HEC MONTRÉAL

Projet d'équipe : Comprendre et détecter le clickbait

Travail présenté à Monsieur Gilles Caporossi

Dans le cadre du cours

Analyse en données textuelles (MATH60621)

Par :

Heddier Alberto SOLER — 11272957

Krystel LAUDON — 11322759

Colin LLACER — 11318002

Hiver 2024

Le vendredi 12 avril 2024

Table des matières

I. Introduction 3
II. Méthodologie 3
III. Prétraitements et analyses 4
1. Prétraitements 4
2. Ajustements 5
3. Analyses 6
4. Jeux de variables et approche hybride 11
IV. Classificateurs 12
1. Modélisation 12
2. Résultats 12
3. Remarques 14
V. Conclusion 15
Bibliographie 17

I. Introduction

Dans l'ère numérique actuelle, le “clickbait” (ou “piège à clics) se dresse comme un outil puissant et redoutable pour capter l'attention en ligne. Comme le définit l’Office québécois de la langue française, le clickbait se caractérise par un « titre provocant ou intrigant qui pique la curiosité des internautes et les incite à cliquer pour en savoir plus, mais qui ne mène qu'à un contenu peu informatif et décevant. ». En regardant de plus près ces titres clickbait, nous remarquons que ces derniers semblent suivre un certain nombre de règles et disposer d’une structure particulière, conçues notamment pour optimiser leur impact sur les lecteurs.

Ainsi, si nous parvenons à identifier ces règles et structures avec suffisamment de précision, nous pourrions potentiellement concevoir des filtres efficaces pour contrer cette pratique. Autrement dit, en comprenant comment le clickbait fonctionne, nous pourrions développer des applications pour mieux naviguer en ligne de sorte que les utilisateurs soient moins susceptibles d'être trompés par des tactiques sensationnalistes.

II. Méthodologie

Pour mener notre étude sur le clickbait et identifier ses règles et structures sous-jacentes nous adoptons la méthodologie suivante. Tout d’abord, notre jeu de données[1] utilisé provient de la plateforme Kaggle. Il comprend 32 000 titres (16 000 pour chaque classe) provenant de divers sites d'information. Pour les titres clickbait, ces derniers proviennent de sites tels que "BuzzFeed", "Upworthy", "ViralNova", "Thatscoop", "Scoopwhoop" et "ViralStories". En ce qui concerne les titres non-clickbait, ils sont tirés de sites d'information plus dignes de confiance comme "WikiNews", "New York Times", "The Guardian" et "The Hindu". Le site de provenance n’est pas indiqué pour chaque titre, nous savons uniquement s’il s’agit d’un titre considéré provenant d'un site clickbait ou non.

Avant d'effectuer des analyses, nous avons procédé tout d’abord à un prétraitement des données pour les nettoyer et les mettre en forme à l’aide de plusieurs niveaux de traitement. Par la suite,

[1] Jeu de données Kaggle : https://www.kaggle.com/datasets/amananandrai/clickbait-dataset

plusieurs analyses ont été effectuées portant notamment sur la sémantique, la grammaire et la syntaxe. En parallèle, à partir d'un modèle sac de mots et de tests statistiques, nous identifierons les mots les plus discriminants entre nos titres clickbait et non-clickbait. Deux modèles de classification seront ensuite évalués suivant chaque jeu de variables à savoir un jeu composé uniquement d'indicateurs, un jeu composé seulement de mots les plus pertinents puis un jeu hybride mélangeant ces indicateurs et mots. Enfin, le modèle le plus performant sera alors retenu comme filtre anti-clickbait.

III. Prétraitements et analyses

1. Prétraitements

Dans le but de mieux comprendre ce qui caractérise les titres clickbait et traditionnels, nous avons décidé de comparer nos titres sur de très nombreux indicateurs. Ces indicateurs peuvent concerner la syntaxe, la grammaire ou encore le sens sémantique. Pour ce faire, nous avons choisi de les diviser en différents blocs, qui dépendent de leur niveau de prétraitement. Chaque traitement est appliqué au niveau du titre, tandis que les moyennes et autres comparaisons seront calculées plus tard.

