Rapport du projet NLP(Toxic comment classification)

Membres du groupe : Landry SANON, Adrien Junior TCHUEM TCHUENTE

Contexte

Dans le cadre du cours de Natural Language Processing, nous avons pour projet de développer un modèle de Machine Learning capable de classifier le type de « toxicité » (toxic, severe_toxic, insult,obscene, threat, identity_hate) présent dans un commentaire.

Jeu de données

Pour réaliser ce modèle, nous avons utilisé un jeu de données 8 colonnes (id, comment_text, toxic, severe_toxic, obscene, threat, insult, identity_hate) et 159571 lignes.

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	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	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0
5	00025465d4725e87	"\n\nCongratulations from me as well, use the ...	0	0	0	0	0	0
6	0002bcb3da6cb337	COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK	1	1	1	0	1	0
7	00031b1e95af7921	Your vandalism to the Matt Shirvington article...	0	0	0	0	0	0
8	00037261f536c51d	Sorry if the word 'nonsense' was offensive to ...	0	0	0	0	0	0
9	00040093b2687caa	alignment on this subject and which are contra...	0	0	0	0	0	0
10	0005300084f90edc	"\nFair use rationale for Image:Wonju.jpg\n\nT...	0	0	0	0	0	0

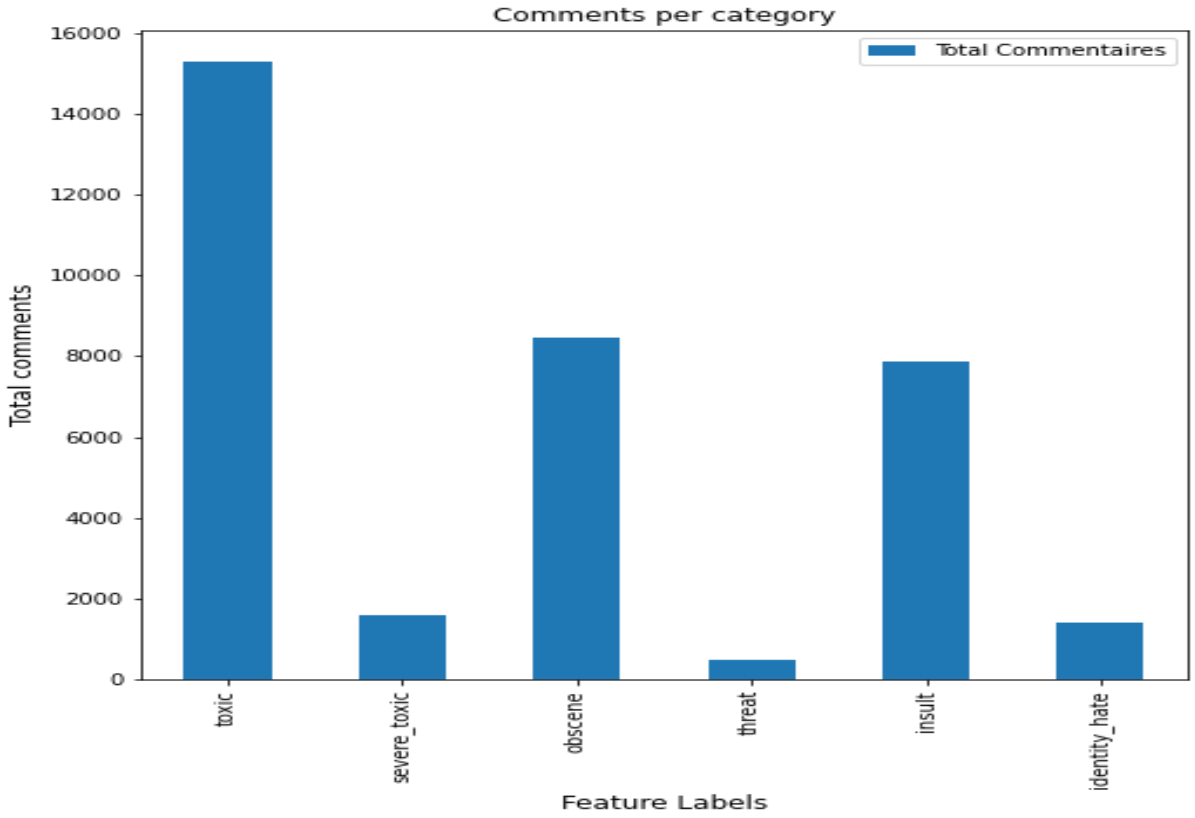
La colonne « id » comprend l'id de chacun des commentaires. Elle est inutile au sein de notre étude. Nous l'avons donc supprimée.

