Baptiste ADAM (ESIEE)
Valentin FOARE (ESIEE)
Matthieu COMOY (D2SN)
Corine TCHEUTGA (D2SN)

Data science et analyse des réseaux sociaux

-

Rapport et analyse du réseau

Liens Reddits

Baptiste ADAM (ESIEE) Matthieu COMOY (D2SN)
Valentin FOARE (ESIEE) Corine TCHEUTGA (D2SN)

Table des matières

Baptiste ADAM (ESIEE) Matthieu COMOY (D2SN)
Valentin FOARE (ESIEE) Corine TCHEUTGA (D2SN)

Introduction

Notre analyse se basera sur un dataset existant (RedditHyperlinks) qui a été mis à disposition sur le site de l'université de Stanford. Le réseau se base sur des posts récoltés sur le réseau social reddit. Les posts récoltés ont tous un lien vers un autre subreddit dans leur titre. Un lien a donc pour origine un message dans une communauté et pointe vers une autre. Chaque lien est accompagné de plusieurs propriétés : un timestamp, l'analyse de sentiment du message qui l'accompagne et plusieurs propriétés textuelles comme le nombre de caractères, la ponctuation, etc. Cela nous permet de voir le type de relation qu'entretiennent deux communautés ou de voir des groupes de subreddits interagissant souvent.

Aussi à partir de ces données, nous pouvons nous poser plusieurs questions : Est-ce que toutes les communautés interagissent entre elles ? Y a-t-il des "communautés de communautés" ? Est ce que certains subreddits sont plus touchés par la négativité ? Les messages négatifs sont-ils différemment construits des messages positifs ?

Baptiste ADAM (ESIEE) Matthieu COMOY (D2SN)
Valentin FOARE (ESIEE) Corine TCHEUTGA (D2SN)

Synthèse du protocole

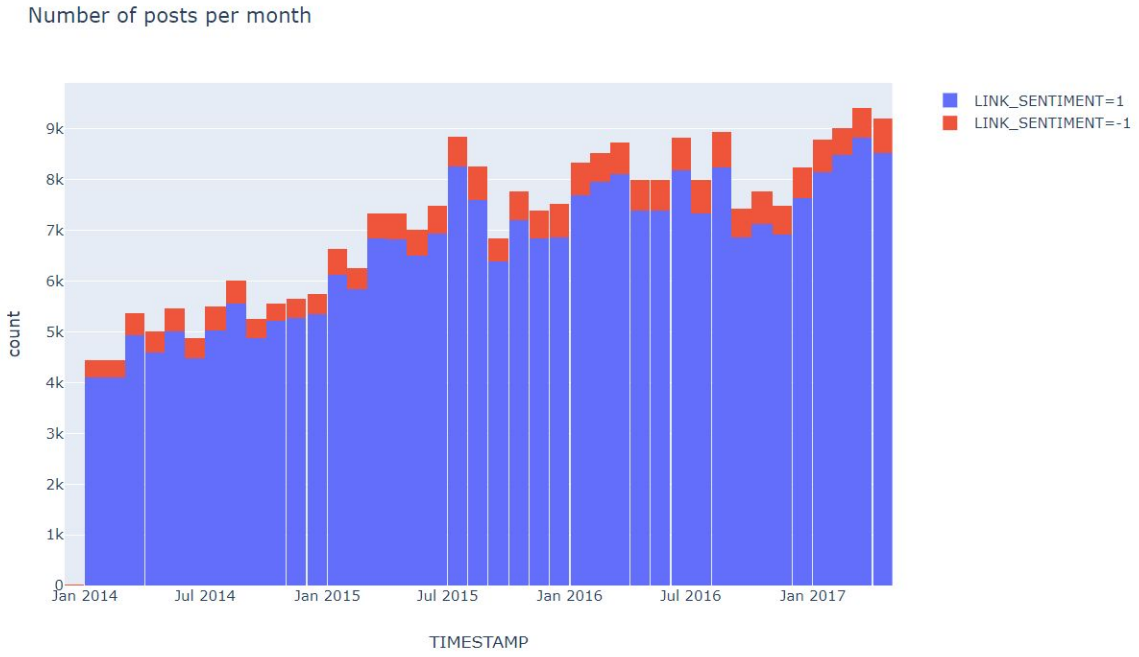
Le dataset s'étend de janvier 2014 à avril 2017 ce qui correspond à une période d'analyse de 31 mois. Nous n'avons donc pas eu besoin d'enrichir le dataset qui comporte suffisamment de données.

