VRAIS OU FAUX AVIS ? EFFETS SUR LES ATTITUDES ET INTENTIONS : RESULTATS D'UNE PREMIERE ETUDE EXPERIMENTALE

Daria Plotkina, professeur associé, EM Strasbourg, Université de Strasbourg – HuManiS, daria.plotkina@em-strasbourg.eu
Jean-Luc Herrmann, Professeur des Universités, Université de Lorraine – CEREFIGE, jean-luc.herrmann@univ-lorraine.fr
Mathieu Kacha, Maître de Conférences HDR, Université de Lorraine – CEREFIGE, mathieu.kacha@univ-lorraine.fr
Gautier Lombard, Maître de Conférences, Université de Lorraine – CEREFIGE, gautier.lombard@univ-lorraine.fr

Résumé : Les avis en ligne constituent une source importante d'information pour réduire l'incertitude des consommateurs, cependant menacée par le développement de la présence de « faux » avis, c'est-à-dire d'opinions délibérément rédigées pour tromper les lecteurs. Alors que de nombreux travaux se sont focalisés sur la détection des faux avis et sur le développement corollaire d'algorithmes performants, la présente recherche en cours de développement s'intéresse au degré de véracité perçue des « vrais » et « faux » avis, et aux effets subséquents sur les attitudes et intentions des lecteurs. Les premiers résultats indiquent notamment que les consommateurs semblent en mesure de percevoir un degré de véracité en moyenne inférieur à l'égard des « faux » avis qu'à l'égard des « vrais », leur accordant moins de crédibilité, conduisant *in fine* à des attitudes et intentions moins favorables, alors même que la valence globale de la page d'avis auxquels ils ont été exposés est neutre. Nos résultats indiquent aussi qu'après avoir reçu une information les alertant sur le caractère « faux » (ou « vrai ») de chaque avis, les consommateurs tendent à corriger plus ou moins justement leurs évaluations dans un sens favorable à l'entreprise concernée, mais défavorable à la plateforme d'avis.
Mots-clés : avis de consommateurs en ligne ; faux avis ; véracité perçue ; évaluations ; intentions

TRUE OR FAKE REVIEWS? EFFECTS ON ATTITUDES AND INTENTIONS: RESULTS OF A FIRST EXPERIMENTATION

Abstract: Online reviews are an important source of information used to reduce consumer uncertainty, which is threatened by the development of "fake" reviews, that is opinions deliberately written to mislead readers. While most research has focused on the detection of

fake online reviews and the corollary development of powerful algorithms, the present research-in-progress focuses on the degree of perceived veracity of "true" and "fake" reviews and the subsequent effects on consumers' attitudes and intentions. The first results mainly indicate that on average consumers appear to be able to perceive a lower degree of veracity of "fake" reviews as compared to "true" ones, giving the former less credibility, leading ultimately to less favorable attitudes and intentions, even though the overall valence of the online reviews page to which they have been exposed is neutral. Our results also indicate that, after receiving information alerting them to the "fake" (or "true") nature of each review, consumers tend to correct their assessments more or less accurately in a way that is favorable to the company concerned, but unfavorable to the online reviews platform.
Key words: online consumers reviews; fake reviews; perceived veracity; evaluations; intention

Introduction

Les avis de consommateurs en ligne (ACL) demeurent une source importante d'information qui aide les consommateurs à prendre de meilleures décisions d'achat (Jindal et Liu, 2008). De multiples recherches antérieures ont montré de façon empirique que les ACL ont une influence sur les intentions de réservation (Sparks et Browning, 2011), les intentions d'achat (Jiménez et Mendoza, 2013 ; Plotkina et Munzel, 2016) et les ventes réelles (Zhu et Zhang, 2010 ; Cui et al., 2012). Cette influence des avis en ligne repose sur la croyance que l'information provient d'une expérience individuelle impartiale et authentique (Rastogi et Mehrotra, 2017), conduisant les consommateurs à percevoir les ACL comme une source d'information diagnostique pour instruire leurs décisions et choix, et non comme une source de persuasion (Watson, 2018).

Cependant, cette croyance est de plus en plus remise en question par la proportion des faux avis sur toutes les plateformes en ligne (p. exemple Xie et al., 2012). Les faux avis peuvent être définis comme « *ceux rédigés dans l'intention d'induire en erreur ou de tromper* » (Watson et Kirmani, 2018, p.138), ou « *des opinions fictives qui ont été délibérément rédigées pour paraître authentiques afin de tromper le lecteur* » (Ott et al., 2011, p.1).

Si de multiples recherches se sont logiquement intéressées à la détection des faux avis, la plupart se sont focalisées sur le développement d'algorithmes raisonnablement performants pour parvenir à distinguer les faux avis des vrais avis (cf. par exemple Jindal et Liu, 2007 ; Ott et al., 2011 ; Harris, 2012). Sur ce plan de l'identification des faux avis, les consommateurs quant à eux apparaissent moins performants que les algorithmes (Plotkina et al, 2020), même s'ils ne sont pas insensibles au langage utilisé dans les ACL (Baker et Kim, 2019 ; Cox et al., 2017 ; Salehan et Kim, 2016). En fait, au regard des résultats existants dans la littérature (cf. par exemple Plotkina et al., 2020), les consommateurs s'avèrent particulièrement peu performants face aux faux avis dès lors qu'il s'agit d'indiquer pour chacun d'eux s'il est faux ou vrai. Observons toutefois que cette tâche d'identification binaire (« faux » vs. « vrai ») ne reflète sans doute guère ce qui se passe réellement lorsque les consommateurs prennent connaissance d'avis d'autres consommateurs dans une perspective ultime de se faire une idée plus ou moins positive d'une produit ou d'un service.

