DÉTECTION DES FAUX AVIS SUR LES PLATEFORMES EN LIGNE

A journey to Data Scientist

Membres d'équipe:
Rudresh MISHRA
Raymond KLUTSE
Runlu QU
Jhonatan TORRES

Contenu

1. Validation de l'authenticité des avis

L’authenticité des avis encourage la crédibilité des plateformes en ligne. Les plateformes en ligne actuellement doivent filtrer les avis qu'ils montrent à leurs clients dans le but d'accroître la confiance et d'éviter les scandales susceptibles d'affecter leur crédibilité. Les grandes plateformes disposent déjà d'algorithmes pour relever ce défi. Cependant, ces algorithmes sont confidentiels et restent une nécessité pour les jeunes plateformes.

1.1 Rôle des avis dans la prise de décision des consommateurs

Les avis des consommateurs sont devenues une partie importante pour la commerce électronique et aussi du comportement de l'acheteur en général. Ces avis ont but d’éclairer la décision d’un nouveau consommateur d’utiliser un service ou non. À l’heure actuel, les consommateurs veulent toujours savoir les expériences des autres avant de visiter un restaurant, de réserver un hôtel, de consulter un médecin ou d'acheter une voiture. Selon [1], 78 % des consommateurs font autant confiance aux avis en ligne qu'aux recommandations personnelles et 86 % des consommateurs lisent les avis en ligne avant de visiter une entreprise. De la même façon, 91 % des consommateurs affirment que les avis positifs les rendent plus susceptibles d'utiliser une entreprise [2].

Au fil des ans, les consommateurs ne dépendent plus uniquement des messages publicitaires comme le marketing traditionnel, mais se tournent vers d'autres sources d'information, en particulier les avis en ligne. À partir de 2019, les acheteurs en ligne aux Etats-Unis s'attendent à un nombre significatif d'avis lorsqu'ils regardent un produit - le nombre moyen d'avis attendus était de 112. Par ailleurs, les plus jeunes s'attendent à plus d'avis que les plus âgés [3].

1.2 Impact des avis sur les entreprises

D'après une étude par Harvard Business School, les notes plus élevé peuvent générer à peu près 9% de revenue pour les restaurants indépendants. De plus, selon une enquête faite par BrightLocal, 57% des consommateurs fréquentent un service s’ils ont 4 notes ou plus [4]. Dans la même article, il indiquent que 68 % des consommateurs sont plus susceptibles d'acheter auprès d'une entreprise dont les avis sont positifs et 40 % des consommateurs n'utiliseront pas une entreprise dont les avis sont négatives.

1.3 Défis actuels des plateformes

Comme les avis en ligne se répandent rapidement, certaines business et gens ont profité du temps pour créer des faux avis afin d'attirer plus de clients. De la même façon, des concurrents d'une entreprise peuvent aussi en créer afin de compromettre leurs affaires. Les utilisateurs pourraient également poster des faux avis afin de gagner le statut dans la communauté [4]. Selon BrightLocal, 74 % des consommateurs ont lu un faux avis au cours de la dernière année, bien qu'ils ne soient pas toujours faciles à repérer. Beaucoup de plateformes d'avis peinent à capturer les faux avis sur leurs plateformes en raison de l'absence de techniques approprié de filtrage. Ils s'appuient sur l'approche traditionnelle de contrôle manuel du profil utilisateur afin d'authentifier les avis [6]. Afin de résoudre le problème, des entreprises comme Yelp, Tripadvisor et Airbnb ont créé leurs propres algorithmes de filtrage pour supprimer les faux avis. Cependant, les détails du fonctionnement de ces algorithmes ne sont pas accessibles au grand public et les nouvelles plateformes sont susceptibles à recevoir tels faux avis.

1.4 Definition des faux avis

Dans le cadre de ce projet, un faux avis est celui qui trompe délibérément les lecteurs ou les systèmes d'exploration de l'opinion en donnant des avis positifs non méritées à certaines entreprises cibles afin de promouvoir ses entreprises et/ou en donnant des avis négatives injustes ou malveillantes à certaines entreprises afin de nuire à leur réputation.