Pour notre analyse nous allons surtout garder les données de temps, de lien pour les subreddit ainsi que toutes les analyses de sentiment de message que comporte le dataset pour donner plus de profondeur à nos questions.

Nous avons trouvé les données sous forme d'un fichier .tsv que nous avons par la suite retravaillé et exporté en .csv à l'aide de pandas sous python. Nous avons ensuite restructuré les données pour pouvoir rendre analysable les derniers paramètres qui étaient rassemblés dans une seul colonne "PROPERTIES" dans le data frame. Sur les 91 colonnes du jeu de donnée de départ nous en avons gardé 27 pour faciliter le traitement des données.

Par la suite nous avons fait quelques analyses de base à l'aide de plotly et Gephi pour mieux caractériser le réseau :

Nombre de noeuds : 35 576 noeuds correspondant chacun à un subreddit
Nombre de relations entre ces nœuds : 286 560 relations correspondant chacune à un message possédant un hyperlien vers un autre subreddit.
Evolution des données dans le temps :



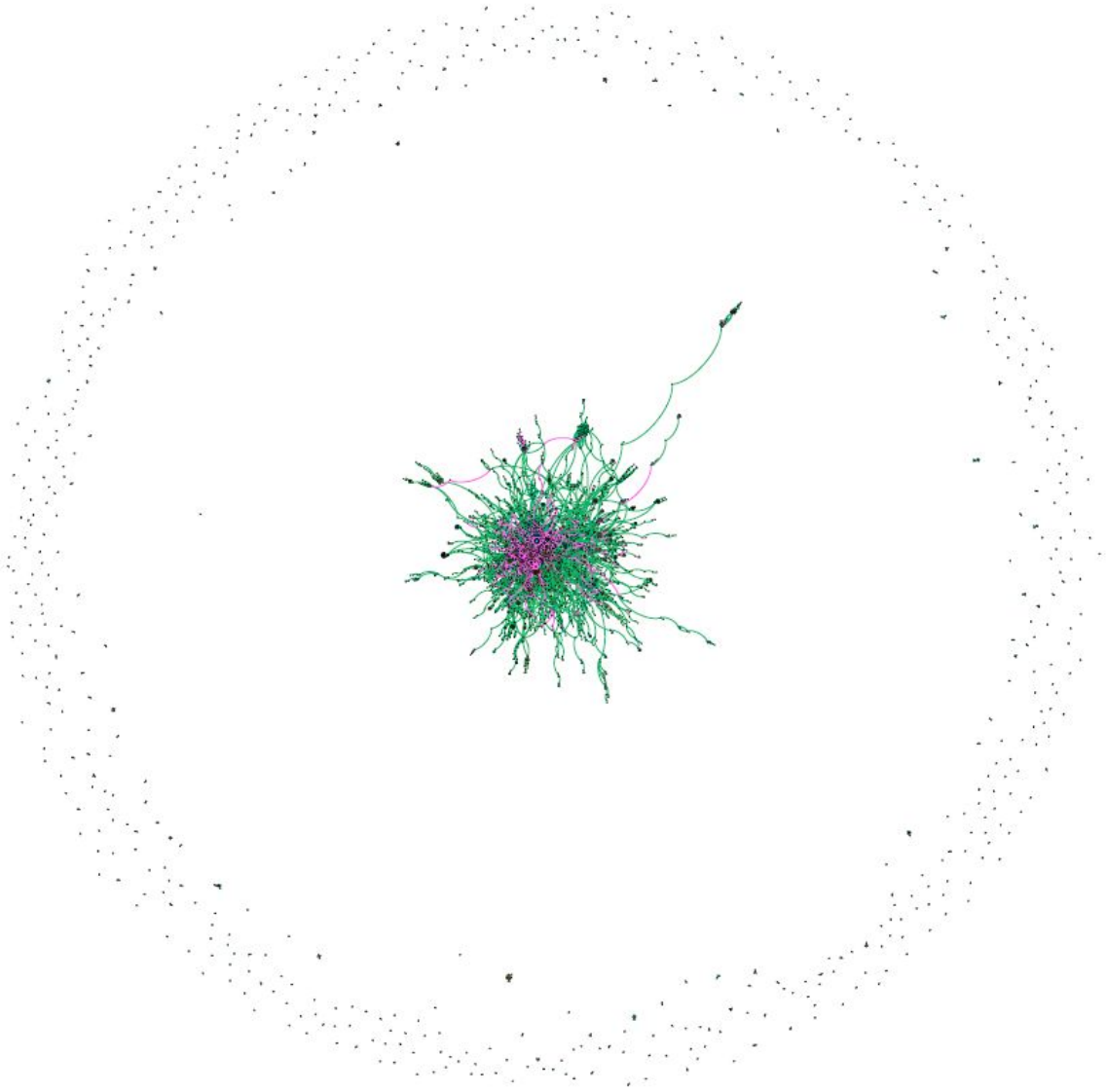
On peut constater une augmentation régulière du nombre de post tout au long de la période d'observation ce qui est attendu pour un réseau social qui a une base d'utilisateur en constante évolution.