Dès lors, l'objectif de cette recherche en cours de développement est d'aller au-delà de l'étude de la seule capacité des lecteurs d'avis à identifier correctement les « faux avis », en portant plutôt intérêt au degré de véracité perçue des avis lus (« vrais » avis et « faux » avis), et aux effets subséquents sur les attitudes et intentions des lecteurs, non seulement envers l'entreprise faisant l'objet des avis, mais aussi envers la plateforme sur laquelle ils sont

publiés. De façon complémentaire, et dans une seconde phase, cette recherche s'intéresse également à l'efficacité des dispositifs d'alerte des lecteurs sur la présence de faux avis (comme l'affichage de labels sur leur qualité ou leur véracité), et plus particulièrement ici, aux effets qu'aurait le fait de leur indiquer explicitement les avis qui sont faux et ceux qui sont vrais.

Littérature existante et effets examinés

La littérature existante sur la problématique des faux avis de consommateurs en ligne s'est majoritairement intéressée à la question de la détection des faux avis. Reposant sur les recherches menées en neurosciences (par exemple Davatzikos et al., 2005) et en psycholinguistique (Depaulo et al., 2003 ; Ott et al., 2011), le postulat qui sous-tend les travaux menés sur cette question importante est que les caractéristiques linguistiques (ou du contenu/texte) des ACL sont porteuses d'information sur le fait qu'un avis soit faux (versus vrai). Les caractéristiques textuelles/linguistiques des avis sont ainsi au cœur de la plupart des approches développées pour détecter les faux avis, certaines les combinant avec des caractéristiques des rédacteurs des avis (pour un récapitulatif, voir Plotkina et al., Table 1, 2020). Comme le relève Watson (2018), des algorithmes plus ou moins complexes ont été développés pour déterminer la véracité des ACL en fonction de certains indices syntaxiques (comme le recours excessif aux majuscules ou des répétitions de ponctuation) et/ou d'indices contextuels (comme l'adresse IP ou la fréquence de rédaction d'ACL) (Feng et al., 2012 ; Akoglu et al., 2013 ; Ott et al., 2013 ; Mukherjee et al., 2013). Au regard de la littérature existante, il est dès lors permis de considérer qu'il existe dans le contenu même des ACL, des caractéristiques textuelles/linguistiques susceptibles d'informer sur leur non véracité (cf. par exemple Ott et al., 2011 ou Zhou et Zhang, 2008).

Pour autant, la littérature est tout aussi unanime à reconnaître que les lecteurs des avis ont beaucoup de mal à percevoir ces caractéristiques (cf. par exemple Bond et DePaulo, 2008), ou à les utiliser correctement (Riggio and Friedman, 1983), pour parvenir à distinguer les « faux » avis des « vrais » avis. En vertu de la tendance générale des individus à n'être guère suspicieux envers les informations reçues (« *truth bias* » ou « *veracity effect* », cf. *Truth Default Theory*, *TDT*, Levine, 2014 ; Levine et al., 1999), qui constitue la pierre angulaire de l'*Interpersonal Deception Theory* (*IDT*, Buller et Burgoon, 1996), les individus tendent à considérer comme « vraies » bon nombre d'informations fausses. Les études existantes sur la capacité humaine à détecter des mensonges s'accordent à l'estimer à un peu plus d'une chance sur deux en moyenne, soit à peine mieux que le hasard (e.g. Landry et Brigham, 1992 ;

Toma et Hancock, 2012 ; Vrij, 2000). Dans le contexte des ACL, Plotkina et ses collègues (2020) rapportent même une estimation plus faible, puisque seuls 43% des faux avis sont identifiés par les lecteurs comme « faux » (contre 69% des vrais avis identifiés comme « vrai »).

Ces résultats indiquent que les lecteurs d'avis ne sont guère performants pour indiquer si un avis est faux (vs. vrai) sachant qu'il est effectivement faux, et *vice versa*. Observons qu'une telle tâche reflète bien l'orientation prédominante de la littérature en informatique et systèmes d'information, à savoir être capable d'identifier les faux avis pour par exemple les éliminer d'une plateforme d'avis. Par contre, cette tâche d'identification binaire (faux vs. vrai) correspond-t-elle réellement à ce qui se passe lorsque les consommateurs prennent connaissance d'avis d'autres consommateurs dans une perspective ultime de se faire idée plus ou moins positive d'une produit ou d'un service ?

