

Prédire l'intensité de contradiction dans les commentaires : faible, forte ou très forte ?

Ismail Badache, Sébastien Fournier, Adrian-Gabriel Chifu

▶ **To cite this version:**

Ismail Badache, Sébastien Fournier, Adrian-Gabriel Chifu. Prédire l'intensité de contradiction dans les commentaires : faible, forte ou très forte ?. 29es Journées Francophones d'Ingénierie des Connaissances, AFIA, Jul 2018, Nancy, France. pp.55-69. hal-01839546

HAL Id: hal-01839546
https://hal.science/hal-01839546

Submitted on 23 Jul 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Prédire l'intensité de contradiction dans les commentaires : faible, forte ou très forte ?

Ismail Badache, Sébastien Fournier, Adrian-Gabriel Chifu

AIX MARSEILLE UNIV, UNIVERSITÉ DE TOULON, CNRS, LIS, MARSEILLE, FRANCE
{ismail.badache, sebastien.fournier, adrian.chifu}@lis-lab.fr

Résumé : Les commentaires sur des ressources Web (ex. des cours, des films) deviennent de plus en plus exploitées dans des tâches d'analyse de texte (ex. détection d'opinion, détection de controverses). Cet article étudie l'intensité de contradiction dans les commentaires en exploitant différents critères tels que la variation des notations et la variation des polarités autour d'entités spécifiques (ex. aspects, sujets). Premièrement, les aspects sont identifiés en fonction des distributions des termes émotionnels à proximité des noms les plus fréquents dans la collection des commentaires. Deuxièmement, la polarité est estimée pour chaque segment de commentaire contenant un aspect. Seules les ressources ayant des commentaires contenant des aspects avec des polarités opposées sont prises en compte. Enfin, les critères sont évalués, en utilisant des algorithmes de sélection d'attributs, pour déterminer leur impact sur l'efficacité de la détection de l'intensité des contradictions. Les critères sélectionnés sont ensuite introduits dans des modèles d'apprentissage pour prédire l'intensité de contradiction. L'évaluation expérimentale est menée sur une collection contenant 2244 cours et leurs 73873 commentaires, collectés à partir de *coursera.org*. Les résultats montrent que la variation des notations, la variation des polarités et la quantité de commentaires sont les meilleurs prédicteurs de l'intensité de contradiction. En outre, J48 est l'approche d'apprentissage la plus efficace pour cette tâche.

Mots-clés : Analyse de sentiments, Détection d'aspects, Évaluation des critères, Intensité de contradiction.

1 Introduction

Au cours des dernières années, le web 2.0 est devenu un espace ouvert où les gens peuvent exprimer leurs opinions en laissant des traces (par exemple, un commentaire, une notation, un j'aime) sur les ressources Web. De nombreux services, tels que les blogs et les réseaux sociaux, représentent une source riche de ces données sociales, qui peuvent être analysées et exploitées dans diverses applications et contextes Badache & Boughanem (2014, 2017a,b). En particulier la détection d'opinion et l'analyse de sentiments Htait *et al.* (2016), par exemple, pour connaître l'attitude d'un client vis-à-vis d'un produit ou de ses caractéristiques, ou pour révéler la réaction des gens à un événement. De tels problèmes nécessitent une analyse rigoureuse des aspects couverts par le sentiment pour produire un résultat représentatif et ciblé.

Une autre problématique concerne la diversité des opinions sur un sujet donné. Certains travaux l'abordent dans le contexte de différents domaines de recherche, avec une notion différente dans chaque cas. Par exemple, Wang & Cardie (2014) visent à identifier des sentiments au niveau d'une phrase exprimée au cours d'une discussion et à les utiliser comme des caractéristiques dans un classifieur qui prédit la dispute dans la discussion. Qiu *et al.* (2013) identifient automatiquement les débats entre des utilisateurs à partir du contenu textuel (interactions) dans les forums, en se basant sur des modèles de variables latentes. Il y a eu d'autres travaux dans l'analyse des interactions avec les utilisateurs, par exemple, l'extraction des expressions de type *agreement* et *disagreement* Mukherjee & Liu (2012) et d'en déduire les relations de l'utilisateur en regardant leurs échanges textuels Awadallah *et al.* (2012).

Cet article étudie les entités (par exemple, les aspects, les sujets) pour lesquelles des contradictions peuvent apparaître dans les commentaires associés à une ressource Web (par exemple des films, des cours) et comment estimer leur intensité. L'intérêt d'estimer l'intensité de la contradiction dépend du cadre d'application. Par exemple, suivre des événements ou des crises politiques controversés tels que la reconnaissance par les États-Unis de Jérusalem comme capitale d'Israël. Cela a généré des opinions (avis) contradictoires, dans les réseaux sociaux, entre différentes communautés à travers le monde. L'estimation de l'intensité de ce conflit peut être utile pour mieux analyser la tendance et les conséquences de cette décision

politique. Dans le cas de la recherche d'information sociale, pour certains besoins d'information, mesurer l'intensité de la contradiction peut être utile pour identifier et classer les documents les plus controversés (par exemple les nouvelles, les événements, etc.). Dans notre cas, connaître l'intensité des opinions contradictoires sur un aspect spécifique (par exemple, *Lecturer, Speaker, Slide, Quiz*) d'un cours en ligne (en anglais) peut être utile pour savoir s'il y a certains éléments à améliorer dans ce cours. Table 1 présente une instance de commentaires contradictoires à propos de l'aspect *Speaker* (conférencier) d'un cours donné.

Ressource	Commentaire (gauche)	Aspect	Commentaire (droite)	Polarité	Notation
Cours [1]	The lecturer was an annoying	speaker	and very repetitive.	-0.9	1
	Passionate	speaker	and truly amazing things to learn	+0.7	4

TABLE *1 – Exemple de deux opinions contradictoires sur le "Speaker" d'un cours coursera*

Afin de concevoir notre approche, des tâches fondamentales sont effectuées. Premièrement, l'extraction automatique des aspects caractérisants ces commentaires. Deuxièmement, l'identification des opinions opposées autour de chacun de ces aspects en utilisant un modèle d'analyse des sentiments. Enfin, nous allons évaluer l'impact de certains critères (par exemple, le nombre de commentaires négatifs, le nombre de commentaires positifs) sur l'estimation de l'intensité de contradiction. Plus précisément, nous tentons de sélectionner les critères les plus efficaces et de les combiner avec des approches d'apprentissage pour prédire l'intensité de contradiction. Les principales contributions abordées dans cet article sont doubles :

- **(C1).** Une contradiction dans des commentaires liés à une ressource Web donnée signifie des opinions contradictoires exprimées sur un aspect spécifique, qui est une forme de diversité de sentiments autour de l'aspect au sein de la même ressource. Mais en plus de détecter la contradiction, il est souhaitable d'estimer son intensité. Par conséquent, nous essayons de répondre aux questions de recherche suivantes :
  - ◇ **QR1.** Comment estimer/prédire l'intensité de la contradiction ?
  - ◇ **QR2.** Quel est l'impact de la prise en compte des polarités et des notations sur la prédiction de l'intensité des commentaires contradictoires ?
- **(C2).** La construction d'une collection de test issue du site Web des MOOC [2] `coursera.org`. Cette collection est utile pour l'évaluation des systèmes mesurant l'intensité de contradiction. Des études expérimentales orientées utilisateurs - *user studies* - ont été menées pour collecter les jugements de l'intensité de contradiction (*Not Contradictory, Very Low, Low, Strong* and *Very Strong*).