Baptiste ADAM (ESIEE) Matthieu COMOY (D2SN)
Valentin FOARE (ESIEE) Corine TCHEUTGA (D2SN)

Visualisation

Pour des raisons de lisibilités, une visualisation contenant les 35 000 nœuds n'était pas possible. Nous avons donc fait le choix de réduire le dataset uniquement pour cette visualion. Nous avons conservé le ratio de messages négatifs (il se trouve qu'il est à 7.8%) et sélectionné les messages aléatoirement pour minimiser l'altération du dataset. Au final, il n'y a qu'environ 4000 nœuds dans cette visualisation.

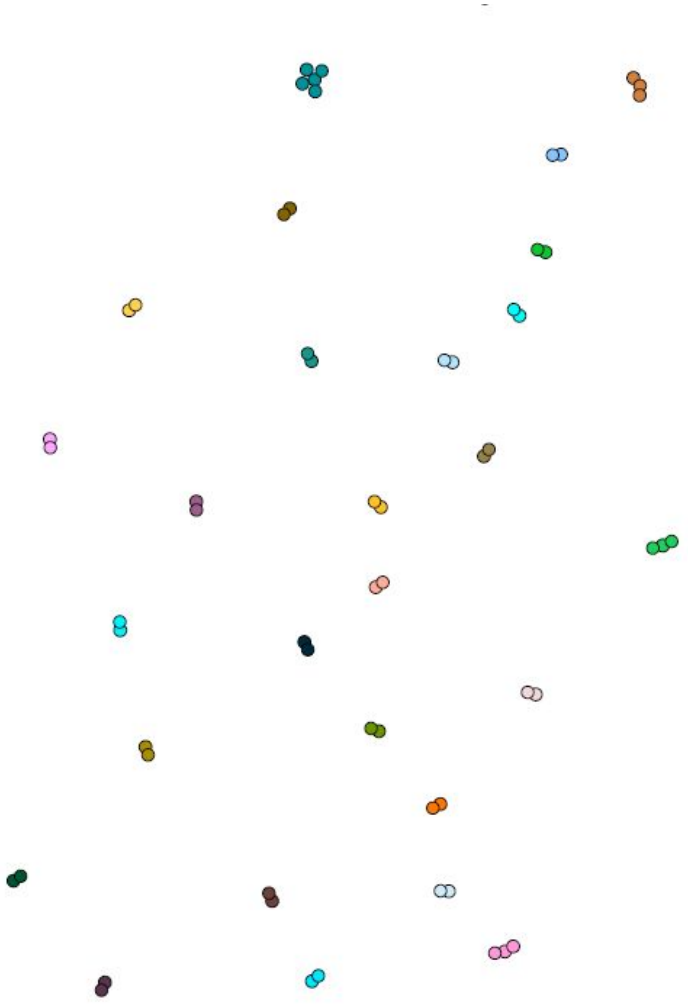
Légende :
Couleur des noeuds : "Communautés" de subreddits
Couleur des liens : Link Sentiment (vert = positif, rouge/rose = négatif)
Taille des noeuds : degré sortant (nombre de message envoyé)