La colonne « comment_text » comprend les commentaires à analyser.

Les colonnes « toxic », « severe_toxic », « obscene », « threat », « insult », « identity_hate » peuvent comporter deux valeurs :

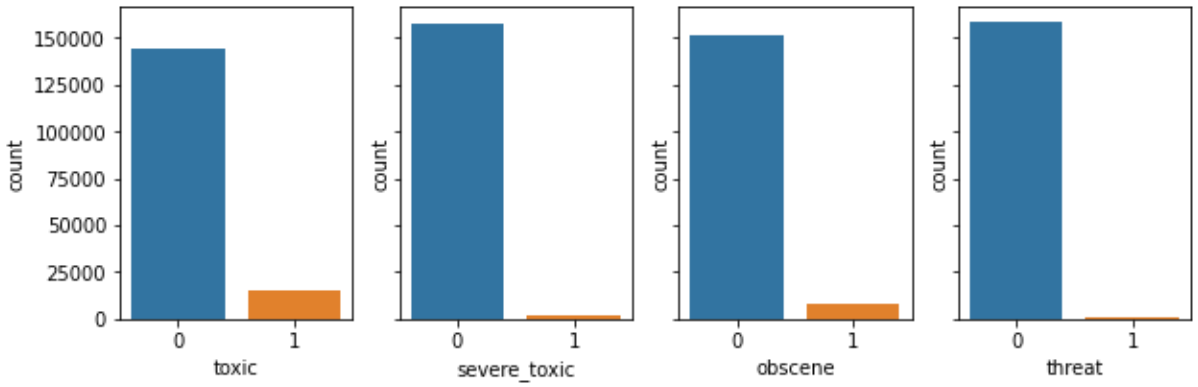
- 1 : si le commentaire appartient au type en question
- 0 : dans le cas contraire

Le schéma ci-dessous représente le nombre de commentaires appartenant à chaque type de toxicité.



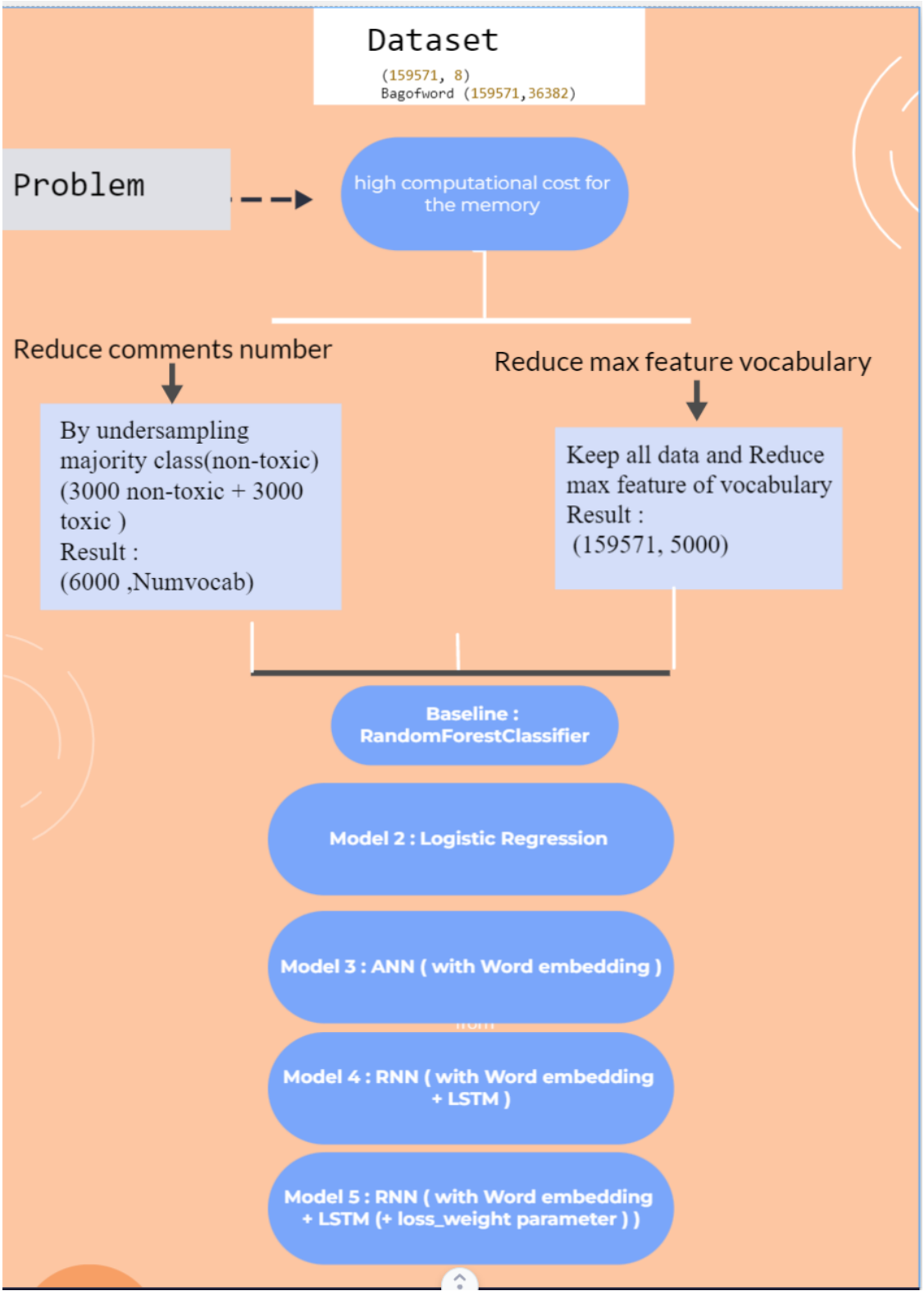
On remarque une prédominance de commentaires de type « toxic » et une sous-représentation des commentaires de type « threat ».

