Analyse des sentiments et des émotions de commentaires complexes en langue française

Stefania Pecore

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ BRETAGNE SUD
COMUE UNIVERSITÉ BRETAGNE LOIRE

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *INFORMATIQUE*

Par

« Stefania PECORE »

« Analyse des sentiments et des émotions de commentaires complexes en langue Française. »

Thèse présentée et soutenue à VANNES le 28 JANVIER 20019
Unité de recherche : IRISA CNRS UMR 6074
Thèse N° : 522

Rapporteurs avant soutenance :

Diana INKPEN	Professeure	Université d'Ottawa
Gudrun LEDEGEN	Professeure	Université Rennes 2

Composition du Jury :

Présidente	Prof. Joana GALLERON	
Dr Berry DE BRUIJN	CNRC Canada	
Dr Farida SAID	Université de Bretagne Sud	
Diana INKPEN	Professeure	Université d'Ottawa
Gudrun LEDEGEN	Professeure	Université Rennes 2

Directeur de thèse
Prof. Pierre-François MARTEAU Université Bretagne Sud

Co-directeur de thèse
Dr Jeanne VILLANEAU

Le présent contrat a pour objet de permettre à l'université de Bretagne Sud de diffuser la thèse soutenue par l'Auteur et mentionnée ci-dessus, dans le respect des dispositions du Code de la Propriété Intellectuelle relatives au droit d'auteur.

Seules les thèses ayant obtenu avis favorable du jury de soutenance et pour lesquelles les éventuelles corrections demandées auront été apportées dans les délais impartis pourront faire l'objet du présent contrat.

Soucieuses de faciliter l'accès au savoir et à la connaissance, d'encourager les contacts et les échanges au sein des communautés scientifiques et universitaires et de contribuer ainsi tant à la renommée de la thèse et de son auteur qu'à celle de l'université de Bretagne Sud, les parties prenantes au présent entendent favoriser la diffusion de l'œuvre sur support électronique selon les modalités précisées ci-après.

Pour les développements qui précèdent et suivent, les termes utilisés sont définis comme suit :

«L'**Auteur** » : La personne signataire de l'œuvre et investie des droits d'auteur s'y rapportant

«Le **Diffuseur** » : L'université de Bretagne Sud

«L'**Œuvre** » : La thèse précitée soutenue par l'auteur.

ET CONVENU CE QUI SUIT :

ARTICLE 1 :

L'Auteur autorise la diffusion électronique gratuite immédiate de l'Œuvre, en totalité ou en partie, par le Diffuseur et ce sans que le Diffuseur ne puisse en retirer un bénéfice financier.

L'Auteur autorise la diffusion : [l'une des deux options doit obligatoirement être choisie]

X De l'Œuvre entière, en intranet, en extranet et sur Internet y compris sur la plateforme TeL (Thèses en ligne) gérée par le Centre pour la Communication Scientifique Directe (CCSD) du CNRS.

☐ De l'Œuvre entière, en intranet exclusivement

Dans le cas où l'Auteur choisit de restreindre la diffusion de l'Œuvre (diffusion en intranet), il s'engage à déposer un exemplaire intégral de l'Œuvre en version papier aux fins de communication de l'Œuvre par le Service Commun de Documentation (SCD) de l'Université dans le cadre du prêt entre bibliothèques.

L'Auteur autorise la diffusion de l'Œuvre :

☐Immédiatement

X Après un délai de 12 mois (L'ajout d'un délai de diffusion est possible seulement pour une diffusion de l'Œuvre entière sur Internet.)

Afin de faciliter la mise en ligne de l'Œuvre, l'Auteur s'engage à respecter les prescriptions techniques minimales communiquées par le Service Commun de Documentation (SCD).

ARTICLE 2 :

L'autorisation donnée au Diffuseur n'a aucun caractère exclusif. L'Auteur conserve par conséquent toutes les possibilités de cession et de diffusion concomitantes de l'Œuvre, notamment dans un cadre éditorial, sous sa propre responsabilité.

ARTICLE 3 :

L'Auteur autorise au Diffuseur la reproduction, la représentation et l'adaptation de l'Œuvre pour le monde entier dans les conditions prévues au présent article, ainsi que l'ajout éventuel d'éléments de description du contenu de l'Œuvre, pouvant prendre la forme notamment d'un résumé, d'un sommaire, d'un avertissement, etc. :

Au titre du droit de reproduction :

- La fixation et la reproduction en nombre illimité de l'Œuvre sur tout support existant ou futur et par tous moyens connus ou inconnus à ce jour,

- L'établissement sans modification aucune de tous duplicata, copies ou photogrammes et en toute langue, pour une diffusion conforme aux autorisations prévues à l'article 1 du présent contrat.

Au titre du droit de représentation :

- La diffusion et la communication de l'Œuvre au public par tous moyens, notamment par réseaux de télécommunications numériques ou analogiques, satellite, câblodistribution, voie hertzienne, etc.

Pour une diffusion conforme aux autorisations prévues à l'article 1 du présent contrat

Les droits d'adaptation comportent, le cas échéant, la faculté de modifier la forme et le format de l'Œuvre en fonction des contraintes techniques imposées par l'archivage, le stockage, la sécurité et la diffusion électronique de l'Œuvre. Ainsi, les modifications imposées par l'état de la technique ne sauraient être considérées comme une dénaturation de l'Œuvre portant atteinte au droit moral de l'Auteur.

ARTICLE 4 :

La présente autorisation est consentie à titre gratuit pour toute la durée légale de protection de la propriété littéraire et artistique offerte par la loi française à l'Auteur, ses ayants droits ou représentants, y compris les prolongations qui pourraient être apportées à cette durée.

ARTICLE 5 :

L'Auteur pourra à tout moment demander au Diffuseur de retirer l'Œuvre des plateformes sur lesquelles elle se trouve et lever l'autorisation de diffusion donnée par lui, à charge pour lui d'en aviser le Diffuseur par lettre RAR adressée au Président de l'Université.

Le Diffuseur procèdera alors au retrait de l'Œuvre dans les meilleurs délais et au plus tard lors de la plus prochaine actualisation des plateformes sur lesquelles celle-ci se trouve.

Le Diffuseur ne pourra pas lui-même procéder au retrait de l'Œuvre de la plateforme TeL. Au cas où l'Auteur souhaiterait ce retrait, il lui incombera d'en faire lui-même la demande au Centre pour la Communication Scientifique Directe (CCSD).

ARTICLE 6 :

La signature du présent contrat n'implique pas l'obligation pour le Diffuseur de faire usage des autorisations qui lui sont données. La diffusion effective, tout comme son éventuelle suppression, n'implique en aucun cas une appréciation, au bénéfice de l'Auteur ou des tiers, du contenu de l'Œuvre diffusée, et ne saurait être source de responsabilité à l'égard des tiers.

De même, l'Auteur demeure responsable sur la base du droit commun, du contenu de l'Œuvre. Il garantit au Diffuseur que l'Œuvre ne fait l'objet d'aucun contrat d'édition ou de diffusion, accordé à un tiers, susceptible de restreindre les dispositions du présent contrat.

Il garantit en outre avoir obtenu les droits nécessaires à la diffusion de l'Œuvre dans les conditions prévues au présent contrat, en particulier toutes les autorisations écrites nécessaires des titulaires des droits sur les œuvres reproduites partiellement ou intégralement (telles que textes, illustrations, extraits multimédia, etc.) et informe, le cas échéant, le Diffuseur des documents contenus dans l'Œuvre pour lesquels il n'aurait pas obtenu les droits et qui ne pourraient pas bénéficier des dispositions de l'article L.122-5 du Code de la Propriété Intellectuelle. Dans ce cas, l'Auteur fournira au Diffuseur une version de l'Œuvre amputée des documents pour lesquels les droits n'auront pas été obtenus qui pourra être utilisée comme version de diffusion. En cas de non-respect de cette clause, le Diffuseur se réserve le droit de refuser, suspendre ou arrêter la diffusion de tout ou partie de l'Œuvre concernée dès connaissance du caractère manifestement illicite du contenu en cause.

Le Diffuseur ne pourra être tenu pour responsable de représentation illégale de documents pour lesquels l'Auteur n'aurait pas signalé qu'il n'en avait pas acquis les droits, ni de la violation d'un éventuel contrat d'édition antérieur non signalé par l'Auteur.

L'Auteur est personnellement responsable tant vis-à-vis des tiers que du Diffuseur du non-respect des stipulations énoncées ci-dessus et s'engage donc à garantir immédiatement et relever indemne le Diffuseur contre toute action, réclamation ou revendication susceptible d'en découler

ARTICLE 7 :

Le Diffuseur ne retire aucun bénéfice de la diffusion de l'Œuvre du fait du présent contrat.

ARTICLE 8 :

Le Diffuseur s'engage dans tous les cas où il ferait usage de l'autorisation de diffusion, à faire figurer au regard du titre de l'Œuvre, le nom de l'Auteur.

Le Diffuseur s'engage également en ces cas à faire apparaître, dans la mesure du possible, sur les pages écrans accompagnant l'Œuvre, l'indication du caractère réservé des droits de l'Auteur et l'interdiction de toute reproduction sans accord express de celui-ci.

L'Auteur est toutefois conscient du fait, au-delà de l'indication de l'interdiction, qu'en l'état des techniques, le Diffuseur ne dispose pas des moyens permettant d'empêcher la consultation et/ou la reproduction matérielle, totale ou partielle, non autorisée de l'Œuvre.

Le Diffuseur ne pourra être tenu pour responsable des agissements illégaux de tiers.

L'Auteur conserve cependant tous ses droits d'ester en justice afin de protéger son droit d'auteur sur l'Œuvre.

ARTICLE 9 :

Conformément aux règles d'accès aux documents administratifs, la diffusion de l'Œuvre par le Diffuseur ne sera effective qu'au terme de l'éventuelle période de confidentialité précitée, prononcée par le Président de l'Université.

ARTICLE 10 :

L'Auteur certifie que la version électronique de l'Œuvre remise au Diffuseur dans le cadre du présent est conforme à la version officielle de son travail, approuvée par le jury de soutenance et pour laquelle les éventuelles corrections demandées auront été apportées, et objet du dépôt légal.

ARTICLE 11 :

Les autorisations données au Diffuseur valent tant pour lui que pour tout établissement à caractère universitaire qui lui serait substitué. Le Diffuseur est notamment autorisé à déposer la version électronique de l'Œuvre sur des archives ouvertes internationales, nationales ou régionales, ainsi qu'à en permettre l'accès au moyen d'un lien vers ses propres archives.

ARTICLE 12 :

En cas de changement de législation concernant la diffusion de contenus pédagogiques, les parties conviennent dès à présent de maintenir les clauses du présent contrat compatibles avec la nouvelle législation.

ARTICLE 13 :

La loi applicable au présent contrat est la loi française.

Les éventuels litiges nés de l'interprétation ou de l'exécution du présent contrat relève, après tentative de conciliation préalable, de la juridiction judiciaire compétente en vertu des règles de droit commun.

Fait à Lorient~~/Vannes~~, le 24 janvier 2019 en trois exemplaires originaux.

Faire précéder de la mention «lu et approuvé»

L'AUTEUR

lu et approuvé

L. Pensec

LE DIRECTEUR DE THÈSE

G.C. Williams

Le DIFFUSEUR, Président de l'université Bretagne Sud

Université Bretagne Sud
Jean PEETERS
Président

"The superior tactic is to never give up!"
To Piccola, Maple, and Ivy

Special thanks to:

Table of Contents

RESUME EN FRANÇAIS DES CHAPITRES

Chapitre 1

Le premier chapitre c'est le chapitre d'introduction où la problématique de la thèse est présentée en relation aux exigences des communautés scientifique et industrielle, mais aux besoins de la population.