Considérés ensemble, ces quelques traits de la littérature existante sur les faux avis invitent à essayer d'aller plus loin que l'étude de la seule capacité à identifier correctement les « faux avis » et les « vrais avis » (*cf. accuracy focus*). Au lieu d'étudier la capacité des lecteurs d'avis à indiquer de façon binaire qu'un avis est faux (vs. vrai) sachant qu'il est faux (et *vice versa*), l'approche que nous proposons ici consiste à s'intéresser davantage au degré de véracité perçue/ressentie, et aux conséquences en termes de crédibilité perçue et *in fine* d'attitudes et d'intentions des consommateurs exposés, non seulement envers la marque ou l'entreprise concernée par les ACL, mais aussi envers la plateforme qui les publie.

Etant donné que les avis contiennent d'éventuelles caractéristiques textuelles/linguistiques de non-véracité (cf. par exemple Ott et al., 2011 ou Zhou et Zhang, 2008), que parallèlement les lecteurs sont devenus de plus en plus suspicieux à l'égard des avis d'autres consommateurs (Larson et Denton, 2014), et que le degré de crédibilité perçue d'un avis (respectivement positif ou négatif) peut inverser les effets subséquents de cet avis (Reimer et Benkenstein, 2016), nous supposons que :

- les consommateurs perçoivent un degré de véracité inférieur à l'égard de « faux » ACL qu'à l'égard de « vrais » ACL ;
- les consommateurs accordent moins de crédibilité aux « faux » ACL qu'aux « vrais » ACL ;
- les consommateurs forment des attitudes et intentions moins favorables après avoir pris connaissance de « faux avis » qu'après lecture de « vrais avis ».

Cette étude destinée à mieux appréhender la capacité des lecteurs d'ACL à percevoir un certain degré de véracité des avis lus, et les effets subséquents sur leurs attitudes et intentions, n'a évidemment pas pour objectif de remettre en cause l'idée de recourir à des dispositifs d'alerte sur les faux avis comme l'affichage sur le site de labels de qualité tels que « avis ou achat vérifié », d'une alerte sur la présence de faux avis après leur suppression (cf. Yelp, Watson, 2018), ou d'outils permettant d'estimer la véracité estimée de chaque avis. Nous proposons ainsi d'examiner également les effets qu'aurait le fait d'indiquer explicitement aux lecteurs les avis qui sont faux et les avis qui sont vrais.

Dans le prolongement des travaux de Watson et Kirmani (2018) sur les alertes relatives à l'existence de faux avis, nous partons de l'idée que ce type d'alerte informe les lecteurs d'ACL que certains rédacteurs d'avis les ont exposés à des informations fausses, ce qui tend à activer leurs connaissances en matière de persuasion (cf. *Persuasion Knowledge*, Friestad et Wright, 1994) et leur suspicion (Darke et Ritchie, 2007 ; Main et al., 2007). Dans ce type de situation, les consommateurs ont en général tendance à corriger leurs impressions initiales afin de tenir compte des informations qu'ils perçoivent comme étant biaisées ou fausses. Cette ligne de raisonnement conduit à supposer que les lecteurs d'ACL recevant une telle alerte corrigent leurs jugements de cette source de biais que constituent les ACL perçus comme « faux ».

S'agissant de corrections de jugements, le *Flexible Correction Model* (cf. par exemple Petty et Wegener, 1993 ; Wegener et Petty, 1995) suggère que les consommateurs ont tendance à recourir à une *Naïve Theory of Persuasion* pour évaluer le sens (la valence) et l'ampleur d'un biais perçu. Les consommateurs corrigent alors leurs jugements dans le sens opposé au biais perçu sans tenir compte correctement de l'ampleur de ce biais.

Dans le contexte des ACL et de la présence d'une alerte informant précisément les lecteurs sur le caractère « faux » ou « vrai » des avis auxquels ils sont exposés, cela conduit à supposer, soit une « juste correction » (i.e. proportionnelle à la valence et à l'ampleur du biais) de leurs jugements, soit une « sur-correction » due à une perception inexacte du biais malgré la présence de l'alerte. En outre, en référence aux travaux de Darke et ses collègues sur le *Defensive Consumer* et les effets de la suspicion à d'autres sources que celle qui en est à l'origine (cf. par exemple Darke et Ritchie, 2007 ; Main et al., 2007), on peut supposer que ces corrections de jugements se propagent à d'autres acteurs que la seule marque ou entreprise concernée par les ACL, et notamment dans le cas présent à la plateforme qui les affiche.

Méthodologie

Nous avons choisi d'exposer les participants à une page comportant plusieurs ACL afin d'être plus en phase avec les conditions réelles d'exposition aux avis et de lever ainsi la limite fréquemment rencontrée dans les travaux existants (cf. par exemple Watson, 2018) que représente l'exposition à un seul et unique ACL. Nous basant sur les choix faits par Lee et al. (2008) et Purnawirawan et al. (2012), nous avons décidé d'exposer les répondants à une page contenant huit avis. Quelle que soit la condition expérimentale, quatre avis étaient positifs et quatre avis étaient négatifs de sorte à ce que la valence globale perçue de la page ne soit ni positive ni négative. Sachant que l'ordre des avis sur une page peut avoir une influence (Huang et al., 2018; Purnawirawan et al., 2012), nous avons randomisé l'ordre d'apparition des huit avis de consommateurs qui la composaient. Les avis étaient toujours présentés au même endroit sur l'écran, et leur taille était constante entre toutes les conditions et tout au long de l'expérimentation.