Certains de nos indicateurs n'ont pas réellement besoin de prétraitement, puisqu'ils se servent des titres dans leur intégralité. C'est notamment le cas de la longueur du titre, de l'analyse de sentiment (avec VADER) ou de la part de majuscule. De même, nous calculons à cette étape une part pour chaque type de ponctuation qui apparaît dans l'échantillon : le point d'exclamation, le point d'interrogation, les virgules, les points et les guillemets.

Par la suite, nous effectuons un POS Tagging. Celui-ci nous permet de mieux comprendre le rôle grammatical de chaque élément du titre et de les compter. Ainsi, nous comptons pour chaque titre le nombre total d'éléments, le nombre de noms, le nombre d'adjectifs, le nombre d'adverbes, le nombre de verbes, le nombre de pronoms ainsi que le nombre de superlatifs. Chacun de ces décomptes est divisé par le nombre de caractères, afin d'obtenir un ratio indépendant de la longueur du titre.

Notre prochaine étape de prétraitement consiste à enlever la ponctuation ainsi qu'à normaliser chacun des titres. Cette étape intermédiaire est nécessaire avant de créer un sac de mots, et correspond au moment parfait pour analyser la longueur moyenne de nos mots. En effet, après cette étape, la ponctuation n'est plus gênante et les mots ont encore leur forme entière pour pouvoir effectuer l'analyse.

Notre dernier prétraitement à l'échelle des titres est une lemmatisation ainsi qu'un retrait des stopwords. Ces étapes sont essentielles pour réduire la dimension de notre jeu de données et retirer le bruit dans nos titres. Le sac de mots créé après lemmatisation pourra ainsi être utilisé pour nos nuages de mots, nos N-grams ainsi que dans notre recherche des mots les plus discriminants.

2. Ajustements

Bien que ces étapes semblent à première vue adéquates, les analyses ultérieures ont également fait ressortir certains problèmes. Par exemple, il semblerait que la liste de ponctuation disponible dans la librairie « string » ne soit pas suffisante, et nous avons dû ajouter certains caractères manuellement. De même, nous nous sommes rendu compte que les contractions n'étaient pas bien traitées par notre modèle. Nous avons ainsi utilisé la librairie « contractions », qui permet de les remplacer par leurs formes entières. Nous effectuons donc ce traitement avant tout autre, ce qui nous assure d'obtenir des phrases qui peuvent être prétraitées adéquatement.

Pour finir, nous nous sommes rendu compte que bien que des nombres apparaissaient dans de très nombreux titres, ils ne ressortaient pas dans nos nuages de mots car ils étaient tous différents. Nous avons donc décidé de les remplacer par un tag unique, qui marque simplement la présence d'un nombre quelconque. Cela n'a pas été trivial, notamment car certains gros nombres peuvent être écrits de manière différente (300000/300 000/300,000/300.000/…) ou être immédiatement suivis de symboles. Nous avons donc dû écrire une expression régulière capable de ne pas les considérer comme plusieurs nombres, et de les remplacer par un unique tag.

3. Analyses

Les analyses représentent le cœur de notre travail. En effet, bien que le projet se termine par un modèle de classification, c'est la volonté de comprendre ce qui caractérise le clickbait qui a motivé la réalisation de ce travail. L'ensemble des analyses qui vont suivre ont été effectuées en divisant le jeu de données en deux groupes selon le statut de la variable « clickbait », puis en calculant la moyenne des différents indicateurs individuels pour le groupe.

- Au niveau de la longueur du titre, nous avons remarqué une légère différence entre les deux groupes. Il semblerait ainsi que les titres clickbait soient en moyenne légèrement plus longs que les titres traditionnels, avec 59 caractères contre 53.

- Le ratio de majuscules est un des indicateurs les plus marquants, puisque nous retrouvons près de deux fois plus de majuscules dans les titres clickbait. Cela s'explique notamment par le fait que les sites clickbait ont tendance à mettre une majuscule au début de chaque mot, pour tenter d'attirer l'utilisateur.