Les définitions des mots « sentiment », « opinion » et « émotion » sont toujours très vagues comme l'atteste aussi le dictionnaire qui semble expliquer un mot en utilisant l'autre. La difficulté réside dans la dimension humaine et très personnelle, d'où la difficulté à donner une définition nette. Tout le monde est affecté par les opinions, tout le monde veut savoir ce que l'autre pense, même si pour des raisons différentes. Les gens font preuve d'une certaine curiosité et au même temps ils utilisent les opinions d'autrui pour faire un choix : politique, personnelle, pour le travail mais aussi les loisirs. Les entreprises ont besoin d'une bonne réputation pour survivre à la concurrence et donc ils ont vraiment besoin de connaître l'opinion des gens. Et encore, grâce à la curiosité humaine de comprendre en profondeur sa nature, mais aussi à cause d'une forte demande du côté industriel, l'intelligence artificielle est toujours sollicitée par les chercheurs.

Une partie de l'intelligence artificielle s'occupe exactement de comprendre la pensée humaine, et en particulier les sentiments qui font bouger les choix de notre existence. Etudier la pensée humaine et les sentiments qui poussent nos choix, ça veut dire vraiment étudier l'homme. En plus on se fait charge de toutes les complexités du cas : une langue qui s'actualise et change du jour au lendemain, des structures linguistiques qui changent d'une langue à l'autre et dont l'origine n'est pas toujours connue ou à la portée de tout le monde. Et encore, et aujourd'hui surtout, on a une quantité d'information disponible jamais vue avant. Malheureusement, bien que l'information soit disponible, ce n'est pas toujours facile à utiliser. Le phénomène des mégadonnées (en anglais « big data ») nous permet d'avoir potentiellement beaucoup de données. Ces données malheureusement ne sont pas organisées, surtout pour certaines langues – d'où la difficulté à les exploiter. La recherche

française souffre d'un manque de ressources « prêt-à-porter » pour conduire des expérimentations, malgré une population francophone très nombreuse. C'était le cas de cette thèse dont l'objectif est d'explorer la nature des sentiments et des émotions, dans le cadre du Traitement Automatique du Langage et des Corpus.

Pendant trois ans on a construit des nouvelles ressources pour l'analyse du sentiment et de l'émotion, on a utilisé plusieurs techniques d'apprentissage automatique et l'on a étudié le problème sous différents points de vue : classification des sentiments positifs et négatifs, classification des commentaires en ligne et des caractéristiques du produit recensé. Enfin, cette thèse présente une étude préliminaire de nature psycholinguistique sur le rapport entre qui juge et l'objet jugé afin toujours d'arriver au but principale dont on a parlé quelques lignes avant : réussir à décrire en partie la complexité de la nature humaine.

Chapitre 2

Le deuxième chapitre présente le contexte de la thèse. La thématique est très variée et les objectifs des chercheurs aussi : classification de la polarité du texte, analyse de l'objectivité et de la subjectivité du texte, classification à plusieurs classes, etc. La classification peut être faite au niveau du document, au niveau de la phrase, et au niveau des expressions.

En outre, il y a aussi une classification qui a comme objectif l'analyse de plusieurs caractéristiques d'un objet, ou attributs d'une entité, ce que en anglais est appelé « **A**spect-**B**ased **S**entiment **A**nalysis » (ABSA). Grace aux compétitions annuelles et aux articles de plusieurs chercheurs, il semble que l'ABSA soit au centre de la recherche en Traitement Automatique du Langage (TAL) depuis quelques années. Cette thèse abordera, dans la deuxième partie, l'ABSA sur les commentaires en ligne et en particulier sur les commentaires complexes par structure et – parfois – longueur.

Il y a des étapes fondamentales lorsqu'on entreprend l'analyse du sentiment. En premier lieu il y a l'extraction des caractéristiques (en anglais « features extraction») qui nous donne une image, une représentation des caractéristiques du problème

qu'on aborde. Ensuite il faut choisir la méthode avec lequel on classera les informations en notre possession. Comme tout dans le TAL, il y a un approche plus linguistique et un approche plus informatique: c'est-à-dire l'approche fondé sur un lexique et l'apprentissage automatique. Avec le premier, nous décrivons le texte selon les mots d'un lexique, mais avant il faut constituer un lexique ! Un lexique peut être constituée d'une façon manuelle, ou en utilisant des dictionnaires qui nous donneront plusieurs mots en fonction d'une liste de mots germes qu'on va constituer auparavant et, enfin, la méthode qui utilise des corpus externes pour cumuler de mots.

Pour l'apprentissage machine dans ce chapitre on a décrit les méthodes principales, les plus utilisées et celles qu'on a utilisées pendant cette thèse : les arbres de décisions et les forêts aléatoires, le Naïve Bayes, les machines à vecteur support et la régression logistique. Comme on a dit dans le premier chapitre, les mégadonnées et des ordinateurs de plus en plus puissants, nous ont permis de créer et tester les méthodes d'apprentissage profond. Malheureusement on ne disposait pas suffisamment de données pour utiliser des méthodes d'apprentissage profond sur notre corpus. Toutefois dans ce chapitre on parlera des réseaux de neurones (convolutifs, récurrents, récursifs) en citant succinctement leur histoire et développement.

Chapitre 3

A partir de ce chapitre on entre dans le cœur de la thèse. Ce chapitre passe en revue les ressources utilisées et celles qu'on a créés.

On commence par les corpus : French Sentiment corpus, ABSA2018, New Corpus.

Le premier est un corpus de commentaires en ligne sur trois sujets : Films, Livres, Hôtels. Il comporte 5 niveaux de satisfaction du client, allant de « 0.5 » - les commentaires les plus négatifs, à « 5 » les commentaires les plus positifs. Ce corpus a été constitué en 2013 par deux chercheurs français et il a été utilisé comme point de départ pour nos expérimentations.

ABSA2017 est un corpus annoté en entités et qualités d'entité, en anglais « Aspect-Based Sentiment analysis ». Les schémas d'annotations et les exemples sont explicités dans les chapitres. ABSA2017 a été constitué à partir de NC Corpus.

NC corpus est une continuation du premier corpus avec plus de 9000 phrases et qui se consacre uniquement sur les Films et les Livres. Il se caractérise essentiellement par des données très fraiches datées 2017 et moins de fautes d'orthographes.

A propos de fautes d'orthographes, le chapitre présente un logiciel préexistant appelé LanguageTool qui a été modifié pendant cette thèse afin d'inclure les corrections en langue française, surtout sur certaines expressions très importantes pour le bon déroulement de la classification.

Ensuite c'est le tour des lexiques. Ces lexiques ont été utilisés pour tester si on peut exploiter indifféremment plusieurs lexiques qui mesurent l'émotion et l'opinion lorsqu'on veut une classification en polarité.

ValEmo et F-POL sont deux lexiques en français qui ont été annotés en émotion et en polarité.

Lex et RLex sont deux lexiques propriétaires et annotés en opinion. RLex présente une liste de mots qui ont été jugés importantes par la régression logistique pendant notre premier test – dont on parle dans le chapitre 4.

Pour finir on aborde l'étude fait sur les motifs linguistiques concernant les mots « plutôt », « bien que », « mais » qui sont très particuliers lorsqu'on fait une classification automatique pouvant donner lieu à une erreur de classification (c'est le cas de « bien » et « bien que »), ou ils peuvent causer un renversement de la polarité de la phrase. L'étude se termine sur des réflexions comportant l'emploi dans un contexte positif et négatif du mode conditionnel et des pronoms.

Chapitre 4

Le chapitre 4 est le chapitre qui présente les tests.

Nos expériences portent sur différents points de vue sur le même sujet en utilisant plusieurs approches : classification en polarité avec une approche uniquement statistique, approche lexicale avec statistique, approche linguistique avec une

classification qui prend en compte les sorties d'un logiciel d'analyse syntaxique de surface et les motifs linguistiques avec les informations concernant la polarité. Enfin, une approche qui utilise les SVM sur nos données annotées pour l'ABSA.

Le premier test s'est consacré à la création d'une baseline, et en particulier la réplique d'un article (Vincent, et al., 2013) qui utilise la regression logistique et les SVM sur un corpus afin de détecter la polarité des commentaires.

Le deuxième test a été l'ajout, par rapport au premier test, de l'approche lexicale en testant les lexiques dont on a parlé dans le chapitre 3 (ValEmo, F-POL, Lex, RLex). Si les résultats étaient un peu inférieurs par rapport au test de référence, on a eu deux retours importants : (A) certains domaines sont plus facilement classés. Par exemple, le domaine hôtellerie a eu un résultat – F1 score – très élevé. (B) L'emploi d'une approche mixte statistique-lexicale donne des résultats suffisamment bons pour continuer dans cette direction et pour prendre en considération l'aspect lexicale dans la classification.

Etant donné que les deux plus mauvais résultats ont été sur les films et les livres, on a donc décidé de se focaliser sur la résolution de ceux deux. Les résultats étaient influencés par : fautes d'orthographes et lexiques trop généralistes. Ce genre de lexiques n'arrive pas à bien détecter les significations spécifiques de mots dans un domaine spécifique. En outre le système prenait en considération l'intégralité du commentaire, non seulement la partie avec l'opinion mais aussi la partie avec la description du produit. Bien évidemment la polarité finale a été influencée par ces problèmes. Pour cette raison on a commencé travailler sur le corpus ABSA 2018, un corpus annoté en entités et en qualités d'entités. Ce corpus aurait dû servir pour ignorer les parties où l'opinion n'était pas exprimé lorsqu'on ne citait ni la qualité ni l'entité.

Pendant ce long travail d'annotation, l'étude des motifs linguistiques a permis de tester un système de classification en régression logistique qui utilisait les motifs, les sorties d'un analyseur syntaxique de surface et les informations sur la polarité des mots. Les résultats étaient suffisamment élevés pour démontrer encore une fois qu'il est possible d'utiliser à la fois des moyens statistiques avec des moyens linguistiques, lexicales et syntaxiques pour bien classifier les commentaires.

Enfin, des arbres de décisions et des SVM ont été utilisés pour classifier les entités et les qualités d'entités dans les phrases à l'intérieur des commentaires. Le résultat a été surprenant : avec des annotations très précises et ciblées – telles quelles des annotations humaines sur les entités et leur qualités – la classification n'a pas obtenu des résultats efficaces, surtout quand les mots liés à une entité étaient aussi partagés parmi les entités.

Chapitre 5

Le chapitre 5 c'est un chapitre extra qui n'était prévu au départ. Ce chapitre naît depuis une expérience et une collaboration avec un centre de recherche étranger (le Conseil National de Recherche du Canada – C-NRC). Dans ce chapitre on reprend la discussion entre émotion et sentiment qu'on avait abordé dans l'introduction. J'ai eu l'occasion de pouvoir tester un lexique qui est indépendant de la langue et de la culture d'origine. Ce lexique s'appelle NRC VAD lexicon et son nom vient de Valence (en français « polarité »), Arousal (en français « excitation »), Dominance (en français « dominance »). Ce lexique est normalement utilisé pour détecter les niveaux de ces trois valeurs entre une personne et une autre personne, ou situation. Il peut être utilisé pour comprendre comment une personne réagit par rapport à une situation ou à la présence d'une personne. Dans le cas de cette thèse on a utilisé ce lexique afin de détecter les différents niveaux de valence mais surtout excitation et dominance, entre le corpus des livres et le corpus de films et pouvoir donc répondre à la question qu'on avait déjà posé pendant le chapitre 3 : est-ce que les évaluateurs des films sont différents par rapport aux évaluateurs des livres ? Pour l'instant il semble qu'il n'y a pas de différences substantielles.

Ensuite on s'est posé la question : est-ce qu'on peut valoriser le rapport évaluateurs-entité évalué ? Pour répondre à cette question on a testé les niveaux de VAD par entité. Pour les films, les niveaux de VAD sont plus élevés lorsque l'entité implique la présence d'une personne (acteur ou directeur). Pour les livres ce sont les entités concernant le style et l'intérêt vers l'œuvre qui ont les valeurs les plus élevées. Et qu'est-ce qu'il se passe lorsqu'on teste les mots uniques par entité, c'est-à-dire les

mots qui apparaissent seulement une fois dans une entité parmi la totalité des entités ? Si pour les films ne change rien par rapport au test précédent, pour les livres l'entité la plus forte en terme de VAD c'est celle des auteurs – ce qui peut nous faire penser à une relation forte entre l'évaluateur et l'entité qui concerne un autre être humain.

Pour terminer, le lexique a été utilisé comme trait dans le système de classification en SVM, sans succès malheureusement.

