# Analyse des Sentiments sur AMAZON

Iscovici Lucas[1] Mokhtari Anis[1] Taleb Adel Ali[1] Diallo Abou Oumar[1]

May 3, 2018

**Abstract**

Dans le contexte numérique actuel, caractérisé par une surabondance d'informations, il apparaît que les capacités humaines ne permettent pas l'analyse exhaustive de telle masse de données. Cette augmentation de la quantité de contenus disponibles est liée à l'évolution des pratiques, les utilisateurs ne se contentent plus d'être consommateurs d'information, ils sont également producteurs de contenus (commentaires, partage de vidéo).De ce fait on a de plus en plus de données qui nous permettent l'analyse de sentiment.

Dans cet article, nous nous concentrons sur l'utilisation d'Amazon, qui est l'un des plus grands fournisseurs en ligne du monde, pour l'analyse de sentiments.On proposera une solution basée sur la combinaison des N-grams qu'on leurs appliquera des méthodes de combinaison de classifieurs afin de tirer profit de l'agrégation des différents modèles, et de pouvoir améliorer sensiblement la qualité de la prédiction finale.

Les évaluations expérimentales montrent que les techniques que nous proposons sont efficaces. Dans notre recherche, nous avons travaillé avec le français, cependant, la technique proposée peut être utilisée avec n'importe quelle autre langue.

## 1 INTRODUCTION

Avec l'avènement du Big data et les réseaux sociaux, des millions de données sont produites chaque jour à travers le web, des données structurés, non structurés ou bien semi structuré. Le traitement de grande masse de données avec les algorithmes classiques n'est plus possible. Pour cela les experts de la data ont développé des outils et algorithmes capable de surmonter cette problématique. De plus en plus d'utilisateurs donnent leurs avis sur des produits qu'ils ont achetés sur internet.

Dans ce travail nous utilisons un ensemble de données formé de commentaires collectées sur le site d'Amazon. Amazon dispose d'un très grand nombre de commentaires créer par les utilisateurs de ce site. Au fur et à mesure, on a de plus en plus de commentaires qui peuvent être utilisées dans des taches d'exploration d'opinion et d'analyse de sentiments. Les entreprises pourraient être intéressé par les questions :

- Que pensent les gens de notre produit (service, entreprise, etc)?

- Dans quelle mesure les gens sont-ils positifs (ou négatifs) à propos de notre produit?
- Pourquoi les gens préfèrent-ils que nos produits ressemblent?

Dans cet article, nous avons collecter 130000 commentaires d'Amazon qu'on a divisé en deux jeux de données.

1. Les commentaires positives sont notés quatre ou cinq.

2. Les commentaires négatifs sont notés un ou deux.

3. Les commentaires avec une notes égales a trois on les ignore.

### 1.1 Contributions

Les contributions de notre article sont les suivantes :

- Nous présentons une méthode pour collecter des sentiments positives et négatives que nous prétraitons.
- Nous utilisons le corpus collecter pour construire un classifieur de sentiments basés sur la combinaison des N-grams et tags.
- Nous effectuons des évaluations expérimentales sur un ensembles de commentaires pour prouver que notre technique présentée est efficace.

### 1.2 Organisations

Le reste du papier est organisé comme suit. Dans la section 2, nous discutons des travaux antérieurs sur l'analyse de sentiment. Dans la section 3, nous décrivons notre processus de collecte et de prétraitement des données. Dans la section 4, nous présentons notre solution. Et dans la dernière section, Nous montrerons nos expérimentations et les résultats que nous avons obtenus.

## 2 Travaux Connexes

Avec l'avènement du big data et les réseaux sociaux, de plus en plus de recherches dans le domaine de l'analyse de sentiments ont vu le jour. En 2008, un travail général résumant la problématique a été présenté par (pang et lee). Dans leur enquête, les auteurs décrivent des techniques et des approches existantes pour une recherche d'information orienté opinion. Cependant ils n'existent pas beaucoup de recherches qui ont traiter les commentaires des utilisateurs sur les différentes plateformes, Amazon, twitter, etc.... Dans (yang et al,2007), les auteurs utilisent des blogs pour construire un corpus pour l'analyse des sentiments et utilisent des icônes d'émotions assignées aux articles de blog comme indicateurs de l'humeur des utilisateurs. En effet l'approche consisté a appliqué des apprenants SVM et CRF pour classer les sentiments au niveau de la phrase. Ensuite ils ont étudié plusieurs stratégies pour déterminer le sentiment général du document. En conséquence, la stratégie gagnante est définie en considérant le sentiment de la dernière phrase comme le sentiment du document. (Read, 2005) a utilisé des émoticônes telles que ":-)" et ":-(" pour former un ensemble d'apprentissage pour la classification des sentiments. Pour ce faire, l'auteur a collecté des textes contenant des émoticônes à partir de groupes de discussion Usenet. L'ensemble de données a été divisé en échantillons "positifs" (textes avec des émoticônes heureuses) et "négatifs" (textes avec