L'article est organisé comme suit. La section 2 présente certains travaux connexes. La section 3 détaille notre approche pour la prédiction de l'intensité des contradictions autour de certains aspects spécifiques. L'évaluation expérimentale est présentée dans la section 4. Enfin, la section 5 conclut l'article en annonçant des perspectives.

2 Vue d'ensemble : État de l'art

La détection de contradictions est un processus complexe qui nécessite souvent l'utilisation de plusieurs méthodes. Plusieurs travaux ont été proposés pour ces méthodes (détection des aspects, analyse de sentiments) mais à notre connaissance, très peu de travaux traitent de la détection et de la mesure de l'intensité de la contradiction. Dans cette section, nous allons brièvement présenter quelques approches de détection de controverses proches de nos travaux puis nous allons présenter les approches liées à la détection des aspects et l'analyse de sentiments, qui sont utiles pour introduire notre approche.

1. `https://www.coursera.org/learn/dog-emotion-and-cognition`
2. Massive Open Online Course

2.1 Approches de détection des contradictions et des controverses

Les études les plus liées à notre approche incluent Harabagiu *et al.* (2006), de Marneffe *et al.* (2008), Tsytsarau *et al.* (2010) et Tsytsarau *et al.* (2011), qui tentent de détecter une contradiction dans le texte. Il y a deux approches principales, où les contradictions sont définies comme une forme d'inférence textuelle (par exemple, *entailment identification*) et analysées en utilisant des technologies linguistiques. Harabagiu *et al.* (2006) ont proposé une approche d'analyse des contradictions en exploitant des caractéristiques linguistiques et sémantiques (ex. typologie de verbes), ainsi que des informations syntaxiques telles que la négation (ex. *I love you - I do not love you*) ou l'antonyme (des mots qui ont des significations opposées, c.-à-d. *hot-cold* ou *light-dark*). Leur travail définit les contradictions comme une implication textuelle (*textual entailment* [3]) qui est fausse, lorsque deux phrases expriment des informations mutuellement exclusives sur le même sujet. L'antonymie peut donner lieu à une contradiction lorsque les gens utilisent ces mots pour décrire un sujet.

Poursuivant l'amélioration des travaux dans ce sens, de Marneffe *et al.* (2008) a introduit une classification des contradictions consistant en 7 types qui se distinguent par les caractéristiques qui contribuent à une contradiction, par exemple, l'antonyme, la négation, les discordances numériques qui peuvent être causées par des données erronées : «*there are 7 wonders of the world - the number of wonders of the world are 9*». Ils ont défini les contradictions comme une situation où il est extrêmement improbable que deux phrases soient vraies lorsqu'elles sont ensemble. Tsytsarau *et al.* (2010), (2011) ont proposé une solution automatique et évolutive pour le problème de détection de contradictions en utilisant l'analyse des sentiments. L'intuition de leur approche est que lorsque la valeur agrégée des sentiments (sur un sujet et un intervalle de temps spécifiques) est proche de zéro, alors que la diversité des sentiments est élevée, la contradiction devrait être élevée.

Un autre thème lié à notre travail concerne la détection des controverses et des disputes. Dans la littérature, la détection des controverses a été abordée à la fois par des méthodes supervisées comme dans Popescu & Pennacchiotti (2010), Balasubramanyan *et al.* (2012) et Wang *et al.* (2014) ou par des méthodes non supervisées comme dans Badache *et al.* (2017), Dori-Hacohen & Allan (2015), Garimella *et al.* (2016) et Jang *et al.* (2016). Pour détecter les événements controversés sur Twitter (par exemple, l'accusation de viol de David Copperfield entre 2007 et 2010) [4], Popescu & Pennacchiotti (2010) ont proposé un classifieur basé sur un apprentissage par arbre de décision et un ensemble de caractéristiques telles que les parties du discours, la présence de mots issus du lexique d'opinion ou de controverse, et les interactions des utilisateurs (*retweet* et *reply*). Balasubramanyan *et al.* (2012) ont étendu le modèle LDA (Latent Dirichlet Allocation) supervisé pour prédire comment les membres des différentes communautés politiques réagiront émotionnellement au même sujet c-à-d. la prédiction du niveau de controverse associé à ce sujet. Des classifieurs de type machine à vecteurs de support et régression logistique ont également été proposés par Wang *et al.* (2014) et par Wang & Cardie (2014) pour détecter les disputes dans les discussions sur la page de Wikipedia. Par exemple dans le cas des commentaires sur les modifications des pages Wikipedia [5].

D'autres travaux ont également exploité Wikipédia pour détecter et identifier des sujets controversés sur le Web Dori-Hacohen & Allan (2015), Jang & Allan (2016) et Jang *et al.* (2016). Dori-Hacohen & Allan (2015) et Jang & Allan (2016) ont proposé d'aligner les pages Web aux pages de Wikipedia en supposant qu'une page traite un sujet controversé si la page Wikipedia décrit un sujet lui-même controversé. La nature controversée ou non controversée d'une page Wikipedia est automatiquement détectée sur la base des métadonnées et des discussions associées à la page. Jang *et al.* (2016) ont construit un modèle de langage des sujets controversés appris sur des articles de Wikipédia et utilisé ensuite pour identifier si une page Web est controversée.

La détection des controverses dans les médias sociaux a également été abordée sans supervision en se basant sur les interactions entre les différents utilisateurs Garimella *et al.*

3. https://en.wikipedia.org/wiki/Textual_entailment

4. http://news.bbc.co.uk/2/hi/entertainment/8456070.stm

5. https://www.wikipedia.org/

(2016). Garimella *et al.* (2016) ont proposé d'autres approches de mesure de contradiction basées sur la topologie du réseau, telles que la marche aléatoire (*random walk*), la centralité intermédiaire (*betweenness centrality*) et le plongement de graphe à faible dimension (*low-dimensional graph embeddings*. Les auteurs ont testé des méthodes simples basées sur le contenu et ont noté leur inefficacité par rapport aux méthodes basées sur un graphe utilisateur. D'autres études tentent de détecter des controverses sur des domaines spécifiques, par exemple dans les news Tsytsarau *et al.* (2014) ou dans l'analyse du débat Qiu *et al.* (2013).

Cependant, à notre connaissance, aucun travail antérieur n'a abordé, de manière explicite et concrète, l'intensité de la contradiction ou de la controverse. Dans cet article, contrairement aux travaux antérieurs, plutôt que d'identifier seulement la controverse autour d'un sujet choisi au préalable (par exemple, aspect lié aux nouvelles politiques), nous nous concentrons également sur l'estimation de l'intensité des opinions contradictoires autour de sujets spécifiques. Nous proposons de mesurer l'intensité de la contradiction en utilisant certaines caractéristiques (par exemple, la notation et la polarité).