1.5 Problématique

Assurer l'authenticité des avis sur les plateformes en ligne de l'industrie hôtelière afin de garder leur crédibilité.

1.6 Objectif

Créer un algorithme pour identifier les faux avis qui peut être mis en œuvre sur les plateformes de l'industrie hôtelière.

1.7 Produit

Notre produit est un algorithme pour identifier les faux avis. Avec la mise en œuvre de notre algorithme, une plateforme peut masquer ou signaler un avis comme faux, ainsi garantissant un niveau de confiance minimum sur les avis indiqués comme vrais.

1.8 Clients potentiels

Notre produit vise à atteindre les jeunes plateformes d'avis en ligne pour l'industrie hôtelière. Lorsqu'une plateforme commence à se développer, il est difficile de suivre toutes les avis d'utilisateurs. En appliquant cet algorithme, ces plateformes peuvent filtrer les faux avis et garantir à ses utilisateurs la fiabilité de leurs plateformes.

2. Ressources

Le jeu des données disponibles a été recueilli auprès de Yelp et avait été utilisé pour valider le fonctionnement de l'algorithme de filtre Yelp [7]. Bien que le filtre antifraude de Yelp ne soit pas parfait (ainsi «proche» ground truth), il semble d'avoir produit des bonnes résultats [8]. Il s'agit d'un modèle qui a été analysé profondément et pourrait être utilisé par les entreprises en croissance dans ce domaine pour offrir des avis plus fiables.

Le jeu des données utilisé est accessible via l'API de Yelp. Cependant, le processus d'exploration de données (pour extraire les données particulières aux hôtels et aux étiquettes correspondantes) a été fait par Bing Liu, membre de l'Université d'Illinois [9], qui les a gentiment mis à notre disposition.

2.1 Description de données

Dans le jeu des données actuel, nous avons 688329 avis pour les hôtels des États Unis. Ce jeu des données peut être fusionné et complété au cas où il serait nécessaire. Des tableaux supplémentaires contenant des renseignements à l'intention des avis (5123 enregistrements) et des hôtels (283086 enregistrements) sont également fournis. Certains enregistrements pourraient ne pas être nécessaires dans ces deux tableaux.

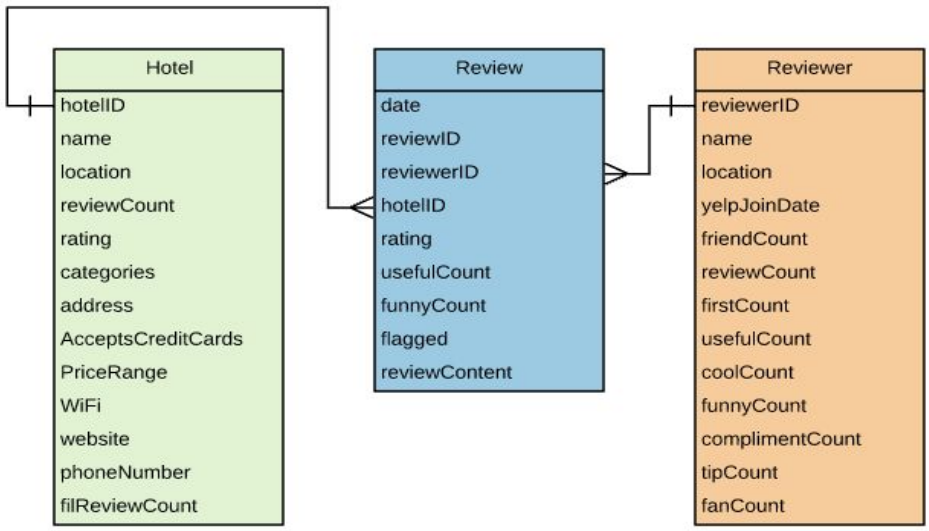

Fig 1. Relations entre les données

Un exemple de chacun des tableaux est présenté ci-dessous indiquant la signification, la pertinence et le type de données de chaque variable. Les tableaux correspondent dans leur ordre aux: 1) avis, 2) utilisateurs et 3) hôtels.