des émoticônes tristes ou en colère). Coté expérimentation, l'auteur a utilisé SVM et naïve bayes et a pu obtenir une précision de 70% sur l'ensemble de test. Dans (Go et al., 2009), les auteurs ont utilisé Twitter pour collecter des données d'entraînement et ensuite pour effectuer une recherche de sentiment. L'approche est similaire à (Read, 2005). Les auteurs construisent des corpus en utilisant des émoticônes pour obtenir des échantillons « positifs » et « négatifs », puis utilisent divers classificateurs. Le meilleur résultat a été obtenu par l'algorithme de Naïve Bayes avec une mesure d'information mutuelle pour la sélection des caractéristiques. Les auteurs ont pu obtenir jusqu'à 81% d'exactitude sur leur ensemble de test. Cependant, la méthode a montré une mauvaise performance avec trois classes ("négatif", "positif" et "neutre"). Dans (par et Pak, 2010), les auteurs ont construit un jeu de données avec trois classes : positive, négative, neutre. L'ensemble de données collectées est utilisé pour extraire des caractéristiques qui seront utilisées pour former le classificateur de sentiments. Ensuite, Les auteurs ont commencé par séparer les commentaires objective et subjective pour ensuite déterminer ceux qui sont positives ou bien négatives. Une fois la phase de prétraitement effectué. Les auteurs ont utilisé les n-grams avec les modèles probabilistes (naïve bayes) et aussi le SVM et CRF, cependant celui qui a donné de bons résultats c'était naïve de bayes. Après, Ils ont aussi proposé une formule pour améliorer la précision de l'algorithme en utilisant la salience.

# 3 Collecte de Données

Nous allons passer à l'une des étapes les plus importantes de notre travail, qui consiste à la collecte de données. Dans l'article que nous avons étudié précédemment, les auteurs utilisent Twitter pour collecter un corpus de textes et formulé un ensemble de données de trois classes. Dans le cadre de notre travail, nous avons utilisé le site d'Amazon pour recueillir nos données.

## 3.1 Le Recueil de données

Amazon est l'un des plus grands fournisseurs en ligne dans le Monde. Les gens regardent souvent les produits et les critiques sur le produit avant de l'acheter sur Amazon lui-même. Chaque avis est très évident, la note d'évaluation fournie par l'utilisateur reflète ce que l'utilisateur écrit comme sa critique, c'est-à-dire si l'utilisateur écrit quelque chose de mal définitivement l'ensemble note que l'utilisateur donne est soit 1 ou 2 sur 5. De ce fait nous allons nous baser sur cela pour former un classificateur à reconnaître les sentiments positifs (note >=4 sur 5) et négatifs (note<=2 sur 5). Pour recueillir les données, nous avons écrit un code PHP, nous permettant de scraper les données. Nous avons exploré l'URL d'Amazon pour en extraire tous les détails requis. Nous avons pris soin du texte afin de satisfaire le format requis. Par exemple, les balises <br/> ont une signification particulière pour le navigateur, c'est-à-dire la lecture interrompue ou la ligne suivante, nous devons convertir explicitement chaque balise <br/> en espaces ou bien le résultat de l'exploration sera incorrect. Nous avons pris un soin particulier

à extraire les données à partir des pages Web. Plus précisément, nous demandons à PHP de nous donner une page, ensuite on parse cette page, on prend les liens et on demande à PHP d'aller chercher les pages de ses liens. En même temps nous construisons nos données, en nous concentrant sur des critiques de livres, de musiques etc. Les sites Web utilisent le jeu de caractères utf-8 pour l'encodage des caractères, mais parfois ce codage peut provoquer des erreurs lors du scrappage Web, car le scrapage implique la mise en correspondance de chaînes et de motifs. La solution consiste simplement à appliquer la chaîne à coder au format utf-8. Avant de pouvoir utilisé ces données dans le cadre de notre travail, nous devons d'abord les nettoyés d'où l'étape de prétraitement.

### 3.2 Prétraitement des données

Les données extraites doivent être nettoyées afin que nous puissions obtenir un examen approprié du texte sur lequel l'analyse peut être effectuée. Le nettoyage des données analysées est effectué en supprimant tous les caractères spéciaux (tels que <": ": /.:'#**$>") afin de récupérer les meilleurs résultats.