- La longueur moyenne des mots a été utilisée comme un potentiel proxy pour la complexité du langage. Les résultats ne sont cependant par très probants, puisque les mots utilisés dans les titres traditionnels ne sont que très légèrement plus longs en moyenne (5,6 contre 5 caractères).

- Au niveau du sentiment moyen, nous nous sommes aperçus que les titres clickbait ont tendance à être légèrement positifs (0,10) tandis que les titres traditionnels sont légèrement négatifs (-0,11). Nous expliquons cela par le fait que les titres traditionnels tendent à rapporter des faits divers, et que ces derniers sont généralement négatifs.

- La diversité du vocabulaire est un autre indicateur ou l'on observe une vaste différence. Ainsi, alors que nous avons le même nombre d'observations pour nos deux groupes (16 000), nous retrouvons plus de 17 500 mots uniques dans les titres traditionnels, contre seulement environ 12 000 pour les titres clickbait. Nous pensons

que cela pourrait provenir d'un niveau de langage un peu plus simple, ou de la répétition de patterns reconnus comme efficaces pour attirer les clics.

– La ponctuation n'est pas toujours présente en grande quantité dans les titres, mais est extrêmement différente selon le groupe lorsque c'est le cas. Ainsi, nous observons 7 fois plus de points et 5 fois plus de virgules dans les titres traditionnels, des marqueurs classiques d'une phrase correctement construite. Au contraire, nous retrouvons 3 fois plus de guillemets dans les titres clickbait, qui semblent citer plus souvent quelque chose que les médias traditionnels. D'autres types de ponctuation ont également été analysés (interrogation, exclamation…) mais sont très rares en pratique.

– Pour finir, l'analyse grammaticale a également fait ressortir de fortes différences entre les groupes. Nous retrouvons ainsi 3 fois plus d'adjectifs dans les titres traditionnels, mais 3 fois plus d'adverbes dans les titres clickbait. La différence la plus marquée se situe au niveau des pronoms, que l'on retrouve 13 fois plus souvent dans les titres clickbait. Cela s'explique notamment par le fait que ces sites font généralement des appels à la personne dans le titre, pour attirer son attention (ex : « Vous Ne Devinerez Jamais… »). Par ailleurs, nous nous attendions à retrouver une quantité démesurée de superlatifs pour les titres clickbait, que l'on pensait très enclins à utiliser des formulations telles que « Le Meilleur (…) ». Cela n'est surprenamment pas le cas, et nous retrouvons une proportion identique à celle des titres traditionnels.

Par la suite, nous avons cherché à dessiner des nuages de mots. Ces derniers nous permettent de voir facilement quels mots apparaissent le plus régulièrement dans chacun des groupes, ainsi que de potentiellement dégager certains sujets de prédilection. C'est aussi grâce aux sacs de mots que nous avons pu itérativement ajuster notre prétraitement, en fonction des problèmes qu'ils faisaient apparaître.