Nom	Signification	Example	Pertinence	Type
date	Date d'avis	6/8/2011	Oui	date
reviewID	Identifiant de l'avis	MyNjnxzZVTPq	Identifier	string
reviewerID	Identifiant de l'utilisateur	IFTr6_6NI4CgCVavIL9k5g	Relationnel	string
reviewContent	Texte de l'avis	Let me begin by saying that …	Oui	string
rating	Évaluation de l'avis	5	Oui	int
usefulCount	Nombre d'utilisateurs qui trouvent l'avis utile	18	Peut-être	int
coolCount	Nombre d'utilisateurs qui trouvent l'avis sympa	11	Peut-être	int
funnyCount	Nombre d'utilisateurs qui trouvent l'avis drôle	28	Peut-être	int
flagged	Vrai (Y) ou faux (N)	N	Oui	string
hotelID	Identifiant de l'hotel	tQfLGoolUMu2J0igcWcoZg	Relationnel	string

Nom	Signification	Example	Pertinence	Type
reviewerID	Identifiant de l'utilisateur	yevHGEUQQmnVlBXIrJ885A	Relationnel	string
name	Nom de l'utilisateur	Kevin T.	Non	string
location	Emplacement de l'utilisateur	Oconomowoc, WI	Peut-être	string
yelpJoinDate	Date d'enregistrement	05/01/11	Peut-être	string
friendCount	Nombre d'amis	4	Non	int
reviewCount	Nombre d'avis écrits	86	Oui	int
firstCount	Nombre de premiers avis pour un hotel	3	Non	int
usefulCount	Nombre d'avis utiles envoyés par l'utilisateur	129	Non	int

coolCount	Nombre d'avis sympas envoyés par l'utilisateur	47	Non	int
funnyCount	Nombre d'avis drôles envoyés par l'utilisateur	31	Non	int
complimentCount	Nombre de compliments reçus par l'utilisateur	12	Non	int
tipCount	Conseils écrits par l'utilisateur	0	Non	int
fanCount	Nombre de fans	1	Non	int

Nom	Signification	Example	Pertinence	Type
hotelID	Identifiant de l'hotel	pSLh_XyV_3QS1hNsBOGHiQ	Relationnel	string
name	Nom de l'hotel	Old Chicago Inn	Non	string
location	Emplacement de l'hotel	Old Chicago Inn - Lakeview - Chicago, IL	Peut-être	string
reviewCount	Nombre d'avis reçus	1	Oui	int
rating	Évaluation moyenne	3	Oui	float
categories	Catégories de l'entreprises	Event Planning & Services, Hotels…	Non	categorie
address	Adresse de l'hotel	3222 N Sheffield Ave (between Belmont…	Non	string
AcceptsCreditCards	Accepte les cartes de crédit	Yes	Non	string
PriceRange	Échelle des prix	$$	Non	object
WiFi	WiFi	Free	Non	string
webSite	Site Internet	http://www.oldchicagoinn.com	Non	string
phoneNumber	Numéro de téléphone	(773) 245-0423	Non	string
filReviewCount	Nombre d'avis filtrés	5	Non	int64