Pour faire cela on a utilisé le langage python, notamment avec les bibliothéques scikit-learn, NLTK. Et nous avons mis en œuvre différentes fonctions nous permettant de supprimer les accents, les emoji et les Backslash. Pour pouvoir faire les unigram qu'on verra précédemment, on a fait un prétraitement de plus qui est la lemminization. Ce prétraitement a pour but de préparer les données pour le training format en dictionnaire Après cette phase de prétraitement, nous avons séparés nos données en deux groupes. On s'est dit qu'un commentaire avec 3 étoile ne veut pas dire que c'est un commentaire neutre. De ce fait on les a supprimés tout simplement. Au final on a, les bonnes critiques sont ceux avec une note de 5 étoiles et 4 étoiles, et les mauvaises critiques sont ceux avec une note de 2 étoiles et 1 étoile. Nous avons mis ces données dans un tableau à trois colonnes : le commentaire, la note donnée et la note max.
La figure suivante nous montre un exemple de données négatives obtenues

| | comments | rate | RateMax |
|---|---|---|---|
| 16 | grand corps malade, c'est le poete qui ne conn... | 1 | 5 |
| 80 | comme tous les chanteurs issus des tele croche... | 2 | 5 |
| 81 | il avait realiser un tres bon premier album,pu... | 1 | 5 |
| 96 | passable produit a un prix raisonnable. je sui... | 2 | 5 |
| 101 | le coffret est interessant pour qui aime ce ch... | 2 | 5 |
| 118 | tres decu par la qualite sonore de presque tou... | 2 | 5 |
| 142 | plusieurs personnes l'avaient dits, j'aurai du... | 2 | 5 |
| 151 | les cinq cd sont tres mal enregistres on reco... | 2 | 5 |
| 190 | le son est pas terrible et des sons enregistr... | 1 | 5 |
| 216 | que dire...je sjis tombee sur un mauvais cd. s... | 1 | 5 |

Exemple de données négatives

## 4 Proposition de Solution

La solution qu'on a élaborée émane de la lecture des différents travaux connexes qui ont était proposé dans le domaine de l'analyse de sentiments. En effet, ce qu'on a mis en place est une solution hybride des méthodes réalisé avant, tout en apportant des améliorations. Nous avons pensé à une solution qui considère que chacun des unigram, bigram, trigram pouvaient amener un plus. Par exemple, si un des N-gram se trompe l'autre aura juste. Autrement dit, on a considéré que dans le cas parfait les erreurs commises par chacune

des méthodes étaient différentes des erreurs commises par les autres.
Dans notre solution nous avons commencé par la collecte des commentaires des utilisateurs d'amazon et on leur a fait un prétraitement comme expliqué dans la section précédente. Par la suite, On a construit les trois n-gram en prenant en considération les grams les plus discriminants.
Une fois les données prétraitées, l'idée est d'obtenir les probabilités des différents sentiments à savoir positive et négative en utilisant un classificateur bayésien. Dans un premier temps, on a commencé par diviser le jeu de données en un ensemble d'apprentissage et un ensemble de test selon la métrique de 75Ensuite, On se verra appliquer ces méthodes sur les différents n-gram puis on récupère les résultats de chaque méthode pour chaque commentaire. Par la suite on appliquera le Voting Hard et Soft, mais aussi le Stacking sur les trois probabilités précédemment calculées.

Pour vous mettre dans le contexte Supposons que nous avons des probabilités comme suit: 0.45 0.45 0.90, le voting hard nous donnera un score de 1/3 (1 vote en faveur et 2 contre), donc il serait classé comme un «négatif», le voting soft nous donnera la moyenne des probabilités, qui est de 0.6, et serait un "positif". Le voting soft tient compte de la certitude de chaque électeur, plutôt que d'une simple entrée binaire de l'électeur.Le stacking (ou dit parfois blending) est un procédé qui consiste à appliquer un algorithme de machine learning à des classifieur générés par un autre algorithme de machine learning.

D'une certaine façon, il s'agit de prédire quels sont les meilleurs classifieurs et de les pondérer. Cette démarche a l'avantage de pouvoir agréger des modèles très différents et d'améliorer sensiblement la qualité de la prédiction finale.

La différence que rapporte notre approche, est qu'on ne se focalise pas sur un seul type de n-gram mais on suppose que chaque n-gram peut donner de bons et de mauvais résultats.
En choisissant un n-gram en fonction de la phrase, on n'est pas contraint par la langue utilisée dans le texte, il suffit juste d'entrainer nos données sur des textes du langage pour qu'il puisse décider du bon n-gram en fonction de la phrase qu'on lui donnera, les résultats qu'on obtient sont très intéressants, on explique ça dans le chapitre suivant.