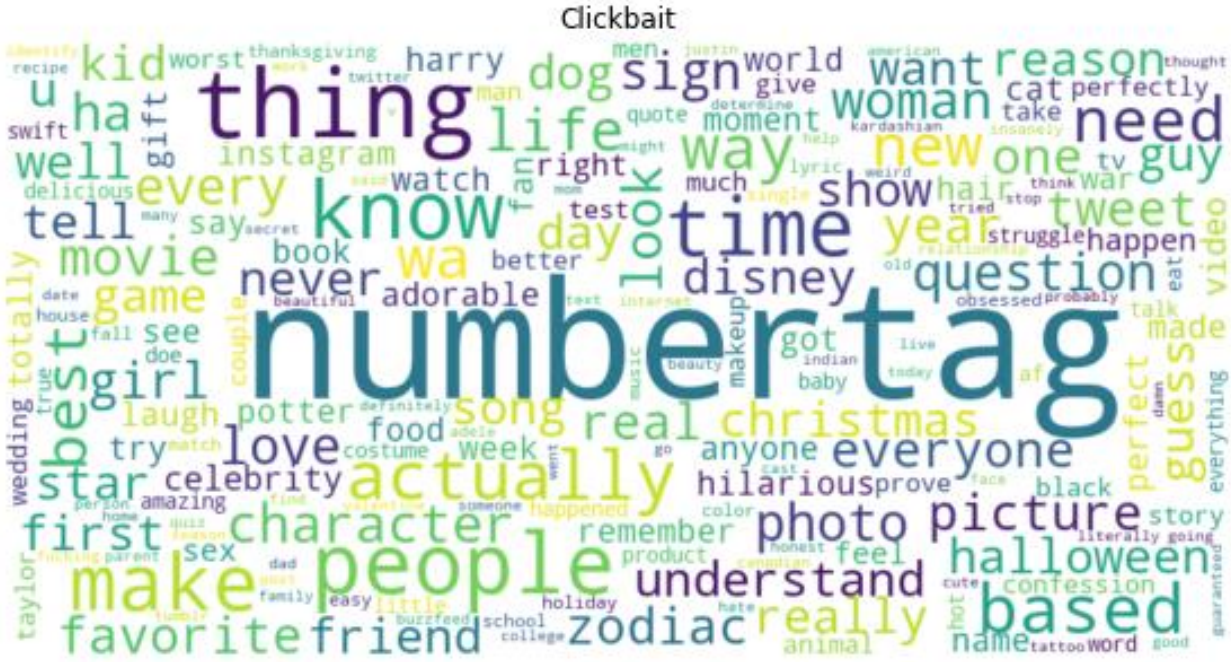

Figure 1 : Nuage de mot pour les titres clickbait

Figure 2 : Nuage de mot pour les titres traditionnels

Par ailleurs, puisque nous ne voulions pas uniquement nous intéresser aux mots seuls, nous avons complémenté cette analyse en nous intéressant également aux bigrammes et trigrammes les plus communs de chaque groupe. Nous avons effectué cette analyse avant et après lemmatisation, et avons trouvé que les N-grams avant retrait des stopwords étaient plus intéressants pour les titres clickbait (« are you » ; « NOMBRE things » ; « based on »). Ces stopwords servent à détecter certaines tournures de phrases très caractéristiques du clickbait. Au contrainre, les N-grams sans stopwords sont plus parlants pour les titres traditionnels, car les stopwords n'y apportent que peu d'information du fait des tournures de phrases plus variées.

L'analyse des nuages de mots ainsi que des N-grams a fait ressortir des sujets très différents entre nos deux groupes. En effet, il semblerait que les titres clickbait se concentrent beaucoup sur des listes numérotées (« 10 choses que » ; « 5 personnes qui... »). De plus, des sujets liés à l'ésotérisme comme l'astrologie semblent régulièrement abordés. On retrouve également un très grand nombre de références à la pop culture, notamment avec les bigrammes « Harry Potter », « Taylor Swift » ou « Star Wars » qui apparaissent régulièrement.

Les titres traditionnels quant à eux contiennent souvent des informations géographiques, qu'il s'agisse de pays ou de villes. On retrouve dans les 5 bigrammes les plus communs « New York » ou « New Zealand » par exemple. On y voit également abondamment le vocabulaire des faits divers ainsi que des avis de décès, notamment avec le trigramme le plus populaire : « die age NOMBRE ». Pour finir, nous retrouvons également beaucoup de mots liés à la politique et à la géopolitique, ce qui est particulièrement visible sur la figure 2.

Afin de pouvoir obtenir une classification plus robuste, et car il s'agit également d'une analyse intéressante, nous avons ensuite décidé de chercher quels mots étaient les plus discriminants entre nos deux groupes. Pour ce faire, nous avons utilisé le test du chi-carré sur chacun des mots observés, puis nous les avons triés par ordre décroissant de pouvoir discriminant.