2.2 Approches de détection des aspects

Les premières tentatives de détection d'aspects ont été basées sur l'approche classique d'extraction d'information (IE) en exploitant les phrase nominales fréquentes Hu & Liu (2004). De telles approches fonctionnent bien dans la détection des aspects qui sont sous la forme d'un seul nom, mais sont moins efficaces lorsque les aspects sont de faible fréquence. Dans le contexte de la détection d'aspects, bon nombre de travaux utilisent les CRF (*Conditional Random Fields*) ou les HMM (*Hidden Markov Models*). Parmi ces travaux, nous pouvons citer Hamdan *et al.* (2015) qui utilisent les CRF. D'autres méthodes sont non supervisées et ont prouvé leur efficacité tel que Titov & McDonald (2008) qui construisent un modèle thématique à grains multiples (Multi-Grain Topic Model). Nous pouvons aussi citer le modèle HASM (*unsupervised Hierarchical Aspect Sentiment Model*) proposé par Kim *et al.* (2013) qui permet de découvrir une structure hiérarchique du sentiment fondée sur les aspects dans les avis en ligne non labellés. Dans nos travaux, nous nous sommes inspirés de la méthode non supervisée développée par Poria *et al.* (2014) basée sur l'utilisation de règles d'extraction pour les avis sur les produits. Cette méthode est en cohérence avec nos données expérimentales issues de *coursera.org*.

2.3 Approches d'analyse de sentiments

L'analyse du sentiment a fait l'objet de très nombreuses recherches antérieures. Comme dans le cas de la détection d'aspects, les approches supervisées et non supervisées ont chacune leurs solutions. Ainsi, dans les approches non supervisées, nous pouvons citer les approches basées sur les lexiques telles que l'approche développée par Turney (2002) ou bien des méthodes basées sur des corpus comme les travaux de Mohammad *et al.* (2013). Au rang des approches supervisées, nous pouvons citer Pang *et al.* (2002) qui comme nombre de travaux perçoivent la tâche d'analyse de sentiments comme une tâche de classification et utilisent donc des méthodes comme les SVM (*Support Vector Machines*) ou les réseaux bayésiens. D'autres travaux récents sont basés sur les RNN (*Recursive Neural Network*) tels que les travaux de Socher *et al.* (2013). Comme le propos de cet article est de mesurer l'intensité de contradiction et que l'analyse de sentiments n'est qu'une étape du processus, nous avons utilisé l'approche proposée par Radford *et al.* (2017), dont son implémentation est publiquement disponible [6]. Nous décrivons cette méthode dans la section 3.1.2.

3 Notre approche : Prédiction de l'intensité des contradictions

Notre approche est basée à la fois sur la détection d'aspects dans les commentaires ainsi que sur l'analyse des sentiments du texte autour de ces aspects. En plus de la détection de

6. `https://github.com/openai/generating-reviews-discovering-sentiment`

contradiction, notre objectif est de prédire le niveau d'intensité de la contradiction en utilisant certains critères et caractéristiques. Ces caractéristiques sont liées à la notation et à la polarité des *commentaires-aspect* (texte autour d'un aspect donné).

3.1 Pré-traitement : Identification des polarités autour des aspects

Le pré-traitement est une étape clé pour l'analyse des commentaires (aspects et sentiments). Le module de pré-traitement se compose de trois étapes principales : d'une part, le marquage des termes (identification des noms, verbes, etc), par une analyse syntaxique, au sein des commentaires. Deuxièmement, les noms les plus fréquents dans l'ensemble des commentaires des différents documents sont extraits. Troisièmement, uniquement les noms entourés par des termes émotionnels sont considérés comme des aspects. Nous détaillons ces étapes dans ce qui suit.

3.1.1 Extraction des aspects

Dans notre étude, un aspect est une entité nominale fréquente dans les commentaires et entourée par des termes émotionnels. Afin d'extraire les aspects à partir du texte des commentaires, nous nous sommes basés sur le travail de Poria *et al.* (2014). Cette méthode correspond à nos données expérimentales (commentaires issus de *coursera*). De plus, les traitements suivants sont appliqués :

1. Calcul fréquentiel des termes constituant le corpus des commentaires,
2. Catégorisation des termes (Part-of-speech tagging) de chaque commentaire en utilisant *Stanford Parser* [7],
3. Sélection des termes ayant la catégorie nominale (NN, NNS) [8],
4. Sélection des noms avec des termes émotionnels dans leur voisinage de 5 mots (en utilisant *SentiWordNet* [9]). Le choix de 5 mots a été fait après plusieurs expérimentations,
5. Extraction des termes les plus fréquents (utilisés) dans le corpus parmi ceux sélectionnés dans l'étape précédente. Ces termes seront considérés comme des aspects.

Exemple : Soit $C = \{c_1, c_2, c_3\}$ un ensemble de 3 commentaires associés à un document D. Nous voulons extraire les aspects à partir de chacun des commentaires en appliquant les étapes décrites ci-dessus.
Nous avons c_1 =*"The lecturer was an annoying speaker and very repetitive. I just couldn't listen to him. . . I'm sorry. There was also so much about human development etc that I started to wonder when the info about dogs would start. . . . I found the formatting so different from other courses I've taken, that it was hard to get started and figure things out. Adding to that, was the constant interruption of the "paid certificate" page. If I answer "no" once, please leave me alone ! I also think it's a bit suspect for a prof to be plugging his own book for one of these courses."*

La table 2 récapitule les 5 étapes. Premièrement, nous calculons les fréquences des termes dans l'ensemble des commentaires (à titre d'exemple, les termes "course", "material", "assignments", "content", "lecturer" apparaissent 44219, 3286, 3118, 2947, 2705, respectivement). Deuxièmement, nous étiquetons grammaticalement chaque mots (par exemple, "NN", "NNS" signifient nom en singulier et nom en pluriel, respectivement [10]). Troisièmement, seul les termes de catégorie nominale sont sélectionnés. Quatrièmement, nous gardons uniquement les noms entourés par des termes appartenant au dictionnaire *SentiWordNet* (The *lecturer* was an annoying *speaker* and very repetitive). Enfin, nous considérons comme aspects utiles uniquement les noms qui figurent parmi les noms les plus fréquents dans le corpus des commentaires (l'aspect utile dans ce commentaire est *lecturer*).

7. http://nlp.stanford.edu:8080/parser/
8. https://cs.nyu.edu/grishman/jet/guide/PennPOS.html
9. http://sentiwordnet.isti.cnr.it/
10. http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Étape	Description
(1)	course : 44219, material : 3286, assignments : 3118, content : 2947, lecturer : 2705, ... terme$_i$
(2)	The/DT **lecturer**/NN was/VBD an/DT annoying/VBG **speaker**/NN and/CC very/RB repetitive/JJ ./. I/PRP just/RB could/MD n't/RB listen/VB to/TO him/PRP .../ : I/PRP 'm/VBP sorry/JJ ./. There/EX was/VBD also/RB so/RB much/JJ about/IN human/JJ **development**/NN **etc**/NN that/IN I/PRP started/VBD to/TO wonder/VB when/WRB the/DT **info**/NN about/IN **dogs**/NNS would/MD start/VB .../ : ./. I/PRP found/VBD the/DT **formatting**/NN so/RB different/JJ from/IN other/JJ **courses**/NNS I/PRP 've/VBP taken/VBN ,/, that/IN it/PRP was/VBD hard/JJ to/TO get/VB started/VBN and/CC figure/VB **things**/NNS out/RP ./. Adding/VBG to/TO that/DT ,/, was/VBD the/DT constant/JJ **interruption**/NN of/IN the/DT "/" paid/VBN **certificate**/NN "/" **page**/NN ./. If/IN I/PRP answer/VBZ "/" no/UH "/" once/RB ,/, please/VB leave/VB me/PRP alone/RB !/. I/PRP also/RB think/VBP it/PRP 's/VBZ a/DT bit/RB suspect/JJ for/IN a/DT **prof**/NN to/TO be/VB plugging/VBG his/PRP$ own/JJ **book**/NN for/IN one/CD of/IN these/DT **courses**/NNS ./.
(3)	lecturer, speaker, development, dogs, formatting, courses, interruption, certificate, page, prof
(4)	lecturer, speaker
(5)	lecturer

TABLE *2 – Les différentes étapes pour extraire les aspects dans un commentaire*

Une fois que nous avons défini la liste des aspects qui caractérisent notre collection de données, nous devons estimer la polarité des sentiments autour de ces aspects. La section suivante présente notre modèle d'analyse de sentiments.

