Nouvelle approche anaphorique pour le résumé automatique des textes d'opinions dans les tweets

Rania Othman[1], **Rami Belkaroui**[1], **Rim Faiz**[2]

1. Institut Supérieur de Gestion de Tunis, LARODEC, Université de Tunis, Tunisia
rania.othman@gmx.com ; rami.belkaroui@gmail.com

2. IHEC Carthage, LARODEC, Université de Carthage, Tunisia
rim.faiz@ihec.rnu.tn

RÉSUMÉ. Fournir un résumé automatique des opinions exprimées via Twitter est un thème émergent ces dernières années. Nous présentons dans cet article une nouvelle approche pour le résumé automatique des opinions sur Twitter basée sur les conversations et non sur le traitement des tweets individuels. Notre approche vise à attribuer à chaque conversation un score, indiquant le niveau de satisfaction de l'utilisateur pour le produit correspondant ainsi que pour ses différentes caractéristiques. Nous avons développé un nouvel algorithme basé sur la relation des réponses dans les conversations qui utilise la résolution anaphorique dans un processus de backtracking pour déterminer efficacement les produits évoqués dans les tweets ainsi que leurs aspects. Les expérimentations montrent des résultats prometteurs. En particulier, nous avons prouvé que l'incorporation de la structure de la conversation pour résumer les opinions contribue à améliorer les performances du système.

ABSTRACT. Summarizing opinions conveyed through Twitter has been an emergent theme over the last several years. In this paper, we present a new approach for customer opinion summarization based on twitter conversations rather than individual tweets. Our approach aims to assign to each conversation, a score indicating the level of user's satisfaction towards the corresponding product as well as its features. We have developed a new algorithm based on the reply links in the conversations which employs the anaphora resolution in a backtracking process to effectively extract the different products involved in the tweets as well as their features. Experimentations show promising results. In particular, we have proved that incorporating conversation structure in the opinion summarization contributes to improving system performance.

MOTS-CLÉS : résumé des opinions, Twitter, conversations, résolution anaphorique.

KEYWORDS: opinion summarization, Twitter, conversations, anaphora resolution.

DOI:10.3166/ISI.22.6.37-51

1. Introduction

En quelques années, le microblogging est devenu une forme de communication incontournable permettant aux utilisateurs distants de publier et de partager des messages courts en temps réel. Twitter est actuellement l'un des services de microblogging les plus populaires avec plus de 320 millions d'utilisateurs actifs mensuels[1]. La plateforme de Twitter offre aux utilisateurs (*twittos*) la possibilité d'interagir avec les autres, d'envoyer et de répondre à leurs messages en créant ainsi des conversations. Une conversation peut être définie comme l'ensemble de messages échangés entre plusieurs utilisateurs sur un sujet donné, suite à un message initial. Les conversations représentent un élément clé dans Twitter. D'ailleurs, près d'un quart de l'ensemble des utilisateurs s'impliquent dans des conversations avec d'autres twittos (Java *et al.,* 2007), de plus, le nombre des tweets appartenant à des conversations devient de plus en plus important (Ritter *et al.,* 2010).

Le contenu des conversations est composé généralement de données textuelles qui sont porteuses d'opinions et de sentiments. Analyser ce contenu peut être très utile pour obtenir un aperçu général sur les avis des twittos sur un sujet donné. Et l'analyse d'opinions est depuis longtemps un outil incontournable des entreprises pour acquérir des connaissances sur les avis des consommateurs sur leurs produits. Cela permet au fournisseur d'un produit ou d'un service de mieux réagir aux demandes de ses clients, et au client de s'inspirer des sentiments et opinions d'autres clients sur le produit auquel il s'intéresse et profiter ainsi d'une aide à la décision. L'analyse sémantique du contenu textuel s'introduit dans le contexte de l'analyse des sentiments qui vise à ressortir les marques d'opinions et de sentiments des documents textuels. Une opinion peut être définie comme l'expression des sentiments d'une personne envers une entité (Liu, 2010).

Les travaux de recherche portés sur l'analyse des sentiments sur Twitter se sont principalement focalisés sur le traitement des tweets individuels évoquant un sujet donné (Agarwal *et al.*, 2011 ; Meng *et al.*, 2012 ; Bora *et al.*, 2012 ; Bahrainian *et al.*, 2013). La totalité de la conversation est généralement négligée. Nous nous intéressons dans cet article à l'analyse des sentiments dans les conversations publiques de Twitter. Notre idée est qu'une conversation est potentiellement plus intéressante qu'un seul tweet comme source d'opinions. Toutefois, la détection automatique d'opinions et l'analyse des sentiments dans les conversations de Twitter sont confrontées à des problèmes qui la distinguent du traitement des tweets individuels. Une des difficultés réside dans la nécessité d'une bonne analyse syntaxique des différentes répliques appartenant à une conversation. Cette analyse peut se révéler particulièrement difficile vue la nature des textes écrits de manière très variée et informelle. Une autre difficulté de l'analyse des conversations réside dans les cas assez fréquents de coordination entre plusieurs répliques, d'anaphore ou de coréférence. L'anaphore peut être définie comme le processus par lequel une