Le schéma ci-dessous démontre pour chaque type de toxicité, la répartition des valeurs 0 et 1.



On remarque une prédominance des commentaires des commentaires non-toxiques dans le jeu de données

Approche utilisée


Dataset
(159571, 8)
Bagofword (159571,36382)
Problem
high computational cost for the memory
Reduce comments number
Reduce max feature vocabulary
By undersampling majority class(non-toxic) (3000 non-toxic + 3000 toxic)
Result :
(6000 ,Numvocab)
Keep all data and Reduce max feature of vocabulary
Result :
(159571, 5000)
Baseline : RandomForestClassifier
Model 2 : Logistic Regression
Model 3 : ANN (with Word embedding)
Model 4 : RNN (with Word embedding + LSTM)
Model 5 : RNN (with Word embedding + LSTM (+ loss_weight parameter))

Détails de la solution

Pour le dataset 1, on remarque que c'est le modèle 5 (RNN with word embedding + LSTM + lost_weight parameter) qui présente une meilleure performance au niveau du f1-score.

ANN 1 :

	precision	recall	f1-score	support
toxic	0.88	0.83	0.86	769
severe_toxic	0.71	0.13	0.22	79
obscene	0.81	0.56	0.66	394
threat	0.00	0.00	0.00	18
insult	0.73	0.65	0.69	382
identity_hate	0.00	0.00	0.00	85
micro avg	0.83	0.65	0.73	1727
macro avg	0.52	0.36	0.40	1727
weighted avg	0.77	0.65	0.69	1727
samples avg	0.39	0.34	0.34	1727

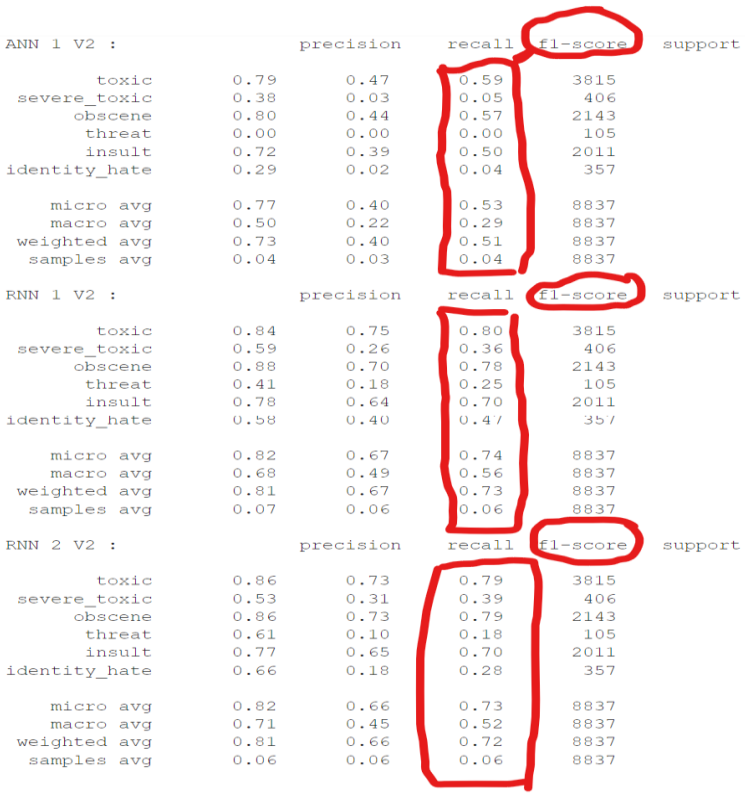
RNN 1 :

