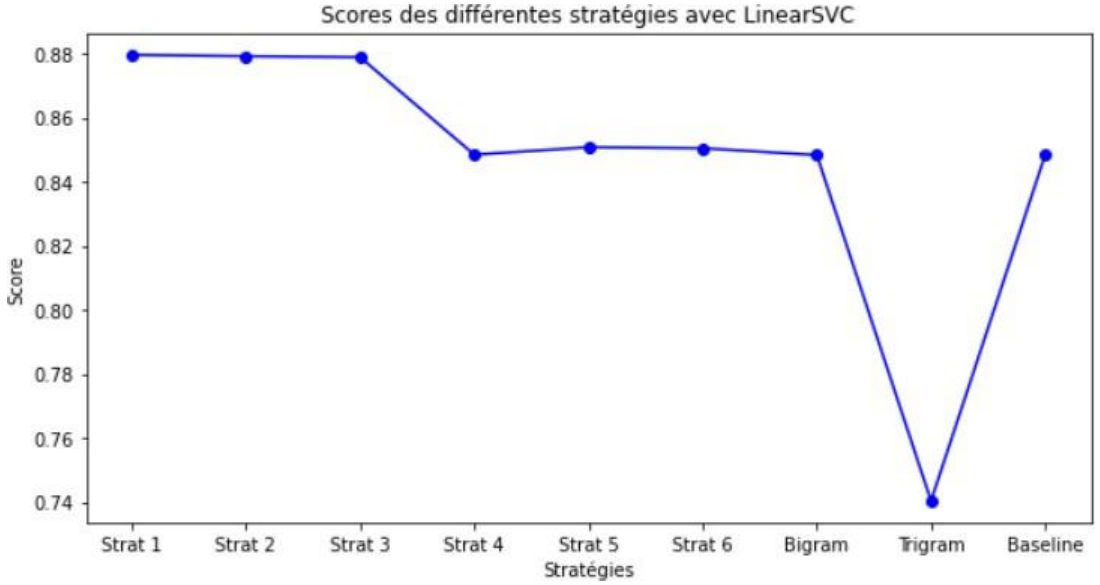
1.4.3 Naive Bayes

Le modèle de classification Multinomial Naive Bayes a également été testé sur les données, avec des résultats globalement satisfaisants. Les scores obtenus varient assez largement entre les différentes itérations de la validation croisée, allant de 0.80988 à 0.82164. Bien que légèrement inférieurs à ceux obtenus avec la régression logistique, les scores obtenus avec ce modèle sont généralement assez élevés. Cependant, il y a eu une itération avec un score particulièrement bas de 0.62692.



1.4.4 SVM Linéaire

Les résultats pour les modèles LinearSVC sont clairement supérieurs aux autres modèles testés. La différence de performance entre ces modèles est assez faible, avec une accuracy oscillant entre 0,87884 et 0,87956. Leur performance est donc très stable et élevée. Cela suggère que le modèle est bien adapté au problème de classification binaire et que l'optimisation des hyperparamètres a été bien menée.