1. https://about.twitter.com/company

expression dans le discours (dite *anaphorique*) renvoie à une autre expression (dite *antécédent*) apparue dans le même contexte (Mitkov *et al*., 2014). C'est le cas dans une phrase comme : « *J'ai acheté aujourd'hui un iPhone7. Il est génial !* » Dans cet exemple, le pronom personnel « *Il* » est une anaphore et son antécédent est « *un iPhone7* ». Si on avait seulement la phrase « *Il est génial !* » dans une réponse au sein d'une conversation, les réponses précédentes devraient être considérées afin d'extraire le bon antécédent correspondant. Ce dernier cas est particulièrement fréquent dans les conversions et les tweets de manière plus générale du fait de la brièveté des messages, ainsi un nombre important des tweets est souvent négligé en raison de leur caractère flou et ambigu.

Nous proposons dans cet article, une nouvelle approche pour le résumé automatique des opinions sur Twitter basée sur les conversations au lieu de traiter séparément des tweets individuels. Notre objectif est d'attribuer à chaque conversation un score qui mesure le niveau de satisfaction de l'utilisateur pour le produit correspondant ainsi que pour ses différents aspects (exemple pour un iPhone, les aspects sont : le prix, la batterie, l'écran, l'image, le son, etc.). Nous avons développé un nouvel algorithme basé sur la relation des réponses dans les conversations qui emploie la résolution anaphorique dans un processus de backtracking afin de déterminer efficacement les produits évoqués dans les tweets ainsi que leurs caractéristiques.

Dans la suite de cet article, nous décrivons en section 2 les travaux relatifs à notre problématique. La section 3 formalise et présente l'approche que nous proposons. La section 4 décrit notre la collection de test sur laquelle nous avons mené nos expérimentations ainsi que les évaluations effectuées.

2. État de l'art

Nous proposons dans cet article un type particulier du résumé automatique qui est basé sur les caractéristiques (*features*) des objets (*feature-based opinion summarization*). En règle générale, il existe deux types d'approches pour le résumé automatique des textes d'opinions, celles qui sont basées l'apprentissage machine (supervisé) et celles qui sont non supervisées basées sur les lexiques. D'autres travaux ont également employé l'approche semi-supervisée. Les approches basées sur l'apprentissage machine consistent à attribuer des données à un classifieur pour l'apprentissage. Ce dernier génère un modèle qui est utilisé par la suite pour la partie test de l'apprentissage. Ce type d'approche comprend deux phases : l'extraction des mesures (*features*) et l'apprentissage du classifieur. Les principaux aspects utilisés sont : la fréquence des mots, les n-grams, le *part of speech* et la polarité. Les classifieurs utilisés sont essentiellement l'arbre de décision, la machine à vecteurs de support (SVM), le voisin le plus proche (KNN), le Naïve Bayes et les réseaux de neurones (Jakob *et al.*, 2010 ; Kessler *et al.*, 2009 ; Toh *et al.*, 2015). Bien que ces techniques fournissent de bons résultats pour la classification de sentiments, elles nécessitent un travail manuel long et laborieux pour la préparation des données

d'apprentissage. Les approches basées sur le lexique (*Lexicon-Based Approach*) utilisent, cependant, un dictionnaire de mots subjectifs. Ce dictionnaire peut être général (SentiWordNet[2], General Inquiry, Wilson lexicon, etc.) ou bien il peut être déduit à partir du corpus étudié. Au sein de ces lexiques, on attribue à chaque mot, un score d'opinions qui est traité différemment par les divers approches pour le calcul du score d'opinion d'un document. La méthode la plus simple est d'attribuer à un document donné un score d'opinion qui est égal à l'agrégation de scores des mots qui contiennent une opinion présents dans le document (Liu *et al.*, 2011 ; Jmal *et al.*, 2013).

Le résumé d'opinion basé sur les caractéristiques (*feature-based summarization*) se base sur deux phases primordiales à savoir l'extraction des objets et leurs caractéristiques commentés par les utilisateurs et l'analyse de sentiments. Une des difficultés rencontrées dans ce type de résumé réside dans la nécessité d'une bonne extraction des objets et leurs caractéristiques qui représentent le noyau de la phase de l'analyse de sentiments. Cette extraction peut se révéler particulièrement difficile dans des cas où l'objet est évoqué implicitement dans le texte. Ce cas est fréquemment utilisé dans les tweets du fait de la brièveté et implique généralement l'emploi d'anaphore ou de coréférence qui consiste en la reprise d'une expression présente plus loin dans le contexte.