On peut voir qu'il y a un gros rassemblement au centre et beaucoup de points isolés qui gravitent à l'extérieur.

Baptiste ADAM (ESIEE) Matthieu COMOY (D2SN)
Valentin FOARE (ESIEE) Corine TCHEUTGA (D2SN)

La ceinture extérieure correspond à tous les subreddits qui ont peu d'interactions, souvent avec un seul autre subreddit.

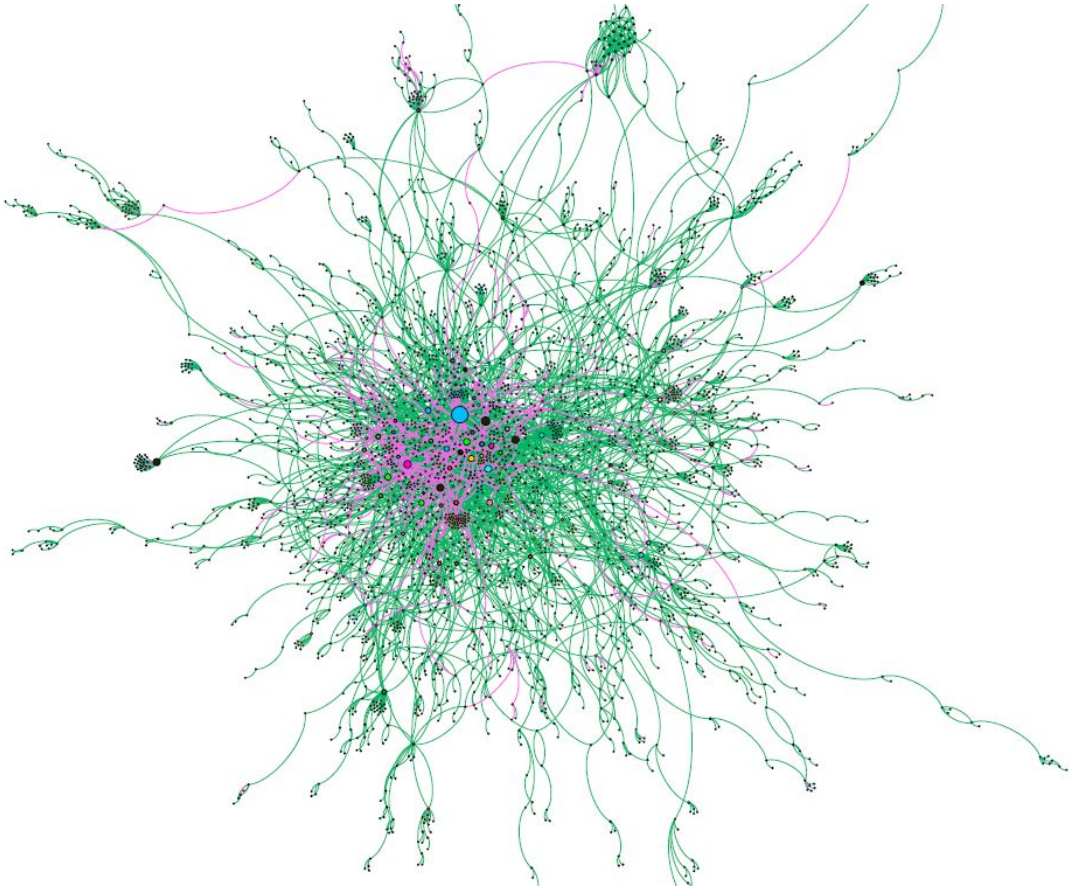


Baptiste ADAM (ESIEE) Matthieu COMOY (D2SN)
Valentin FOARE (ESIEE) Corine TCHEUTGA (D2SN)

Au contraire, au centre se trouve les subreddits avec beaucoup de lien. On peut voir que les communautés sont assez mélangées au sein de ce tout, indiquant l'hyper connection de ses subreddits. Néanmoins, nous remarquons bien sur les extrémités des groupes de subreddits favorisant les échanges entre eux.