2.2 Distributions de données

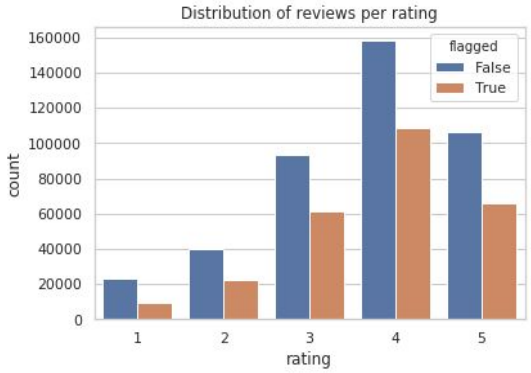

Fig 2. Répartition des avis par note

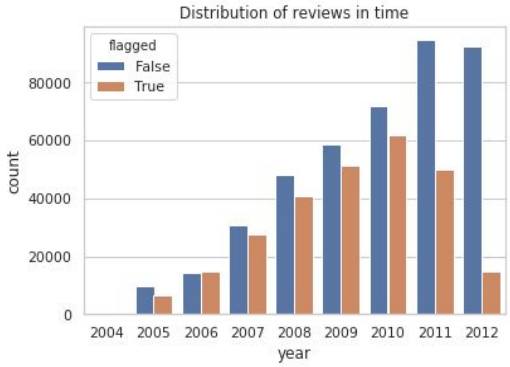

Fig 3. Répartition des avis dans le temps

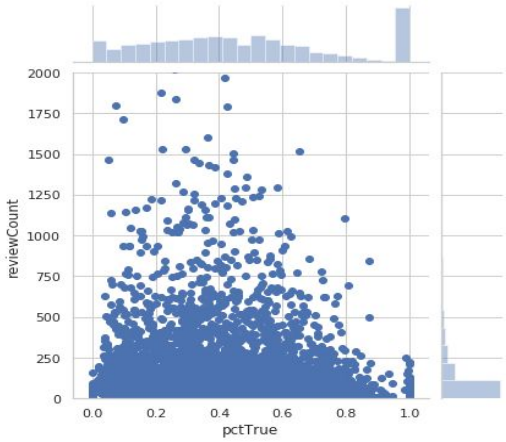

Fig 4. Scatterplot de le nombre d'avis par rapport au pourcentage d'avis réels par utilisateur

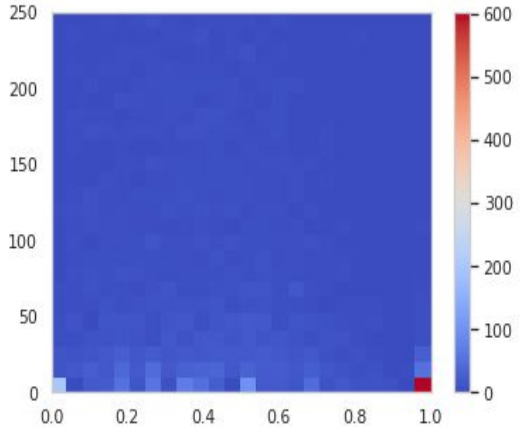

Fig 5. Heatmap de le nombre d'avis par rapport au pourcentage d'avis réels par utilisateur

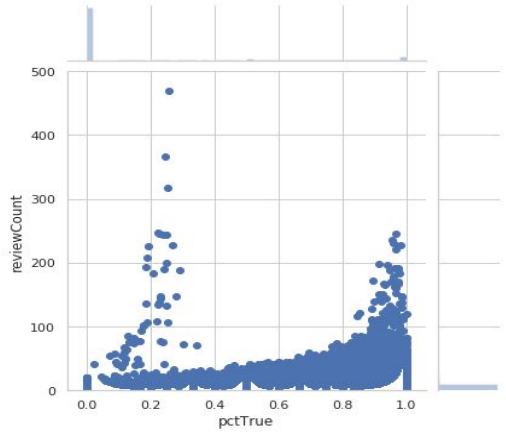

Fig 6. Scatterplot de le nombre d'avis par rapport au pourcentage d'avis réels par hôtel

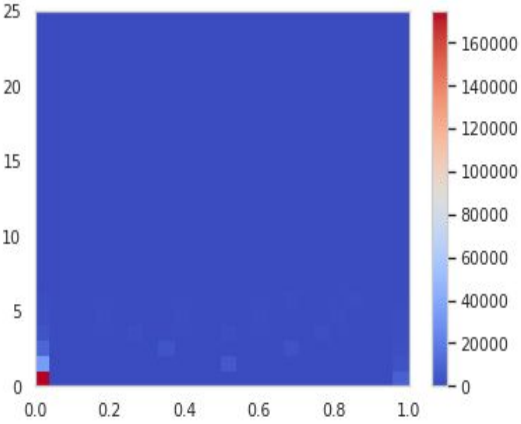

Fig 7. Heatmap de le nombre d'avis par rapport au pourcentage d'avis réels par hôtel