Afin d'examiner dans un premier temps les effets supposés de la présence sur la même page de vrais et de faux avis, puis dans un second temps l'impact de l'indication du caractère « faux » (« vrai ») de chaque avis de la page, nous avons mis en œuvre un plan factoriel 3 (facteur « nombre de faux avis », inter-sujets) x 2 (facteur « indication du caractère faux (vrai) », intra-sujets i.e. avant-après). Les 3 modalités du facteur « nombre de faux avis » sont « 2 avis faux et 6 avis vrais » (25% de faux avis), « 4 avis faux et 4 avis vrais » (50%), et « 6 avis faux et 2 avis vrais » (75%). L'annexe 1 précise l'origine des « faux » et « vrais » avis utilisés, l'annexe 2 décrit le pré-test réalisé pour s'assurer que la valence globale perçue de la page d'avis demeurait neutre d'une condition à l'autre, et l'annexe 3 présente la page d'avis utilisée pour la condition « 2 avis faux et 6 avis vrais », au moment de l'indication du caractère faux (vrai).

Un échantillon de 143 étudiants âgés de 18 à 21 ans a pris part à cette étude expérimentale, chaque étudiant ayant été affecté aléatoirement à l'une des trois conditions du facteur inter-sujets (25% vs. 50% vs. 75% de faux avis). Cette catégorie de répondants est appropriée au contexte de l'étude. Les jeunes adultes sont très familiers avec le bouche-à-oreille électronique et les avis de consommateurs en ligne (Choi et al., 2008 ; Reimer et Benkenstein, 2018). De surcroît, les avis portaient sur un restaurant fictif « *Cuisines du monde* », catégorie de services en adéquation avec les expériences de consommation des jeunes adultes (Reimer et Benkenstein, 2018). D'ailleurs, ces derniers cherchent et lisent fréquemment des avis de consommateurs afin de sélectionner un restaurant (Zhang et al., 2014).

L'expérimentation était structurée en trois étapes au cours desquelles la page d'avis restait continuellement affichée sur la moitié gauche de l'écran (tandis que, les questions

apparaissaient sur la partie droite). Dans la première étape, nous avons demandé aux participants d'imaginer la note moyenne (sur 5 étoiles, puis sur 20) du restaurant faisant l'objet des avis présentés, note qui faisait défaut sur la page présentée. Ont ensuite été mesurées l'attitude envers le restaurant (échelle de Yin et al. (2014) composée de trois items de -3 à +3, alpha de Cronbach = 0,948), l'attitude envers la plateforme (même échelle adaptée, alpha de Cronbach = 0,923), l'intention d'aller manger dans ce restaurant (échelle adaptée de Pavlou et Fygenson (2006) et Zhang et al. (2014), à 2 items, coefficient de corrélation = 0,793, p<.001), et l'intention de le recommander (un item de Sela, Wheeler, Sarial-Abi (2012)). Ces mesures ont été complétées par l'évaluation pour chaque avis, de la crédibilité perçue (un item de Cheung et al. (2009)), du degré de véracité perçue et du degré de certitude associé.

Dans une deuxième étape, les participants étaient explicitement informés de la non-véracité des faux avis qui se trouvaient sur la page comme le font des applications Web telles que Fakespot. Nous leur indiquions précisément quels avis étaient faux parmi les huit affichés (i.e. deux, quatre ou six selon la condition). Dans la condition 25% par exemple, les participants lisaient : « *Parmi les avis de consommateurs en ligne que vous voyez sur cette page, les deux avis suivants sont faux : [un avis faux positif et un avis faux négatif étaient expressément identifiés]* ».

Enfin, la troisième étape a consisté à mesurer la perception d'une tromperie et d'une manipulation, puis à nouveau les notes moyennes, les attitudes et intentions envers le restaurant et la plateforme, ainsi que plusieurs variables de contrôle[1].

Résultats

Dans un premier temps, nous avons examiné les effets du nombre de faux avis avant que les participants soient informés explicitement de la non-véracité de certains avis en l'occurrence « faux ». De la série d'analyses de variance à un facteur (3 modalités, 2, 4 ou 6 faux avis) réalisées, il ressort tout d'abord que l'augmentation du nombre de faux avis apparaissant sur la page a un effet négatif sur l'ensemble des évaluations du restaurant et de la plateforme concernée, et ce alors même que la valence globale neutre de la page est maintenue constante quel que soit le nombre de faux avis contenus. Ainsi la note moyenne sur 5 étoiles attribuée

[1] Ces variables de contrôle indiquent que le besoin en cognition (Cotte, June & Wood, 2004 ; p=.674), la part de faux avis estimée de manière générale sur les sites d'avis (p=.345), le scepticisme à l'égard des faux avis en général (Zhang et al., 2016 ; dimension véracité p=.751, motivation p=.942, identité p=.304), le degré de recherche d'information (Chu et Kim, 2001 ; p=.657) et de connaissance du contexte (Martin et al., 2001 ; p=.492) ne sont pas différents selon les différentes conditions du facteur « nombre de faux avis».