Ce ne sont que de tests préliminaires et le travail sera amélioré dans l'avenir en cherchant de vérifier si tous les mots très pertinents pour la classification sont bien représentés et en ajoutant de mots qui – dans un lexique qui est traduit depuis l'anglais – ne trouvent pas forcément des correspondances exactes, tels quels « bouquin », « bouquin de gare » ou « navet ». Enfin, il serait aussi intéressant de tester la polarité en utilisant le paramètre V du lexique.

Chapitre 6

Ce chapitre c'est le chapitre conclusif de la thèse. Après un petit rappel de chaque chapitre on arrive aux conclusions.

En ce qui concerne les tests avec les entités et leur classification on a confirmé qu'un système SVM simple ou une combinaison d'autres systèmes d'apprentissage automatique ne peut pas atteindre des résultats optimaux comme ceux que nous avons obtenus lors de la phase de détection de polarité. Cela est vrai également lorsque nous utilisons des annotations très précises et qu'on n'évalue que les phrases contenant un seul type d'entité. Même en présence des échantillons normalisés, les classes les plus générales ont absorbé toujours les plus spécifiques. Plus le nombre de classes impliquées est grand, plus le système est spécifique et il en résulte un nombre accru d'exemples mal classés. L'utilisation d'arbres de décision et SVM dans un système de classification hiérarchique n'a pas été suffisant pour obtenir un F1-score élevé.

On suppose qu'il faut trouver pour l'avenir un moyen de peser davantage certains mots spécifiques à l'entité et un système qui peut s'alimenter en continu de ce type

de mots.

Au cours des derniers mois de cette thèse, grâce à une collaboration entre des centres de recherche, nous avons lancé une analyse psychologique des commentaires et de leurs entités via un lexique qui mesure précisément cet aspect. C'est une étude préliminaire et à l'heure actuelle, nous savons seulement qu'il existe une différence entre certaines entités pour chaque domaine. Nous considérons cette partie étant importante parce que certains articles récents ont souligné comment l'utilisation du langage est directement liée à l'état émotionnel de la personne qui choisit et prononce certains mots.

Enfin, nous espérons que cette thèse a apporté sa contribution à la communauté du TAL en décrivant les problèmes liés à l'analyse des sentiments, à l'analyse des émotions et à l'analyse des sentiments basée sur les entités. En plus, nous avons créé des nombreuses ressources (corpus, lexiques, etc.) pour le français et les avons testées sur différents systèmes de classification.

Pour l'avenir, on pense qu'il pourrait être très intéressant de poursuivre le travail présenté dans le chapitre 5, en exploitant mieux le lexique psychologique - en particulier pour les paramètres de dominance et d'excitation. Il pourrait être intéressant d'insérer des nouvelles dimensions telles quelles de l'abstraction et de mots concrèts pour évaluer les aspects évoquant l'imaginaire en tant qu'aspect médiatique.

Des travaux plus récents (Mehl, et al., 2017) ont montré qu'il est possible d'identifier certains marqueurs (la plupart d'entre eux sont des adverbes et des pronoms) liés directement au niveau de stress d'une personne.

Pour cette raison, nous espérons que tout notre travail, et en particulier le dernier chapitre, pourra servir à d'autres travaux qui se focalisent sur l'émotion véhiculé par les mots que nous prononçons tous les jours.

CHAPTER 1 – INTRODUCTION

During this thesis we will use three words several times: sentiment, emotion and opinion. The concepts behind these words are strictly linked and frankly sometimes it can be hard to distinguish one word from the others. Nevertheless, sometimes they are used in different branches of Natural Language Processing. We had the chance to work with the three during these years. We used Sentiment and Opinion during the phase of polarity detection. Finally we used Emotion when trying to detect the different psychological attitude of writers' reviews. We will start this chapter giving some examples of these three words and how they are used (or confused) in NLP literature.

What is Sentiment? and Emotion? and Opinion?

Definition of "sentiment" (from *Merriam Webster online*)

1.
    a. an attitude, thought, or judgment prompted by feeling : predilection
    b. a specific view or notion : **opinion**
2.
    a. **emotion**
    b. refined feeling : delicate sensibility especially as expressed in a work of art
    c. emotional idealism
    d. a romantic or nostalgic feeling verging on **sentimentality**
3.
    a. **an idea coloured by emotion**
    b. the emotional significance of a passage or expression as distinguished from its verbal context

What is sentiment? What are emotions? What are opinions? Are they describing the same phenomena? As it is possible to see from Merriam Webster online dictionary – one of the most reknown dictionary, it is difficult to define the boundaries of this "thing" called sentiment. The famous dictionary explains the definition using three different concepts or,

better, nuances all together. Sentiment can be something close to a *judgment* as an opinion, but also a *state of feeling* as an emotion, or something more nostalgic as *sentimentality*. Finally, sentiment is also something that draws a line almost always clear between a neutral statement and an *emotional expression*. It is evident that the distinction among these words is not so clear even for a human being.

This is nothing new; in the research environment we face the same problem. Some researchers as (Hovy 2015), underline that even the research community finds it difficult to standardize and identify in exact words what they are analysing. Flipping through the pages of Natural Language Processing (from now abbreviated as NLP) literature, it is easy to find several methods, ideas, and procedures to analyse the triplet opinion/sentiment/emotion in several languages, but it is harder to find a real definition of them. According to (Pang and Lee 2008) "*This proliferation of terms reflects differences in the connotations that these terms carry, both in their original general-discourse usages and in the usages that have evolved in the technical literature of several communities.*". The research community definitely agrees.

As stated in (Pang and Lee 2008), 2001 is the year labelled as the beginning of the sentiment analysis age for NLP community. In this context, sentiment analysis is "*the process of computationally identifying and categorising opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product is positive, negative, or neutral*", as Oxford dictionary explains. Through the years, many research papers have been produced, using different keywords and names to express interesting common research points. We can then agree that to multiple names for the research problem analysed corresponds a mirror of different points of view of the problem itself. This is why we can encounter several different names: subjectivity analysis (Wiebe 2004), opinion mining (Pang and Lee 2008), sentiment extraction (Das, et al., 2007).

All these authors share one common point, the root of their work: the analysis of what has been said or written by people, the study of a sentiment linked to something in particular (a product, a topic) in the text. Using different techniques, the goal is to assign a final score to the text that should correspond to the value of the opinion (or sentiment/emotion). The score is usually something like *good/bad* with different (usually numeric) nuances.

While the subject is mature, as proved by many published surveys (Pang and Lee 2008), (LIU 2012), there is still room for improvement, demonstrated by the interest for the yearly NLP conferences and workshops as SemEval (Pontiki, et al., 2014), (Pontiki, et al., 2015), (Pontiki, et al., 2016) or Wassa (Mohammad and Bravo-Marquez 2017) and by the challenges that opinion analysis still offers (Breck and Cardie 2017), (Mohammad 2016).

Is language so predictable?

The problem in itself is not so simple: language is the mirror of people's thought, personal situation, story, environment, background and education (Malt Barbara C. 2013). For example, one of the most spoken language, English, has a large variety of dialects and expressions that sometimes create misunderstandings also between people speaking the same language since birth.

For example the expression " *(to use the) washroom*" used in Canadian English to indicate the use of "*a room that is equipped with washing and toilet facilities*" – quoting Merriam-Webster, is not recognised as an English expression from people who come from UK, even though they don't speak the famous Queen's English and they usually prefer a "less noble" English variety. UK people use in fact the expression "*go to the toilet*", that it is perceived as too much sincere and personal by Canadians.

It is not possible to count exactly the number of existing English dialects but it is known that they are very numerous (Hickey, 2005). In addition to this, humans change, adapting themselves to the environment and the historical period. They create new words, especially new meanings for words already existing.

An example can be easily taken from words describing colours. We can think that colours are universal and they have always had that meaning and lexical form, but it is not true. First of all, colours are not perceived in a uniform manner by each culture - and by each person. In addition to this, the *orange* colour, for example, has been accepted by English only during the 17th century, thanks to the increasing importance given to dyeing technologies. Before that, it was just an indirect loan from Sanskrit word "*naranga*" that became "*naranj*" in Arabic. Finally, thanks to Arabic trade it was exported to Spain, Sicily and France where it became "*orenge*" and finally welcomed in English as "*orange*".

The connection between words and history is there for all to see, and sometimes it is strictly related to the support we use to express ourselves. Many words from today, such as "*youturn*" (on Twitter, to follow another person on social media with the intention of unfollowing them once they have you followed back), "*wallflower*" (when someone consumes other people's social network without actively participating), "*hash-browning*" (the excessive use of hashtags in a post) are used now, because of social networks, and maybe they will not exist in 10 years.

As a consequence of this, language and opinion expression - that would be non-existent without words - change. Is it really so easy to limit, to describe people's thought as a value and a topic? Does sentiment really correspond to emotion? Are they interchangeable? You will find some experiments in Chapter 4 - Experimentations. In fact, one of the first experiment that has been carried out during this thesis wanted to answer the question "are opinions and emotions the same?"

The central role of opinions in daily life and in enterprises' business

> ***"Man is by nature a social animal; an individual who is unsocial naturally and not accidentally is either beneath our notice or more than human. Society is something that precedes the individual."***
>
> ***(Aristotle - Politics, 1253a.).***

Why are opinions so important for us? Without entering in complex psychological and philosophical considerations, the answer can be at least predicted. Everyone wants to know something more before buying a product. Everyone likes staying informed about the latest trends. Everyone, even the most misanthropic person, wants to know what the others think.

Enterprises are very interested in sentiment analysis, too. Imagine being an enterprise launching a new product. First of all, enterprises try to understand if the product is innovative and can be appreciated before investing money and time. In any case, the goal is to produce something enticing.

Once the enterprises launch their product, they obviously want to know if it is well received by customers. Maybe they get several reviews, and they think that many people are, as a consequence, buying the product. They can automatically infer that the product is good and

that people love it. Without an analysis on the sentiment expressed on the review itself, enterprises will not know if people are reviewing it positively or negatively. Moreover, gaining knowledge about people's sentiment and needs is their key to be able to create the useful product that everyone wants. With a good sentiment analysis tool, enterprises can improve their business: determine marketing strategy, improve campaign success and customer service, and so on.

Apart from the fact that both enterprises and individuals want to know others' opinions - that can be easily understood, because enterprises are made by people anyway, there is something more that they have in common: the use of social networks.

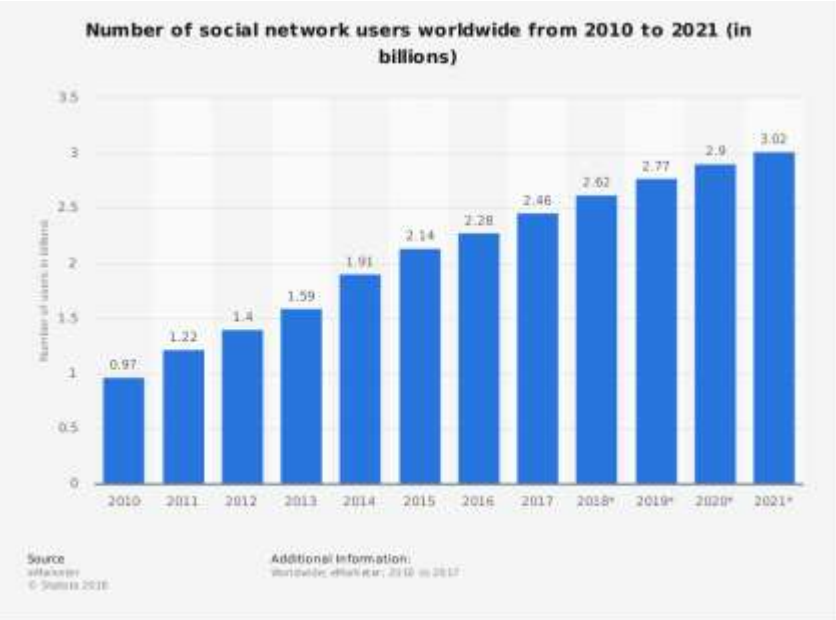

Figure 1 number of social network users worlwide from 2010 to 2021 (source eMarketer,2016)

To give an idea: in 2010, 970 million people used social networks for several reasons: buy products, write a personal journal/blog, and stay informed. In 2017 the number of people using social networks has almost triplied: 2.46 billions of people. It is estimated that in 2020, there will be over 3 billion people using social networks. (source: eMarketer[1], 2016). One of the reasons that can explain this phenomenon is that people are invited to use social networks to do everything: from expressing opinion for a product to order food and let the others know that it was delicious (Bolton, Parasuraman et al. 2013). In addition to this, it is

[1] https://www.emarketer.com

easier and easier to get a mobile app connected to a given social network or service. As of 2017, daily social media usage of global internet users amounted to 135 minutes per day, up from 126 daily minutes in the previous year (source: Nielsen[2], 2017).

