

TABLE DES MATIERES

TABLE DES FIGURES	3
PARTIE 1 - EXPLICATIONS DES METRIQUES CHOISIES.....	5
I. Matrice de confusion	5
II. Accuracy	6
III. Recall	6
IV. Précision	6
V. F1-Score	7
PARTIE 2 - EXPLICATIONS DES RESULTATS OBTENUS	8
I. Récapitulatif de tous les résultats	8
II. Interprétation des résultats du S1	10
K-means.....	10
SVM.....	10
Random Forest.....	11
Conclusion générale sur le S1	12
III. Interprétation des résultats du S2	12
Introduction.....	12
Modèle de prédiction entraîné sur Wikipédia	12
Modèle de prédiction entraîné sur Wikipédia utilisé sur des tweets	13
Modèle de prédiction entraîné sur Wikipédia utilisé sur des posts Reddit.....	14
Perspectives	15
Explication du site web créé	17
CONCLUSION GENERALE SUR LE S2.....	19
CONCLUSION GENERALE SUR L'INTERPRETATION DES RESULTATS.....	20
CONCLUSION GENERALE SUR LE PROJET	21

TABLE DES FIGURES

Figure 1 : Formule de l’accuracy. Source : OpenClassroom	6
Figure 2 : Formule du recall. Source : OpenClassroom	6
Figure 3 : Formule de la précision. Source : OpenClassroom	6
Figure 4 : Formule du F1-score. Source : OpenClassroom.....	7
Figure 5 : Matrice de confusion du modèle Bert entraîné et testé sur les corpus de Wikipédia (S2) ...	13
Figure 6 : Matrice de confusion du modèle Bert entraîné sur les corpus de Wikipédia et testé sur les corpus de Twitter (S2).....	14
Figure 7 : Matrice de confusion du modèle Bert entraîné sur les corpus de Wikipédia et testé sur les corpus de Reddit (S2).....	15
Figure 8 : Matrice de confusion du modèle Bert entraîné et testé sur les corpus de Reddit (S2)	16
Figure 9 : Matrice de confusion du modèle Bert entraîné sur les corpus de Reddit et testé sur les corpus de Twitter (S2)	17
Figure 10 : Test de l’application avec un message suicidaire (S2)	18
Figure 11 : Test de l’application avec un message non suicidaire (S2)	18

Après avoir initialisé et programmé nos modèles, nous avons obtenu des premiers résultats. Ne sachant pas la véracité des scores obtenus, nous avons fait quelques modifications des hyperparamètres jusqu'à obtenir des résultats qui nous paraissaient convaincants. Evidemment, il est coûteux en temps et en calcul d'améliorer les hyperparamètres, surtout sur une période plutôt courte que celle du TER.

Pour comprendre ces résultats, il faut comprendre la problématique de base : est-ce que l'observation correspond à une donnée suicide ? En effet, nous cherchons à prédire des comportements suicidaires parmi les données que nous avons. Comme expliqué précédemment, le fait d'avoir ajouté une étiquette « suicide » et « non suicide » dans nos corpus nous permet de pouvoir faire de la prédiction.

Dans la première partie de l'année, l'objectif était de tester des algorithmes classiques par classification binaire avec le SVM, le random forest et le K-means. Nous entraînons et testons les mêmes observations avec la validation croisée sur un jeu de même nature. Les résultats que nous devrions obtenir doivent donc être excellents. L'objectif du second semestre est de tester l'hypothèse du transfert. On entraîne le modèle sur un jeu de données où on est sûrs des étiquettes labellisées avec Wikipédia, et on observe si le modèle fonctionne bien quand on le teste sur des sources différentes. Nous avons aussi pour ce second semestre utiliser un réseau de deep learning : Bert.

Pour comprendre notre raisonnement, nous allons commencer par vous faire une brève explication des métriques de classification que nous avons utilisées. Puis nous présenterons un récapitulatif des résultats obtenus aux cours de l'année, avant de nous intéresser plus précisément aux résultats du S1 puis à ceux du S2, ainsi que des possibles perspectives envisageables.

PARTIE 1 - EXPLICATIONS DES METRIQUES CHOISIES

I. Matrice de confusion

Afin de comprendre les indicateurs de prédictions que nous allons vous détailler ensuite, il est important d'avoir une bonne compréhension des résultats de la prédictions des modèles.

C'est pour cela que la matrice de confusion est un bon indicateur de résultats. La matrice de confusion permet de connaître les différentes erreurs commises par un modèle de prédiction.

Elle compare les données réelles de la variable cible à celles qui ont été prédites par le modèle.

Les résultats d'une matrice de confusion sont classés en 4 catégories comme on peut le voir sur le tableau ci-dessous :

- vrai positif (TP) : cas où les prédictions et les valeurs réelles sont positives,
- faux positif (FP) : cas où les prédictions sont positives mais que leur valeurs réelles sont négatives,
- vrai négatif (TN) : cas où les prédictions et les valeurs réelles sont négatives,
- faux négatif (FN) : cas où les prédictions sont négatives mais que leurs valeurs réelles sont positives.