3.1.2 Analyse de sentiments

Le sentiment porté par un commentaire sur un aspect donné (commentaire-aspect) est estimé en utilisant la méthode appelée *SentiNeuron* [11]. *SentiNeuron* est un modèle non supervisé proposé par Radford *et al.* (2017) pour détecter les signaux de sentiment dans les commentaires. Cette approche est basée sur les réseaux de neurones récurrent (*recurrent neural network*) de type mLSTM (multiplicative Long Short-Term Memory). Radford *et al.* (2017) ont également trouvé qu'une unité dans le mLSTM correspond directement au sentiment de la sortie. Les auteurs ont mené une série d'expérimentations sur plusieurs collections de tests telles que les collections des commentaires issues d'Amazon McAuley *et al.* (2015) et d'IMDb [12]. Cette approche fournie une précision de 91.8%, et surpasse de manière significative plusieurs approches de l'état de l'art telles que celles présentées dans Looks *et al.* (2017). Nous notons que le terme polarité signifie sentiment, c'est une valeur comprise entre -1 et 1.

3.2 Collection de test : coursera.org

3.2.1 Données collectées

A notre connaissance, il n'existe pas à ce jour de collection de test standard, contenant des informations comme les aspects, les notations et les polarités des commentaires, pour évaluer l'efficacité des systèmes de détection de contradictions dans les commentaires. De ce fait, dans le but d'expérimenter l'efficacité de notre approche, nous avons collecté 2244 ressources en anglais extraites du site "coursera.org" via son API [13], entre le 10 et le 14 octobre 2016.

11. https://github.com/openai/generating-reviews-discovering-sentiment
12. https://www.cs.cornell.edu/personnes/pabo/film-review-data/
13. https://building.coursera.org/app-platform/catalog

Chaque ressource décrit un cours et est représentée par un ensemble de métadonnées. Pour chaque cours, nous avons collecté également ses commentaires et ses notations via le *parsing* des pages web des cours (voir les statistiques sur la table 3).

Champ	Nombre
Cours	2244
Cours notés	1115
Commentaires	73873
notations	298326
Commentaires avec notation ★☆☆☆☆	1705
Commentaires avec notation ★★☆☆☆	1443
Commentaires avec notation ★★★☆☆	3302
Commentaires avec notation ★★★★☆	12202
Commentaires avec notation ★★★★★	55221

TABLE *3 – Les chiffres des données de la collection de Coursera.org*

Nous avons pu capturer automatiquement 22 aspects utiles à partir de l'ensemble des commentaires (voir table 4). La table 4 présente les statistiques sur les 22 aspects détectés, par exemple, pour l'aspect *Slide* nous avons enregistré : 56 notations d'une étoile, 64 notations de deux étoiles, 81 notations de trois étoiles, 121 notations de quatre étoiles, 115 notations de cinq étoiles, 131 commentaires avec une polarité négative, 102 commentaires avec une polarité positive ainsi que 192 commentaires et 41 cours concernant cet aspect.

Aspects	Not1	Not2	Not3	Not4	Not5	Négatif	Positif	Comment	Cours
Assignment	204	208	333	840	1726	1057	1763	2384	186
Content	176	179	341	676	1641	505	1496	1883	207
Exercise	29	46	94	290	693	195	531	673	58
Information	100	123	238	523	1389	299	1165	1359	143
Instructor	129	106	122	302	1514	295	1107	1322	140
Knowledge	74	72	121	400	1604	905	791	1243	178
Lecture	185	206	290	613	1762	763	1508	1988	208
Lecturer	32	41	48	85	461	55	193	236	39
Lesson	40	59	75	224	712	187	420	554	84
Material	191	203	328	722	2234	784	1693	2254	237
Method	19	23	40	125	404	53	187	224	31
Presentation	46	50	75	142	413	93	196	274	54
Professor	76	74	129	452	3001	331	2234	2369	151
Quality	55	53	51	110	372	113	170	262	54
Question	94	98	172	284	356	311	289	502	104
Quiz	151	155	221	401	581	481	475	824	128
Slide	56	64	81	121	115	131	102	192	47
Speaker	17	15	34	70	170	34	72	103	24
Student	140	105	171	383	1035	519	709	1066	172
Teacher	62	46	82	293	2180	248	1481	1642	119
Topic	67	89	176	437	1154	236	951	1066	130
Video	228	238	356	707	1614	941	1421	2058	245
Nombre total : 22 aspects détectés									

TABLE *4 – Statistiques sur les aspects issus des commentaires de Coursera.org*

3.2.2 Jugements par les utilisateurs (contradictions et sentiments)

Afin d'obtenir des jugements de contradictions et de sentiments pour un aspect donné : 1) nous avons demandé à trois utilisateurs d'évaluer la classe de sentiment pour chaque

commentaires-aspect de 1100 cours ; 2) trois autres utilisateurs ont évalué le degré de contradiction entre les commentaires-aspect. En moyenne 60 commentaires-aspect par cours sont jugés manuellement pour chaque aspect (totalement : 66104 commentaires-aspect de 1100 cours, c'est-à-dire 50 cours pour chaque aspect). Nous notons que chaque aspect a été jugé par 3 utilisateurs.

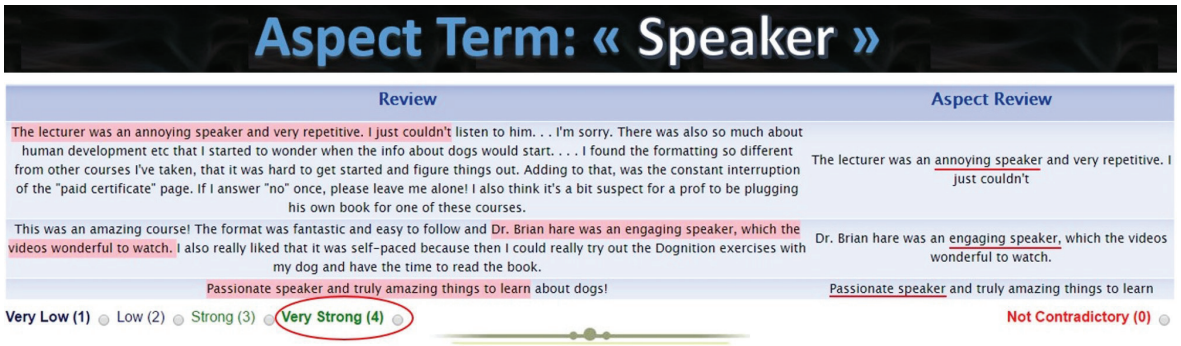


FIGURE *1 – Interface du système d'évaluation*

Pour évaluer les sentiments et les contradictions dans les commentaires-aspect de chaque cours, nous utilisons une échelle de notation de 3 points pour les sentiments : (*Negative, Neutral, Positive*) ; et une échelle de 5 points pour les contradictions : *Not Contradictory, Very Low, Low, Strong* et *Very Strong* (voir la figure 1).