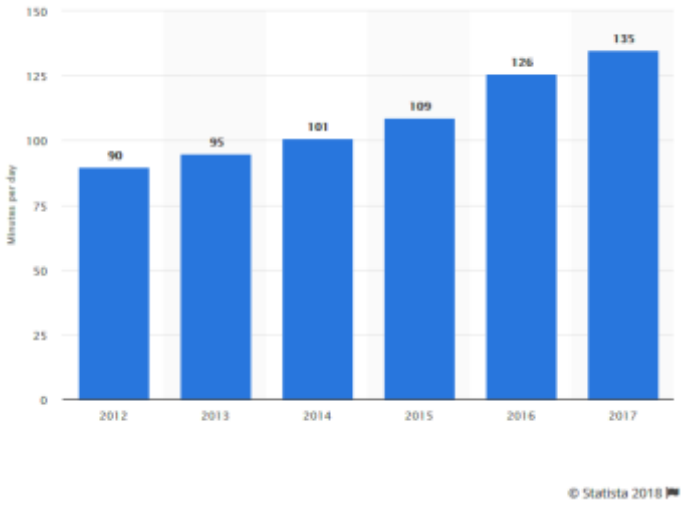

Figure 2 Daily social media usage of global internet (source Nielsen,2017)

Recent studies underlined that 92% of marketers have increased their exposure through social media (*brand awareness*) and 80% had positive results in terms of website traffic (SmartInsights.com[3]).

Does Big Data involve better analysis?

Needless to say: there is a huge amount of data on the net that can revolutionize not only business but also scientific research, public administration, security. According to (Chen, et al., 2014), (Manyika, Chui et al. 2011), the volume of business data worldwide doubles every 1.2 years. Enterprises use a ginormous amount of data: Walmart's 6000 stores produce around 267 million transactions per day. Recently, the company asked for a data warehouse of 4 petabytes for their transactions. And again, it seems that machine learning can improve business by exploiting the knowledge hidden in high volume of data, improving pricing strategies and advertising campaigns.

[2] https://www.nielsen.com

[3] https://www.smartinsights.com/social-media-marketing/social-media-governance/social-media-marketing-effectiveness-2014/

But is big data easily available? Are we able to analyse and take advantage of that data? The answer is negative. Or better, the answer is just “maybe”. One of the most common problems on the net, despite the language used, is that data is not organised. Individuals, researchers and enterprises still have some difficulties when searching for an exact type of data. That is why the research community is trying to organise data by classifying it in different ways: from the point of view of sentiment analysis (Turney, 2002), (Yi, Bunescu et al. 2003), (Francisco, et al., 2006), (Mohammad, et al., 2011) to the opinion analysis point of view (Rabelo, Prudêncio et al. 2012), (Miao, Li et al. 2009), (Bellegarda 2010), (Liu, Lieberman et al. 2003), (Mohammad, 2012).

Languages unluckier than others for resources

Another problem that doesn’t concern all languages is the lack of resources to analyze big data. Many resources have been created for English: ANEW (Bradley, et al., 1999), WordNet-Affect (Strapparava, et al., 2004), Balanced Affective Liste Word (Siegle 1994), SentiWordnet (Esuli, et al., 2006), (Baccianella, et al., 2010), NRC Emotion Lexicon (Mohammad, et al., 2010), Bing Liu's Lexicon (Hu, et al., 2004), the MPQA Subjectivity Lexicon (Wilson, et al., 2005) are some examples. Unfortunately, it is not possible to state that for French.

When starting the thesis, in fact, I faced several problems: lack of resources as per what concerns opinion lexicons on movie and book domains, together with a lack of corpora big enough for my experiments. Even though there are some resources like (Lark, et al., 2015), (Syssau, et al., 2005), (Pak, et al., 2010), (Vincze, et al., 2011), (Sajous, et al., 2013), they are not oriented to customers’ reviews. And still, I couldn’t find a spellchecker to correct misspelled words and slang.

For these reasons it was interesting to start from these problems and try to solve them one by one. This represents just one problem, more is described in the next section.

Complex contexts, complex texts

How to handle complex texts and the related linguistics and semantics? How to find opinion expressions that are unknown? And how to exclude n-grams related to the description of a

product and not directly an opinion? These are only a few of the questions that this thesis has worked on.

The idea behind this thesis was to work on French language, improve resources and tools, and create a system able to handle the complexity of such reviews.

The achievements and contributions

At present, the system is indeed able to extract reviews from the web and extract most of the opinions from them, distinguishing different aspects that are listed in the review. Unfortunately, the system still lacks of flexibility because it can't recognise new entities and qualities.

The interesting point of this thesis is the particular attention that has been given to the linguistic dimension of the problem during the whole process. In addition to this, it has been interesting to work on French language, creating new resources and facing some challenges that sometimes were harder to solve than the same ones in other languages that can benefit from a larger community and more resources, such as English for instance. This thesis gave some contributions to the research community thanks to resources such as NCABSA17 Movie and Book database and corpus, linguistic descriptions of some French part of speech (PoS) that may be useful for sentiment analysis system based on shallow parsers, a linguistic comparison of two domains from the point of view of grammar and style, a deep reflection on the role of linguistics and statistics when used to solve tasks such as opinion analysis and finally, thanks to some tests that will be continued in the future, a preliminary study on how a psycholinguistic analysis can influence sentiment analysis systems.

Structure of the thesis

Chapter 2 describes the background, while chapter 3 gives a detailed description of the created resources. In chapter 4 there are the experimentations involving some machine learning techniques and a study about the different impact of emotional or opinion words in a sentiment analysis system, a mixed (statistical and linguistic) approach using a shallow parser modified *ad hoc* to analyse some of the particularities of the language used in these

domains. Finally chapter 5 is an extra chapter created during a collaboration with another research centre, exploring the psycho-linguistic side of the reviews.

Summary of Chapter 1

This chapter is an introduction of the thesis and it explains the reasons behind the choice of this subject and the goals. We define the sentiment analysis problem as something known in both research community and everyday life. In fact, people seem interested to know more about something, enterprises want to analyse customers' thoughts in order to improve or create their products. Research community finds difficult to establish just one name for these kind of problem, because there are different points of views to explore: subjectivity/objectivity, sentiment/emotion. It is recognised as a difficult task and unfortunately the presence of non-organized (big) data and few resources - especially in languages other than English - can potentially pose a problem. The objective of this thesis is to explore sentiment and emotion analysis, to give a contribution to the research community by creating new freely available resources for French language, and to progress in classification systems able to classify complex reviews involving alternate sequences of opinions and descriptions of an object.

CHAPTER 2 - BACKGROUND

Sentiment Analysis has been tackled in different ways according to the main task, goal, and point of view. Therefore, we will not enter too much in depth in the mathematics and statistics behind because, as a Computer Science oriented to languages thesis, we have used these algorithms to analyse languages.

A classification problem

If sentiment analysis is defined as the "field of study analysing people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organisations, individuals, issues, events, topics, and their attributes" (Liu, 2012), it is easy to say that the problem is faced - and described, using different keywords and points of view as well. According to theories about sentiment analysis, for some researchers opinion analysis involves some opinion holders that have a claim about a topic associated to a sentiment (Kim and Hovy 2004). For other researchers (Liu 2010), opinions are identified by five other factors instead: object, features, holder, time and (sentiment) orientation. In this case the opinion holders are interested in some features of a certain object; then, they associate a certain (sentiment) orientation about these features during a specific lapse of time.

Researchers in Natural Language Processing address these problems as a classification task. The **classification task** is not unique: there are a lot of approaches, and interesting and different goals. For example we could be interested in classifying texts as positive or negative: this is called **polarity classification** (Pang B. Lee, et al., 2002), (Turney, 2002). We could also want to classify texts based on the features they contain, and their values: in this case we have a **multi-class (or multi-categorisation) task**. A very common example is represented by that kind of websites describing, for instance, mobile phones, using a star ranking system to describe each feature (screen, RAM, audio, touch sensitivity, etc.).

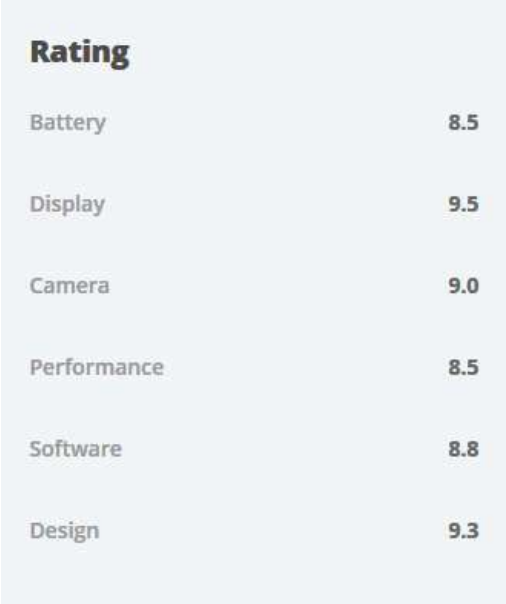

Figure 3 Rating system of aspects of a mobile phone (source tomsguide.com).

Another way to classify texts using sentiment/opinion analysis is the **subjective vs the objective** classification of the document (Wiebe, 2004), (Wiebe, et al., 2005). Texts can, in fact, express an objective fact or express a subjective opinion. In some cases, we can have speech informalities (such as slang and language inaccuracy), sarcasm, subjective language, emoticons, and co-text (external references) that makes the opinion detection harder. User-generated content such as blogs, tweets, Facebook pages, are plenty of opinionated text, and it is essential sometimes to analyse them, as we said in the first chapter. For many years, research focused on factual texts because they are easier to analyse (Liu 2010), but today the new trend is to analyse the subjective part of the web content to retrieve opinions about several themes. Classification can be executed at different levels: document, sentence, phrase, aspect. We will start from the coarsest granularity to the finest one. We will now take the same text as an example for the different levels of analysis.

Classification Levels

Document Level

> ***"I have been using this case for over a year now and love it. Before this one, I kept switching cases to find something I liked. This case is just the perfect combination of protection and style. The black, simple design is perfect for me, and the case provides great protection! The sides overlap the front of the phone to protect the screen. Only the installation was a bit hard. Would definitely recommend, the price cannot be beaten!" - Anonymous***

In **document** level, the entity to be classified is the whole document. Given a document *D*, the goal is to define the overall sentiment it carries. The sentiment can be positive, negative or neutral. Usually this is done by using the average of the sentiment orientations of all sentences in the document. In the text above we can say that generally the review is very good, but we can't specify the reasons why. Many researchers, in different fields, have worked on detecting sentiment in documents (Pak, et al., 2010), (Moraes, et al., 2013), (Das, et al., 2001), (Huettner, et al., 2000). One of the aspect that has been criticized is that sometimes a document can contain more than one opinion. This makes it hard to give just one sentiment label to a whole document; the classification will be done for each sentence (McDonald, et al., 2007).