Baptiste ADAM (ESIEE) Matthieu COMOY (D2SN)
Valentin FOARE (ESIEE) Corine TCHEUTGA (D2SN)

Si on se concentre sur la nature des liens, on peut voir que les liens négatifs se concentrent dans une grosse moitié de cette hyper communauté.

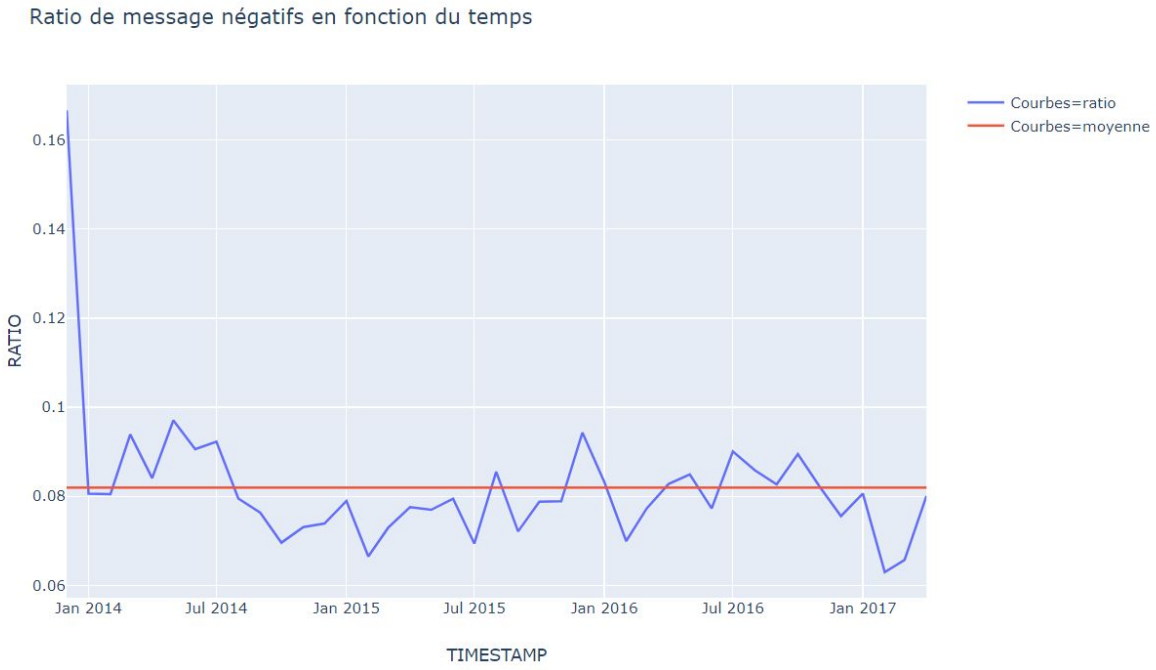


Baptiste ADAM (ESIEE) Matthieu COMOY (D2SN)
Valentin FOARE (ESIEE) Corine TCHEUTGA (D2SN)

Analyse et mesures

1. Les commentaires négatifs

Nous avons vu au début que le nombre de messages par mois augmentait avec le temps, mais est-ce que la proportion de messages négatifs augmente-t-elle aussi ?