au restaurant (F(2,140)=6,86, p=0,001, η_p^2=0,089) est plus faible lorsque la page contient 6 faux avis ($\overline{x}$=2,59) que lorsqu'elle en contient 2 ($\overline{x}$=2,96, p<0,001), ou 4 ($\overline{x}$=2,80, p=0,046). Idem pour la moyenne sur 20 points (F(2,140)=5,37, p=0,006, η_p^2=0,071). L'attitude envers le restaurant suit la même tendance défavorable (F(2,140)=7,74, p=0,001, η_p^2=0,100) lorsque la page contient 6 faux avis ($\overline{x}$=-0,55) comparativement à 2 ($\overline{x}$=0,15 p<0,001) et à 4 ($\overline{x}$=-0,03, p=0,008). La même tendance est observée au niveau des intentions envers le restaurant, qu'il s'agisse d'y aller (F(2,140)=4,28, p=0,016, η_p^2=0,058) ou de le recommander (F(2,140)=5,50, p=0,005, η_p^2=0,073). Quant à l'attitude envers la plateforme d'avis, elle s'avère elle aussi moins favorable (F(2,140)=2,59, p=0,078, η_p^2=0,036) dans la condition « 6 faux avis » ($\overline{x}$=-0,33) que dans la condition 2 faux avis ($\overline{x}$=0,78, p=0,028).

Parallèlement, alors qu'aucune information n'a été donnée aux participants sur le caractère « faux » vs. « vrai » de chaque avis apparaissant sur la page présentée, il est intéressant d'observer que le degré de véracité perçu des « faux » avis ($\overline{x}$=4,27) est en moyenne inférieur à celui perçu pour les « vrais » avis ($\overline{x}$=4,82), la taille de l'effet étant toutefois plus grande dans le cas des avis négatifs que dans celui des avis positifs. Relevons toutefois que le degré de véracité perçu des « faux avis » est très proche de la valeur médiane (i.e. 4) de l'échelle de mesure utilisée, indiquant que les « faux » avis ne sont pas perçus strictement comme « faux », seulement comme moins vrais que les « vrais » avis. Suivant la même tendance que le degré de véracité perçue, notons également que tant le degré de certitude associé à la véracité perçue que la crédibilité perçue sont en moyenne inférieurs dans le cas des « faux » avis (comparativement au cas des « vrais » avis).

Forts de ces premiers résultats, nous avons alors analysé dans un second temps les effets produits par le fait d'indiquer aux participants très explicitement les avis de la page présentée qui sont faux (soit 2, 4 ou 6 en vertu du design expérimental adopté).

Les résultats d'une première ANOVA à un facteur indiquent tout d'abord qu'après avoir indiqué aux participants que la page d'avis présentée contenait 2, 4 ou 6 « faux » avis, ces derniers ne ressentent pas davantage de tromperie en fonction du nombre de « faux avis » ((F(2,140)=1,16, p=0,316, η_p^2=0,016). Ceci étant, les participants ont toutefois tendance à ressentir de la tromperie ($\overline{x}$=4,56, σ=1,76, valeur moyenne supérieure à la valeur centrale 4 de l'échelle de 1 à 7, t(142)=3,79, p<0,001), mais indépendamment du nombre de faux avis contenus dans la page présentée.

Nous avons ensuite mené une série d'analyses de variance à deux facteurs (nombre de faux avis, et avant-après indication) sur chacune des principales variables dépendantes. Les résultats les plus intéressants sont les suivants :

- lorsque l'attitude est mesurée avec l'échelle de Yin et al. (2014, trois items, alpha de Cronbach = 0,979), les participants tendent à corriger leur attitude envers le restaurant consécutivement à l'indication des faux avis (F(1,140)=6,89, p=0,010, η_p^2=0,047). Quel que soit le nombre de faux avis, ils évaluent systématiquement plus favorablement le restaurant après ($\overline{x}$=0,21) avoir eu connaissance de la présence de faux avis, qu'avant ($\overline{x}$=-0,17) ;
- par contre, lorsque les participants ont été invités à imaginer la note moyenne (sur 5 étoiles, puis sur 20) du restaurant (note qui faisait défaut sur la page présentée), c'est-à-dire lorsqu'une mesure indirecte est utilisée, cette tendance à la (sur-)correction en faveur du restaurant entre avant et après indication des faux avis n'est pas retrouvée (F(1,140)=0,08, p=0,375, η_p^2=0,006 pour la note sur 5 étoiles, et F(1,140)=1,01, p=0,317, η_p^2=0,007) ;
- quant à l'attitude envers la plateforme d'avis (mesurée avec une échelle adaptée de celle utilisée pour le restaurant (3 items, alpha de Cronbach = 0,962), les participants opèrent une correction après indication des faux avis (F(1,140)=79,93, p<0,001, η_p^2=0,363), mais cette fois dans le sens opposé (i.e. en défaveur de la plateforme), et ce quel que soit le nombre de faux avis révélé (F(2,140)=1,85, p=0,161, η_p^2=0,026).