Sentence Level

In **sentence level** each sentence *S* has to be analysed and classified according to the sentiment orientation of the words and phrases in S. When the analysis concerns more than one sentence in the same paragraph, it can be called passage level analysis. In the example above, phrases such as: "*Would definitely recommend, the price cannot be beaten*" and "*I have been using this case for over a year now and love It.*" are classified as positives. When the analysis is carried out at the sentence level, it is important to distinguish the opinionated content versus a factual content in the sentence (subjectivity classification). This is due to the fact that for this type of classification it is fundamental to assume that for each sentence we have none or at least one opinion about something (Kim, et al., 2005), (Yu, et al., 2003). According to (Liu, 2010), most opinionated sentences are subjective, but objective (factual)

sentences imply indirectly opinions too. One way to study the subjective part of a sentence is searching for desires, beliefs, suspicions, and other human's sentiment (Wiebe, 2000), (Wilson, et al., 2005). Even in this case there can be some problems. For example, if we search only for some verbs because we think that we can be sure to associate a sentiment to them, we can have bad surprises. Some verbs can or not convey an opinion: "*I believe that he's coming home"* and "*I believe that's amazing!"*.

Once we are sure that there is something to be classified, the goal is to define it, and of course find its orientation. We can have sentences expressing a clear and direct opinion about something ("I don't like this stuff"), or express the same concept using conditional structures (Narayanan, et al., 2009). Conditional sentences are sentences involving two parts: a hypothetical situation and its consequence. An example is: *"If your Nokia phone is not good, buy this Samsung (phone)*".

Another type of sentence that is hard to handle is the comparative one. In this case the opinion is expressed by comparing the analysed object to other objects. ("I preferred the other movie about robots than this one") (Pecòre, et al., 2018), (Jindal, et al., 2006).

Some questions arise from the use of this approach: what should we do when we are unable to identify a specific opinion about something? How should we carry out our analysis when we have more than one opinion in a sentence?

Sometimes the research community has wondered if there is really a difference between document and sentence analysis. This is due to the fact that we can consider a sentence like a small document (Liu, 2012).

Are there finer levels of analysis? Yes, and they are called phrase and aspect-sentiment analysis.

Phrase Level

A phrase is a collection of words that may contain nouns or verbs, but it does not have an acting subject. Phrase is a level of study in Sentiment Analysis (Wilson, et al., 2005), (Hatzivassiloglou, et al., 1997), (Esuli, et al., 2006). As underlined in this last paper, sometimes it is important to identify the polarity of some expressions inside a sentence. An example is represented by the question-answering systems that have to handle expressions of positive and negative sentiments to successfully answer questions about people's

opinions (*"We don't hate⁺ the sinner," he says, "but we hate⁻ the sin.*"). Is it sufficient to create a lexicon of positive/negative expressions? Unfortunately not always. Sometimes it can be useful to study words inside a context, in a phrase-level sentiment analysis. Some concepts ("to sleep well") seem to be always positive ("I slept well in this hotel"), but a change on the context can give a different meaning to the same phrase and change the polarity ("I slept well during the movie"). Others difficulties concern also negations. Negation could be only local ("*not good*"), or it could be a negated proposition ("*does not* look very good"), or subject ("*no one* thinks that it's good"). There are also negation words that intensify rather than change polarity ("*not only* good but amazing"). Contextual polarity may also be influenced by modality, tenses and modes of verbs, fixed expressions involving words apparently positive but neutral in reality ("*Environmental Trust"* - the name of a statutory body vs "you gain my *trust"* - an expression of appreciation), diminishers, and so on.

Aspect Level

Aspect level is suitable for reviews. Thanks to this kind of analysis, we are able to describe a product by using its features and by expressing a judgment for each (or some) feature (s). The classification can be harder for blogs or forums, where entities and features are unknown, and information is even less structured than reviews.

In general, we can say that the expressions and their related entities are used and categorised according to some characteristics, which are normally the features of the reviewed product(s). Aspects can be explicit if they have been already mentioned, they are known or can be easily supposed. Aspects are implicit when they are not directly mentioned nor already known and they need to be discovered before being analysed. Some examples of explicit aspects can be found when analysing a phone. Some features of the product are well known (and then expected to be mentioned) such as screen, weight, battery, etc. In our example the sentence "*The black, simple design is perfect for me, and the case provides great protection!*" is describing the design as well as the strength of the case giving a very positive value.

According to (Hu, et al., 2004) an entity E can be a product, person, event, organization or topic. The entity E can be a hierarchy of components, sub-components, and so on. Each

component has some attributes or features or facets (Canon > Lens, Battery; Battery > life, size, capacity; etc.).

Among the several competitions and challenges in Sentiment Analysis field (Nakov, 2016), (Rosenthal, et al., 2015), (Pontiki, et al., 2016), (Ghosh, et al., 2015), (Recupero, et al., 2014), Semantic Evaluation (SemEval) is one of the most important competition for the evaluation of aspect-based sentiment analysis systems. The intuition behind this type of competition is that the exploration of language's nature of meaning is very important and difficult, especially when it is handled by computers that cannot perceive meaning as intuitively as humans.

Aspect Based Sentiment Analysis was introduced as a shared task for the first time in SemEval-2014 (Pontiki, et al., 2014), with datasets in English language for two domains: laptops and restaurants. The datasets were annotated at the sentence level and aspect terms as well as coarser aspect categories were provided with polarities. In SemEval-2015, the task was repeated and extended, with the use of opinion sentence-level tuples adding the hotel domain with a dataset of whole reviews and not just isolated sentences (Pontiki, et al., 2015). There were several sub-tasks too: detecting the specific topics an opinion refers to (slot1); extracting the opinion targets (slot2), combining the topic and target identification (slot1&2) and, computing the polarity of the identified word/targets (slot3). In SemEval-2016 (Pontiki, et al., 2016), new and multilingual datasets were provided: restaurant reviews in six languages (English, French, Dutch, Russian, Spanish and Turkish), hotel reviews in Arabic, consumer electronics reviews in three languages (English, Dutch and Chinese), telecom reviews in Turkish and museum reviews in French.

After analysing the different levels of classification, it is now time to understand which are the most used techniques of classification in Sentiment Analysis and why.

Feature Extraction

When dealing with classification, the first things to look to at are features, such as: opinion words and phrases, negations, Parts Of Speech, chunks, n-grams presence and frequency.

Once the features are established, the next step is the selection and the methods that can be used. Feature selection is important to remove irrelevant or redundant information

before training the algorithm. This improves both data visualisation and understanding and usually computational efficiency (Forman, 2003), (Guyon, et al., 2003).

There are two families of methods addressing sentiment analysis, and they will be the object of the next section: lexicon-based methods and statistical methods. The lexicon-based methods can be extended in an unsupervised way: using some reliable “seed” words, it is possible then to bootstrap this set with synonyms, antonyms, or other resources (depending on the final goal) to obtain a larger lexicon. Is it always possible? The answer is no. For example during the first tests for this thesis, Wordnet was used to extend opinion words by translating the synonyms from English. Unfortunately, and it seems true for every language, synonyms do not convey the same meaning. In addition to this, when dealing with emotions and sentiment and translation too, it seems very hard to convey the same sentiment nor the same intensity.

Another test - the next chapter will explain it in more details - was to merge two lexicons of opinion and emotion, and also in this case the result was not so good for opinion classification.

Sentiment classification and regression techniques

This section covers several approaches used in sentiment analysis field to study sentiment and classify them. We will start from the most linguistic approach, the lexicon-based one, to reach the most mathematical one, the deep learning approach.

Lexicon-Based Approach

What is a lexicon? It is a collection of words covering one or several topics or domains. One of the most used lexicon in Sentiment Analysis is the opinion lexicon, i.e. a lexicon made of opinion words, in other words, words carrying a positive or negative opinion. Sometimes it is very difficult to create a lexicon that can fit at the same time more than one domain, because words can have a different polarity according to the context of expression. How is a lexicon created? There are three main approaches: manual, dictionary-based and corpus-based.

Manual Approach

It is a very precise way to create a lexicon, but also the most time consuming. This approach requires many people to work on it and many controls to ensure that the lexicon is consistent and accurate.

Dictionary-based approach

In this case, there is usually a small set of words - called “seed” list - that is then expanded using dictionaries / thesaurus (Mohammad, et al., 2009), or resources such as WordNet, (Fellbaum, 1998) or SentiWordNet (Esuli, et al., 2006), (Baccianella, et al., 2010) to take into account synonyms and antonyms. It is an iterative process; for each cycle, a list of new words is added to the seed list. The expanded seed list is then used to search and add other words, and the cycle restarts. The main disadvantages of this method are:

a. it is hard to find opinion words with domain and context specific orientation;
b. Using resources such as Wordnet for synonyms and antonyms might not be an ideal solution for sentiment/emotion analysis, because it is nearly impossible to find two words really sharing the same meaning and intensity.

Corpus-based approach

The corpus-based approach can be a good solution to have a lexicon that can fit the domain and the context of the application. One method involves always a seed list of opinion words that is extended with other similar opinion words seen in the corpus. A possible approach consists of starting with a list of seed opinion adjectives and use some linguistic and syntactic constraints to identify potential additional adjectives and their orientation. (Hatzivassiloglou, et al., 1997).

One example is given by two adjectives connected by a conjunction such as “and”, “or” and so on. In this case the sentiment consistency is analysed. Some words such as “but”, “however”, “rather than” can be exploited thanks to their adversative orientation in order to predict two adjectives or, in general, words that are one the opposite of each other in terms of opinion orientation. Usually the corpus-based approach relies on statistical methods, such as Conditional Random Fields (Lafferty, et al., 2001) that can be used as a sequence learning model to extract opinion expressions. Other work: (Kanayama, et al., 2006) first use clause

level context coherency to find candidates, then use a statistical estimation method to determine whether the candidates are appropriate opinion words.

Statistics is also used to create opinion lexicon. For example patterns or seed opinion words can be found using co-occurrence of adjectives in a corpus and then deriving posterior polarities (Fahrni, et al., 2008).

Another possibility is to identify the polarity of a word by studying the occurrence frequency of the word in several texts. If the word is found in a negative context, it will be probably a word with negative polarity. If it is found in a positive context, then it will be probably positive. (Read, et al., 2009), (Turney, 2002). A way to infer polarity of an unknown word is to use pointwise mutual information by calculating the relative frequency of co-occurrence with another word. Another statistical approach is Latent Semantic Analysis that lets us analyse the relationships between documents and the terms on them in order to produce patterns related to both (Deerwester, et al., 1990).

Finally, another way to create a lexicon is using a semantic approach. The semantic approach aims at computing semantic similarity between words. If two words are semantically near, they will share the sentiment. Following this principle, resources such as WordNet have been created.

Machine Learning Approach

Machine Learning sees Sentiment Analysis like a text classification problem using syntactic, linguistic, semantic, pragmatic features.

The classification problem can be divided into:

a) hard classification, when only one label is assigned to the feature;
b) soft classification, when a probabilistic value referring to a set of possible labels is assigned instead of a single label.

Classification can be human-assisted or not. It can feature a pre-labelled training set, but it's not compulsory. Also, classification can be supervised or unsupervised; the choice depends on the problem we want to solve, and the available data.

If we want to predict a predefined target value, the solution will be supervised learning.

Unsupervised learning learns from test data that has not been labelled, classified or categorized.

Supervised Learning (Machine Learning)

The supervised learning methods depend on the existence of labelled training documents. In a supervised task in fact, we tell the algorithm what to predict. In this section we will present some of the used methods with some references as examples.

Decision trees

The decision tree has been one of the most common used techniques in classification (Seni, et al., 2010), (Quinlan, 1986). The metaphor behind a decision tree is the representation of a set of decisions as an expert could do while choosing the correct answer to a problem. The idea behind a decision tree is that we break classification down into a set of problems and we make choices about each feature in turn, starting at the root (base) of the tree and going along the leaves, where we have the classification decision.

Decision trees use a greedy heuristic to perform search, evaluating the possible options and choosing the one that seems optimal at that point. They exploit the idea of "divide and conquer": they divide recursively a problem in many sub-problems and then they solve them. Decision trees are so popular because they are easy to implement, transparent and similar to a set of logical disjunctions (*if* … *then* rules):

Should I go outside?

- **if** it's raining **and** you have to write a thesis, **then** stay at home
- **if** it is not raining, **then** go out

Figure 4 - An example of Decision Trees

The algorithm building the tree chooses the most informative feature for each step.

The most informative feature is the one that brings out the final answer faster than the other features.

Let's think about the game of the 20 questions: we have more chances to win when we ask promptly the right questions in order to be the first to solve the mystery. A decision tree works in the same way.

