ISEN M1 2023-2024

Rapport de projet

Classification de commentaires toxiques

Natural Language Progamming

LANGRAND Mathis – LECOUTURIER Louis

Introduction

Avec la montée en puissance des médias sociaux et des plateformes en ligne, la libre expression a trouvé un nouveau terrain fertile. Cependant, cette liberté s'accompagne souvent d'un revers : l'émergence de commentaires toxiques. Ces commentaires, qu'ils prennent la forme de discours haineux, de harcèlement, de désinformation ou de toute autre forme d'expression nuisible, peuvent entraîner des conséquences graves sur les individus, les communautés en ligne et même sur la société dans son ensemble. Les plateformes numériques se retrouvent ainsi confrontées à un défi de taille : comment maintenir un environnement en ligne sécurisé et respectueux tout en préservant la liberté d'expression ?

La modération manuelle des commentaires s'avère être une tâche fastidieuse, coûteuse et souvent insuffisante face au volume toujours croissant de contenus publiés chaque jour. C'est dans ce contexte qu'intervient la nécessité de développer des outils d'intelligence artificielle capables de détecter et de classifier automatiquement les commentaires toxiques. Ces outils promettent non seulement d'améliorer l'expérience utilisateur en ligne, mais aussi de contribuer à la lutte contre les discours de haine, le cyberharcèlement et autres formes de comportements nuisibles sur internet. Ainsi, le développement d'un système de classification de commentaires toxiques représente un enjeu crucial pour la sécurité et le bien-être des utilisateurs dans l'espace numérique contemporain.

Description du sujet

Le projet s'attache à concevoir un système de classification sophistiqué pour les commentaires en ligne, avec pour objectif principal la détection et la catégorisation des commentaires toxiques. Les commentaires seront soumis à un processus de classification, les séparant en deux catégories principales : toxiques et non-toxiques. Cette première phase permettra d'identifier rapidement les commentaires potentiellement préjudiciables qui nécessitent une attention particulière de la part des modérateurs. L'algorithme classera également dans un second temps les commentaires jugés toxiques qui seront analysés en profondeur afin de les catégoriser en fonction de diverses formes de toxicité, telles que severe_toxic, obscene, threat, insult et identity_hate. Ce processus de classification plus détaillé permettra de mieux comprendre la nature spécifique des commentaires toxiques, facilitant ainsi leur traitement et leur modération ultérieurs. L'objectif final est de mettre au point un modèle de classification robuste et précis, intégrant des techniques avancées de traitement du langage naturel et d'apprentissage automatique, afin de détecter efficacement les commentaires toxiques et de les catégoriser de manière appropriée. Ce système de classification automatisé contribuera ainsi à renforcer la sécurité et la convivialité des plateformes en ligne, offrant aux utilisateurs un environnement virtuel plus sûr et plus sain.

Données et approche utilisée

Les données issues du jeu de données sont sous la forme d'un fichier CSV en 8 colonnes comme ci-dessous :

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	Explanation\r\nWhy the edits made under my use...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\r\nMore\r\nI can't make any real suggestions...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

La première colonne précise l'ID du commentaire, la deuxième le commentaire en lui-même et les autres sont sous forme binaire, 1 si le commentaire entre dans la catégorie précisée par le nom de la colonne, 0 si elle n'y entre pas.

```
Length of dataset: 159571
%\ of toxic commentaries: 9.584448302009765 %

Name            Occurrence

severe_toxic    1595
obscene         8449
threat          478
insult          7877
identity_hate   1405
```

Le jeu de données est ainsi constitué de 159 571 commentaires dont environ 9.5% catégorisés comme toxiques. Dans ces toxiques, l'occurrence de catégorie supplémentaire est spécifiée à gauche. Notons qu'un commentaire peut appartenir à plusieurs catégories

L'analyse du jeu de données indique qu'aucune valeur est manquante et que l'intégralité des valeurs situées dans les colonnes permettant de classifier les commentaires sont uniquement des valeurs correspondantes à 0 ou 1.

Suite à cette analyse, les commentaires sont « nettoyés » : abréviations et sigles/contractions sont complétés, la ponctuation et stopwords (issus de la librairie NLTK) sont retirés.
Les commentaires sont finalement tokenisés afin d'être traités.

```
Number of missing values in each column :
id              0
comment_text    0
toxic           0
severe_toxic    0
obscene         0
threat          0
insult          0
identity_hate   0
dtype: int64

0s or 1s in dataset except comments:

toxic           True
severe_toxic    True
obscene         True
threat          True
insult          True
identity_hate   True
dtype: bool
```

Notre objectif est de déterminer ici quel algorithme de Natural Language Programming (NLP) serait le plus efficace en termes de classification de ces commentaires afin de faire cohabiter vitesse et précision. Nous estimerons qu'un algorithme atteignant un ratio de commentaires correctement classifiés témoigne d'une accuracy d'au moins 90%.

Les algorithmes traités dans ce projet sont : TF-IDF, RNN simple, RNN utilisant un réseau LSTM et un GRU.

Chacun utilisant une méthode de classification différente avec des modalités changeantes, la méthode le plus efficace pourra donc être ressortie.