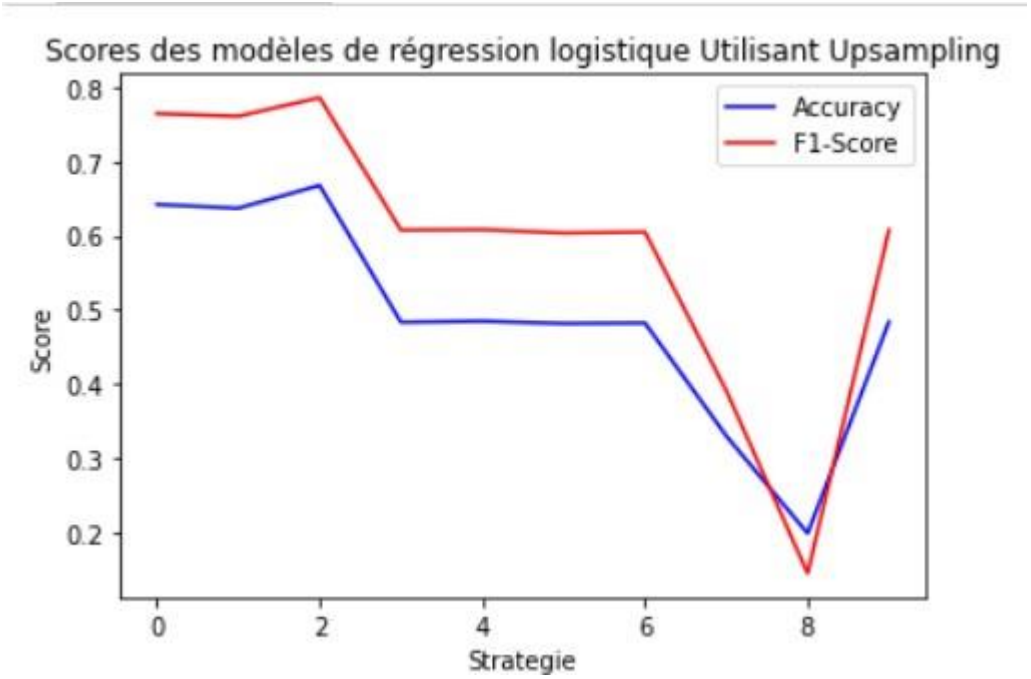
Base de données présidents

La base de données contenant les échanges entre les présidents de la France souffre d'un déséquili- bre de classes important, où les messages de Chirac sont prédominants par rapport à ceux de Mit- terrand (86% contre 13%). Cette disparité peut entraîner un biais dans l'apprentissage des mod- èles de classification et restreindre leur capacité à généraliser correctement. Afin de pallier cette

situation, nous avons opté pour des techniques de sous-échantillonnage et sur-échantillonnage telles que SMOTE, l'ajustement de la fonction coût et une autre méthode impliquant la courbe ROC et le biais.

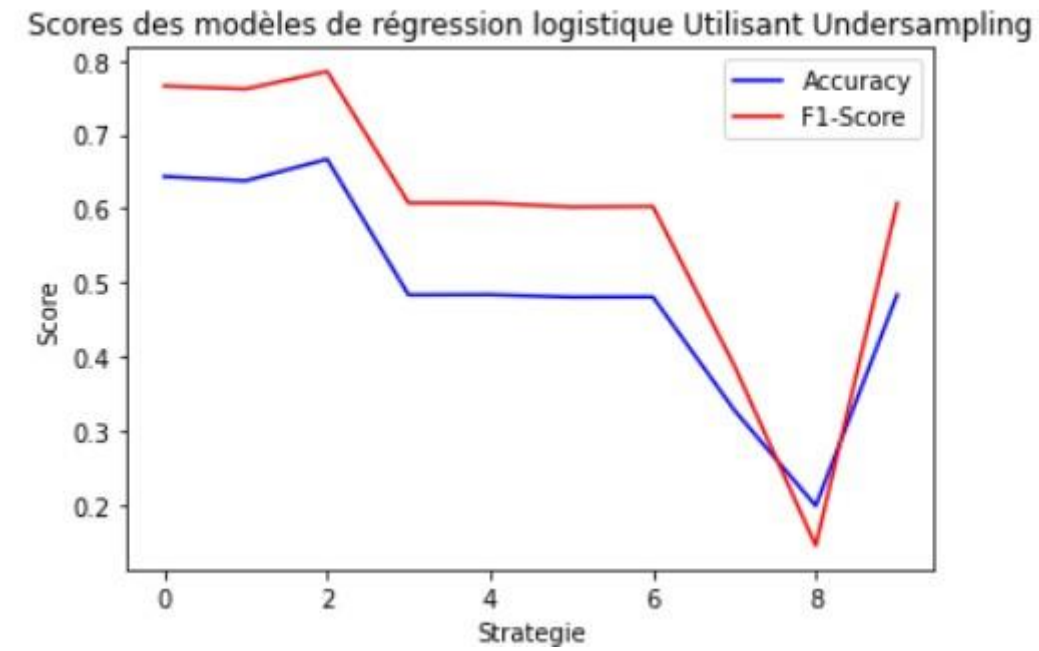
1.5.1 Upsampling

Le comparatif ci-dessus présente les performances de la méthode Logistic Regression pour notre dataset, en utilisant la méthode SMOTE pour corriger le déséquilibre de classes. Les métriques utilisées pour évaluer les performances sont l'accuracy et le F1 score.
La méthode SMOTE semble moyennement efficace en terme de généralisation sur la classe minoritaire.