À notre connaissance, seuls les travaux de Stoyanov et Cardie (Stoyanov et Cardie, 2008) et de Jakob et Gurevych (Jakob et Gurevych, 2010) intègrent la résolution anaphorique (RA) dans l'extraction des objets pour la fouille d'opinions. Stoyanov et Cardie (Stoyanov et Cardie, 2008) développent un algorithme qui identifie les cibles (objets) référents (les antécédents dans une anaphore) dans les articles de journaux. Ils s'appuient sur des cibles annotées manuellement, une phase de sélection des cibles candidats n'est ainsi pas nécessaire. Les auteurs se focalisent essentiellement sur la reconnaissance automatique de chaînes de coréférence. La coréférence se décrit, comme la relation existant entre plusieurs expressions référant à une même entité. Contrairement à l'anaphore qui distingue strictement ses deux parties en antécédent et anaphore, la relation de coréférence est symétrique (Désoyer *et al.*, 2015). Pour leur part, Jakob et Gurevych (Jakob et Gurevych, 2010) adaptent l'algorithme de résolution anaphorique basé sur les règles CogNIAC pour extraire les cibles sur un corpus de critiques des utilisateurs sur une collection de films. Ils ont démontré que l'extension d'un algorithme de fouille d'opinions avec la résolution anaphorique pour l'extraction des objets cibles d'opinions peut apporter des améliorations significatives sur le système. Dans cet article, nous proposons une approche pour le résumé des opinions basé sur la résolution anaphorique en utilisant les conversations publiques dans Twitter. À notre connaissance, c'est la première recherche utilisant une telle technique pour le résumé des opinions.

2. http://sentiwordnet.isti.cnr.it/

3. Approche proposée

Comme illustré sur la figure 1, notre approche est basée sur trois modules essentiels, à savoir, le pré-traitement, la construction des conversations et le résumé des opinions qui consiste en deux phases essentiels : l'extraction des caractéristiques (*features*) des produits commentés par les utilisateurs et l'identification des sentiments évoqués pour chaque caractéristique. Dans notre travail, les objets sont des produits électroniques notamment, des Smartphones et des caméras numériques.

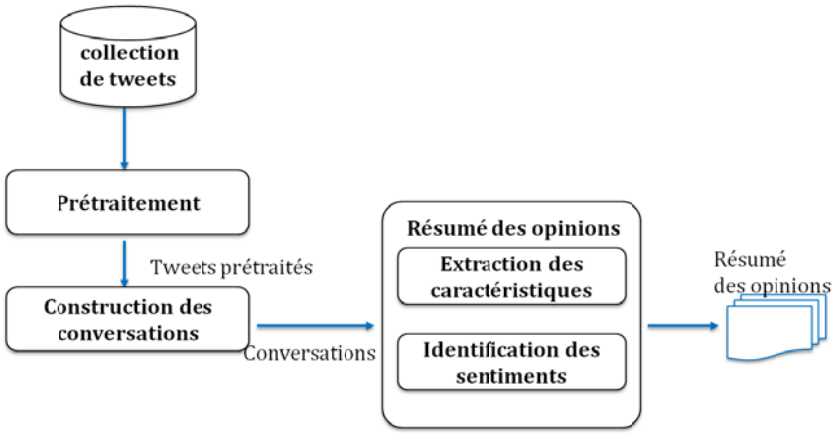

Figure 1. Approche proposée

3.1. Prétraitement

Avant de procéder à la construction des conversations, nous procédons à un ensemble de prétraitements afin de filtrer les textes des tweets et enlever toute information inutile. Pour chaque URL, nous analysons la page web à laquelle elle renvoie, nous utilisons un service API[3] pour l'extraction de texte HTML qui supprime les commentaires, les liens, les publicités et toute partie inutile dans une page web et renvoie les contenus pertinents en texte brut. Nous passons par la suite au filtrage du texte. Nous considérons à la fois le texte des tweets et le texte récupéré à partir des URL. Nous nettoyons le texte en supprimant les caractères ASCII, les chiffres et la ponctuation. Nous convertissons finalement le texte en minuscules et procèdons à la *tokenization* qui permet de découper le texte des tweets en un ensemble de mots. Les attributs liés à chaque tweet tels que l'ID de l'auteur et d'autres informations sociales sont également extraites à l'aide de l'API et stockés.

3. http://www.alchemyapi.com/

Nous indexons également la collection de tweets avec la bibliothèque de traitement de texte Apache Lucene[4].