Conclusion – discussion

Ces premiers résultats de cette recherche en cours de développement sont encourageants et globalement plutôt conformes à nos attentes. Ils indiquent tout d'abord que les consommateurs semblent en mesure de percevoir un degré de véracité différent entre les « faux » avis et les « vrais avis », en moyenne inférieur à l'égard des « faux » avis qu'à l'égard des « vrais », et ce en l'absence d'information sur la véracité (ou non) réelle des avis. Ils semblent dès lors accorder moins de crédibilité aux avis dont le degré de véracité perçu est moindre, et forment *in fine* des attitudes et intentions moins favorables envers le restaurant qui fait l'objet de ces avis perçus comme moins vrais, et ce alors même que la valence globale de la page d'avis auxquels ils ont été exposés est toujours neutre (autant d'avis positifs que négatifs). Il en est de même pour l'attitude envers la plateforme d'avis.

Nos résultats indiquent ensuite qu'après avoir reçu une information les alertant sur le caractère « faux » (ou « vrai ») de chaque avis de la page à laquelle ils ont été exposés, les

participants paraissent corriger de manière « juste » (i.e. proportionnelle à la valence et à l'ampleur du biais consécutif à la présence de faux avis) leurs évaluations du restaurant, tout au moins si la mesure est suffisamment indirecte (i.e. mise en situation de note moyenne manquante sur la plateforme d'avis). Cette correction est toutefois surfaite dès lors qu'il s'agit d'une mesure d'attitude plus directe via une échelle traditionnelle de 3 items ; les participants tendent alors à sur-corriger favorablement leur attitude envers le restaurant, comme s'ils éprouvaient une forme de compassion envers le restaurant, victime de « faux » avis, et ce alors même que les faux avis sont pour moitié positifs et pour moitié négatifs. De façon tout aussi intéressante, lorsqu'il s'agit de l'attitude envers la plateforme d'avis, on retrouve cette sur-correction, mais cette fois dans le sens opposé, c'est-à-dire d'une dégradation de son attitude ; la plateforme semble ainsi être tenue pour responsable de la présence de faux avis et dès lors sanctionnée.

Prometteurs, ces premiers résultats souffrent toutefois de limites importantes. Par-delà les précautions prises par exemple au niveau de la démarche de génération des « vrais » et « faux » avis, du design expérimental ou du pré-test pour contrôler que la valence globale perçue de la page d'avis utilisée demeure neutre et constante d'une condition expérimentale à l'autre, il reste que d'une condition expérimentale à l'autre les participants ne sont pas exposés strictement aux mêmes avis. En effet, eu égard au design de l'expérimentation, si 4 avis sont communs (1 « vrai » et 1 « faux » positifs, 1 « vrai » et 1 « faux » négatifs) aux 3 pages de 8 avis utilisées, les 4 autres ne le sont pas. Ajoutons que la proportion de faux avis apparaissant sur la page d'avis utilisée n'est pas également « réaliste » d'une condition à l'autre (2/8, 4/8, 6/8).

Une autre limite importante provient du fait que les résultats présentés ici concernent le seul cas d'une page d'avis de valence globale neutre, et qu'il n'est pas possible d'examiner dans cette première étude les effets observés pour les cas d'une page de valence globalement plutôt positive, à l'inverse plutôt négative.

Enfin, l'information communiquée aux participants pour leur révéler les avis de la page qui sont faux, est trop tranchée. Afin de pouvoir progresser sur la voie d'informer les consommateurs pour réduire la situation d'incertitude créée par la présence de faux avis, il convient sans doute d'examiner les effets de dispositifs d'alerte plus nuancés indiquant par exemple une probabilité de non-véracité de chaque avis, estimée à partir d'un algorithme de détection de faux avis ou de la perception d'un échantillon de consommateurs formés à cela, plutôt qu'en recourant à une information binaire (« faux » vs. « vrai »).

En un mot comme en cent, cette proposition de communication est avant tout une communication d'étape sur les résultats d'une première étude d'une recherche en cours de développement, souffrant d'importantes limites à ce stade, et ne portant en outre que sur une seule catégorie de produit/service (les restaurants), une marque et une plateforme d'avis fictives, et un seul type de population (les étudiants) ; les commentaires et suggestions qu'elle suscitera seront dès lors très appréciés et utiles aux études qui suivront.

Références bibliographiques

Akoglu L, Chandy R et Faloutsos C (2013, Juin) Opinion fraud detection in online reviews by network effects. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 7, No. 1).

Baker MA et Kim K (2019) Value destruction in exaggerated online reviews. *International Journal of Contemporary Hospitality Management* 31(4): 1956-1976.

Bond Jr CF et DePaulo BM (2008) Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin* 134(4): 477-492.

Buller DB et Burgoon JK (1996) Interpersonal deception theory. *Communication theory* 6(3): 203-242.

Cheung MY, Luo C, Sia CL et Chen H (2009) Credibility of electronic word-of-mouth : Informational and normative determinants of on-line consumer recommendations. *International Journal of Electronic Commerce* 13(4): 9-38.

Choi YK, Hwang J-S et McMillan SJ (2008) Gearing up for mobile advertising : A cross-cultural examination of key factors that drive mobile messages home to consumers. *Psychology et Marketing* 25(8): 756-768.

Cox D, Cox JG et Cox AD (2017) To Err is human? How typographical and orthographical errors affect perceptions of online reviewers. *Computers in Human Behavior* 75: 245-253.