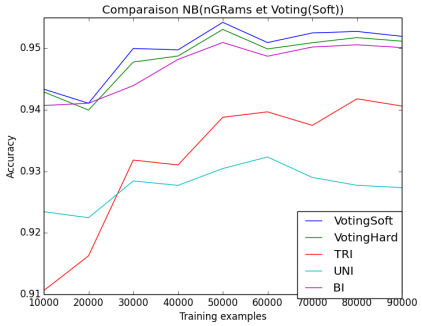
# 5 Expérimentations et Résultats

## 5.1 Expérimentations

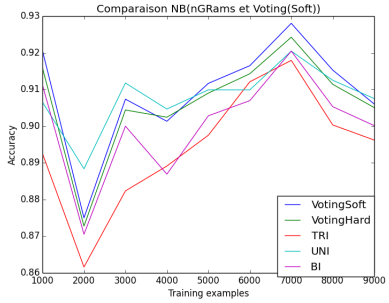
On a commencé par diviser notre jeu de données avec la métrique de 75% pour le jeu de train et 25% pour l'ensemble de test. On a utilisé un classifieur bayésien pour pouvoir avoir nos trois probabilités. Une qui représentera les uni-gram, une deuxiéme pour le bigram et la dernière pour le trigram.
Comme on l'a dit dans la section précédente, c'est sur ses trois probabilités qu'on va appliquer le stacking, le voting soft et le voting hard pour essayer d'avoir un résultat meilleur et une précision supérieure.
Durant nos tests on a fait varié le nombre de données d'apprentissage pour voir à quel points le nombre de données pouvais changer les résultats.

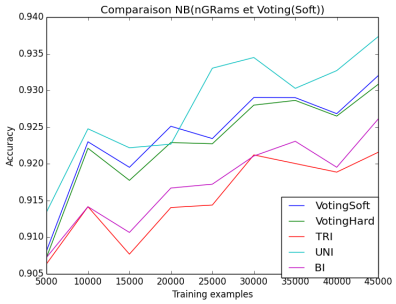
## 5.2 Résultats

### 5.2.1 Comparaison N-gram Voting Soft et Hard

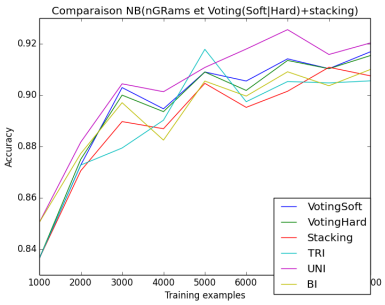



Dans la figure précédente on a comparé les n-gram, le voting soft et hard avec un min train de 2 sur des quantités de données d'apprentissages entre 10000 et 90000.
On s'apperçoit que le voting soft a une meilleure précision que les autres. Et que le bi-gram est meilleur que le uni et le tri sur ces jeux de données.




Dans ces deux figures précédentes on a pris un a minimun occurence égale à 0 et max fréquence égale à 1. On a essayé de faire varié le jeux de données aussi. Dans la première on a pris entre 1000 et 9000 jeux de données et on s'aperçoit que le uni gram est meilleur que les deux autres et qu'il dépasse même le voting soft entre 3000 et 4000 données d'apprentissages. Dans la seconde, on s'apperçoit que l'uni-gram est meilleur partout.




En ajoutant le stacking aussi, on s'appercoit que l'unigram a une meilleure précision que les autres.
Mais aussi que le voting hard et soft donne une meilleure précision que les autres.

Pour finir avec l'analyse des résultats on peut en déduire que le voting soft apporte sa marque sur une grande quatité de données avec une min occurence à 2, avec cette configuration on a atteint les 97% de précision sur 100000 données d'apprentissages.

# 6 CONCLUSION

L'analyse de sentiment pose de nouveaux défis et enjeux dans le développement de la nouvelle génération des systèmes de filtrage de l'information. Dans ce travail, nous avons présentés une méthode basée sur l'utilisation des N-grams, l'approche proposé repose sur des techniques ensemblistes de combinaisons de classifieurs appliquées sur les commentaires des utilisateurs. En particulier, étant donné que nous nous sommes concentrés sur le domaine de la vente en ligne, nous avons exploité les données d'Amazon pour recueillir des avis utilisateurs. L'aspect novateur de cette approche est la combinaison des différents N-Gram afin d'améliorer la qualité de prédiction. Les résultats obtenus sur le jeu de données d'Amazon satisfont l'aspect pertinence et qualité de filtrage d'information.

# References


[1] [pang et lee, 2008] Opinion mining and sentiment analysis.

[2] [yang et al, 2007] Twitter as a Corpus for Sentiment Analysis and Opinion Mining

[3] [Read, 2005]Using emoticons to reduce dependency in machine learning techniques for sentiment classification.In ACL. The Association for Computer Linguistics.

[4] [par et Pak, 2010] Alexander Pak, Analyse de sentiments automatique, adaptative et applicative Université Paris-Sud, Lab. LIMSI-CNRS, Bâtiment 508, F-91405 Orsay Cedex, France.

[5] [Go et al., 2009] Twitter Sentiment Classification using Distant Supervision.