		Classe réelle	
		-	+
Classe prédite	-	Vrais négatifs	Faux négatifs
	+	Faux positifs	Vrais positifs

II. Accuracy

L'accuracy est une métrique qui sert à l'évaluation des performances des modèles de classification : elle indique le pourcentage de bonnes prédictions. En d'autres termes, elle permet de décrire la performance du modèle sur les individus positifs et négatifs.

Elle mesure le taux de prédictions correctes sur l'ensemble des individus. L'équation de l'accuracy est écrite dans la figure 1 :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Figure 1 : Formule de l'accuracy. Source : OpenClassroom

III. Recall

Le recall indique le pourcentage de positifs bien prédit par notre modèle. En d'autres termes, il s'agit du nombre de positifs bien prédit divisé par l'ensemble des positifs. L'équation du recall est écrite dans la figure 2 :

$$Recall = \frac{TP}{TP + FN}$$

Figure 2 : Formule du recall. Source : OpenClassroom

On peut interpréter le recall de la manière suivante : plus il est élevé, plus le modèle en question maximise le nombre de vrai positif. C'est-à-dire qu'il ne ratera aucun positif.

IV. Précision

La précision indique le nombre de prédictions positives bien effectuées. En d'autres termes c'est le nombre de positifs bien prédit divisé par l'ensemble des positifs prédit. L'équation de la précision est écrite dans la figure 3 :

$$Precision = \frac{TP}{TP + FP}$$

Figure 3 : Formule de la précision. Source : OpenClassroom

On peut interpréter la précision de la manière suivante : plus il est élevé, plus le modèle en question minimise le nombre de faux négatifs. C'est-à-dire qu'une grande majorité des prédictions positives sont bien prédits.

V. F1-Score

Le F1 Score permet d'effectuer une bonne évaluation de la performance de notre modèle. En effet, il combine à la fois le recall et la précision, ce qui permet d'avoir un indicateur optimal. L'équation du F1 Score est écrite dans la figure 4 :

$$F1-score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 4 : Formule du F1-score. Source : OpenClassroom

On peut interpréter le F1-score de la manière suivante : plus il est élevé, plus le modèle sera performant car il combine deux indicateurs complémentaires.

PARTIE 2 - EXPLICATIONS DES RESULTATS OBTENUS

I. Récapitulatif de tous les résultats

Dans un premier temps, on peut s'intéresser au tableau ci-dessous qui résume tous les résultats obtenus au cours des deux semestres. On remarque que nos meilleurs résultats ont été obtenus avec les méthodes random forest et SVM.

Récapitulatif des scores obtenus au cours de l'année					
Corpus	Modèle	Accuracy	Recall	Précision	F1-score
Wikipédia	SVM	0,88	0,88	0,88	0,88
	Random Forest	0,88	0,88	0,88	0,88
	Bert (entraîné sur Wikipédia)	0,86	0,86	0,86	0,86
Reddit	SVM	0,95	0,95	0,95	0,95
	Random Forest	0,93	0,93	0,93	0,93
	Bert (entraîné sur Wikipédia)	0,53	0,98	0,52	0,68
	Bert (entraîné sur Reddit)	0,85	0,83	0,87	0,85
Twitter	SVM	0,99	0,99	0,99	0,99
	Random Forest	0,99	0,99	0,99	0,99
	Bert (entraîné sur Wikipédia)	0,56	0,95	0,53	0,68
	Bert (entraîné sur Reddit)	0,79	0,64	0,90	0,75

II. Interprétation des résultats du S1

K-means

K-MEANS				
	Accuracy	Recall	Précision	F1-Score
Wikipédia	0,55	0,55	0,55	0,55
Reddit	0,59	0,59	0,59	0,59
Twitter	0,52	0,52	0,75	0,61

Dans le cas des prédictions avec la méthode des K-means (cf. tableau ci-dessus), qui a pour but d'identifier des groupes de données inconnus à partir d'ensembles de données, on se rend compte que les résultats des indicateurs de prédictions sont assez bas, même s'ils sont au-dessus de 50%.

Ces résultats assez faibles ne correspondent pas avec l'hypothèse de base, on aurait dû obtenir des résultats plus importants.

On peut donc considérer que le modèle ne prédit pas efficacement les observations qui coïncident avec un comportement suicidaire. Le modèle a environ 1 chance sur 2 de se tromper de résultat et de prédire malencontreusement un résultat faux.

Prenons l'exemple de Twitter, interprétons les résultats obtenus :

- Un score d'accuracy de 0.52 signifie qu'uniquement 52% des observations ont été correctement prédites,
- Un recall de 0.52 signifie que le modèle oublie des prédictions positives,
- Une précision de 0.75, un score plus élevé que les autres métriques, signifie quant à lui que le modèle prédit en partie bien les positifs,
- Un F1-score de 0.61 signifie que le modèle est moyennement performant.