Cui G, Lui HK et Guo X (2012) The effect of online consumer reviews on new product sales. *International Journal of Electronic Commerce* 17(1): 39-58.

Darke PR et Robin JB (2007) The defensive consumer: Advertising deception, defensive processing, and distrust. *Journal of Marketing Research* 44(1): 114-127.

Davatzikos C, Ruparel K, Fan Y, Shen DG, Acharyya M, Loughead JW et Langleben DD (2005) Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage* 28(3): 663-668.

DePaulo BM, Lindsay JJ, Malone BE, Muhlenbruck L, Charlton K et Cooper H (2003) Cues to deception. *Psychological Bulletin* 129(1): 74-118.

Feng S, Banerjee R et Choi Y (2012, July) Syntactic stylometry for deception detection. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) pp. 171-175.

Friestad M et Wright P (1994) The persuasion knowledge model: How people cope with persuasion attempts *Journal of Consumer Research*, 21(1): 1-31.

Harris CG (2012, July) Detecting deceptive opinion spam using human computation. In Proceedings Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence, Palo Alto, CA: AAAI, 2012, pp. 87-93.

Huang Y, Li C, Wu J et Lin Z (2018) Online customer reviews and consumer evaluation: The role of review font. *Information et Management* 55(4): 430-440.

Jiménez FR et Mendoza NA (2013) Too popular to ignore: The influence of online reviews on purchase intentions of search and experience products. *Journal of Interactive Marketing* 27(3): 226-235.

Jindal N et Liu B (2007, Mai) Review spam detection. In Proceedings of the 16th international conference on World Wide Web, pp. 1189-1190.

Jindal N et Liu B (2008, Février) Opinion spam and analysis. In Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 219-230.

Landry KL et Brigham JC (1992) The effect of training in criteria-based content analysis on the ability to detect deception in adults. *Law and Human Behavior* 16(6): 663-676.

Larson LR et Denton LT (2014) eWOM watchdogs: Ego-threatening product domains and the policing of positive online reviews. *Psychology et Marketing* 31(9): 801-811.

Lee J, Park DH et Han I (2008) The effect of negative online consumer reviews on product attitude: An information processing view. *Electronic commerce research and applications* 7(3): 341-352.

Levine TR (2014) Truth-default theory (TDT) a theory of human deception and deception detection. *Journal of Language and Social Psychology* 33(4): 378-392.

Levine TR, Park HS et McCornack SA (1999) Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communications Monographs* 66(2): 125-144.

Main KJ, Dahl DW et Darke PR (2007) Deliberative and automatic bases of suspicion: Empirical evidence of the sinister attribution error. *Journal of Consumer Psychology* 17(1): 59-69.

Martin IM et Stewart DW (2001) The differential impact of goal congruency on attitudes, intentions, and the transfer of brand equity. *Journal of Marketing Research*, 38(4): 471-484.

Mukherjee A, Venkataraman V, Liu B et Glance N (2013, Juin) What yelp fake review filter might be doing?. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 7, No. 1).

Ott M, Cardie C et Hancock JT (2013, Juin) Negative deceptive opinion spam. In Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies, pp. 497-501.

Ott M, Choi Y, Cardie C et Hancock JT (2011) Finding deceptive opinion spam by any stretch of the imagination. In Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA: ACL, pp. 309-319.

Pavlou PA et Fygenson M (2006) Understanding and predicting electronic commerce adoption: An extension of the theory of planned behavior. *MIS Quarterly*, 115-143.

Petty RE et Duane TW (1993) Flexible correction processes in social judgment: Correcting for context-induced contrast. *Journal of Experimental Social Psychology* 29(2): 137-165.

Plotkina D et Munzel A (2016) Delight the experts, but never dissatisfy your customers! A multi-category study on the effects of online review source on intention to buy a new product. *Journal of Retailing and Consumer Services* 29: 1-11.

Plotkina D, Munzel A et Pallud J (2020) Illusions of truth-experimental insights into human and algorithmic detections of fake online reviews. *Journal of Business Research* 109: 511-523.

Purnawirawan N, De Pelsmacker P et Dens N (2012) Balance and sequence in online reviews: How perceived usefulness affects attitudes and intentions. *Journal of Interactive Marketing* 26(4): 244-255.

Rastogi A et Mehrotra M (2017) Opinion spam detection in online reviews. *Journal of Information et Knowledge Management* 16(04): 17-36.

Reimer T et Benkenstein M (2016) When good WOM hurts and bad WOM gains: The effect of untrustworthy online reviews. *Journal of Business Research* 69(12): 5993-6001.

Reimer T et Benkenstein M (2018) Not just for the recommender : How eWOM incentives influence the recommendation audience. *Journal of Business Research* 86: 11-21.

Riggio RE et Friedman HS (1983) Individual differences and cues to deception. *Journal of personality and Social Psychology* 45(4): 899-915.

Salehan M et Kim DJ (2016) Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems* 81: 30-40.

Sela A, Wheeler SC et Sarial-Abi G (2012) We are not the same as you and I: Causal effects of minor language variations on consumers' attitudes toward brands. *Journal of Consumer Research* 39(3): 644-661.

Sparks BA et Browning V (2011) The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management* 32(6): 1310-1323.

Toma CL et Hancock JT (2012) What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication* 62(1): 78-97.