3.2. Construction des conversations

Dans cette section, nous décrivons le processus de construction de conversations à partir d'une collection de tweets individuels. Nous adoptons la définition de Belkaroui *et al*. (2014) qui définissent une conversation comme un ensemble de messages échangé entre un ensemble d'utilisateurs sur un sujet donné. Nous appliquons la méthode développée par (Cogan *et al*., 2012) qui a proposé un modèle d'arborescence (*reply tree*) basé sur les relations de réponses *(reply*) entre utilisateurs afin de récupérer les conversations à partir des tweets initiaux.

Nous commençons par collecter tous les tweets en réponse à de précédents tweets. Une réponse à un tweet commence toujours avec « @username », une requête de recherche renvoie ainsi toutes les réponses à un tweet donné. Afin de vérifier si les tweets obtenus appartiennent réellement à une même conversation, nous utilisons le champ « in reply to status id » accessible *via* les fichiers statuses/lookup.json *via* Twitter API. Cela permet de construire un arbre de réponse pour chaque conversation en identifiant la racine qui constitue le message initial déclenchant toute la conversation ayant le champ « in reply to status id » vide et en identifiant les autres tweets liés à la racine.

3.3. Résumé des opinions

Le module du résumé des textes d'opinions est composé de deux phases : une phase d'extraction des produits et leurs caractéristiques (*features*) et une phase de détection de sentiment relatif à chaque caractéristique.

3.3.1. Extraction des caractéristiques des produits

Notre objectif, à ce niveau, est d'extraire les caractéristiques des produits commentés et critiqués par les utilisateurs. Par exemple, si nous traitons des textes d'opinions portés sur des Smartphones, les caractéristiques seraient « image », « batterie », « son », « utilisation », « prix », etc. Ayant une collection de conversations, notre système procède à la segmentations des tweets en phrases, et supprime les caractères spéciaux au début et à la fin de chaque mot (ex : « # IPhone # » devient « iphone »). En outre, Hu et Liu (Hu et Liu, 2004) ont démontré que les syntagmes nominaux et les substantifs dans les phrases peuvent être les caractéristiques des produits sur lesquels les utilisateurs commentent. Par ailleurs, les adjectifs véhiculent l'opinion et le jugement. Nous avons ainsi effectué

4. https://lucene.apache.org/

l'étiquetage (*part-of-speech*) de l'ensemble du corpus en utilisant TreeTagger[5] pour identifier la classe grammaticale de chaque mot. Nous construisons une liste de mots vide (*stop-words*) composée des mots fréquents qui n'apportent aucune information utile pour l'analyse du texte et nous supprimons chaque mot appartenant à la liste.

Nous extrayons tous les noms dans les tweets puis nous passons au filtrage des mots obtenus. Nous procédons par la suite à la construction des termes composés (*noun* phrases) qui sont formés de deux noms successifs, comme par exemple « Battery life», « Click Weel », etc. Nous construisons ainsi tous les termes composés et nous ne gardons que les termes fréquents apparaissant au moins 3 fois dans le corpus. Nous éliminons la redondance si elle existe dans les phrases, si un mot apparaît plus qu'une fois dans la même phrase, nous le gardons qu'une seule fois. Nous calculons ensuite la fréquence d'apparition des différents noms extraits dans les tweets, nous ne gardons que ceux dont la fréquence est supérieure à 0,02. Toutes les étapes précédentes menant à la reconstruction de la liste des caractéristiques des produits sont modélisées sur la figure 2.

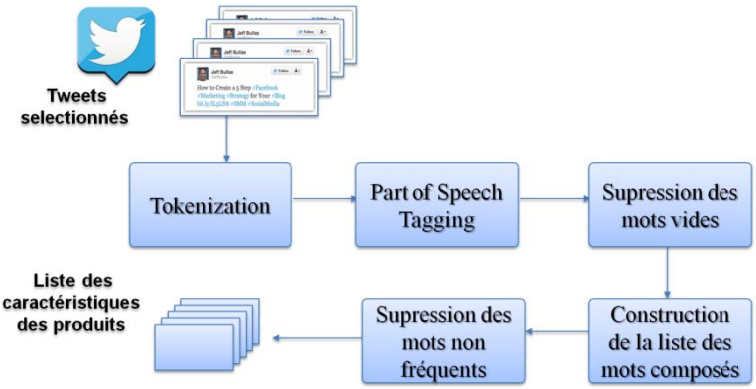

Figure 2. Différentes étapes menant à l'extraction de la liste des caractéristiques

Pour décider si les mots composés que nous avons recueillis sont significatives, nous appliquons une mesure de similarité appelée information mutuelle ponctuelle (Turney *et al.*, 2001) qui utilise des pages web retournées par un moteur de recherche pour extraire les synonymes. Comme Jmal *et al.* (Jmal *et al.*, 2013) nous détectons la compacité d'un mot composé en utilisant le nombre de tweets concernant un produit donné au lieu d'une page web. Ayant deux mots w1 et w2, PMI est défini comme suit :