Cela est dû au fait que le k-means est une classification non supervisée et donc qu'elle n'a pas de labels. Or dans notre cas, nous en avons, donc nous n'aurions pas dû utiliser ce classifieur dans notre raisonnement.

SVM

SVM				
	Accuracy	Recall	Précision	F1-Score
Wikipédia	0,88	0,88	0,88	0,88
Reddit	0,95	0,95	0,95	0,95
Twitter	0,99	0,99	0,99	0,99

Dans le cas du SVM (cf. tableau ci-dessus), destinées à résoudre des problèmes de discrimination, on se rend compte que les scores sont tous au-dessus de 80%. Le SVM est un algorithme d'apprentissage très utilisé dans les cas d'affiliation à des groupes fixés, comme dans notre cas, où nous cherchons à prédire si l'observation est un suicide ou non.

Ces résultats sont en concordance avec notre hypothèse, le modèle réussit dans l'ensemble à faire de bonnes prédictions. Une observation étant considérée comme un suicide dans le monde réel est aussi prédite comme un suicide.

Prenons encore une fois l'exemple de Twitter, on observe ici des scores de 0.99 pour toutes les métriques de prédictions, cela signifie que :

- 99% des observations sont correctement prédites,
- le modèle ne rate qu'1% de positif,
- 99% des prédictions positives sont bien prédites,
- et que le modèle est extrêmement performant.

Dans tous les cas, nos corpus de données sont adéquats à l'algorithme. Néanmoins, comme nous n'obtenons pas un score de 100% dans les indicateurs, nous pourrions quand même effectuer quelques modifications dans nos données, comme faire un nettoyage de données plus minutieux, ou encore récupérer encore plus de données pour avoir une population plus représentative de la réalité.

Random Forest

RANDOM FOREST				
	Accuracy	Recall	Précision	F1-Score
Wikipédia	0,88	0,88	0,88	0,88
Reddit	0,93	0,93	0,93	0,93
Twitter	0,99	0,99	0,99	0,99

Pour finir, dans le cas du Random Forest, utilisé dans des problèmes de classification, on se rend compte également que les résultats récoltés sont très bons.

Le Random Forest est utile lorsque l'on cherche à faire des prise de décision, dans notre cas, on doit décider si l'observation est considéré comme un suicide ou non.

Dans le même cas que le SVM, le Random Forest est adapté à nos corpus, ses prédictions sont assez représentatives de nos données et permettent de confirmer notre hypothèse. Cela signifie qu'une observation considérée comme un suicide est correctement prédite dans l'ensemble.

Prenons encore une fois l'exemple de Twitter, on observe ici des scores de 0.99 pour toutes les métriques de prédictions, cela signifie que :

- 99% des observations sont correctement prédites,
- le modèle ne rate que 1% de positif,
- 99% des prédictions positives sont bien prédites,
- et que le modèle est extrêmement performant.

La conclusion que nous pouvons faire sur cet algorithme est exactement la même que celle du SVM, il est adéquat à notre problématique. Et pour améliorer le modèle, on pourrait récupérer encore plus de données par exemple.

Conclusion générale sur le S1

Nous pouvons conclure que nos recherches et nos modélisations effectuées lors du semestre 1 nous ont aidé à affirmer l'hypothèse, « Est-ce que l'observation correspond à une donnée suicide ? », que nous avons émise dès le début de l'année.

Nous avons pu répondre à cette question de recherche grâce aux algorithmes de classification supervisées, le Random Forest et le SVM. Nous avons eu des indicateurs très performants sur des modèles qui nous ont permis d'affirmer la problématique.

Il ne faut pas oublier que nous avons fait de la prédiction sur des données qui ont été entraînées préalablement avec la validation croisée, d'où nos résultats très élevés.

Donc, nous pouvons répondre à la problématique certes, mais ces modélisations vont en réalité nous permettre d'introduire la seconde partie de l'année.

III. Interprétation des résultats du S2

Introduction

Dans cette seconde partie de l'année, nous avons effectué des classifications et des prédictions grâce au modèle de BERT. Le modèle de BERT est un modèle d'apprentissage profond qui découle de l'analyse de texte.

Comme expliqué précédemment, le principe d'utilisation du modèle est le suivant : c'est un algorithme qui est déjà entraîné sur une grande quantité de données, suivant les tâches que nous souhaitons effectuer nous le modifions, puis on procède à un réentraînement sur nos données Wikipédia et nous testons nos données sur nos corpus Reddit et Twitter.

Pour comprendre ces résultats, il faut comprendre la problématique de base : est-ce que l'observation correspond à une donnée suicide ? En effet, nous cherchons à prédire des comportements suicidaires parmi les données dont nous disposons.