First of all we should know and measure the consequences of choosing a feature over another one. In other words we should consider how much information we gain or lose for each split according to the feature. For this reason, some information gain estimators have been used during these years:

- **Gini:** this is the first type of measurement used for decision trees. In 1984, Breiman used the Gini coefficient that measures the dispersion of data in a tree structure.
- **Information Gain and Entropy**: in 1986 Quinlan, the creator of ID3 algorithm, used the information gain to measure the reduction of the entropy level for each choice, i.e. reduction of noise in the chosen data.

Entropy and Information Gain

In order to define information gain precisely, we begin by defining a measure commonly used in information theory, called *entropy*, that characterizes the (im)purity of an arbitrary collection of examples.

The entropy H of a set of discrete probability distribution p_i is

$$H = -\sum_i p_i log_2 p_i$$

Equation 1 - Entropy

where p_i is the proportion of examples belonging to class i. The logarithm is still base 2 because entropy is a measure of the expected encoding length measured in bits. The entropy is 0 if all members of p belong to the same class. For example, if all members are positive then p_+ is 1, and p_- = 0, and Entropy (H) = $1\log_2(1)+0\log_2(0)=0$. The entropy is 1 when the collection contains an equal number of positive and negative examples. If the collection is mixed, Entropy is between 0 and 1: the upper bound $-\log_2[1/\#cat]$, where #cat is the number of categories; if two categories, then Entropy (H) = 1

In other words, if all of the examples are positive, then we don't get any extra information from knowing the value of the feature for any particular example, since whatever the value of the feature, the example will be positive. When the entropy is at a maximum, i.e. we have the most information possible, is very useful to know about that feature. After using that feature, we re-evaluate the entropy of each feature and again pick the one with the highest entropy.

The next step is to apply the feature in our tree. We should keep in mind that our goal is to know the quantity of entropy of the whole training set that could decrease if we choose a particular feature or another one. This is known as the *information gain*, and it is defined as the entropy of the whole set minus the entropy when a particular feature is chosen

$$Gain(S,F) = H(S) - \sum_{f \in values\ (F)} \left|\frac{Sf}{s}\right| H\ (Sf)$$

Equation 2 - Information Gain equation

where S is the set of examples, F is a possible feature out of the set of all possible ones, and |Sf| is a count of the number of members of S that have value f for feature F). The function Entropy (S_f) is similar, but only computed with the subset of data where feature F has values f. The decision trees can use different type of algorithms.

The *ID3 algorithm* (Iterative Dichotomiser) (Quinlan, 1986) computes this information gain for each feature and chooses the one that produces the highest value at each stage. ID3 is a greedy algorithm that grows the tree top-down, at each node selecting the attribute that best classifies the local training examples. This process stops when the tree classifies correctly the training examples or none of the attributes is available anymore.

C4.5

C4.5 (Quinlan, 1993) is the improved version of ID3 because it accepts both continuous and discrete features and handles incomplete data points. In addition to this, it tries to solve over-fitting problem using "rule post-pruning" technique (described below), and it accepts different weights applied to the features.

On the other hand, C4.5 can have empty branches and if the data is too noisy, it can pick up weird features that are uncommon.

CART (or Classification and Regression Trees)

CART (Breiman, et al., 1984) uses binary trees for classification. The idea behind CART is that every problem can always be translated in a binary sub-problem. The real difference is that CART uses another type of measurement: the Gini Impurity. When each leaf node doesn't have data points of the same class, it is considered as impure. The algorithm loops over the different features and checks how many points belong to each class. If the node is pure, then N (i) = 0 for all values of i except a particular one. So for any particular feature k:

$$G_k = \sum_{i=1}^{c} \sum_{i \neq 1} N\,(i) N\,(j)$$

Equation 3 - Gini Impurity equation

where c is the number of classes. The Gini impurity is equivalent to computing the expected error rate if the classification was picked according to the class distribution.

The information gain can then be measured in the same way, subtracting each value G_i from the total Gini impurity.

With CART is possible to add a weight to the misclassifications. The idea is to consider the cost of misclassifying an instance of class i as class j (called also risk) and add a weight that says how important each data point is.

$$G_i = \sum_{j \neq i} \lambda_{ij} N(i) N\ (j)$$

Equation 4 - CART equation

Pros/Cons of Decision Trees

There are some problems when using decision trees:

- over-fitting
- existence of different trees that can describe the same problem
- instability: a small change can lead a large change in the structure of the tree

How to avoid over-fitting?

- Limit the size of the tree
- Use a validation set and measuring the performance of the tree against it and decide whether to stop or not.

Another and most used approach is pruning. Basically the act of pruning consists in computing the full tree and reducing it, evaluating the error on a validation set. The error of the pruned tree is evaluated on the validation set, and the pruned tree is kept if the error is the same as or less than the original tree, and rejected otherwise.

C4.5 uses a different method called rule post-pruning:

1. Create the tree and allow over-fitting to occur.
2. Convert the learned tree into an equivalent set of rules by creating one rule for each path from the root node to a leaf node.

3. Prune (generalize) each rule by removing any preconditions that result in improving its estimated accuracy.
4. Sort the pruned rules by their estimated accuracy, and consider them in this sequence when classifying subsequent instances.

Decision trees are human-understandable and are used for several goals: spam detection, part-of-speech tagging (Quinlan, 1986), (Schmid, 2013), customer satisfaction (Yussopova, 2015), and so on. The advantages of the decision trees are that they are simple to understand, they can be combined with other decision techniques, and they can simplify and better describe a problem.

Random Forest

Speaking of decision trees: If one tree is good, a forest of many trees should be better as long as their data vary enough. When using Random Forest (Breiman, 2001), we end up with randomness caused by two factors:

- different trees trained on slightly different data
- limit of the choices that the decision tree can make: a random subset of the features is given to the tree at each node, and it can only pick from that subset rather than from the whole set.

The effect of these two forms of randomness is to reduce the variance without affecting the bias. In Random Forest there is no need to prune the trees but the number of the trees to put into the forest is an important parameter: this can be chosen measuring each time the error until it stops decreasing. Once the set of trees are trained, the output of the forest is the majority vote for classification or the mean response for regression.

Naive Bayes

Bayesian reasoning and theorem (Price, 1763) is a probabilistic approach to inference. The assumption behind bayesian reasoning is that we can use probability distributions and observe data to infer optimal decisions (Hand, et al., 2001). Some studies underline the power of a classifier based on this algorithm. (Mitchell, 1997) provides a detailed study

comparing the naive Bayes classifier to decision trees and neural network algorithms, showing its competitiveness. In some cases, in fact, a naive Bayes can outperform other methods. Bayesian methods is famous due to:

- Flexibility: each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct.
- Mixed knowledge: prior knowledge and observed data can be combined to find the final probability of a hypotheses.
- Produce of probabilistic predictions: e.g. hypotheses such as "the success of your PhD is about **99%**".
- Incremental classification: new instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities

Bayesian methods usually use initial knowledge of many probabilities or estimation about them based on background knowledge, previously available data and assumptions about the form of the underlying distributions.

Given a hypothesis H and an observed training data D, we would like to determine the best hypotheses from H. A "best hypothesis" is the most probable hypothesis given the data D and any initial knowledge about the prior probabilities of the various hypothesis in H. The prior probability of h - P (h) - may reflect any background knowledge we have about the chance that h is a correct hypothesis. When we lack prior probabilities, we assign the same uniform prior probability to each candidate hypothesis. Prior probability applies to the data D, too - P (D) denotes the prior probability that training data D will be observed. When we want to study the relation between D given h, we will write P (D|h) to denote the probability of observing data D given some world in which hypothesis h holds.

In machine learning we are interested in the probability P (h|D), the posterior probability of h or the confidence that h holds after we have seen the training data D. The posterior probability p (h|D) reflects the influence of the training data D, in contrast to prior probability P (h), which is independent of D.

Bayes theorem let us to calculate the posterior probability p (h|D) from the prior probability P (h):

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Equation 5 - Posterior probability equation

Most of the time, the goal of the learning is to find the most probable hypothesis h(E) given the observed data D: the *Maximum A Posteriori* hypothesis (MAP)

$$h_{MAP} = \text{argmax } P(D|h)P(h)$$

Equation 6 - Maximum A posteriori hypothesis

If we assume that every hypothesis in h is equally probable a priori ($P(h_i) = P(h_j)$ for all h_i and h_j in H) the equation is simplified because we need only consider the term $P(D|h)$ to find the most probable hypothesis, i.e the *likelihood* of the data D given h, and any hypothesis that maximizes $P(D|h)$ is called a *maximum likelihood* (ML) hypothesis, h_{ML}:

$$h_{ML} = argmax_h P(D|h)$$

Equation 7 - Maximum Likelihood hypothesis

One highly practical Bayesian learning method h is the naive Bayes learner, often called the naive Bayes classifier.

The Naive Bayes classifiers are a family of probabilistic classifiers based on Bayes theorem, with independence assumptions between the features. What does it mean? It means that the method considers "naively" that the features have the same importance and they are used in the same contexts. As underlined by (Harrington, 2012) "*The word bacon is as likely to appear next to unhealthy as it is next to delicious. We know this assumption isn't true; bacon almost always appears near delicious but very seldom near unhealthy. This is what is meant by naive in the naive Bayes classifier.*" Another good example can be the classification of the item "apple". The features of the apple are: *red, round, 10 cm in diameter*. A Naive Bayes classifier will take these features as equally important and independent to classify the

apple. Naive Bayes Classifier uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

$$P(\text{label}|\text{features}) = \frac{P(label) * P(features|label)}{P(features)}$$

Equation 8 - Bayes Theorem

P (label) is the prior probability of a label or the likelihood that a random feature sets the label. P (features|label) is the prior probability that a given feature set is being classified as a label. P (features) is the prior probability that a given feature set is occurred.

We can use the naive Bayes classifier when each instance x is described by attribute values and where the target function *f (x)* can take on any value from some finite set V. For the instance described by the tuple of attribute values (a_1, a_2.. .a_n) and given a set of training examples of the target function, the Bayesian approach to classifying the new instance is to assign the most probable target value, V_{MAP} given the attribute values c that describe the instance.

$$V_{MAP} = \text{argmax}_{vj} P(a_1, a_2,\ldots,a_n|v_j)P(v_j)$$

Equation 9 - Equation to assign the most probable target value given the attribute values c

As said before, this type of classifier is naive because it assumes that the attribute values are independent given the target value: the probability of observing the attributes (a_1, a_2.. .a_n) is just the product of the probabilities for the individual attributes $P(a_1, a_2,\ldots,a_n|v_j) = \prod_{i=1} P(\text{ai}|\text{vj})$.

This, gives us the final equation for the Naive Bayes Classifier:

$$v_{NB} = \text{argmax}_{vj} P(a_1, a_2,\ldots,a_n|v_j) = \prod_{i=1} P(\text{ai}|\text{vj})P(\text{vj})$$

Where v_{NB} is the target value output by the naive Bayes classifier: a classifier that has been used, among all, as sentiment classifier in microblogging systems such as Twitter. (Pak, et al., 2010), (Pang B. Lee, et al., 2002), (Bhayani, et al., 2009).

Support Vector Machines

Sometimes to classify data, we need to change its representation. It is always possible to transform any set of data so that the classes within it can be separated linearly. The problem is to use a right number of dimensions and an adapted kernel.

Support Vector Machines (Cortes, et al., 1995) is one of the most popular algorithms to use this method. We will start from an easy example. We have two classes to classify and our goal is to draw a line (i.e. the *separating hyperplane*), a linear decision boundary that can visibly separate these two classes. As you can see from the figures, with the same representation of data, we can draw different lines to separate the two classes. So the question is: which is the best line among A-B-C-D?