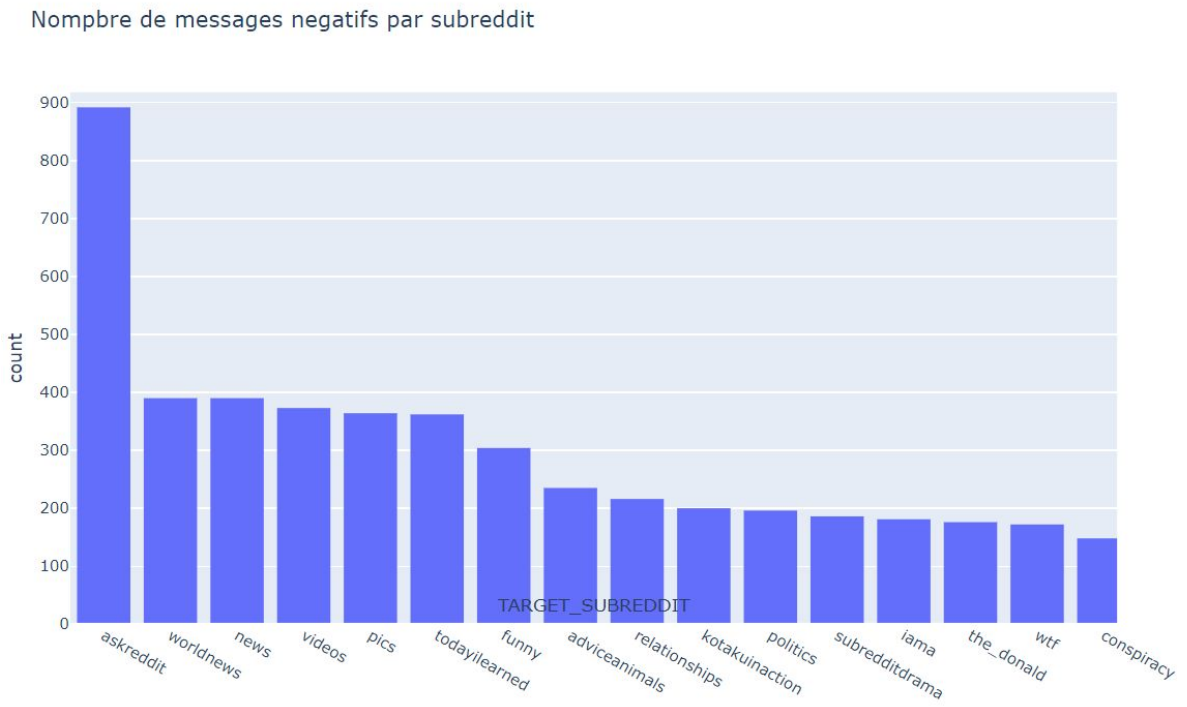
On peut voir ici que (à part pour le premier point qui ne correspond pas à un mois entier), le ratio de commentaire négatif varie très faiblement autour de 0.8, et ce alors même que le nombre de messages total augmente. Nous pouvons donc affirmer dans un premier temps que la communauté de Reddit ne se détériore pas au fur et à mesure que des gens la rejoignent.

Maintenant que nous savons que les messages négatifs ne se concentrent pas dans le temps, nous pouvons nous demander s' ils se concentrent sur certains subreddits. C'est effectivement le cas, les 3 subreddits avec le plus de messages hyperlinks négatifs sont *askreddit*, *worldnews* et *news*. Le premier en totalisant plus que les deux autres réunis.

Nous ne pouvons pas réellement apporter de preuves concrètes sur le pourquoi du comment de ce phénomène, mais il est assez facile de supposer.

En ce qui concerne *worldnews* et *news*, c'est plus que facile. Les médias ont de plus en plus tendance à se concentrer sur ce qui buzz pour faire du l'audience plutôt que de relayer des informations importantes rapidement. Et même si c'est loin d'être une vérité générale, cela a grandement diminué la confiance des gens et leur volonté à vouloir écouter les news.

Pour *askreddit*, c'est finalement tout aussi simple, la description du reddit est "a place to ask and answer thought-provoking questions". Il est maintenant assez facile de comprendre pourquoi des messages négatifs apparaissent au sujet de sujets qui sont justement là pour faire réagir.