5. http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

Étant, tweets (w1, w2) représentant le nombre de tweets impliquant w1 et w2 comme un mot compacté (p. ex. *battery life*), tweets (w1), tweets (w2) représentent le nombre de tweets contenant séparément w1, w2, respectivement. Nous gardons seulement les mots composés ayant un PMI < 0.

$$\text{PMI}\ (w_1, w_2) = log\ \frac{tweets\ (w1,w2)}{tweets\ (w1)\ tweets\ (w2)} \quad (1)$$

Ayant notre liste de caractéristiques des produits commentés par les utilisateurs, nous passons à la prochaine phase qui vise à affecter chaque tweet à la caractéristique sur laquelle l'utilisateur a commenté. Pour ce faire, nous appliquons notre algorithme d'identification des caractéristiques qui est basé sur les relations de réponse au sein d'une conversation. En effet, un grand nombre de caractéristiques évoquées apparaissant comme des mots composés et même des noms simples dans les commentaires sont généralement référencés par des pronoms et des expressions anaphoriques dans les messages postérieurs (Kamal, Abulaish, 2013 ; Jakob, Gurevych, 2010) L'expression anaphorique est appelée anaphore. Par exemple, dans la phrase : « Je suis ravi de mon nouvel iPhone, il est incroyable », le pronom « il » représente l'anaphore de l'antécédent « mon nouvel iPhone ». En effet, la plus grande majorité des cibles d'opinion (*i.e.* les caractéristiques du produit) se sont référés par des pronoms, des abréviations, ou des jargons dans les corpus des microblogs (Kamal, Abulaish, 2013).

Ainsi, le recours à la résolution anaphorique est crucial pour identifier correctement les pairs tweet-objet (*feature-tweet*), sans cette phase, un grand nombre de tweets sera négligé en raison de leur aspect ambigu. Afin d'identifier correctement les associations de ces tweets avec les caractéristiques commentés, nous avons développé un nouveau algorithme basé sur les relations de réponses (*reply*) entre utilisateurs au sein d'une de conversation. À notre connaissance, c'est le premier algorithme employant la résolution anaphorique sur les conversations publiques de Twitter pour l'identification des caractéristiques des produits.

Notre algorithme procède comme suit, pour un tweet donné $t_{i,l}$, qui ne contient aucun mot existant dans la liste des caractéristiques mais implique des mots exprimant une opinion (*i.e.* des adjectifs, des adverbes, etc.) ainsi que certains pronoms, tous les pronoms anaphoriques présents dans ce tweet sont extraits, et une liste des pronoms anaphoriques candidats $P=\{p_1, p_2, p_3, ..., p_n\}$ est construite afin de chercher par la suite ceux qui sont liés à un antécédent. Nous employons un mécanisme de *backtracking* au cours duquel les tweets dans une conversation sont accessibles dans l'ordre inverse en fonction des liens de réponse afin d'extraire les répliques précédentes $t_{i-1,j}$ sur lesquelles $t_{i,l}$ réponds. Pour chaque pronom anaphorique $p_i \in P$, les répliques précédentes sont déterminées pour extraire une liste $A=\{a_{k,1}, a_{k,2}, a_{k,3}, ..., a_{q,m}\}$ constituée d'antécédents candidats. Le meilleur antécédent $a_{k,t} \in A$ est sélectionné pour être associée à a_o en utilisant l'algorithme CogNIAC (Baldwin, 1997), qui est une algorithme pour la résolution anaphorique qui utilise une approche fondée sur des règles (*rule based approach*) pour

l'identification des antécédents. Cette approche convient dans notre cas, vu que dans notre corpus, un pourcentage réduit du nombre total de pronoms font référence à des caractéristiques d'un produit (seulement 6 %). Nous désignons le l^{th} tweet à l'itération i, $t_{i,l}$, et le j^{th} tweet à l'itération i, $t_{i,j}$. Comme chaque pronom de anaphorique est remplacé par l'antécédent correspondant, le processus de *backtracking* s'arrête avec l'itération *i-1* pour chaque tweet. À la fin de cette phase, nous obtenons une liste de tweets subjectifs et chaque tweet est associé à la caractéristique correspondante commenté par un utilisateur donné.