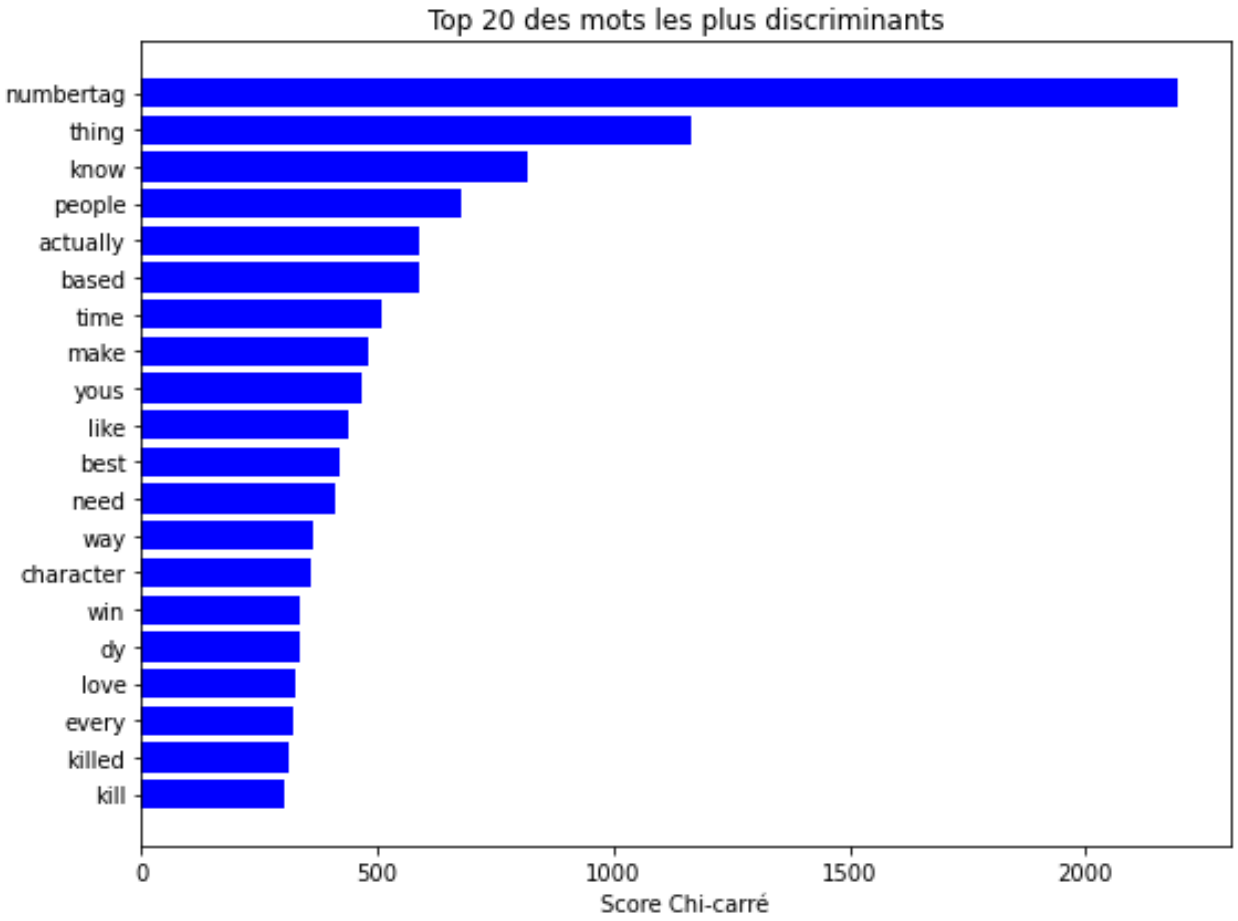

Figure 3: Liste des mots les plus discriminants

Bien que nous affichions ici seulement les 20 premiers, nous avons décidé de conserver les 100 mots les plus significatifs pour notre classification ultérieure. Notre analyse des mots les plus discriminants nous a conduit à effectuer trois observations :

- Malgré le fait que le tag pour le nombre soit aussi le token le plus commun dans les titres traditionnels, il est tellement surreprésenté dans les titres clickbait qu'il reste extrêmement discriminant dans notre cas.
- Les mots les plus discriminants tendent à être ceux que l'on retrouve régulièrement dans les titres clickbait (« thing », « know », « people », « based »). Cela est probablement une conséquence du vocabulaire moins varié observé précédemment.
- Le vocabulaire lié aux drames (« killed, « die », « crash »), souvent présent dans les titres traditionnels est plus secondaire, bien que tout de même discriminant.