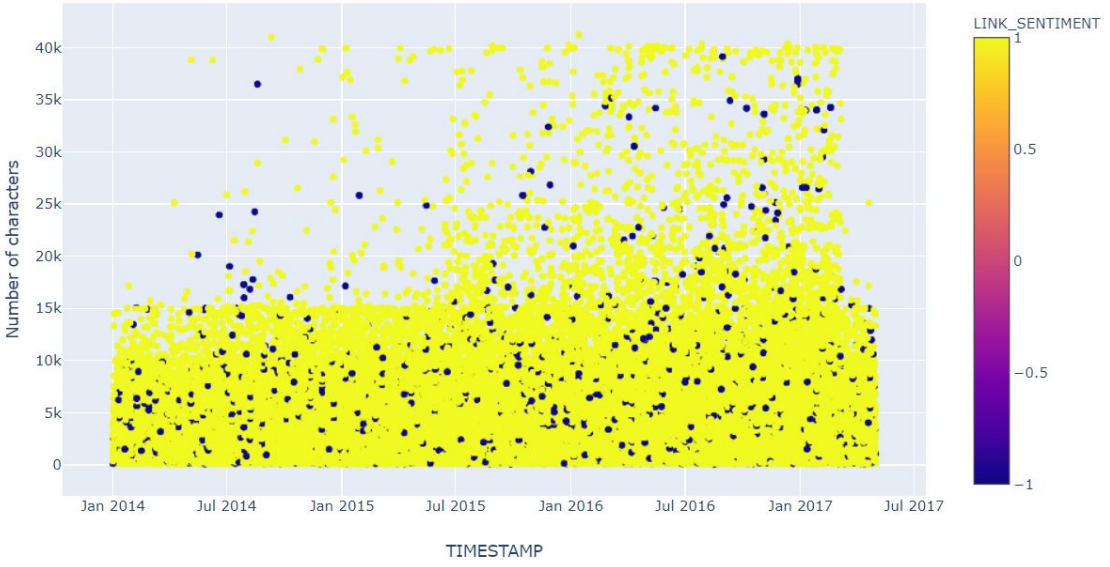


La nature des messages négatifs dans les deux cas sont très différents. Dans un cas, ce sont des sujets volontairement provocateurs alors que dans l'autre, c'est de la perte de confiance et de crédibilité d'un média. Une question subsidiaire que nous pourrions nous poser est donc la suivante : est-ce que ces messages négatifs ont une forme différente en fonction du sujet auquel ils s'adressent ?

Baptiste ADAM (ESIEE) Matthieu COMOY (D2SN)
Valentin FOARE (ESIEE) Corine TCHEUTGA (D2SN)