Pour comprendre nos recherches sur la question, nous allons interpréter les résultats fournis par le modèle lors de l'entraînement des corpus Wikipédia, puis les résultats obtenus lorsque nous avons testé notre modèle sur nos données Reddit, puis sur nos données Twitter. Nous finirons cette seconde partie sur des perspectives potentielles de recherches sur lesquelles nous sommes posé des questions : *si nous entraînons notre modèle sur Reddit, puis nous testons notre modèle sur Twitter, que se passerait-il ?*

Modèle de prédiction entraîné sur Wikipédia

Comme expliqué dans la fin du rapport 7, le modèle de deep learning créé à partir de BERT à été entraîné sur nos corpus Wikipédia. Lorsqu'on utilise ce modèle sur les corpus de test de Wikipédia on obtient de plutôt bon résultats comme le montre la matrice de confusion (figure 5).

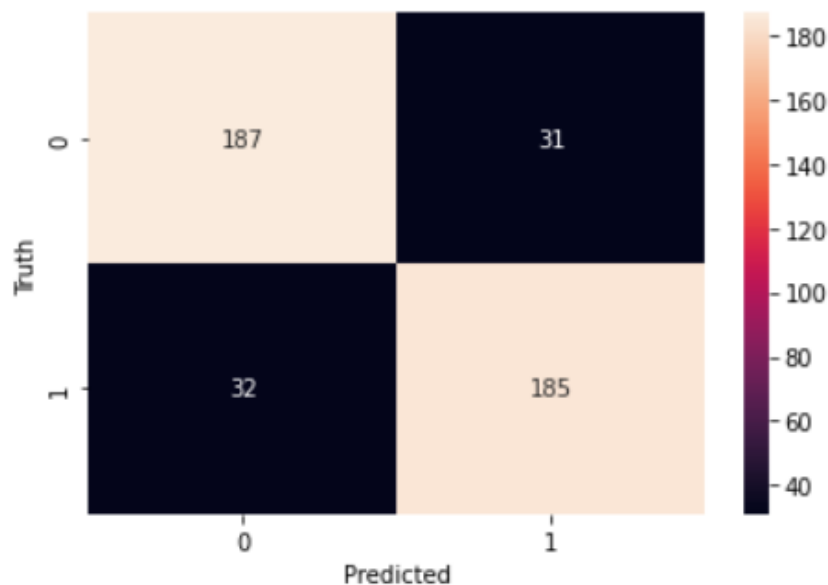


Figure 5 : Matrice de confusion du modèle Bert entraîné et testé sur les corpus de Wikipédia (S2)

On obtient les métriques de prédiction du tableau suivant :

BERT sur Wikipédia				
	Accuracy	Recall	Précision	F1-Score
Wikipédia	0,86	0,86	0,86	0,86

On obtient dans ce cas-ci que :

- 86% des observations sont correctement prédites,
- le modèle rate 14% de positif,
- 86% des prédictions positives sont bien prédites,
- et que le modèle est moins performant que certains de ceux utilisés au premier semestre.

On peut supposer que ces résultats plus faibles que ceux obtenus pendant le précédent semestre sont dus au fait que BERT est un modèle est déjà pré-entraîné sur les pages Wikipédias ou encore parce qu'il n'y a qu'un nombre assez faible de pages Wikipédia pour un algorithme de deep learning. Il peut notamment peut-être être encore amélioré en augmentant le nombre d'epochs.

Dans un deuxième temps, nous avons mis en place ce que notre commanditaire nous avait demandé, c'est-à-dire de voir si on pouvait utiliser un algorithme entraîné sur Wikipédia pour prédire si un tweet est une tentative de suicide ou non. L'intérêt est d'avoir des étiquettes dont on est sûr.

Modèle de prédiction entraîné sur Wikipédia utilisé sur des tweets

Les corpus de Wikipédia étant écrits de façon très différentes d'un tweet de Twitter, on obtient des résultats médiocres comme vu dans la matrice de confusion ci-dessous (voir figure 6). Pour rappel, dans le rapport 8 on voit dans les nuages de mots et les arbres de similitudes que les corpus Wikipédia sont centrés sur les parcours de vie alors que les tweets et les posts Reddit étaient plus centrés sur la personne et ses sentiments.