Vrij A (2000) *Detecting lies and deceit: The psychology of lying and implications for professional practice*. Wiley.

Watson Jared J (2018) *Aspects of Online Reviews and their effects in consumer decisions*, Thèse de Doctorat, Université de Maryland, College Park.

Watson Jared J and Kirmani A (2018) *Red Flag! the Consequences of Alerting Consumers to Fake Reviews*, in NA-Advances in Consumer Research Volume 46, eds. Andrew Gershoff, Robert Kozinets, and Tiffany White, Duluth, MN: ACR, 138-143.

Wegener DT et Petty RE (1995) Flexible correction processes in social judgment: The role of naive theories in corrections for perceived bias. *Journal of Personality and Social Psychology* 68(1): 36.

Xie S, Wang G, Lin S et Yu PS (2012) Review spam detection via temporal pattern discovery. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 823-831.

Yin D, Bond SD et Zhang H (2014) Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quarterly* 38(2): 539-560.

Zhang XJ, Ko M et Carpenter D (2016) Development of a scale to measure skepticism toward electronic word-of-mouth. *Computers in Human Behavior*, 56: 198-208.

Zhang KZ, Zhao SJ, Cheung CM et Lee MK (2014) Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model. *Decision Support Systems* 67: 78-89.

Zhou L et Zhang D (2008) Following linguistic footprints: Automatic deception detection in online communication. *Communications of the ACM* 51(9): 119-122.

Zhu F et Zhang X (2010) Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing* 74(2): 133-148.

Annexe 1 : Génération des « faux » et « vrais » avis utilisés

Afin de disposer de faux et de vrais avis, nous avons sollicité des étudiants d'une promotion niveau Master 1 (~200 étudiants) via un questionnaire en ligne. Nous leur avons demandé de réfléchir à une expérience récente dans un café ou un restaurant. Tout d'abord, nous avons mesuré leur jugement à l'égard de l'expérience à l'aide d'affirmations comme « Je pense que ce restaurant est vraiment bon », « Cet endroit m'attire », « J'ai apprécié mon expérience dans ce restaurant ». Ensuite, les répondants ont été dirigés au hasard vers l'une des deux conditions suivantes : (1) la demande de rédiger un avis correspondant à leur expérience réelle, et (2) la demande de rédiger un avis de valence opposée à celle de leur expérience réelle (i.e. rédiger un avis positif pour une expérience négative ou un avis négatif pour une expérience positive). N'ont été conservés que les avis relatifs à des expériences ayant eu lieu depuis moins d'un mois avant la collecte des données, tandis que les avis très courts et peu informatifs (i.e. « *c'était bien* »; « *rien de spécial* ») ont été éliminés. Sur 110 avis reçus initialement (53 « vrais » et 57 « faux »), nous avons ensuite sélectionné les avis les plus complets qui auraient pu être laissés sur le même restaurant (c'est-à-dire pas sur une entreprise ou une cantine universitaire, ni sur un endroit spécifique comme un bar à sushis). Cette démarche a permis au final de retenir 6 « vrais » avis et 6 « faux » avis, tous modérément positifs ou négatifs, utilisés pour composer les trois pages correspondant aux trois conditions expérimentales retenues (2, 4, ou 6 faux avis sur un total de 8, la moitié des faux (vrais) avis étant positive, l'autre négative).

Annexe 2 : Pré-test de la valence globale de la page d'avis utilisée

Comme pour chaque condition la moitié des avis apparaissant sur la page était respectivement positive, l'autre négative, nous avons cherché à nous assurer que la valence globale perçue de la page d'avis de consommateurs était neutre quelle que soit la condition. Afin d'administrer le questionnaire, nous avons eu recours au panel en ligne ClickWorker. Les participants ont été indemnisés à hauteur de 0.9€ qui correspond à une indemnisation sur la base du SMIC pour 5 minutes (durée estimée de remplissage du questionnaire ; base de 10,25€/ heure). L'échantillon finalement utilisable était composé de 123 répondants.

Comme pour l'étude principale, la page d'avis restait continuellement affichée sur la moitié gauche de l'écran. La mesure de la valence globale perçue de la page était composée de deux items affichés sur deux pages distinctes. Il était indiqué : « *Nous aimerions savoir ce que vous pensez globalement des avis que vous voyez à l'écran. Globalement, vous diriez que les avis que vous voyez à l'écran sont plutôt : » de « Très négatif » à « Très positif »* (page 1) *et de « Très défavorable » à « Très favorable »* (page 2 ; coefficient de corrélation de 0,854, $p<.001$).

Les résultats de l'analyse de variance montrent l'absence d'effet du nombre de faux avis sur la valence globale perçue ($F(2,120)=1.07$, $p=.346$, $\eta_p^2=.018$; condition « 2 faux avis » : $\bar{x}=.11$, $\sigma=1.16$; condition « 4 faux avis » : $\bar{x}=-.04$, $\sigma=1.29$; condition « 6 faux avis » : $\bar{x}=-.29$, $\sigma=1.16$).

Annexe 3 : Page d'avis affichée dans la condition 6 avis « vrais » et 2 avis « faux » (capture d'écran prise au moment de l'indication du caractère faux)