- A partir de la figure 2, nous identifions une forte tendance de notes élevées. La relation entre les vrais et les faux avis est cependant plus forte dans les notes inférieures.
- D'après la figure 3, on observe qu'il y a une quantité croissante des avis et que le nombre des faux avis augmente par rapport aux vrais avis au fil du temps.
- Les figures 4 et 5 montrent qu'un utilisateur peut faire de faux et vraies avis. Il existe un grand nombre d'utilisateurs qui ne font que des avis vrais ou faux, mais la plupart sont situés dans une position intermédiaire.
- Les figures 6 et 7 montrent un comportement similaire à celui des utilisateurs. On peut également observer que la présence d'un seul avis sur un hôtel est associée au fait qu'il est faux.

2.3 Validité de données et hypothèses

Dans [7], certains algorithmes habituellement utilisés pour identifier le comportement de spam ont été testés en utilisant ces données. Les résultats montrent qu'il existe une corrélation directe. Il est possible d'extraire et d'utiliser différentes variables des avis en appliquant des différentes algorithmes d'apprentissage automatique pour identifier l'authenticité de l'avis.

Nous nous concentrerons sur le comportement des utilisateurs et l'historique de l'hôtel pour compléter les caractéristiques de chaque avis. Les algorithmes seront appliqués au nouvel ensemble de données généré.

2.4 Risques

Nous considérons les labels fournies comme le «ground truth». Cependant, ces labels sont le résultat d'un modèle proposé par Yelp qui peut lui-même ne pas accomplir la tâche proposée.

3. Extraction de caractéristiques

3.1 Sélection et nettoyage des variables

Les variables identifiées comme pertinentes dans les tableaux précédents ont été sélectionnées et les autres ont été éliminées. Certaines variables ont été modifiées pour faciliter le traitement, par exemple les dates.

3.2 Construction de la base de données des travaux

Dans les tableaux d'utilisateurs et d'hôtels, l'historique des avis est généré sous forme d'un tableau. Étant donné que la forme change pour chacun en fonction de son activité, des variables caractéristiques de ces séquences sont prises, telles que les quartiles, l'écart type et la moyenne.

Enfin, les différentes variables sont intégrées dans un tableau unique qui alimente nos modèles et reflète le comportement de l'utilisateur qui fait l'avis et de l'hôtel qui la reçoit.

Le résultat final de notre traitement donne un tableau qui comprend les variables suivantes:

3.3 Perspectives

Étant donné que nos données étaient fortement biaisées en faveur des faux avis, nous avons échantillonné au hasard des vrais et faux avis à partir du jeu de données afin de l'équilibrer. Des

algorithmes comme les forêts de décisionnels et l'arbre de décision s'attendent au jeu de données équilibrés.