4. Jeux de variables et approche hybride

Une fois nos analyses effectuées, nous avons divisé nos variables en trois jeux, qui servent notamment à vérifier la pertinence de notre approche. Ainsi, nous définissons un premier jeu de variable qui contient l'ensemble des indicateurs disponibles au niveau individuel : le nombre de mots, la ponctuation, le nombre de majuscules, la longueur moyenne des mots, le sentiment moyen, ainsi que les différents tags issus de l'analyse morphosyntaxique. Le second jeu de variables contient lui les 100 mots les plus discriminants selon le chi-carré. Nous nous sommes restreints à 100 variables afin de limiter la dimensionnalité, mais aussi afin d'éviter de surajuster nos modèles à certains sujets de prédilections que pourrait avoir un seul site du groupe.

Notre troisième et dernier jeu de variables est une combinaison des deux premiers. L'idée sous-jacente est de nous assurer que le modèle est robuste, en ne nous basant ni seulement sur les mots, ni seulement sur les indicateurs. En effet, comme évoqué précédemment, le jeu de données n'est composé que de 6 sites internet pour les titres clickbait et de 4 pour les titres normaux. Il y a ainsi un vrai risque de nous surajuster à des sujets propres à un site plutôt qu'à son groupe. Incorporer les indicateurs permet de mitiger cet effet, et améliore notre capacité à généraliser.

Par ailleurs, nous avons vu que certains indicateurs sont très différents selon le type de titre. On a ainsi un risque de classifier comme clickbait un article provenant d'un journal réputé qui reprendrait certaines caractéristiques présentes dans les titres clickbait (guillemets, présence d'un nombre, majuscules…). Utiliser les mots en plus des indicateurs permet de mitiger ce problème en incorporant un autre type d'information et fonctionne très bien, comme nous le verrons plus tard au moment des tests. Nous optimiserons également le seuil de classification de nos modèles, afin de minimiser autant que raisonnable le nombre de faux positifs prédits par notre modèle.

Nous assurer de trouver des variables suffisamment variées et discriminantes pour lutter contre le faible nombre de sources a été une des principales difficultés de ce projet, notamment car la courte taille des titres ne permet pas d'utiliser certains indicateurs à l'échelle du titre (taille du vocabulaire, niveau de lecture…). Nous ne pouvions pas non plus utiliser certaines techniques, comme le TF-IDF à cause de la faible longueur d'un titre.

IV. Classificateurs

1. Modélisation

Notre démarche de modélisation a été aussi exhaustive que possible. Bien que de multiples modèles aient été essayés, nous avons fini par comparer deux modèles principaux : la régression logistique pour son explicabilité et la forêt aléatoire pour sa capacité à modéliser des relations plus complexes entre nos variables. Ces modèles ont ensuite été entraînés sur nos trois jeux de variables distincts : un ensemble basé sur des indicateurs clés, un autre sur les 100 mots les plus discriminants obtenus par le test du chi-carré et un dernier jeu combinant les deux pour tenter d'améliorer notre capacité de généralisation. Ces modèles ont ensuite été évalués sur un échantillon de validation, tout en veillant à maintenir une répartition équilibrée entre les classes.