3.3.2. Identification des sentiments

Dans cette section, nous cherchons à calculer le score d'opinion pour chaque caractéristique extraite. Nous expliquons tout d'abord notre méthode pour l'évaluation de la polarité de l'opinion exprimée pour chaque caractéristique, et nous détaillons par la suite le calcul du score d'opinion correspondant. L'un des objectifs de notre approche est de détecter les expressions indiquant une opinion dans les tweets, puis déterminer leur polarité et mesurer l'intensité de l'opinion exprimée. En fait, les utilisateurs utilisent les expressions d'opinion qui se situent généralement avant ou après la caractéristique dans le texte. Ainsi, en utilisant la liste des tweets associés avec les caractéristiques déjà détectées, nous extrayons tous les tweets dont le texte contient au moins un terme d'opinion (adjectif, adverbe, verbe, etc.). Par exemple : le terme d'opinion dans la phrase « Mon nouveau iPhone est génial » est « génial », cependant la phrase suivante : « Cette photo est prise par mon iPhone 4S » ne contient aucun mot d'opinion. Elle est considérée comme une « phrase objective ». Dans le cas où nous avons des tweets pareils, nous proposons un deuxième algorithme de détection de polarité qui emploie les liens de favoris pour une meilleure détection des tweets subjectifs. Dans le cas où on a un (tweet) $t_{i,j}$ qui ne contient pas une expression d'opinion mais a été mis en favoris par un autre blogueur b_l qui a déjà interagi dans la conversation c avec un tweet $t_{l,m}$, $t_{i,j}$ prend la même expression d'opinion attribuée au tweet $t_{l,m}$. Pour calculer le score d'opinion pour chaque tweet, nous utilisons SentiWordNet 3.0, un lexique basée sur WordNet 3.0, dans leque chaque mot w de WordNet est associé à trois scores numériques ObjScore(w), PosScore(w) et NegScore(w) décrivant à quel point le mot w est objectif, positif ou négatif, où :

$$\text{ObjScore(s)} + \text{PosScore(s)} + \text{NegScore(s)} = 1 \qquad (2)$$

Les scores sont compris entre 0 et 1. Le score négatif appartient à l'intervalle [0,0,5], et le score positif à [0,5,1]. Pour l'identification de la force d'opinion, nous adoptons l'hypothèse suivante : plus on est proche de 0, plus l'opinion est négative est et *vice versa*. Considérant un mot w, et n le nombre de ses synonymes, le score correspondant est calculé en utilisant la formule suivante :

$$score_w = \frac{\sum_{i=1}^{n} scoreS_{w_i}}{n} \quad (3)$$

Ayant un score d'opinion pour chaque tweet mentionnant une caractéristique, nous calculons le score de tous les mots d'opinion pour chaque caractéristique. Le score d'opinion lié à chaque caractéristique est calculé par la formule (4), où le score f*j* est le score de la caractéristique (*feature*) et n est le nombre total de tweets mentionnant la caractéristique f_j.

$$score_{f_j} = \frac{\sum_{i=1}^{n} score_{w_i}}{n} \quad (4)$$

Ayant les résultats de la détection des caractéristiques ainsi que les scores des sentiments pour chaque caractéristique, nous procédons à la génération des résumés d'opinion permettant aux utilisateurs d'avoir un aperçu rapide des opinions des utilisateurs sur les caractéristiques du produit en question. Un extrait de notre résultat pour le produit iPhone 5s est fourni dans la figure 3. Ces statistiques montrent que le niveau de satisfaction de la clientèle pour le produit iPhone 5s atteint 60 %. On peut constater que les utilisateurs donnent plus d'intérêt à la durée de vie de la batterie et auprix pour lequel on voit des bas niveaux de satisfaction.


```
{
"Product": "iPhone 5s",
"Customer Satisfaction": 0.6,
"Features": [
{
"Battery Life":[
{"Popularity": 0.7,
"Customer Satisfaction": 0.3
}]},

{
"Price":[
{"Popularity": 0.6,
"Customer Satisfaction": 0.4
}]},

{
"Screen":[
{"Popularity": 0.4,
"Customer Satisfaction": 0.8
}]}]
}
```


Figure 3. Les scores d'opinions pour le produit iPhone 5s

4. Expérimentations et résultats

4.1. Description du corpus

Par manque de collection de test spécifique sur les conversations de tweets, nous avons créé notre propre collection. Nous avons collecté 221 663 tweets anglais en utilsant l'API Twitter[6]. La collection de tweet a été explorée sur une période de 4 mois du 25 avril au 25 juillet 2015. Nous collectons uniquement des tweets populaires parlant d'un produit donné impliquant une description du produit, des informations de promotion et des commentaires sur les nouveaux produits. Nous avons gardé 211 350 tweets après avoir éliminé les tweets redondants. Nous avons, par la suite, construit 8 720 conversations contenant 64 370 tweets, 13 827 utilisateurs et cinq produits électroniques, à savoir : 2 caméras numériques et 3 smartphones. Nous avons utilisé l'ensemble des fichiers statuses/lookup.json accessibles par l'API Twitter, qui contiennent l'ensemble des informations relatives aux tweets (id_str, user_id, in_reply_to_status_id, etc.). Le tableau 1 présente quelques statistiques sur notre collection de conversations. Comme le montre le tableau, notre corpus comporte plus de 120 000 pronoms personnels et environ 11,13 % des caractéristiques évoquées dans les tweets sont mentionnées par des pronoms personnels.