Nom	Signification	Example
usefulCount	Nombre d'utilisateurs qui trouvent l'avis utile	18
coolCount	Nombre d'utilisateurs qui trouvent l'avis cool	11
funnyCount	Nombre d'utilisateurs qui trouvent l'avis amusant	28
flagged	Étiquette	FALSE
reviewContent_compound	note de sentiment compound sur l'avis des textes	6.881
reviewContent_pos	note de sentiment positif sur l'avis des textes	0.09
reviewContent_neg	note de sentiment négatif sur l'avis des textes	83
reviewContent_neu	note de sentiment neutral sur l'avis des textes	827
log_wordsCount	logarithmique du nombre de mots	604.025
rating_1	colonne binaire indiquant la note 1	0
rating_2	colonne binaire indiquant la note 2	0
rating_3	colonne binaire indiquant la note 3	0
rating_4	colonne binaire indiquant la note 4	0
rating_5	colonne binaire indiquant la note 5	1
weekDay_categorized_0	note la colonne binaire faite le premier jour de la semaine	0
weekDay_categorized_1	note la colonne binaire faite le deuxième jour de la semaine	0
weekDay_categorized_2	note la colonne binaire faite le troisième jour de la semaine	0
weekDay_categorized_3	note la colonne binaire faite le quatrième jour de la semaine	0
weekDay_categorized_4	note la colonne binaire faite le cinquième jour de la semaine	0
weekDay_categorized_5	note la colonne binaire faite le sixième jour de la semaine	0
weekDay_categorized_6	note la colonne binaire faite le septième jour de la semaine	1
hote_rating_overall	Note moyenne de l'hôtel	3
hote_count_overall	Nombre d'avis effectués par l'hôtel	15
hote_pct_true	Pourcentage d'avis par note pour hôtel	333.333
hotel_pct0	Pourcentage d'avis effectués le même jour	285.714
hotel_gap_zero_counts	Nombre d'avis effectués le même jour	238.629
hotel_ratings_counts1	colonne binaire indiquant la note 1	6
hotel_ratings_counts2	colonne binaire indiquant la note 2	0
hotel_ratings_counts3	colonne binaire indiquant la note 3	2
hotel_ratings_counts4	colonne binaire indiquant la note 4	2
hotel_ratings_counts5	colonne binaire indiquant la note 5	5
hotel_meanTime_norm	Moyenne de la distribution normale des avis de l'hôtel	180.454
hotel_stdTime_norm	Écart type de la distribution normale des avis de l'hôtel	197.142
rating	Évaluation de l'avis	483.333
reviewCount	Nombre d'avis effectués pour l'utilisateur	6
pctTrue	Pourcentage de vrais avis pour utilisateur	0
reviewer_pct0	Pourcentage d'avis effectués le même jour	0.8
reviewer_gap_zero_counts	nombre d'avis effectués le même jour	238.629

ratings_counts1	colonne binaire indiquant la note 1 pour utilisateur	0
ratings_counts2	colonne binaire indiquant la note 2 pour utilisateur	0
ratings_counts3	colonne binaire indiquant la note 3 pour utilisateur	0
ratings_counts4	colonne binaire indiquant la note 4 pour utilisateur	1
ratings_counts5	colonne binaire indiquant la note 5 pour utilisateur	5
meanTime_norm	Moyenne de la distribution normale des avis de l'utilisateur	0.2
stdTime_norm	Écart type de la distribution normale des avis de l'utilisateur	0.4

4. Modélisation

4.1 Description

Nous voulons rendre notre produit transparent afin de pouvoir justifier les résultats à l'entreprise (si demandé). Nous remarquons qu'il est important de savoir pourquoi l'algorithme a marqué un avis comme faux ou vrai en fonction des caractéristiques du comportement de l'utilisateur et l'hôtel.

Ayant considéré cela, nous avons choisi d'utiliser la forêt aléatoire et l'arbre de décision. Ils sont capables de travailler avec une quantité considérable de variables comme celle que nous avons dans ce problème et d'extraire un sens du résultat en identifiant les variables les plus importantes dans la prise de décision.

4.2 Exigences

L'entrée du modèle est un vecteur composé des caractéristiques comportementales de l'utilisateur, l'hôtel et l'avis écrit par l'utilisateur à l'hôtel. Il serait préférable d'avoir toutes les variables mentionnées dans le dernier tableau mentionné.

La sortie de l'algorithme sera une valeur binaire avec *false* pour indiquant les faux avis et *true* indiquant les vrais avis (0 ou 1).