Un point important de notre modélisation a été l'optimisation des seuils de classification. Nous avons ainsi cherché un seuil (avec notre échantillon de validation) qui maximise le score F-beta (beta = 0,5). L'idée derrière cette décision est de ne pas nous concentrer uniquement sur l'exactitude, notamment car toutes les erreurs n'ont pas nécessairement le même impact dans notre cas. L'utilisation du score F-beta plutôt qu'un simple score F1 nous permet de mettre moins de poids sur le rappel par rapport à la précision, et donc de donner une importance plus grande aux faux positifs que nous voulons éviter tant que possible. Par ailleurs, nous avons préféré l'utilisation du score F-beta à la précision seule, car le seuil sélectionné devient sinon trop élevé et le filtre inutile car pas assez strict. Nous avons préféré un compromis.

2. Résultats

Modèle	Exactitude	F-beta (beta=0.5)	Taux de faux positifs
Forêt aléatoire	0,930	0,938	5,5%
Régression logistique	0,858	0,887	7,2%

Tableau 1 : Résultat pour le jeu de variables "indicateurs"

Les résultats de nos modèles sont à la fois prometteurs et révélateurs. Avec notre jeu de variable basé sur les indicateurs, la forêt aléatoire a excellemment performé, atteignant une exactitude de 93%, et un score F-beta de 0,94. Elle surpasse ainsi la régression logistique, qui a toutefois montré des performances plus que respectables. Nous pouvons cependant voir que l'utilisation des seuls indicateurs tend à s'accompagner de taux de faux positifs relativement élevés (7,2% pour la forêt et 5,5% pour la régression logistique).

Modèle	Exactitude	F-beta (beta=0.5)	Taux de faux positifs
Forêt aléatoire	0,824	0,886	2,8%
Régression logistique	0,818	0,885	2,1%

Tableau 2 : Résultat pour le jeu de variables "mots discriminants"

Pour notre jeu de variable content les 100 mots les plus discriminants, nous pouvons voir que la forêt aléatoire et la régression logistique ont affiché des performances comparables. La performance globale est moins bonne que pour le jeu contenant les indicateurs (F-beta d'environ 0,88 pour les deux modèles) mais le taux de faux positifs est également bien plus faible (2,8% pour la fôret, 2,1% pour la régression logistique).

Modèle	Exactitude	F-beta (beta=0.5)	Taux de faux positifs
Forêt aléatoire	0,952	0,961	3,2%
Régression logistique	0,911	0,943	2,7%

Tableau 3 : Résultat pour le jeu de variables "combiné"

Finalement, c'est le jeu de données combiné qui a révélé la véritable efficacité de la forêt aléatoire avec une exactitude de 95%, un score F-beta de 0,96 et un taux de faux positifs d'environ 3%. Nous pouvons voir dans le tableau ci-dessus que la performance de la régression logistique avec le jeu de variables combiné reste tout de même plus qu'honorable, et qu'elle pourrait être utilisée par les utilisateurs pour qui l'explicabilité du modèle est importante.

Cette performance élevée confirme ainsi la pertinence de notre approche hybride, puisqu'elle nous permet une performance supérieure tout en maintenant un taux de faux positifs plus raisonnable

qu'avec les indicateurs seuls. Ainsi, après analyse de ces résultats, nous avons décidé de retenir la forêt aléatoire avec le jeu de données combiné comme modèle privilégié pour la détection de clickbait.

3. Remarques

Nous avions initialement souhaité explorer les modèles k-NN et Naive Bayes vu en cours, afin de pouvoir prendre en compte l'asymétrie d'information. Nous avons dû écarter le k-NN car nous avons 32 000 observations et 114 variables. Le grand nombre d'observations rendrait l'algorithme beaucoup trop lent au moment de l'inférence, tandis que la haute dimensionnalité risquerait de le rendre peu performant. Concernant le Naive Bayes, il aurait pu facilement être utilisé dans le cas où seuls les mots auraient été utilisés. Nous avons cependant également les indicateurs dans notre modèle final, et deux modèles distincts auraient dû être effectués selon le type des variables (binaires ou continues). La distribution gaussienne semblait également difficile à justifier pour certaines de nos variables continues. Ainsi, par élimination et dû à la très bonne performance/flexibilité de nos forêts aléatoires, nous avons préféré nous contenter de ces modèles, malgré qu'ils ne prennent pas en compte l'asymétrie d'information.