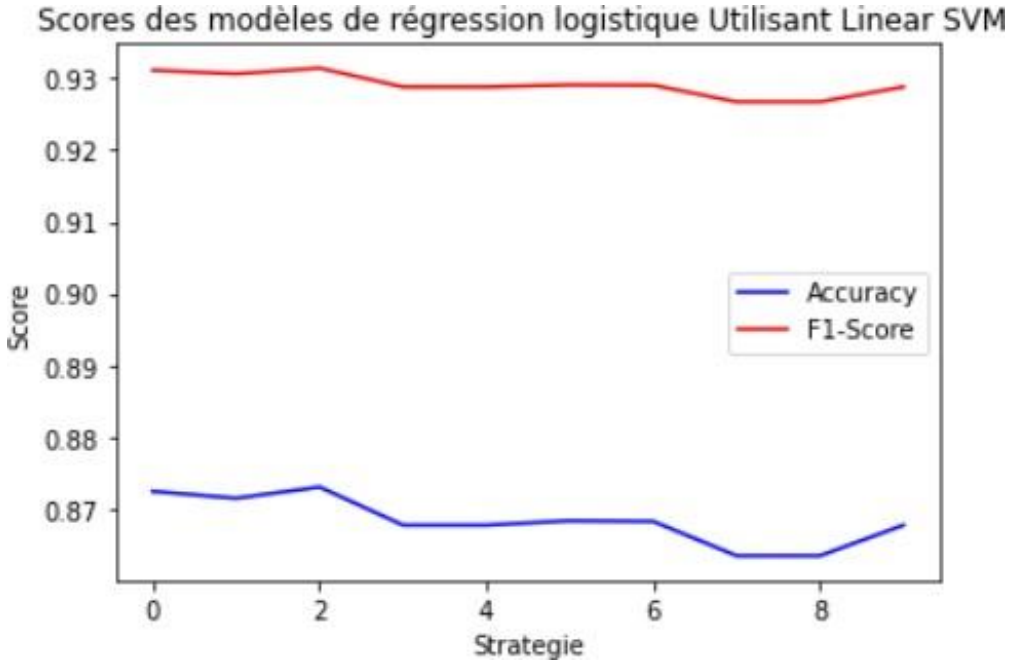
1.5.2 Undersampling

Le comparatif ci-dessus présente les performances de la méthode Logistic Regression pour notre dataset, en utilisant la méthode SMOTE pour corriger le déséquilibre de classes. Les métriques utilisées pour évaluer les performances sont l'accuracy et le F1 score.
La méthode SMOTE semble moyennement efficace en terme de généralisation sur la classe minoritaire.