Nous avons analysé le degré d'accord entre les évaluateurs des contradictions pour chaque aspect avec la mesure Kappa Cohen k Cohen (1960). Cet indicateur prend en compte la proportion d'accord entre les évaluateurs et la proportion de l'accord attendu entre les évaluateurs par hasard. La mesure de Kappa est égale à 1 si les évaluateurs sont complètement d'accord, 0 s'ils ne sont d'accord que par hasard. k est négatif si l'accord entre évaluateurs est pire que l'aléatoire. Comme nous avons trois évaluateurs par aspect, la valeur Kappa a été calculée pour chaque paire d'évaluateurs, puis leur moyenne a été calculée.

La figure 2 montre la distribution de la mesure kappa pour chaque aspect. Nous constatons que la mesure de l'accord varie de 0.60 à 0.91. La mesure moyenne d'accord entre les évaluateurs est de 80%, ce qui correspond à un accord fort. Concernant l'analyse du degré d'accord entre les évaluateurs des sentiments, nous avons trouvé un accord de Kappa $k = 0.78$, qui correspond aussi à un accord fort.

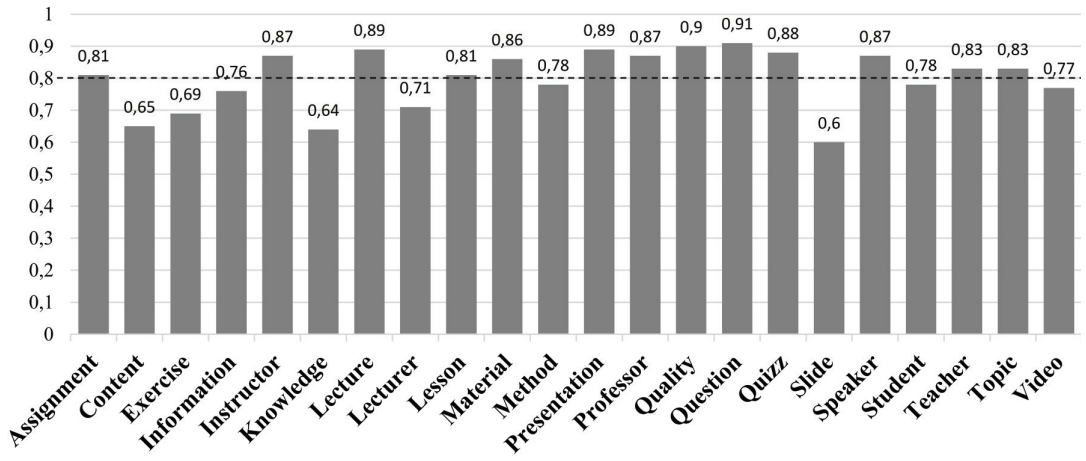


FIGURE *2 – Répartition de la mesure Kappa k par aspect. < 0 désaccord, $0.0 - 0.2$ accord très faible, $0.21 - 0.40$ accord faible, $0,41 - 0,6$ accord modéré, $0,61 - 0,80$ accord fort, $0,81 - 1$ accord parfait.*

3.3 Identification des critères les plus efficaces

Dans cette étude, nous nous sommes appuyés sur des algorithmes de sélection d'attributs pour déterminer les critères les plus importants pour la tâche de prédiction d'intensité de contradiction. Les algorithmes de sélection d'attributs visent à identifier et supprimer le maximum d'information inutile, redondante et non pertinente en amont d'un processus à base d'apprentissage Hall & Holmes (2003). Ils permettent également de sélectionner de manière automatique les sous ensembles de critères permettant d'avoir les meilleurs résultats. Nous avons utilisé Weka [14] (dernière version stable 2018 : 3.8.2), un outil open-source écrit entièrement en Java et qui rassemble un bon ensemble de techniques d'apprentissage et des techniques de sélection d'attributs.

c_i	Critère	Description
c_1	NegCom	Nombre de commentaires négatifs sur le document
c_2	PosCom	Nombre de commentaires positifs sur le document
c_3	TotalCom	Nombre total des commentaires sur le document
c_4	Not1	Nombre de commentaires avec notation ★☆☆☆☆
c_5	Not2	Nombre de commentaires avec notation ★★☆☆☆
c_6	Not3	Nombre de commentaires avec notation ★★★☆☆
c_7	Not4	Nombre de commentaires avec notation ★★★★☆
c_8	Not5	Nombre de commentaires avec notation ★★★★★
c_9	VarNot	Variation des notations (écart type selon Pearson & Stephens (1964))
c_{10}	VarPol	Variation des polarités (écart type selon Pearson & Stephens (1964))

TABLE *5 – Liste des critères exploités*

La table 5 présente les 10 critères que nous avons considérés pour prédire l'intensité de contradiction dans les commentaires. La nature des critères c_1 jusqu'au critère c_8 est un simple comptage, par exemple les critères c_1 et c_2 liés à la polarité représentent le nombre de commentaires négatifs et positifs sur le document, respectivement. Les critères c_4, c_5, c_6, c_7 et c_8 sont liés à la notation. La notation est une note sur une échelle de 1 à une valeur max de 5, où 3 signifie "moyen" et 5 signifie "excellent". Concernant les deux derniers critères c_9 et c_{10}, ils représentent la variation des notations et des polarités des commentaires pour un aspect donné associés à un document (un cours dans notre cas). Ces deux critères sont calculé en se basant sur la formule de l'écart type suivante, proposée par Pearson & Stephens (1964) :

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x - \bar{x})^2}{n}} \tag{1}$$

Où x est la valeur du critère (notation, polarité), $\bar{x}$ est la moyenne de l'échantillon du critère concerné, et n est la taille de l'échantillon.

Dans cette étude, nous avons procédé ainsi : 50 cours avec leurs commentaires pour chaque aspect (22 aspects) de la collection *coursera* ont été extraits aléatoirement. Ensuite, nous avons considéré l'échelle des 4 points comme des classes reflétant l'intensité des contradictions autour d'un aspect spécifique : *Very Low, Low, Strong* et *Very Strong*, selon les jugements des évaluateurs. L'ensemble résultant contient 1100 cours (instances) répartis selon leur classe d'intensité de contradiction comme suit :

— 230 Very Low
— 264 Low
— 330 Strong
— 276 Very Strong

14. `http://www.cs.waikato.ac.nz/ml`

Les classes de ces ensembles sont déséquilibrées, or lorsque le nombre d'éléments d'une classe dans une collection d'apprentissage dépasse considérablement les autres échantillons des autres classes, un classifieur tend à prédire les échantillons de la classe majoritaire et peut ignorer complètement les classes minoritaires Yen & Lee (2006). Pour cette raison, nous avons appliqué une approche de sous-échantillonnage (en réduisant le nombre d'échantillons qui ont la classe majoritaire) pour générer des collections équilibrées composées de :

— 230 Very Low
— 230 Low
— 230 Strong
— 230 Very Strong

Les classes *Low, Strong* et *Very Strong* ont été sélectionnées aléatoirement. Enfin, nous avons appliqué les algorithmes de sélection d'attributs sur les quatre ensembles obtenus, pour 5 itérations de validation croisée (5-folds cross-validation).