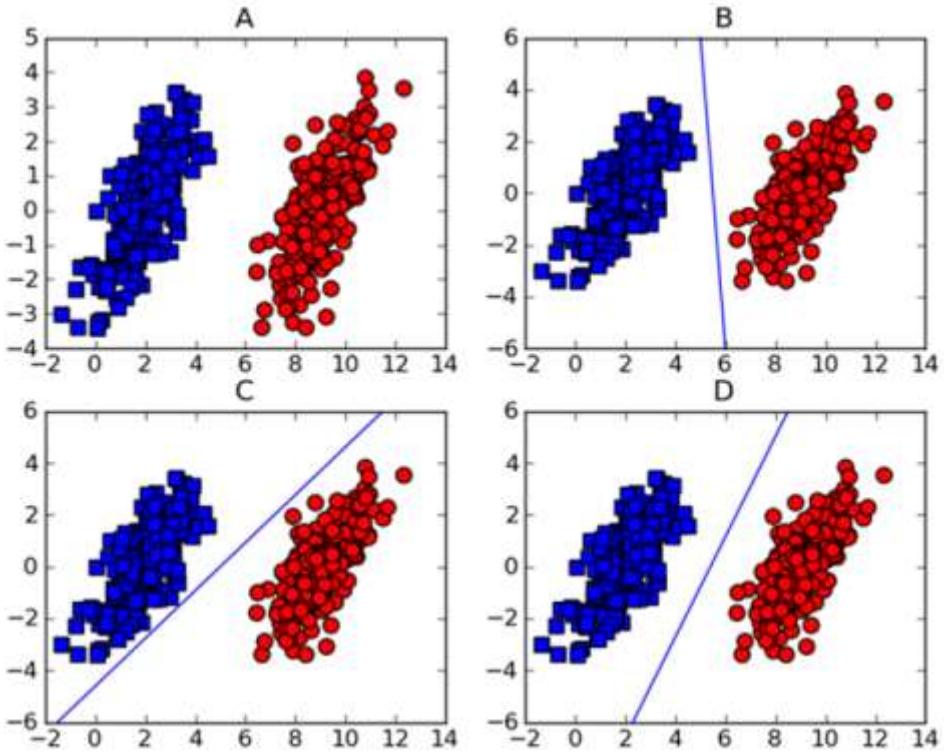

Figure 5 - Data and separating hyperplanes

In order to choose we need to know the *margin*. A margin decides the closest point (i.e. the *support vectors*) to the separating hyperplane and make sure this is as far away from the separating line as possible. Greater the margin, more robust the system. Figure 6 shows the maximized margins and the functional margin.

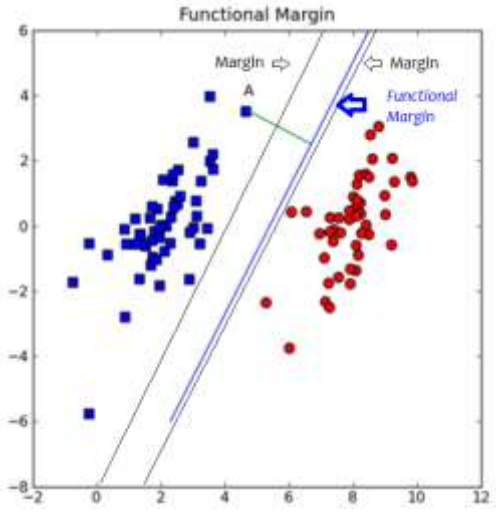

Figure 6 - Maximize margins and functional margin

The separating hyperplane has the form w^Tx+b. If we want to find the distance from a point (A) to the separating plane, we should find the perpendicular to the line w^Tx+b / $||w||$. Then we will use the SVM expecting to obtain from a function f (u) = -1 when u < 0, and f (u) = 1 otherwise.

The margin is then calculated by *label* (w^Tx+b)*. This is where the -1 and 1 class labels help out. If a point is far away from the separating plane on the positive side, then w^Tx+b will be a large positive number, and label* (w^Tx+b) will give us a large number. If it is far from the negative side and has a negative label, label* (w^Tx+b) will also give us a large positive number.

The goal now is to find the w and b values that will define our classifier. To do this, we should find the points with the smallest margin: the support vectors. Once get them, we should maximize that margin. Only the closest values to the separating hyperplane will have a label* (w^Tx+b) equal to 1.

The optimization problem we now have is a constrained optimization problem because we must find the best values, provided they meet some constraints. The constraint is that label* (w^Tx+b) should be 1.0 or greater. Using the Lagrange multipliers we can write the problem in terms of our constraints. Because our constraints are our data points, we can write the values of our hyperplane in terms of our data points.

c is a constant argument to our optimization code that has been introduced because we know that the data is not always 100% linearly separable and, as a consequence of this, we

need to introduce some slack variables to allow examples to be on the wrong side of the decision boundary. c controls weighing between our goal of making the margin large and ensuring that most of the examples have a functional margin of at least 1.0.

Sometimes we need something more to solve our problems using SVMs. This takes the name of kernel. A kernel transform our data into a form that is easily understood by the classifier. The typical example to understand the power of SVMs and kernel is this one:

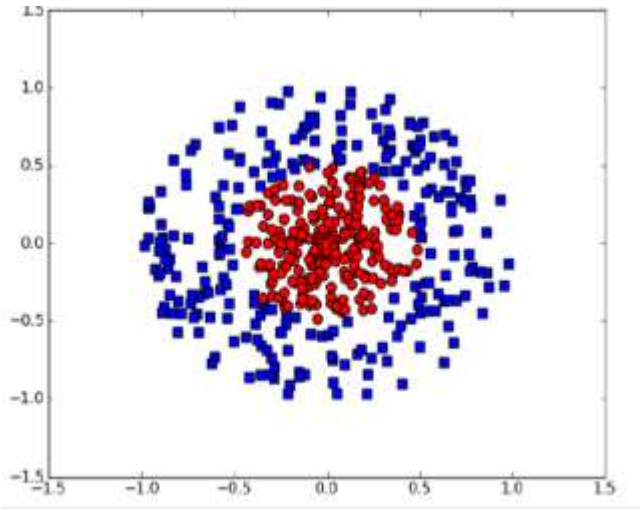

Figure 7- Non linearly separable data for Kernel Method

Now, it is even more complicated to draw a line to separate the two classes. We can change point of view: we will use more dimensions that will capture each class and solve them as a low-dimensional space problem.

In SVM optimization all operations can be written in terms of inner products. In that case we do a kernel trick: i.e. replace the inner products with our kernel functions without making simplifications. Four types of kernels for SVMs are usually exploited: Linear, Polynomial, Radial basis Function, Sigmoid. Most of the times users prefer to use automatic procedures to choose the best kernel parameters ("grid search") once they know if their problem is a linear or non-linear one.

Support Vector Machine are used in sparse data, such as the data from texts. It defines the correlation between features (words). Some works have used hybrid-SVM systems to combine several features of different types: several measures for phrases and adjectives

and, where available, knowledge of the topic of the text (Mullen, et al., 2004). SVM classifiers have been created also to detect the sentiment of messages such as tweets and SMS (message-level task) and sentiment of a term within a message - term-level task - (Mohammad, et al., 2013).

Logistic Regression

The goal of binary logistic regression is to find the best-fitting model to describe the dichotomous characteristic of interest (i.e. 0 or 1, True/False, etc.) and a set of independent variables. In a two-class case, the function will output a 0 or a 1. Unlike linear regression that can give us a result, for example, on a scale of 0-100, logistic regression can predict only binary/multi/ordinal outcomes.

To map predicted values to probabilities we can use different functions. One of them is the Heaviside step function. The step function has just one flaw: when it changes from 0 to 1 it does instantly.

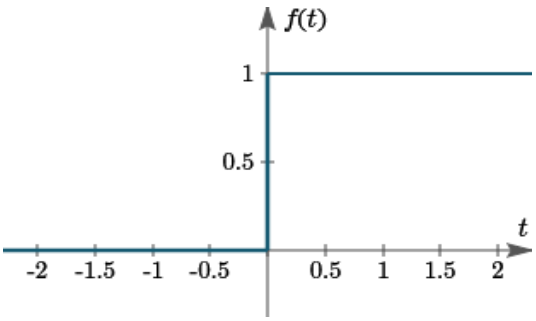

Figure 8 - The Step Function

Another function that is similar to this one is called sigmoid. Mathematically speaking, a sigmoid is given by

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Equation 10 - Sigmoid equation

At 0 the value of the sigmoid is 0.5 and it can reach the 1 for increasing values of x. On the other side, for decreasing values of x, it can approach 0. In a very large scale, a sigmoid looks like a step function.

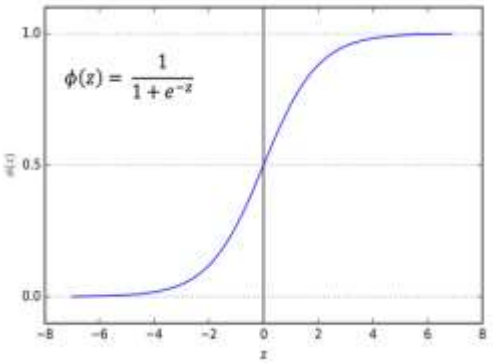

Figure 9 - A sigmoid function

For the logistic regression classifier we'll take our features and multiply each one by a weight - i.e. *regression coefficients*, and then add them up. This result will be put into the sigmoid, and we'll get a number between 0 and 1. Anything above 0.5 we'll classify as a 1, and anything below 0.5 we'll classify as a 0.

The input z to the sigmoid is given by the multiplication of two vectors and the addition of each element to get one number:

$$z = w_0x_0+w_1x_1+w_2x_2+...+w_nx_n$$

$$z = w^Tx$$

Equation 11 - input z - regression coefficients

While x is known because is the input data, we should find the best coefficients w in order to have a successful classification. We will use some optimization algorithms. In this chapter, two of the most used optimization algorithms, the gradient ascent and the stochastic gradient ascent are shown.

(Stochastic) Gradient Ascent

Gradient ascent is based on the idea that if we want to find the maximum point on a function (i.e. 1), then the best way to move is in the direction of the gradient

$$\nabla f(w) = \begin{pmatrix} \frac{\partial f\ (w)}{\partial w_1} \\ \frac{\partial f\ (w)}{\partial w_n} \end{pmatrix}$$

Equation 12 - Gradient operator

The gradient operator above will always point in the direction of the greatest increase. The step size or magnitude of movement is described as:

$$w := w + \alpha \nabla w f\ (w)$$

Equation 13 - step size

The step is repeated until we reach a stopping condition: a specified number of steps or the algorithm is within a certain tolerance margin.

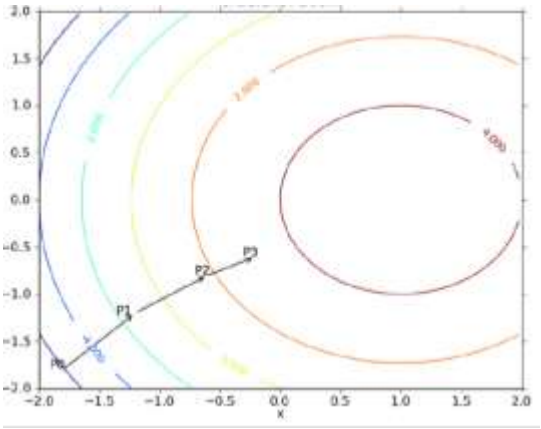

Figure 10 - Gradient Ascent

This algorithm let us find the best-fit line that better describe data. When the dataset is large, using gradient ascent can be expensive.

For this reason it can be used the stochastic version. *Stochastic gradient ascent* is an example of an online learning algorithm. This is known as online because we can incrementally update the classifier as new data comes in rather than all at once.

Deep Learning Approach

Since a decade ago, deep learning has emerged as a powerful machine learning technique exploiting the current strong computing power and the huge data available on the net. Deep learning has been inspired by the structure of the biological brain and it derives from the application of artificial neural network. It has been used in the last 10 years for computer vision applications and speech recognition (Goodfellow, et al., 2016).

How does a deep learning system work? We have several layers: the lower ones, the closest to the data input, learn simple features, while the highest layer learn more complex features derived from lower layers. It is a hierarchical architecture.

After a short presentation of the types of deep learning systems - feedforward networks (Rosenblatt, 1957), (Marvin, et al., 1969), (Grossberg, 1973), backpropagation method (Linnainmaa, 1970), (Linnainmaa, 1976), (Werbos, 1974), (Dreyfus, 1973) and the type of input features they usually use (word embeddings (Mikolov, 2013)), we will show some neural networks such as: Convolutional Neural Networks - CNN (Fukushima, 1980), (LeCun, et al., 1998), Recurrent Neural Networks and Recursive Neural Networks – RNN (Hopfield, 1982).