1.5.3 Modification de la fonction coût

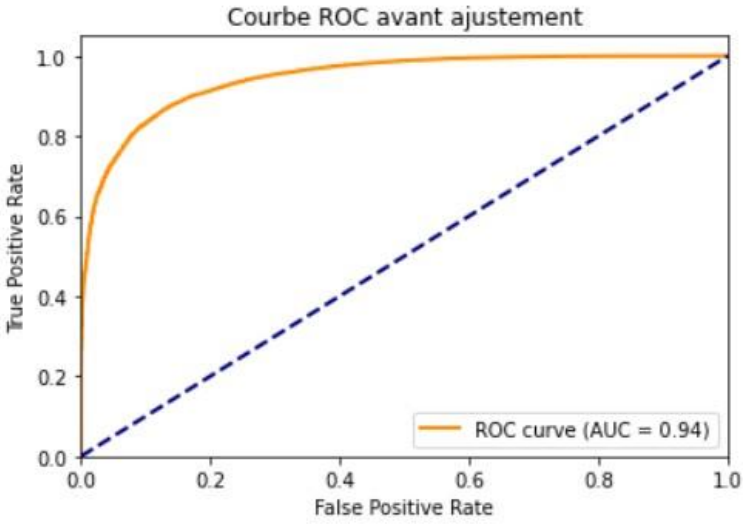
Le comparatif des performances entre les deux modèles montre que le deuxième modèle avec la fonction de coût modifiée a obtenu de meilleures performances en termes d'accuracy et de f1-score. En effet, le modèle avec la fonction de coût modifiée a obtenu une accuracy moyenne de 0.868 par rapport à une accuracy moyenne de 0.547 pour le premier modèle. De même, le f1-score moyen du deuxième modèle était de 0.929, tandis que celui du premier modèle était de 0.675. Il est donc évident que la modification de la fonction de coût a permis de mieux prendre en compte la classe minoritaire, ce qui a conduit à des performances nettement supérieures.



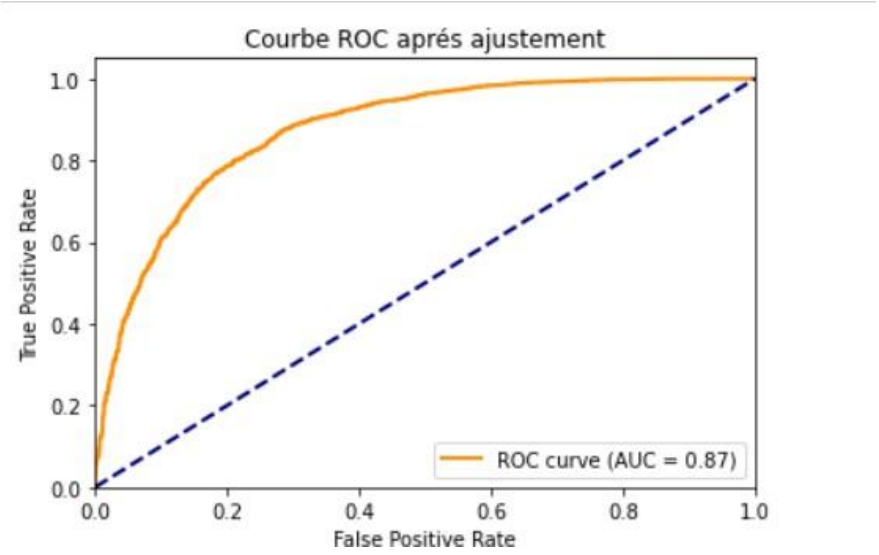
1.5.4 Courbe ROC et modification du biais

Le fait d'ajuster le biais pour obtenir une courbe ROC est une technique utile pour améliorer les performances de notre modèle. Dans notre cas, il semble que l'ajustement du biais ait légèrement amélioré les performances du modèle, avec une augmentation du score F1 de 0,01.ainsi nous avons constaté que cette amélioration est significative elle se généralise plutôt bien à de nouvelles données.

```
F1-Score Avant ajustement :  0.9388913790527416
```




```
F1 - Score aprés ajustement  :  0.946190522526029
```



2. conclusion

En conclusion, le prétraitement des données est une étape cruciale dans l'utilisation de techniques de machine learning notamment dans le traitement automatique du language. Les techniques de prétraitement peuvent grandement influencer les performances des modèles, et une mauvaise manipulation des données peut conduire à des résultats imprécis.. En outre, ce projet nous a permis de comprendre l'importance de l'évaluation de la performance des modèles de classification et la nécessité d'utiliser des mesures appropriées telles que la précision, le rappel, la F1-score et la courbe ROC. Nous avons également exploré l'impact de la modification du biais sur les performances des modèles et avons constaté que cela peut améliorer les résultats mais doit être utilisé avec précaution. En fin de compte, ce projet nous a donné une solide compréhension des concepts clés de la classification et de l'évaluation des performances des modèles, qui peuvent être appliqués à une variété de problèmes dans différents domaines.