Dans notre cas, les algorithmes de sélection d'attributs consistent à attribuer un score à chaque critère en fonction de sa signification vis-à-vis la classe d'intensité de contradiction (*Very Low, Low, Strong* et *Very Strong*). Ces algorithmes fonctionnent différemment : certains retournent un classement d'importance des critères (par exemple, *FilteredAttributeEval*), tandis que d'autres retournent le nombre de fois qu'un critère donné a été sélectionné par un algorithme dans une validation croisée (par exemple, *FilteredSubsetEval*). Nous notons que nous avons utilisé pour chaque algorithme le paramétrage par défaut fourni par Weka.

Nous avons appliqué une validation croisée à 5 itérations pour 10 critères, c'est-à-dire $n = 10$. La table 6 présente les critères sélectionnés par les algorithmes de sélection d'attributs. Nous avons utilisé deux types de ces algorithmes : a) ceux qui utilisent des méthodes de classement pour ordonner les critères sélectionnés (la métrique dans la table est [Rank]) ; et b) ceux qui utilisent des méthodes de recherche qui indiquent combien de fois le critère a été sélectionné pendant la tâche de la validation croisée (la métrique dans la table est [Folds]). Un critère fortement préféré (choisi) par l'algorithme de sélection est un critère bien classé, c'est-à-dire *Rank* $= 1$ et fortement sélectionné, c'est-à-dire *Folds* $= 5$.

Algorithm	Metric	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
CfsSubsetEval	[Folds]	5	5	2	0	0	0	0	0	5	5
WrapperSubsetEval	[Folds]	4	4	4	2	0	0	0	2	5	5
ConsistencySubsetEval	[Folds]	5	5	4	2	1	1	2	2	5	5
FilteredSubsetEval	[Folds]	5	5	4	3	2	2	3	3	5	5
	Moyenne	4.75	4.75	3.5	1.75	0.75	0.75	1.25	1.75	5	5
ChiSquaredAttributeEval	[Rank]	3	4	5	7	9	10	8	6	2	1
FilteredAttributeEval	[Rank]	4	3	5	7	9	10	8	6	2	1
GainRatioAttributeEval	[Rank]	3	4	5	7	9	10	8	6	2	1
InfoGainAttributeEval	[Rank]	3	4	5	7	9	10	8	6	1	2
OneRAttributeEval	[Rank]	4	3	5	7	9	10	8	6	2	1
ReliefFAttributeEval	[Rank]	4	3	6	8	9	10	7	5	1	2
SVMAttributeEval	[Rank]	4	3	5	7	9	10	8	6	2	1
SymetricalUncertEval	[Rank]	3	4	5	7	9	10	8	6	2	1
	Moyenne	3.5	3.5	5.12	7.12	9	10	7.87	5.87	1.75	1.25

TABLE *6 – Les critères sélectionnés par les algorithmes de sélection d'attributs*

La table 6 montre que les critères c_{10} : *VarPol*, c_9 : *VarNot*, c_1 : *NegCom* et c_2 : *PosCom* sont les plus sélectionnés et les mieux classés par rapport aux autres critères. Les critères c_3 : *TotalCom*, c_4 : *Not1* et c_8 : *Not5* sont modérément favorisés par les algorithmes de sélection d'attributs, à l'exception de l'algorithme *CfsSubsetEval* qui n'a pas sélectionné c_4 et c_8. Les critères c_5, c_6 et c_7 ne sont pas sélectionnés à la fois par les algorithmes *CfsSubsetEval* et *WrapperSubsetEval*. Enfin, les critères les plus faibles et les plus désavantagés sont c_5 : *Not2* et c_6 : *Not3*, ils sont ordonnés au rang 9 et 10, respectivement.

3.4 Apprentissage des critères pour prédire l'intensité de contradiction

D'autres expérimentations ont été menées en exploitant ces critères dans des approches supervisées basées sur des modèles d'apprentissage. Nous avons utilisé les instances (les cours) des 22 aspects de la collection *coursera.org* comme ensembles d'apprentissage. Nous avons ensuite utilisé trois algorithmes d'apprentissage. Ce choix s'explique par le fait qu'ils ont souvent montré leur efficacité dans les tâches d'analyse de texte : SVM Vosecky *et al.* (2012), J48 (implémentation C4.5) Quinlan (1993) et Naive Bayes Yuan *et al.* (2012). L'entrée de chaque algorithme est un vecteur de critères (voir table 5), soit tous les critères ou seulement les critères sélectionnés par un algorithme de sélection précis. Les algorithmes d'apprentissage prédisent la classe d'intensité de contradiction pour les cours (*Very Low, Low, Strong* et *Very Strong*). Enfin, nous avons appliqué une validation croisée pour 5 itérations (5-folds cross-validation). La figure 3 illustre le processus d'apprentissage que nous avons mis en place pour l'évaluation des critères.

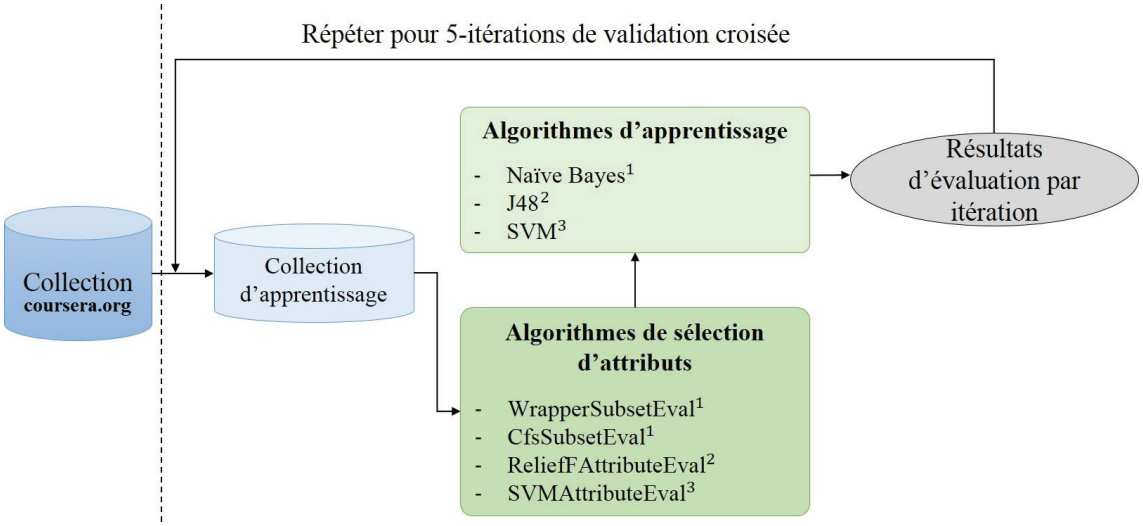


FIGURE *3 – Processus d'apprentissage en utilisant les algorithmes de sélection*

Nous rappelons que la phase des algorithmes de sélection d'attributs a fait ressortir les ensembles de critères suivants (voir la table 7).