Algorithmes et détails

TF-IDF

Le TF-IDF est une méthode utilisant la fréquence d'apparition des mots dans les commentaires considérés comme toxiques pour déterminer quels mots sont les plus présents dans ces commentaires, permettant ainsi de localiser ceux qui font pencher la balance de la toxicité.

```
confusion matrix :
[[2667   38]
 [ 150  145]]

classification report :     precision    recall  f1-score   support

                    0          0.95      0.99      0.97      2705
                    1          0.79      0.49      0.61       295

             accuracy                               0.94      3000
            macro avg          0.87      0.74      0.79      3000
         weighted avg          0.93      0.94      0.93      3000


accuracy : 0.9373333333333334
```

L'accuracy atteinte a été de 93.7%. Le modèle est donc considéré comme fiable. Cependant, un modèle basé sur la fréquence de mots pourrait confondre des phrases comme « Ce plat est mauvais » et « Ce plat n'est pas mauvais » juste par la simple présence de « mauvais », sans prendre en compte le contexte de la négation grammaticale de la phrase. Une solution serait d'apporter un contexte qui serait pris en compte.

RNN simple

Le réseau de neurones récurrent (RNN) utilise quant à lui des données séquentielles qui passent dans une boucle. Il tient compte des données actuelles et précédentes

afin de pouvoir « retenir » les mots de la phrase au fur et à mesure qu'il la parcourt, lui permettant d'en capturer le sens.

```
Epoch 1/3
77/77 [==============================] - 24s 292ms/step - loss: 0.4248 - accuracy: 0.8516 - val_loss: 0.3089 - val_accuracy: 0.9080
Epoch 2/3
77/77 [==============================] - 24s 317ms/step - loss: 0.3198 - accuracy: 0.9022 - val_loss: 0.3083 - val_accuracy: 0.9080
Epoch 3/3
77/77 [==============================] - 21s 273ms/step - loss: 0.3208 - accuracy: 0.9021 - val_loss: 0.3071 - val_accuracy: 0.9080
Training Complete
```

Ici, l'accuracy obtenue est de 90.8%. Le modèle est donc considéré comme fiable. L'une des raisons pour laquelle le score est moins élevé serait la présence de commentaires longs. En effet, avec la mémoire du modèle prenant en compte chaque mot, il est possible sur des commentaires longs qu'à l'arrivée à la fin de la séquence, la mémoire des premiers mots ait été effacés, écrasés par la mémoire des mots suivants. Une plus grande mémoire pourrait résoudre ce problème.

RNN avec LSTM

LSTM est un type de neurone spécifique en adéquation avec le RNN car celui-ci permet de « retenir » plus longtemps les mots de la phrase (souvent utilisé pour les séquences longues).

```
Epoch 1/3
77/77 [==============================] - 111s 1s/step - loss: 0.2597 - accuracy: 0.5764 - val_loss: 0.1561 - val_accuracy: 0.9980
Epoch 2/3
77/77 [==============================] - 117s 2s/step - loss: 0.1442 - accuracy: 0.9939 - val_loss: 0.1555 - val_accuracy: 0.9980
Epoch 3/3
77/77 [==============================] - 120s 2s/step - loss: 0.1438 - accuracy: 0.9939 - val_loss: 0.1557 - val_accuracy: 0.9980
Training Complete
```

L'accuracy obtenue est montée à 99.8%. Le modèle est donc considéré comme fiable. Cependant, sur un grand nombre d'epochs ou de séquences, le temps de calcul est extrêmement long, à cause de toutes les données retenues. Certaines d'entre elles pourraient être jugées inutiles de par leur neutralité dans la phrase. Nous pourrions donc nous débarrasser régulièrement de cette donnée.

GRU

Les unités récurrentes refermées (GRU) utilise le même principe que le LSTM en ayant également une « porte de l'oubli » utilisée pour les informations jugées inutiles afin de gagner en efficacité.

```
974/974 [==============================] - 608s 624ms/step - loss: 0.3309 - accuracy: 0.9507 - val_loss: 0.3143 - val_accuracy: 0.9802
```

Avec ce modèle, une accuracy de 98.02% a été relevée. Certes elle a été légèrement amoindrie cependant le nombre de données calculée a été considérablement augmenté, entraînant le modèle de façon légèrement fiable mais bien plus efficace.

Améliorations possibles des modèles

Ici ont été traités des commentaires uniquement en anglais. Une technologie apparue il y a maintenant quelques années pourrait aider à améliorer ce projet et lui donner une dimension avec une vision macro bien plus étendue.

Les transformers sont un type d'architecture de réseaux de neurones dont le potentiel en traduction n'est plus à prouver. Combinés avec ce projet ils pourraient offrir une classification de commentaires toxiques internationale avec une possibilité de détection de commentaires toxiques n'étant pas affectée par les barrières de la langue, donnant ainsi un modèle de détection de la toxicité des commentaires bien plus avancé et utile sur les réseaux sociaux et Internet en général.

De plus, les mécanismes d'attentions des transformers permettraient de traiter de manière plus fiable les séquences et de pouvoir traiter des séquences beaucoup plus longues.