Tableau 1. Des statistiques sur le corpus utilisé

Tweets	64 370
Tokens	2 568 160
Target + Opinion Pairs	7 960
Targets that are Pronouns	886
Pronouns	>120 350

4.2. Évaluation des résultats

Vue la nécessité d'un travail manuel laborieux afin d'évaluer l'identification des caractéristiques des produits, nous avons aléatoirement réduit l'ensemble des conversations de 8K à seulement 4K. Pour chaque produit, les premières 800 conversations ont été extraites et prétraitées. Nous exécutons par la suite notre système afin d'extraire les caractéristiques des produits. Pour l'évaluation, toutes les conversations extraites ont été évaluées manuellement. Pour chaque tweet, s'il comporte une opinion exprimée par l'utilisateur, toutes les caractéristiques évoquées ont été identifiées. Pour chaque produit, nous avons produit manuellement une liste de caractéristiques associée. Le nombre de caractéristiques identifié manuellement

6. https://dev.twitter.com/overview/api

pour chaque produit est indiqué dans la colonne « Nbr de caractéristiques » du tableau 3. Les résultats fournis par notre système sont comparés aux résultats fournis manuellement. Nous avons utilisé la précision et le *recall* comme mesures d'évaluation vue leur efficacité prouvée pour le *task* de l'identification des caractéristiques (*features*) (Kim *et al.*, 2011).

Les valeurs TP, TP + TN TP + FP correspondent respectivement au nombre de caractéristiques identifiées pertinentes, au nombre de caractéristiques pertinentes et au nombre de caractéristiques identifiées. Le tableau 2 illustre les résultats d'évaluation pour les 5 produits dans la phase d'extraction des caractéristiques. Les valeurs les plus élevées atteintes par notre système sont une précision de 77,81 % et un *recall* de 82,62 % alors que les scores moyens sont respectivement une 74,90 % précision et 80,03 % *recall*. Notre méthode présente un niveau élevé de précision et de *recall*. Cependant, nous pouvons constater que les scores de *recall* sont généralement supérieurs au scores de précision ce qui prouve que la majorité des caractéristiques correctes ont été bien reconnues par le système. Cela démontre l'efficacité de l'utilisation des interactions de conversation dans l'extraction des caractéristiques des produits. En effet, un nombre important de tweets qui sont très courts ou considérés comme ambigus s'ils sont traités séparément, peuvent être traités convenablement si nous prenons en considération leurs messages précédents. La valeur de précision est inférieure à la valeur du *recall* indiquant que certaines caractéristiques identifiées ne sont pas correctes. Cela peut être justifié par le fait que la plupart des tweets ne respectent pas strictement les règles grammaticales, l'analyseur (*i.e. the parser*) n'arrive pas à affecter les bonnes classes grammaticales (*i.e. part-of-speech*) et les relations de dépendance entre les mots.

Nous comparons par la suite les caractéristiques générées par notre méthode avec les caractéristiques obtenues en utilisant la même méthode mais appliquée à la collection de tweets pris séparément sans construire les conversations et sans l'emploi de notre algorithme basé sur la résolution anaphorique entre les messages. Cette méthode est similaire au processus de détection des caractéristiques appliqué par (Jmal, Faiz, 2013) qui effectue le résumé des textes d'opinions pour les commentaires des internautes sur Twitter ainsi que sur des sites Web de commerce électronique. Le *recall* moyen est d'environ 63 % et la précision moyenne est de l'ordre de 57 %. Nous pouvons constater que le niveau de *recall* et le niveau de précision moyens atteints par la deuxième méthode sont remarquablement inférieurs à ceux obtenus par notre méthode. Suite à notre analyse des résultats, nous avons constaté que notre collection de tweets comporte 13 824 pronoms anaphoriques, parmi lesquels 886 pronoms se réfèrent correctement à des caractéristiques des produits. En appliquant la deuxième méthode, nous notons que seulement 40 % des pronoms ont été traités, le reste des pronoms ont été négligés par le système ou bien extraits de manière erronée. Les résultats présentés dans le tableau 2, montre que notre méthode proposée est significativement plus efficace pour l'identification des caractéristiques.