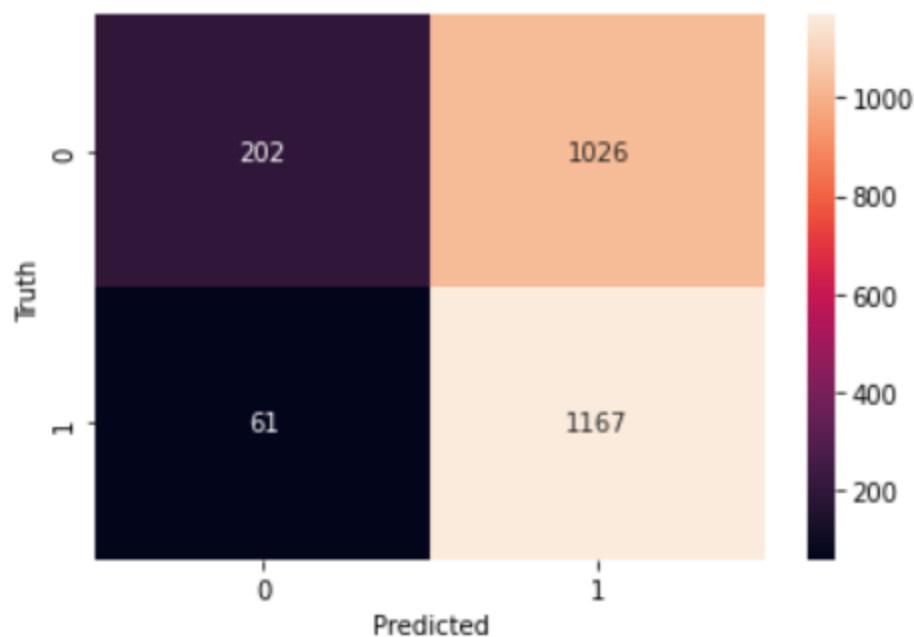


Figure 6 : Matrice de confusion du modèle Bert entraîné sur les corpus de Wikipédia et testé sur les corpus de Twitter (S2)

On obtient les métriques de prédiction du tableau suivant :

Modèle appliqué sur Twitter				
	Accuracy	Recall	Précision	F1-Score
Twitter	0,56	0,95	0,53	0,68

On obtient dans ce cas-ci seulement 56% de bonnes prédictions, l'efficacité du modèle sur ce type de corpus est pratiquement équivalente à une prise de décision après un pile ou face. On voit de plus que le modèle rate seulement 5% de positifs. Cette anomalie s'explique par le fait que le modèle classe la plupart des tweets dans la catégorie suicide. On peut conclure que ce modèle n'est absolument pas adapté à des tweets.

On va maintenant tester ce modèle sur des corpus de Reddit pour essayer de prédire des tentatives de suicide.

Modèle de prédiction entraîné sur Wikipédia utilisé sur des posts Reddit

Tout comme Twitter, Reddit a un style de rédaction très différent de celui de Wikipédia. Cela explique les très mauvais résultats obtenus lorsqu'on réapplique ce même modèle aux posts Reddit pour prédire les tentatives de suicide. On obtient la matrice de confusion ci-dessous (figure 7) :

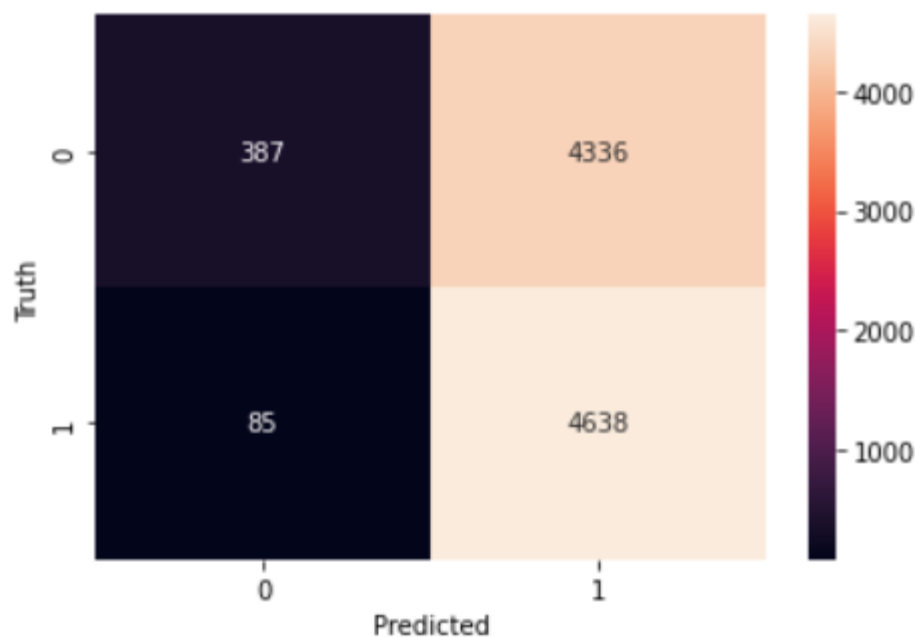


Figure 7 : Matrice de confusion du modèle Bert entraîné sur les corpus de Wikipédia et testé sur les corpus de Reddit (S2)

On obtient les métriques de prédiction du tableau suivant :

Modèle appliqué sur Reddit				
	Accuracy	Recall	Précision	F1-Score
Twitter	0,53	0,98	0,51	0,68

Dans ce cas-ci aussi, on obtient un taux de bonnes prédictions proche d'une chance sur deux, avec énormément de faux positifs. Ce modèle n'est pas concluant, il faut chercher un meilleur modèle.

Perspectives

L'une des principales explications des mauvais résultats obtenus vient des grandes différences entre les styles et les structures d'écriture des corpus de Wikipédia face à ceux de Reddit et Twitter. Une des perspectives qui respecterait la volonté d'utiliser un modèle déjà entraîné sur un type de corpus pour prédire des tentatives de suicides sur des corpus d'une autre source est d'entraîner le modèle sur les posts Reddit et de prédire les tentatives de suicide sur les tweets avec.