Algorithmes de sélection	Critères
CfsSubsetEval	c_1, c_2, c_3, c_9, c_{10}
WrapperSubsetEval	c_1, c_2, c_3, c_4, c_8, c_9, c_{10}
Other algorithms	c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}

TABLE *7 – Ensembles des critères sélectionnés*

La question à ce stade est liée à la spécification du vecteur de critères d'entrée pour les algorithmes d'apprentissage, soit on prend tous les critères, soit on garde uniquement ceux sélectionnés par les techniques de sélection d'attributs. Dans ce cas, avec quels algorithmes d'apprentissage ces derniers seront combinés.

Afin de prendre en compte les critères choisis par les algorithmes de sélection dans des modèles d'apprentissage, nous nous sommes basés sur les travaux de Hall & Holmes (2003). Ils ont étudié l'efficacité de certaines techniques de sélection d'attributs en les confrontant avec les techniques d'apprentissage. Étant donné que la performance des critères diffère d'une technique d'apprentissage à une autre, ils ont identifié les meilleures techniques de sélection d'attributs permettant de retrouver les critères les plus performants en fonction des techniques d'apprentissage à utiliser. En se basant sur leur étude, nous avons utilisé les mêmes couples des techniques d'apprentissage et des techniques de sélection d'attributs :

— L'ensemble des critères sélectionnés par *CfsSubsetEval* (CFS) et *WrapperSubsetEval* (WRP) sont appris par le modèle Naïve Bayes.
— L'ensemble des critères sélectionnés par *ReliefFAttributeEval* (RLF) sont appris par le modèle J48 (C4.5 implementation).
— L'ensemble des critères sélectionnés par *SVMAttributeEval* (SVM) sont appris par le modèle SVM à multi-classes (appelé *SMO function* sur Weka).

Afin de vérifier la significativité des résultats par rapport aux résultats de Naïve Bayes (considérés comme références - résultats de base), nous avons effectué le test de Student Student (1908). Nous avons attaché * (forte signification) et ** (très forte signification) aux résultats de la table 8 quand *p-value <0.05* et *p-value <0.01*, respectivement.

Classifieurs	Classes (Niveaux d'intensité)	Techniques de sélection	Tous les critères
Naïve Bayes (Baseline)	Very Low	0.81 (CFS)	0.71
	Low	0.38 (CFS)	0.34
	Strong	0.75 (CFS)	0.66
	Very Strong	0.78 (CFS)	0.69
	Moyenne	0.68 (CFS)	0.60
	Very Low	0.86 (WRP)	0.72
	Low	0.46 (WRP)	0.38
	Strong	0.76 (WRP)	0.63
	Very Strong	0.80 (WRP)	0.67
	Moyenne	0.72 (WRP)	0.60
SVM	Very Low	0.88* (SVM)	0.88*
	Low	0.72** (SVM)	0.72**
	Strong	0.78* (SVM)	0.78*
	Very Strong	0.90** (SVM)	0.90**
	Moyenne	0.82** (SVM)	0.82**
J48	Very Low	0.97** (RLF)	0.97**
	Low	0.92** (RLF)	0.92**
	Strong	0.97** (RLF)	0.97**
	Very Strong	0.98** (RLF)	0.98**
	Moyenne	0.96** (RLF)	0.96**

TABLE *8 – Les résultats de précision Weka pour les techniques d'apprentissage automatique*

La table 8 présente les résultats des trois algorithmes d'apprentissage des critères ressortis de l'étude utilisant les techniques de sélection d'attributs. Les résultats sont discutés ci-dessous pour chaque algorithme d'apprentissage.

3.4.1 Résultats obtenus par Naïve Bayes (Baseline)

Les résultats en termes de précision obtenus en utilisant des algorithmes de sélection CFS et WRP avec NaiveBayes, sont de 0.68 et 0.72, respectivement. Ces résultats dépassent ceux obtenus en utilisant tous les critères (précision : 0.60). En effet, nous avons enregistré des taux d'amélioration moyens de 14% et 20% pour Naïve Bayes en utilisant seulement les critères sélectionnés par CFS (0.68) et WRP (0.72), respectivement, par apport au résultat obtenu en utilisant tous les critères (0.60). Par conséquent, les approches d'apprentissage automatique peuvent donner une meilleure efficacité (précision) quand ils sont combinés avec des approches de sélection d'attributs. Les meilleures précisions sont obtenues pour les classes *Very Strong, Strong* et *Very Low*. Il semble que la classe *Low* est difficile à prédire avec Naïve Bayes, tout en utilisant à la fois les deux algorithmes de sélection CFS (0.38) et WRP (0.46).

3.4.2 Résultats obtenus par SVM

Les résultats obtenus par SVM en utilisant l'algorithme de sélection *SVMAttributeEval*, où tous les critères ont été sélectionnés, sont meilleurs par rapport à ceux obtenus par Naïve Bayes. Nous avons enregistré des taux d'amélioration moyens de 21% et 14% pour SVM par rapport à Naïve Bayes en utilisant CFS et WRP, respectivement. Nous avons également remarqué que SVM était capable de prédire la classe *Low* avec une meilleure précision que celle fournie par Naïve Bayes. Même si l'algorithme SVM est un peu coûteux en termes de temps d'exécution par rapport à Naïve Bayes, il reste favoriser pour obtenir des résultats significatifs en termes de précision.

3.4.3 Résultats obtenus par J48

Les résultats confirment que l'arbre de décision J48 est le modèle le plus approprié, il prend en compte tous les critères de manière plus efficace que les autres configurations. Les taux d'amélioration moyens par rapport à Naïve Bayes (en utilisant CFS et WRP) et SVM sont 41%, 33% et 17%, respectivement. En outre, les améliorations sont également fortement significatives pour chaque classe par rapport à SVM et Naïve Bayes. La classe *Low*, difficile à prédire avec les configurations précédentes, a été prédite avec une très forte précision de 92%. Comparées à Naïve Bayes (en utilisant CFS et WRP) et SVM, les améliorations enregistrées concernant la classe *Low* sont de 142%, 100% et 28%, respectivement.

Enfin, tous ces résultats expérimentaux montrent clairement que l'approche proposée permet de détecter de manière significative l'intensité de la contradiction dans les commentaires. Nous avons constaté que les résultats obtenus, par les deux algorithmes CFS et WRP, confirment l'hypothèse lancé par Hall et Holmes. C'est en effet les deux seuls cas pour lequel les résultats de précision obtenus avec la sélection d'attributs, soient 0.68 (CFS) et 0.72 (WRP), dépassent l'utilisation de tous les critères, 0.60 en termes de précision. Ces améliorations montrent l'intérêt de combiner les algorithmes de sélection d'attributs avec les modèles d'apprentissage. En outre, le modèle J48 a donné les meilleures améliorations par rapport à toutes les autres configurations. Nous concluons que les ressources (cours) ayant des opinions plus diversifiées (commentaires positifs et négatifs), sont susceptibles d'avoir des contradictions avec différents niveaux d'intensité.