Par ailleurs, nous avons voulu tester la robustesse de nos modèles finaux. Nous voulions notamment savoir si aborder un thème « typé clickbait » risquait de classifier un article légitime comme aguicheur. Nous voulions également tester la même chose avec un titre traditionnel qui reprendrait exceptionnellement certaines caractéristiques surreprésentées dans les titres clickbait (beaucoup de majuscules par exemple). Le tableau ci-dessous rend compte de nos résultats.

Titre essayé	Classification par le modèle
« 10 Things To Know About Road Accidents »	Clickbait
« Ten dead in road accident »	Non-clickbait
« 10 Dead In Road Accident? »	Non-clickbait
« Why are zodiac signs so popular amongst younger generations? »	Non-clickbait
« What You Should Know About Zodiac Signs »	Clickbait

Ces essais semblent confirmer la validité de notre approche hybride. Nous pouvons voir dans le troisième exemple que la présence de caractéristiques surreprésentées dans les titres clickbait (nombre, majuscules, interrogation) ne suffit pas à faire passer un titre que l'on considèrerait légitime comme clickbait. De la même manière, le premier exemple nous montre que la présence d'un sujet plutôt associé aux titres légitimes (accidents de la route) n'est pas une condition suffisante pour qu'un titre aguicheur passe au travers de notre filtre.

Pour ce qui est des sujets très abordés dans les titres clickbait, nous voyons qu'ils ne suffisent pas seuls à faire classifier un titre comme clickbait. Un journal traditionnel peut parler des signes du zodiaque sans être détecté comme clickbait (4ème essai) tandis que le même thème abordé avec une structure type clickbait sera lui filtré (5ème essai). Ces résultats semblent ainsi valider la robustesse de l'approche hybride et d'un seuil de classification visant à limiter le nombre de faux positifs.

V. Conclusion

En conclusion, notre étude approfondie sur les titres clickbait nous a permis de mieux comprendre les mécanismes sous-jacents à cette pratique. Elle a révélé des distinctions significatives entre les titres clickbait et traditionnels, notamment sur leur structure grammaticale et syntaxique où de très fortes disparités ont pu être soulignées.

Nous avons également pu remarquer que les sujets abordés sont également très différents entre les deux groupes, et que certains mots peuvent donc servir à distinguer les deux groupes. Nous avons par exemple remarqué que les titres clickbait portent souvent sur des listes, sur l'ésotérisme et la pop culture, tandis que les titres traditionnels abordent fréquemment des faits divers et parlent majoritairement de politique et de géopolitique.

Grâce à cette compréhension approfondie, nous avons pu déterminer un modèle de classification performant (forêt aléatoire) résultant d'une approche hybride qui combine à la fois l'analyse des mots et des indicateurs spécifiques. Cette approche nous permet d'obtenir un filtre de classification de clickbait robuste et généralisable, qui pourrait aider les utilisateurs à mieux naviguer en ligne et réduire leur susceptibilité d'être trompés par des tactiques sensationnalistes.

Finalement, nous pouvons noter que bien que notre approche semble fonctionner lors de nos essais avec des titres créés de toutes pièces, il serait préférable de diversifier le jeu de données si un tel filtre venait à être implanté en pratique. En effet, la plupart des difficultés rencontrées lors de ce projet étaient liées à notre volonté de généraliser pour tous les sites clickbait, et ce malgré un faible nombre de sources dans notre jeu de données. Nous pensons avoir correctement identifié les grandes tendances du clickbait, mais sommes convaincus que la performance de notre filtre pourrait encore être améliorée en introduisant plus de diversité dans les sites utilisés pour la création du jeu de données. Un tel filtre devrait également être actualisé régulièrement, au cas où les tendances dans les sujets abordés viendraient à changer.

Bibliographie

Anand, A. (2020, April 18). *Clickbait dataset*. Kaggle. https://www.kaggle.com/datasets/amananandrai/clickbait-dataset