Lorsqu'on utilise le modèle entraîné sur Reddit sur le jeu de test, on obtient la matrice de confusion de la figure 8. Les résultats sont plutôt bons comme on le voit avec le tableau des métriques de prédiction.

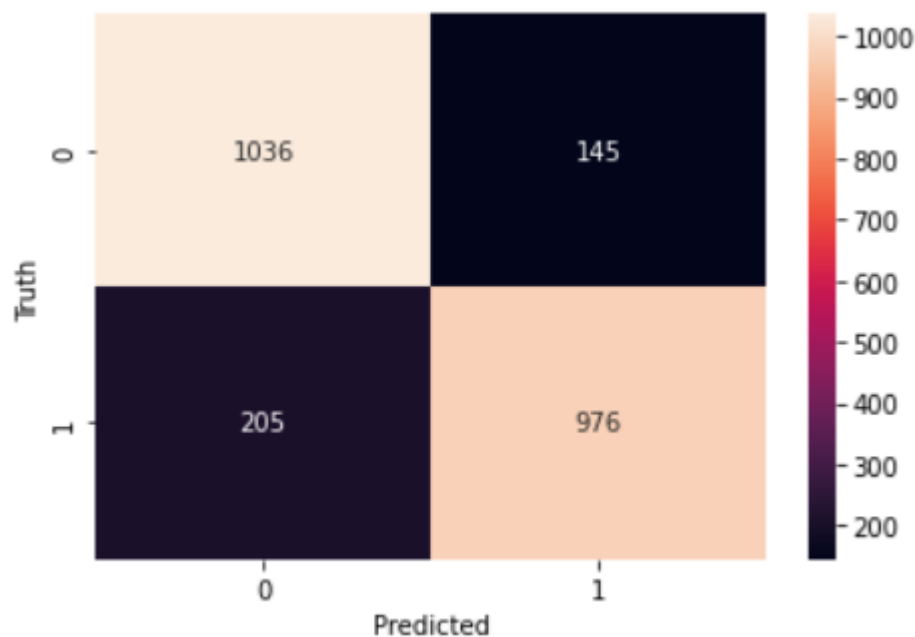


Figure 8 : Matrice de confusion du modèle Bert entraîné et testé sur les corpus de Reddit (S2)

On obtient les métriques de prédiction du tableau suivant :

Modèle entraîné sur Reddit				
	Accuracy	Recall	Précision	F1-Score
Reddit	0,85	0,83	0,87	0,85

On obtient dans ce cas-ci que :

- 85% des observations sont correctement prédites,
- le modèle rate 17% de positif,
- 87% des prédictions positives sont bien prédites,
- et que le modèle est moins performant que certains de ceux utilisés au premier semestre.

Encore une fois, les modèles du premier semestre étaient beaucoup plus justes pour prédire les tentatives de suicide sur des posts Reddit. Maintenant en appliquant ce modèle sur les tweets pour prédire les tentatives de suicides on obtient la matrice de confusion de la figure 9.

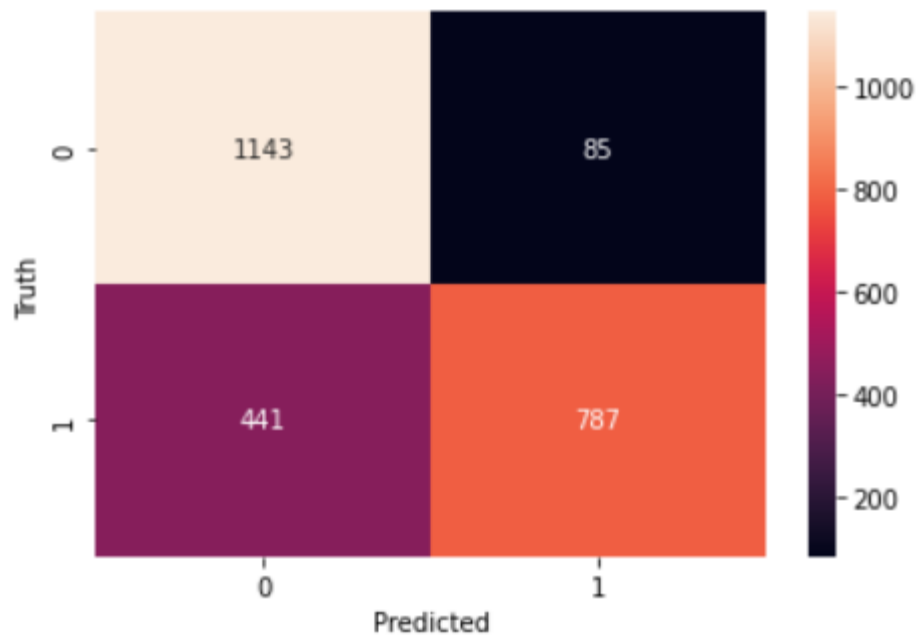


Figure 9 : Matrice de confusion du modèle Bert entraîné sur les corpus de Reddit et testé sur les corpus de Twitter (S2)