4.3 D'autres modèles tentés

Nous avons pensé que la structure du texte pourrait inclure des relations significatives pour identifier la véracité d'un avis. La différence de taille des textes et la grande taille de certains exemplaires ont rendu la tâche difficile. Nous avons mis en œuvre l'apprentissage supervisé basée uniquement sur le texte en utilisant le modèle de réseau de convolution de graphes, mais la précision du modèle était faible. Le modèle GCN texte a donné un taux de recall de 64 %, ce qui montre que le résultat obtenu sur le texte est pire que celui d'un classificateur basé sur le comportement. Nous avons également mis en œuvre l'arbre de décision et le classificateur SVM sur le jeu de données comportementales, le

score F1 des deux modèles était inférieur à celui du modèle de forêt aléatoire, et le temps d'exécution pour le modèle de SVM était trop long en raison de l'énorme jeu de données. En conclusion, les résultats montrent que le comportement de l'hôtelier, de l'examinateur et de l'avis a une contribution importante à la décision de classer l'avis comme vrai ou faux.

4.4 Avantages par rapport aux autres modèles

- Le modèle proposé permet d'interpréter des variables d'entrée par rapport au résultat obtenu (interprétation des résultats).

- La complexité de l'algorithme n'est pas assez élevée, ce qui nous permet de traiter un grand volume de données.
- Le diagramme fig 8 montre que l'arbre de forêt aléatoire prend la décision principale des vrais et faux avis en fonction de la variable hotel_pct_true(pourcentage de vrais avis faites par l'hôtel).
- La deuxième meilleure variable qui aide à la prise de décision est pct_true(pourcentage de vraies critiques faites par le critique)

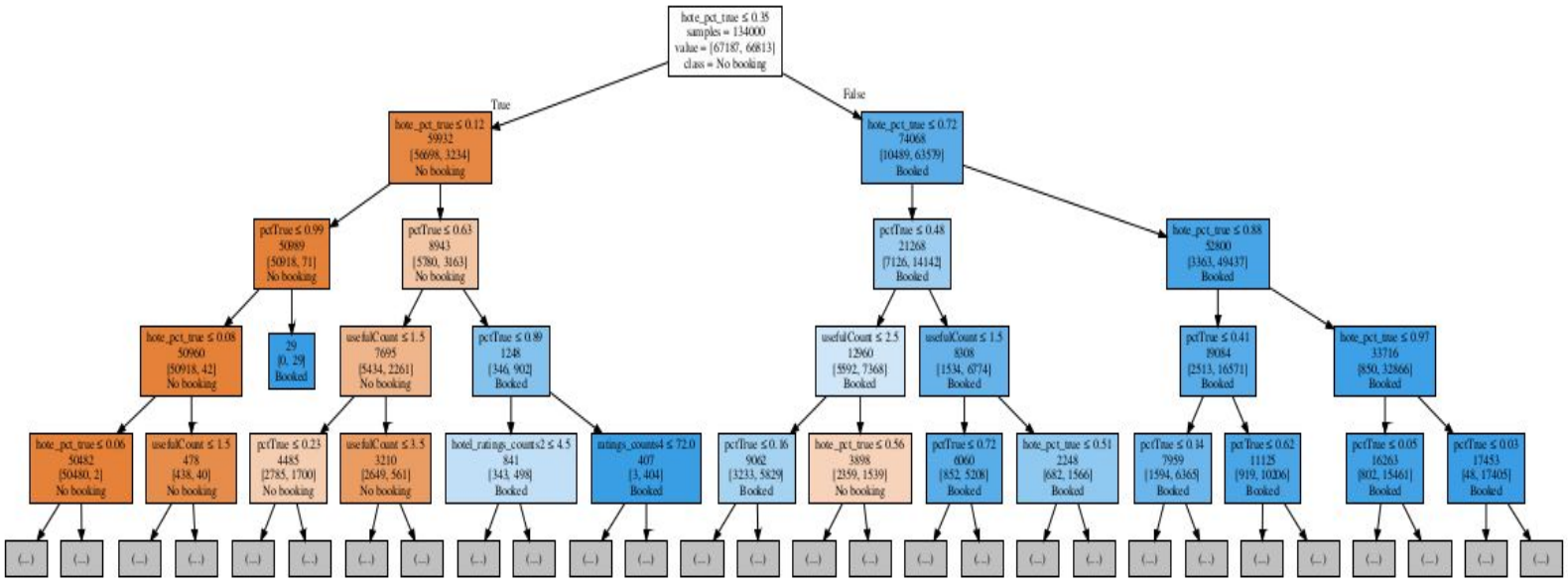

Fig 8. Arbre de décision du modèle de forêt aléatoire.