	precision	recall	f1-score	support
toxic	0.85	0.91	0.88	769
severe_toxic	0.00	0.00	0.00	79
obscene	0.72	0.69	0.70	394
threat	0.00	0.00	0.00	18
insult	0.72	0.63	0.67	382
identity_hate	0.00	0.00	0.00	85
micro avg	0.79	0.70	0.74	1727
macro avg	0.38	0.37	0.38	1727
weighted avg	0.70	0.70	0.70	1727
samples avg	0.42	0.38	0.37	1727

RNN 2 :

	precision	recall	f1-score	support
toxic	0.89	0.86	0.88	769
severe_toxic	0.53	0.11	0.19	79
obscene	0.77	0.71	0.74	394
threat	0.00	0.00	0.00	18
insult	0.76	0.62	0.69	382
identity_hate	0.00	0.00	0.00	85
micro avg	0.83	0.69	0.75	1727
macro avg	0.49	0.38	0.41	1727
weighted avg	0.76	0.69	0.72	1727
samples avg	0.41	0.36	0.36	1727

Pour le dataset 2, on remarque que c'est le modèle 4 (RNN with word embedding + LSTM) qui présente une meilleure performance au niveau du f1-score.



ANN 1 V2 :

	precision	recall	f1-score	support
toxic	0.79	0.47	0.59	3815
severe_toxic	0.38	0.03	0.05	406
obscene	0.80	0.44	0.57	2143
threat	0.00	0.00	0.00	105
insult	0.72	0.39	0.50	2011
identity_hate	0.29	0.02	0.04	357
micro avg	0.77	0.40	0.53	8837
macro avg	0.50	0.22	0.29	8837
weighted avg	0.73	0.40	0.51	8837
samples avg	0.04	0.03	0.04	8837

RNN 1 V2 :

	precision	recall	f1-score	support
toxic	0.84	0.75	0.80	3815
severe_toxic	0.59	0.26	0.36	406
obscene	0.88	0.70	0.78	2143
threat	0.41	0.18	0.25	105
insult	0.78	0.64	0.70	2011
identity_hate	0.58	0.40	0.47	357
micro avg	0.82	0.67	0.74	8837
macro avg	0.68	0.49	0.56	8837
weighted avg	0.81	0.67	0.73	8837
samples avg	0.07	0.06	0.06	8837

RNN 2 V2 :

	precision	recall	f1-score	support
toxic	0.86	0.73	0.79	3815
severe_toxic	0.53	0.31	0.39	406
obscene	0.86	0.73	0.79	2143
threat	0.61	0.10	0.18	105
insult	0.77	0.65	0.70	2011
identity_hate	0.66	0.18	0.28	357
micro avg	0.82	0.66	0.73	8837
macro avg	0.71	0.45	0.52	8837
weighted avg	0.81	0.66	0.72	8837
samples avg	0.06	0.06	0.06	8837

Conclusion

On remarque que le modèle 4 combiné au dataset 2 présente de meilleures performances de prédiction pour la globalité des types de toxicité.

Cependant, il est préférable d'utiliser le modèle 5 combiné au dataset 1 dans le cas où on veut prédire en priorité les commentaires de type " toxic", "obscene", "insult" puisque que ce modèle présente de très mauvaises performances pour les commentaires "threat" et "identity_hate".

Améliorations possibles de notre solution

- Il faut une plus grande capacité de GPU/ CPU
- On pourrait équilibrer la répartition entre les commentaires toxiques et non-toxiques. Pour cela, on pourrait par exemple récolter davantage de commentaires toxiques du jeu de données.
- On pourrait faire davantage de preprocessing au niveau des commentaires. Par exemple, on pourrait nettoyer les commentaires.