On voit que les résultats sont bien meilleurs que ceux obtenus avec le modèle entraîné sur Wikipédia. Le tableau des métriques de prédictions ci-dessous nous le confirme puisqu'on obtient les valeurs suivantes :

Modèle entraîné sur Reddit				
	Accuracy	Recall	Précision	F1-Score
Twitter	0,79	0,64	0,90	0,75

On obtient dans ce cas-ci que :

- 79% des observations sont correctement prédites,
- le modèle rate 34% de positif,
- 90% des prédictions positives sont bien prédites,
- et que le modèle est moins performant que certains de ceux utilisés au premier semestre mais tout de même meilleur que celui entraîné sur Wikipédia.

En conclusion, ce modèle est beaucoup plus juste pour prédire si un tweet est une tentative de suicide ou non que le modèle entraîné sur Wikipédia. Cela s'explique sans doute par le format et le style d'écriture plus proche entre Twitter et Reddit qu'entre Wikipédia et Twitter.

Explication du site web créé

À l'aide d'une application créée sur python, on peut tester si un texte est probablement une tentative de suicide ou non. L'application utilise le modèle entraîné sur les corpus de Wikipédia pour respecter la volonté du commanditaire.

Après plusieurs tests sur l'application, on se rend compte que lorsque l'on cherche à prédire la tendance suicidaire d'un tweet qui s'écrit à la première personne du singulier 'I', les prédictions sont très mauvaises. Le modèle n'arrive pas à détecter la pertinence du tweet. Cela s'explique par le fait que les données Wikipédia s'inscrivent dans un registre biographique, donc rédigés à la 3e personne du singulier ou du pluriel, mais jamais avec un « Je ».

On peut prouver cela par deux exemples : dans le premier cas (figure 11), nous écrivons un tweet à la première personne du singulier, et dans le deuxième cas (figure 10), nous écrivons le même tweet mais à la troisième personne du singulier. On se rend rapidement compte que les scores de prédictions sont différents, et donc que selon l'écriture verbale, notre prédiction sera plus ou moins faussée.

Prévention suicidaire à l'aide de tweet

Prédisez si une personne est potentiellement suicidaire rien qu'en inscrivant un tweet 🚨

Tweet:



He want to die

AI travaille ...

Nous sommes 0.5860000252723694% sûr que cette personne est suicidaire 🤖

Figure 10 : Test de l'application avec un message suicidaire (S2)

Prévention suicidaire à l'aide de tweet

Prédisez si une personne est potentiellement suicidaire rien qu'en inscrivant un tweet 🚨

Tweet:



I am happy

AI travaille ...

Cette personne va bien, nous en sommes 99.505% sûr ! 😊

Figure 11 : Test de l'application avec un message non suicidaire (S2)

CONCLUSION GENERALE SUR LE S2

Pour conclure sur les résultats du second semestre, on voit que réutiliser un modèle entraîné sur les corpus d'une autre source de données pour essayer de déterminer si un tweet est probablement une tentative de suicide n'est pas toujours concluant en fonction du corpus choisi pour entraîner le modèle. C'est notamment dans le cas du modèle entraîné sur Wikipédia. À l'inverse, on obtient de plutôt bons résultats avec un modèle entraîné sur Reddit et appliqué sur les corpus de Twitter.

Nous avons certes, obtenu de meilleurs résultats avec les algorithmes du premier semestre, mais on voit qu'il est néanmoins possible d'utiliser un bon modèle entraîné sur Reddit pour faire de la prédiction sur des tweets. C'est un avantage puisque Reddit nous permet de classer beaucoup plus facilement les posts entre tentative de suicide ou non, puisque les posts sont divisés par forums. A contrario, sur Twitter, il est plus compliqué de classer les tweets et c'est fait de façon beaucoup plus subjective.

CONCLUSION GENERALE SUR L'INTERPRETATION DES RESULTATS

Cela fait 3 ans, que ce sujet sur la prédiction des comportements suicidaires a déjà été donné aux étudiants MIASHS. L'objectif a dans un premier temps été de chercher à construire des trajectoires dans les biographies Wikipédia pour retrouver des événements similaires chez des individus qui se sont suicidés. C'était l'objectif de l'équipe TER d'il y a 3 ans. L'année dernière, l'objectif était de connaître l'humeur des gens avant et après la pandémie de COVID-18 sur les tweets. Cette année, notre objectif était de créer un algorithme de classification pour permettre de distinguer si un texte était susceptible d'être écrit par une personne suicidaire ou non.