4 Conclusion

Cet article propose une approche supervisée exploitant un ensemble de critères permettant de prédire l'intensité de la contradiction, en attirant l'attention sur les aspects dans lesquels les utilisateurs ont des opinions contradictoires. L'intuition derrière l'approche proposée est que les notations et les sentiments associés aux commentaires sur un aspect spécifique peuvent être considérés comme des critères (ex. diversité des sentiments et des notations en fonction de l'écart-type) pour mesurer l'intensité de contradiction. L'évaluation expérimentale menée sur la collection issue de *coursera.org* montre que les critères *NegCom, PosCom, VarNot* et *VarPol* sont les plus fructueux pour prédire l'intensité de la contradiction. De plus, les algorithmes d'apprentissage basés sur les critères les plus pertinents selon les algorithmes de sélection d'attributs sont généralement mieux comparés à ceux obtenus lorsque les algorithmes de sélection d'attributs sont ignorés. L'algorithme J48 apporte les meilleurs résultats par rapport à Naïve Bayes et SVM. Enfin, nous notons que nous sommes conscients que l'évaluation de notre approche est encore limitée. La principale faiblesse de notre approche est sa dépendance à la qualité des modèles de sentiments et d'extraction d'aspect. D'autres expérimentations à plus grande échelle sur d'autres types de collections sont également envisagées. Ceci étant même avec ces éléments simples, les premiers résultats obtenus nous encouragent à investir davantage cette piste.

Références

AWADALLAH A. H., ABU-JBARA A. & RADEV D. R. (2012). Detecting subgroups in online discussions by modeling positive and negative relations among participants. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, p. 59–70.

BADACHE I. & BOUGHANEM M. (2014). Harnessing social signals to enhance a search. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 01*, WI-IAT '14, p. 303–309, Washington, DC, USA.

BADACHE I. & BOUGHANEM M. (2017a). Emotional social signals for search ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, p. 1053–1056, New York, NY, USA : ACM.

BADACHE I. & BOUGHANEM M. (2017b). Fresh and diverse social signals : Any impacts on search ? In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, p. 155–164, New York, NY, USA : ACM.

BADACHE I., FOURNIER S. & CHIFU A. (2017). Finding and quantifying temporal-aware contradiction in reviews. In *Information Retrieval Technology - 13th Asia Information Retrieval Societies Conference, AIRS 2017, Jeju Island, South Korea, November 22-24, 2017, Proceedings*, p. 167–180.

BALASUBRAMANYAN R., COHEN W. W., PIERCE D. & REDLAWSK D. P. (2012). Modeling polarizing topics : When do different political communities respond differently to the same news ? In *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46.

DE MARNEFFE M., RAFFERTY A. N. & MANNING C. D. (2008). Finding contradictions in text. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, p. 1039–1047.

DORI-HACOHEN S. & ALLAN J. (2015). Automated controversy detection on the web. In *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, p. 423–434.

GARIMELLA K., MORALES G. D. F., GIONIS A. & MATHIOUDAKIS M. (2016). Quantifying controversy in social media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*, p. 33–42.

HALL M. A. & HOLMES G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans. Knowl. Data Eng.*, **15**(6), 1437–1447.

HAMDAN H., BELLOT P. & BÉCHET F. (2015). Lsislif : CRF and logistic regression for opinion target extraction and sentiment polarity analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, Colorado, USA, June 4-5, 2015*, p. 753–758.

HARABAGIU S. M., HICKL A. & LACATUSU V. F. (2006). Negation, contrast and contradiction in text processing. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, p. 755–762.

HTAIT A., FOURNIER S. & BELLOT P. (2016). LSIS at semeval-2016 task 7 : Using web search engines for english and arabic unsupervised sentiment intensity prediction. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, p. 469–473.

HU M. & LIU B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, p. 168–177.

JANG M. & ALLAN J. (2016). Improving automated controversy detection on the web. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, p. 865–868.

JANG M., FOLEY J., DORI-HACOHEN S. & ALLAN J. (2016). Probabilistic approaches to controversy detection. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, p. 2069–2072.

KIM S., ZHANG J., CHEN Z., OH A. H. & LIU S. (2013). A hierarchical aspect-sentiment model for online reviews. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA*.

LOOKS M., HERRESHOFF M., HUTCHINS D. & NORVIG P. (2017). Deep learning with dynamic computation graphs. *CoRR*, **abs/1702.02181**.

MCAULEY J. J., PANDEY R. & LESKOVEC J. (2015). Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, p. 785–794.

MOHAMMAD S., KIRITCHENKO S. & ZHU X. (2013). Nrc-canada : Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14-15, 2013*, p. 321–327.

MUKHERJEE A. & LIU B. (2012). Mining contentions from discussions and debates. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, p. 841–849.

PANG B., LEE L. & VAITHYANATHAN S. (2002). Thumbs up ? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, Philadelphia, PA, USA, July 6-7, 2002*.

PEARSON E. & STEPHENS M. (1964). The ratio of range to standard deviation in the same normal sample. *Biometrika*, **51**(3/4), 484–487.

POPESCU A. & PENNACCHIOTTI M. (2010). Detecting controversial events from twitter. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, p. 1873–1876.

PORIA S., CAMBRIA E., KU L., GUI C. & GELBUKH A. F. (2014). A rule-based approach to aspect extraction from product reviews. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media, SocialNLP@COLING, Dublin, Ireland, August 24, 2014*, p. 28–37.

QIU M., YANG L. & JIANG J. (2013). Modeling interaction features for debate side clustering. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, p. 873–878.

QUINLAN J. R. (1993). *C4.5 : Programs for Machine Learning*. Morgan Kaufmann.

RADFORD A., JÓZEFOWICZ R. & SUTSKEVER I. (2017). Learning to generate reviews and discovering sentiment. *CoRR*, **abs/1704.01444**.

SOCHER R., PERELYGIN A., WU J., CHUANG J., MANNING C. D., NG A. & POTTS C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, volume 1631, p. 1631–1642.

STUDENT (1908). The probable error of a mean. *Biometrika*, **6**(1), 1–25.

TITOV I. & MCDONALD R. T. (2008). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*, p. 111–120.

TSYTSARAU M., PALPANAS T. & CASTELLANOS M. (2014). Dynamics of news events and social media reaction. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, p. 901–910.

TSYTSARAU M., PALPANAS T. & DENECKE K. (2010). Scalable discovery of contradictions on the web. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, p. 1195–1196.

TSYTSARAU M., PALPANAS T. & DENECKE K. (2011). Scalable detection of sentiment-based contradictions. *DiversiWeb, WWW*, **1**, 9–16.

TURNEY P. D. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, p. 417–424.

VOSECKY J., LEUNG K. W. & NG W. (2012). Searching for quality microblog posts : Filtering and ranking based on content analysis and implicit links. In *Database Systems for Advanced Applications - 17th International Conference, DASFAA 2012, Busan, South Korea, April 15-19, 2012, Proceedings, Part I*, p. 397–413.

WANG L. & CARDIE C. (2014). A piece of my mind : A sentiment analysis approach for online dispute detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2*, p. 693–699.

WANG L., RAGHAVAN H., CARDIE C. & CASTELLI V. (2014). Query-focused opinion summarization for user-generated content. In *COLING 2014, 25th International Conference on Computational Linguistics, August 23-29, 2014, Dublin, Ireland*, p. 1660–1669.

YEN S.-J. & LEE Y.-S. (2006). Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. p. 731–740.

YUAN Q., CONG G. & MAGNENAT-THALMANN N. (2012). Enhancing naive bayes with various smoothing methods for short text classification. In *Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume)*, p. 645–646.