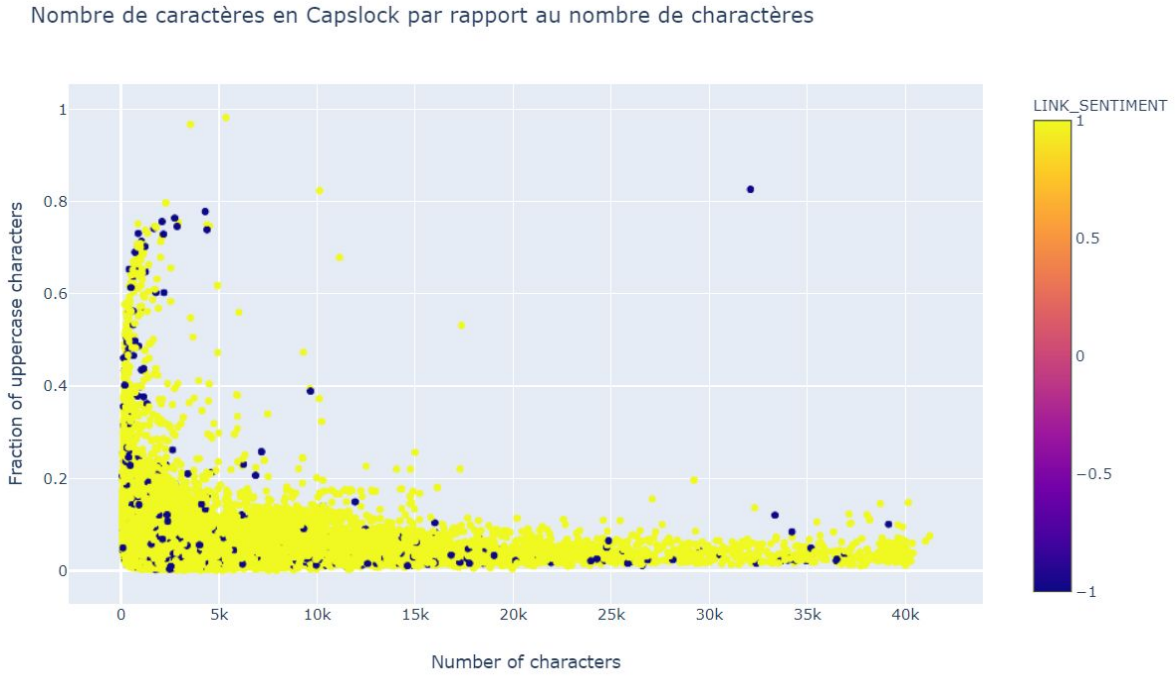
2. La nature de ses messages négatifs

Avant de décortiquer les messages négatifs entre eux, il serait bien de vérifier s' ils sont en eux même différents des messages positifs, s' ils sont plus agressifs, s' ils sont plus long, plus ponctué. Déterminer si la nature du message influence son contenu.



Un premier graph simple à regarder est le nombre de caractères. Ici, nous voyons qu'il y a beaucoup plus de messages a moins de 15 000 caractères mais que les messages négatifs sont plus ou moins répartis de façon homogène.

On peut ensuite se demander si les messages négatifs ont plus de caractères en capital, simulant à l'écrit l'action de crier. Cela se traduirait par des messages agressifs voire insultants.



On peut voir que les messages les plus courts ont tendance à avoir un plus gros ratio de caractères en capital, allant jusqu'à 80%. Cependant c'est aussi le cas des messages positifs donc ce n'est pas un signe distinctif. Il est ainsi possible d'assumer que les messages négatifs ne sont pas du type agressif.

De manière générale, il est assez difficile d'isoler les messages négatifs. Cela veut dire qu'ils viennent sous toutes les formes, des longs, des courts, avec des chiffres, avec beaucoup de phrases. Et surtout, cela veut dire que ses formes sont les mêmes pour tous les messages quelle que soit leur nature.

Baptiste ADAM (ESIEE) Matthieu COMOY (D2SN)
Valentin FOARE (ESIEE) Corine TCHEUTGA (D2SN)

Conclusion

Tous les subreddits ne sont pas en lien avec les autres, certains ont très peu d'interactions, certains ont des interactions répétées avec certains autres subreddits, créant des communautés de subreddits qui échangent entre eux. Mais il y a aussi un centre névralgique autour de *subredditdrama* qui crée une espèce de super communauté où les frontières entre les différentes communautés s'amoindrissent pour laisser place à une hyper connectivités entre les subreddits.

Il est tout naturel que les messages négatifs se concentrent donc dans cette super communauté. Mais malgré cela, ils se concentrent autour de quelques subreddits aux sujets controversés comme *askreddit*. Néanmoins, ces messages ne sont pas agressifs et gardent la forme qu'un message positif adopterait. Cela dénote un certain savoir vivre des *redditors* qui discutent de façon civilisée même pour donner un avis dépréciatif au lieu de céder à la tentation de l'anonymat d'internet et d'insulter ceux qui ne sont pas du même avis.

En plus de cela, nous avons vu que le nombre d'utilisateurs augmentait avec le temps, et malgré tout, le ratio de commentaire négatif (que nous avons établi civilisé) n'augmente pas (ni ne diminue). La communauté de reddit ne deviens donc pas plus toxique au fur et à mesure qu'elle grandit.

Annexes

Lien du dataset : **SNAP: Social network: Reddit Hyperlink Network**

PDF *Visualisation* : pour voir le réseau un peu plus en détail que les screenshots présentés.
PDF *Visualisation - Labels* : Le réseau avec le nom des subreddits.