Nous nous démarquons des autres années par la présence des 3 types de corpus : Twitter, Wikipédia et Reddit. Lors du S1 nous avons cherché à faire de la prédiction simple sur les données. Nous avons utilisé plusieurs algorithmes dont les plus concluants étaient SVM et Randoms Forest. Ces prédictions se sont avérées utiles dans le questionnement de notre recherche. À la suite de cela, durant le S2, nous nous sommes tournés vers un modèle d'apprentissage profond, le modèle de Bert. L'objectif était un peu différent cette fois-ci puisqu'il était d'entraîner le modèle sur le corpus de textes Wikipédia sur lesquels les labels sont sûrs pour ensuite utiliser ce modèle sur des tweets ou des posts Reddit pour déterminer s'il y a tentative de suicide ou non. Les modèles du S1 se sont avérés très performants contrairement à ceux faits au S2, qui se trompent pratiquement une fois sur deux, bien que ces faibles résultats étaient déjà attendus.

Nous avons rencontré plusieurs difficultés lors de cette année, sur lesquelles nous avons tous travaillé de concert pour les dépasser autant que possible. Notamment, sur le nettoyage des données qui était plutôt complexe à réaliser, un nettoyage encore plus approfondi nous permettrait sans doute d'améliorer nos résultats. Le deep learning était une nouveauté pour nous et a nécessité beaucoup de travail de tri sur ce qu'on pouvait trouver sur internet et de mise à niveau.

Le modèle entraîné sur Reddit au lieu de Wikipédia à l'aide de Bert obtient de plutôt bons résultats quand on l'applique sur des tweets. C'est un avantage puisqu'il est plus facile de se procurer des posts Reddits avec des labels que des tweets. Mais comme perspective, on pourrait imaginer améliorer nos résultats avec des ensembles de modèles comme vu en cours de Régularisation et optimisations des modèles en fin du second semestre. On pourrait aussi imaginer travailler avec plus de données. Ou par exemple, créer un type de modèle appris par type de data. Enfin on pourrait chercher une façon de retranscrire les tweets de la première personne au singulier à la troisième personne au singulier pour mieux correspondre au style d'écriture de Wikipédia (exemple : "Je suis très heureux" deviendrait "Il est très heureux"). Cela nous permettrait potentiellement d'avoir de meilleurs résultats sur le modèle entraîné sur les corpus de Wikipédia.

CONCLUSION GENERALE SUR LE PROJET

Le projet que nous devions faire au cours de cette année consister à développer un algorithme de prédiction des risques suicidaires à partir de corpus de textes. Ces corpus de textes proviennent des plateformes Reddit, Twitter et Wikipedia. C'est grâce aux différentes visualisations et modélisations que nous avons pu au mieux exploiter le sujet et ses données.

Après cette première partie consacrée au projet en lui-même nous allons maintenant vous présenter notre gestion du projet ainsi que nos ressentis.

Nous nous sommes réparti les tâches en fonction de nos compétences pour optimiser au maximum notre projet, certains étant plus à l'aise sur la partie code, d'autre sur la partie rédactionnelle, ou encore sur la partie visualisation.

Puis nous avons partagé nos connaissances entre nous pour pouvoir avancer plus rapidement. En effet, certains membres du groupe sont plus à l'aise dans certaines matières et ont donc pu partager leurs connaissances avec les autres afin de les former.

Ensuite, nous avons dû réaliser une mise à niveau dans certaines matières de façon autonome, en effet nous n'avons pas toutes les compétences nécessaires individuelles pour pouvoir réaliser tout ce que l'on souhaitait. C'est pourquoi tous les membres du groupe ont à un moment ou un autre fait une formation en ligne, visité plusieurs sites internet pour se former sur des sujets où ils avaient besoin d'apprendre pour pouvoir continuer là où ils bloquaient.

Tout cela nous a permis de développer de nouvelles compétences tel que le travail en équipe, la gestion de projet (planifier des tâches, respecter des délais, gérer les priorités, suivre l'avancée du projet...), des compétences relationnelles, et des compétences MIASHS avec l'apprentissage du Deep Learning et de l'analyse textuelle.

Nous sommes aussi devenus flexibles et adaptables en toutes circonstances. En effet, il est essentiel de savoir réagir et rebondir face aux imprévus que l'on rencontre dans un projet. Toutes ces compétences nous ont permis de mener notre projet vers la réussite.

Ce projet nous aura apporté beaucoup tant sur le plan éducatif que relationnel. Nous avons tous eu de bons ressentis, ce projet fût intéressant et concret. Nous avons pris plaisir à travailler en groupe. On espère que notre application servira dans le futur dans le domaine médical ou bien dans la recherche.

Puis nous avons tous tiré un avantage de ce travail de groupe, en effet la bonne entente et la cohésion de groupe nous ont fait avancer efficacement et nous ont permis de mieux gérer les difficultés que l'on a rencontré dans ce projet.

Enfin nous tenons à remercier nos encadrants Sandra BRINGAY, Florian LOMBARDO et Pierre LAFAYE DE MICHEAUX pour leur accompagnement tout au long de ce projet et leur disponibilité qui nous ont permis d'avancer dans de bonnes conditions.