Tableau 2. Évaluation des résultats pour l'identification des caractéristiques des produits

Produit	Nbr de caractéri-stiques	Sélection des caractéristiques sur des tweets individuels		Sélection des caractéristiques sur des conversations (notre approche)	
		Précision (%)	Recall (%)	Précision (%)	Recall (%)
Digital camera1	27	68,33	62,13	75,26	79,61
Digital camera 2	31	63,52	**64,40**	76,78	77,98
Smartphone 1	52	55,97	51,83	75,22	78,23
Smartphone 2	25	**71,54**	57,74	69,46	**82,62**
Smartphone 3	43	57,88	53,73	**77,81**	81,73
Moyenne		63,44	57,96	74,90	80,03

5. Conclusion

Nous avons présenté dans cet article une nouvelle approche pour le résumé des opinions qui utilise les conversations publiques de Twitter plutôt que les tweets individuels. Notre contribution majeure est d'adopter une nouvelle méthode anaphorique qui utilise les interactions au sein d'une conversation, notamment les relations de réponses pour extraire efficacement les caractéristiques des produits commentés par les utilisateurs. Nous avons également exploité les relations de favoris entre les tweets pour une meilleure détection de la polarité des messages. Nos résultats montrent que notre méthode proposée est prometteuse. En particulier, nous avons prouvé que l'intégration de la structure de la conversation pour le résumé des textes d'opinions contribue à améliorer le système performance. Dans notre futur travail, nous comptons ajouter l'aspect sémantique dans le processus de la détection des polarités des messages en prenant en considération le contexte des messages vue que certains termes changent de polarité selon le contexte.

Bibliographie

Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau R. (2011). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, p. 30-38. Association for Computational Linguistics.

Baldwin B. (1997). Cogniac: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, p. 38-45. Association for Computational Linguistics.

Bahrainian S. A., Dengel A. (2013). Sentiment analysis and summarization of twitter data. In *Computational Science and Engineering (CSE)*, p. 227-234. IEEE.

Belkaroui R., Faiz R. (2015). Towards events tweet contextualization using social inuence model and users conversations. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics,* ACM.

Belkaroui R., Faiz R., Elkhlifi A. (2014). Conversation analysis on social networking sites. *In Signal-Image Technology and Internet-Based Systems (SITIS), IEEE*, p. 172-178.

Bora N. N. (2012). Summarizing public opinions in tweets. *International Journal of Computational Linguistics and Applications*, vol. 3, n° 1, p. 41-55.

Feldman R., Fresko M., Goldenberg J., Netzer O., Ungar L. (2007). Extracting product comparisons from discussion boards. In *Data Mining, ICDM 2007, 7th IEEE International Conference on Data Mining*, p. 469-474.

Ferreira L., N. Jakob, and I. Gurevych. (2008). A comparative study of feature extraction algorithms in customer reviews. In *Semantic Computing, 2008 IEEE International Conference on*, p. 144-151. IEEE.

Flesch R. (1948). A new readability yardstick. *Journal of applied psychology*, vol. 32, n° 3, p. 221.

Jakob N., Gurevych I. (2010). Using anaphora resolution to improve opinion target identification in movie reviews. In *Proceedings of the ACL 2010 Conference.*

Jmal J., Faiz R. (2013). Customer review summarization approach using twitter and sentiwordnet. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantic*s, p. 33. ACM.

Kamal A., Abulaish M. (2013). Statistical features identification for sentiment analysis using machine learning techniques. *Computational and Business Intelligence (ISCBI),* IEE, p. 178-181.

Kessler J. S., Nicolov N. (2009). Targeting sentiment expressions through supervised ranking of linguistic configurations. In *ICWSM.* AAAI Press.

Kim H. D., Ganesan K., Sondhi P., Zhai C. (2011). Comprehensive review of opinion summarization. Technical report, University of Illinois at Urbana-Champaignl.

Liu (2011). Opinion mining and sentiment analysis. *Web Data Mining*. Springer, p. 459-526.

Liu X., Li Y., Wei F., Zhou M. (2012).Graph-based multi-tweet summarization using social signals. *COLING*, p. 1699-1714.

Meng X., Wei F., Liu X., Zhou M., Li S., Wang H. (2012). Entity-centric topic-oriented opinion summarization in twitter. In Proceedings of the *18th international conference on Knowledge discovery and data mining*, ACM, p. 379-387.

Mitkov R. (2014). *Anaphora resolution.* Routledge.

Popescu A.-M., Etzioni O. (2007). Extracting product features and opinions from reviews. *Natural language processing and text mining,* Springer, p. 9-28.

Ritter A., Cherry C., Dolan B. (2010). Unsupervised modeling of twitter conversations. In *HLT-NAACL*. The Association for Computational Linguistics, p. 172-180.

Steinberger J., Poesio M., Kabadjov M. A., Jezek K. (2007). Two uses of anaphora resolution in summarization. *Information Processing and Management,* vol. 43, n° 6, p. 1663-1680,

Stoyanov V., Cardie C. (2008). Topic identification for for fine-grained opinion analysis. In Proceedings of the *22nd International Conference on Computational Linguistics,*vol. 1, p. 817-824.

Turney P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. *In Machine Learning: ECML*, p 491-502.

Toh Z., Su J. (2015). Nlangp: Supervised machine learning system for aspect category classification and opinion target extraction. In *Proceedings of the 9th International Workshop on SemEval*, p. 719-724,