5. Validation

En estimant la performance des deux modèles, la précision du modèle de forêt aléatoire surpasse celle du modèle de l'arbre de décision. Nous avons donc choisi modèle de forêt aléatoire.

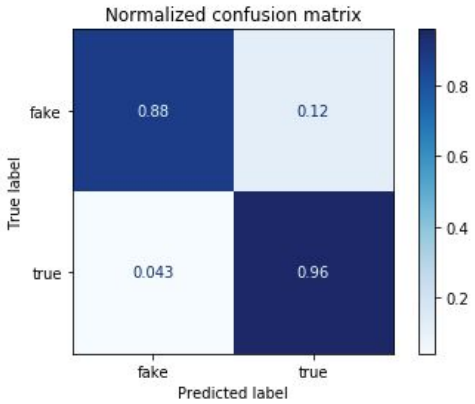

Fig 9. Matrice de confusion du modèle final

Selon la table ci-dessous, 95 % des faux avis détectées ont été correctement classées comme faux et 5 % des vrais avis ont été classées comme faux avis. Cependant, il montre également que l'algorithme n'a pu prédire que 88 % des faux avis réelles comme faux avis et qu'il n'a pas été en mesure de classer les 12 % restants comme faux avis.

	Precision	**Recall**
False	0.95	0.88
True	0.89	0.96

5.1 Commentaires

- L'algorithme peut détecter 88 % des faux avis réelles à partir d'une plateforme en ligne. La plateforme peut garantir un niveau de qualité des avis.
- L'algorithme affecte seulement 5 % des vraies avis de la plateforme. Les utilisateurs et les hôtels qui ont les vrais avis ne seront pas très affectés.
- Globalement, le résultat montre que l'algorithme est capable de répondre avec succès aux besoins soulevés.

6. Recommandation

L'algorithme permet d'identifier les vrais et faux avis. Avec ces informations, il est recommandé à l'entreprise de masquer ou de signaler les faux avis. Étant donné que cela peut entraîner des plaintes de certains utilisateurs, il est également recommandé d'indiquer au moment de la création de l'utilisateur que le contenu suspect peut être traité selon les considérations de l'entreprise.

RÉFÉRENCES

[1] 2018 Online Review Statistics You Need to Know, Accessed on : 16 December 2019 [Online], Available :https://boast.io/2018-online-review-statistics-you-need-to-know

[2] Local Consumer Review Survey 2019, Accessed on : 16 December 2019 [Online], Available:https://www.brightlocal.com/research/local-consumer-review-survey/#influence-of-reviews

[3]Online reviews - Statistics & Facts, Accessed on : 16 December 2019 [Online], Available:https://www.statista.com/topics/4381/online-reviews/

[4] 2018 Online Review Statistics You Need to Know, Accessed on : 16 December 2019 [Online], Available:https://boast.io/2018-online-review-statistics-you-need-to-know

[5] Banerjee, S., Chua, A. Y., & Kim, J. J.. *Using supervised learning to classify authentic and fake online reviews*. In Proceedings of the 9th International Conference on Ubiquitous

[6] How to spot a fake review,Accessed on : 16 December 2019 [Online], Available: https://www.yotpo.com/blog/amazon-fake-reviews

[7] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, *What Yelp fake review filter might be doing?* ,ICWSM, 2013.

[8] Weise,K.,*A Lie Detector Test for Online Reviewers*, 2011.

[9] Opinion Spam Detection: Detecting Fake Reviews and Reviewers. Accessed on : 5 November 2019 [Online], Available : https://www.cs.uic.edu/~liub/FBS/fake-reviews.html .