General structure of a Deep Learning system using Neural Networks

Neural networks can be divided into two categories: feedforward and recurrent/recursive neural networks.

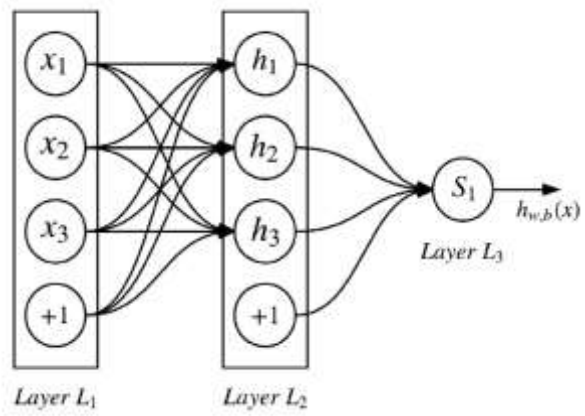

Figure 11 - The structure of a feedforward neural network

Referring to the figure above, a feedforward neural network can consist of three (or more) layers (L1, L2, L3) where L1 is the input layer, L2 is the hidden layer and L3 is the output layer. In the input layer there are vectors and an intercept term (+1). In the hidden layer there are the neurons (h1,h2,h3) called also activation functions. The activation function f can be: a sigmoid function, a tanh - hyperbolic tangent function or a ReLU - rectified linear function.

$$f(W^t x) = sigmoid(W^t x) = \frac{1}{1+\exp(-W^t x)}$$

$$f(W^t x) = \tanh(W^t x) = \frac{e^{W^t x} - e^{-W^t x}}{e^{W^t x} + e^{-W^t x}}$$

$$f(W^t x) = ReLU(W^t x) = \max(0, W^t x)$$

Equazione 14 - activation functions: sigmoid, tahn, ReLU

In the output layer there is the output vector. Each connection has a weight that controls the signal between two neurons. Weights are very important because they play a central role for the learning phase, managing the information through neurons. At the end of the training process the system will obtain a complex form of hypotheses fitting the data. The sigmoid function takes a number and transforms it to a value between 0 and 1. Sometimes its activation can easily saturate and the information would be cut. Tanh function is often more preferred because its output is zero-centered (-1, 1). ReLU function can be also implemented because when the input is less than 0, its activation function is simply thresholded at zero. In addition to this ReLU is easy to compute and faster to converge in training and yields equal/better performance than tanh.

In L3, the last layer, we can use a softmax function as the output neuron for final classification. Softmax is used to minimize result values that are less important than the greatest result value that has to be underlined.

For example given a vector V (1; 2; 3; 4; 1; 2), the softmax function will give as a result (0.024; 0.064; 0.175; **0.475**; 0.024; 0.064). The result will give a more important weight to the number 4 that has a value 20 times bigger than the smallest value 1.

To train a neural network, stochastic gradient descent via backpropagation is usually employed, to minimize the cross-entropy loss, which is a loss function for softmax output. "*Gradients of the loss function with respect to weights from the last hidden layer to the output layer are first calculated, and then gradients of the expressions with respect to weights between upper network layers are calculated recursively by applying the chain rule in a backward manner. With those gradients, the weights between layers are adjusted accordingly*". (Zhang, et al., 2018).

Word embedding

Word embeddings are used in deep learning models as input features. Words in a vocabulary are "translated" (represented) in vectors of continuous real numbers (e.g., word "hello" is represented by the vector (..., 0.11, …, 0.28, …, 0.45, …)).

Vectors may encode linguistic regularities and patterns. One word embedding system is **Word2Vec**, a neural network prediction model that learns word embeddings from text. It can use *Continuous Bag-of-Words* CBOW model (Mikolov, 2013), that predicts the target word from its context words, or Skip-Gram model that learns context words from the presence of target word.

Another system is ***Glo**bal **Ve**ctor* (Pennington, et al., 2014) that uses non-zero entries of a global word-word co-occurrence matrix (Maas, 2011) learned word embeddings that can capture both semantic and sentiment information.

Two examples of word embedding in sentiment classification are (Bespalov, 2011) that showed that an n-gram model, combined with latent representation would produce a more suitable embedding for sentiment classification, and (Labutov, et al., 2013) that re-embed existing word embeddings with logistic regression by regarding sentiment supervision of sentences as a regularization term.

Convolutional Neural Network

The Convolutional Neural Network is a feedforward neural network used for the first time in computer vision. The human visual cortex contains small and overlapping receptive fields that receive light. These fields are a filter of the input space. Convolutional Neural Network consists of multiple convolutional layers, each of which does the same as the receptive fields with light. An image is decomposed in several parts by a filter that scans the image. When the filter slides (or convolves) it produces a number for each part of the image that has been scanned. At the end of the process it produces an array of numbers, called activation/feature map. Each convolutional layer is composed by several filters. To reduce the computational complexity of the network, a subsampling layer is usually used to reduce the spatial size of the representation. Each output from the first stage becomes the input to the second stage until every possible feature is extracted. The central role of Convolutional Neural Network is that of feature extractor; one of the application is topic learning.

Convolutional Neural Network can, in fact, find a sequence of words that are interesting to distinguish topics, regardless of their position in a document.

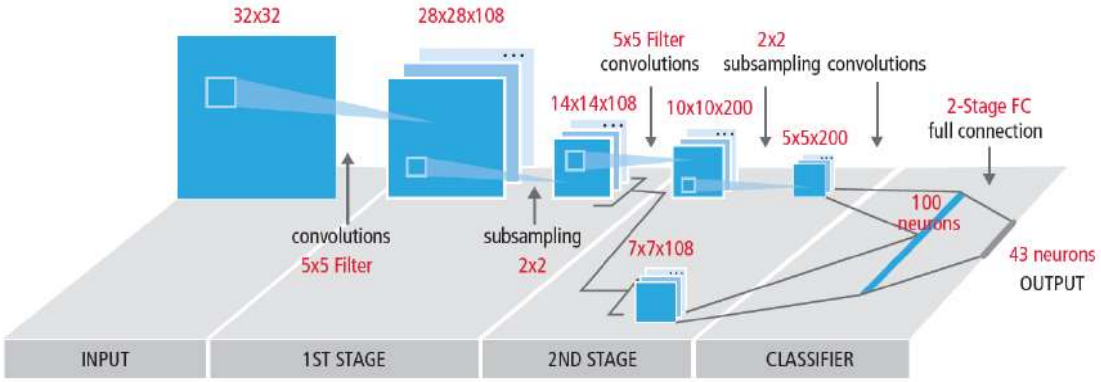

Figure 12 - An example of Convolutional Neural Network

(Wang, et al., 2016) described a joint CNN and RNN architecture for sentiment classification of short texts, which takes advantage of the coarse-grained local features generated by CNN and long-distance dependencies learned via RNN.; (Guan, 2016) employed a weakly-supervised CNN for sentence (and also aspect) level sentiment classification. It contains a two-step learning process: it first learns a sentence representation, supervised only by overall review ratings and then uses the sentence (and aspect) level labels for fine tuning.

Recurrent Neural Network

When we watch a movie, we unconsciously classify each event at every point in a movie. When we talk/read/think/listen we (normally should) understand each word based on the understanding of previous words. Each concept/image is connected to the previous ones. Traditional Neural Networks were not able to do these connections: they were unable to use reasoning about previous events to inform later ones. Recurrent Neural Networks can do this by using a structure of networks with loops in them, allowing information to persist.

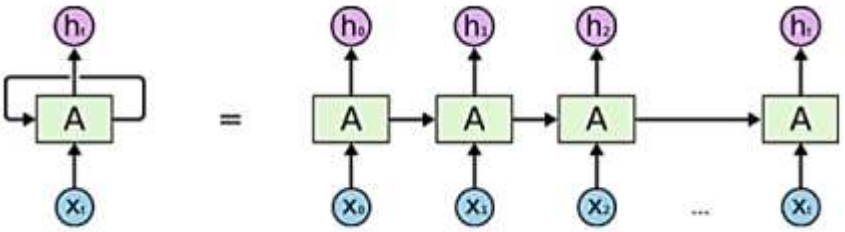

Figure 13 - A simple image of Recurrent Neural Network

Recurrent Neural Networks use their memory to process a sequence of inputs in cycle of neurons. The recurrent neural network perform the same task for every element of sequence and each output is dependent on all previous actions. In the figure we have a recurrent neural network with a left graph (unfolded network with cycles) and three folded sequence layers. One layer corresponds to a word. The number of layers is equal to the number of words of the sentence that is analysed.

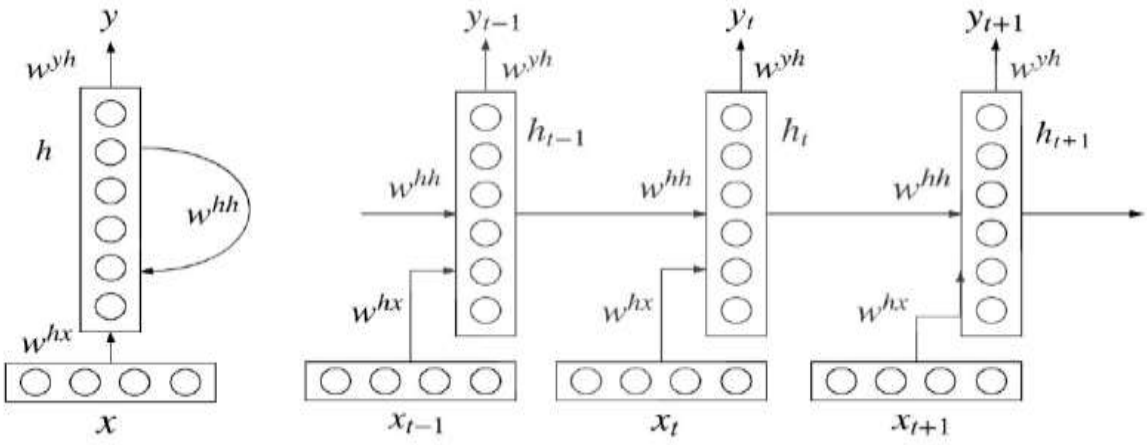

Figure 14 - Unrolled Recurrent Neural Networks

In figure 14, x_t is the input vector at time step t. h_t is the hidden state at time step t.

$$h_t = f\ (w^{hh}h_{t-1}w^{hx}x_t)$$

Equation 15 - Hidden state at time step t equation

The activation function f is usually a tanh or ReLU function. w^{hx} and w^{hh} are weight matrix for respectively the input x_t and the hidden state h_{t-1}.

y_t is the probability distribution over the vocabulary at a given time step t. If we want to predict the next word in a sentence, it would be a vector of probabilities across the word vocabulary.

$$y_t = softmax\ (w^{yt}h_t)$$

Equation 16 - the probability distribution over the vocabulary at a given time step t equation.

(Chen, et al., 2017) used a recurrent network to capture sentiment in complex contexts. (Zhao, et al., 2017) studied via recurrent network the deep semantic representation of user posted tweets and their social relationships.

Recursive Neural Network

Recursive Neural Network is used to learn a tree structure type from data. It is seen as a generalization of the recurrent neural network. Given for example a parse tree, a recursive neural network is able to generate parent representations in a bottom-up style. This type of neural network is interesting to use to produce representation for phrases, or whole sentence. The sentence level representation can then be used to make a final classification for a given input sentence (sentiment classification). (Dong, et al., 2014) used recursive neural network for sentiment classification through a specific target, exploiting context and syntactic structure. Their model uses the representation of the root node as the features, and feeds them into the softmax classifier to predict the distribution over classes.

Summary of chapter 2

This chapter showed the literature background behind this thesis.

Researchers in NLP study sentiment with different goals in mind: polarity classification, subjective vs objective analysis and multi-categorisation tasks.

The analysis concerns every layer of the language: the whole document (document level), the sentences of the document (sentence level), only some collection of words without acting subject (phrase level). Finally there is the aspect level that describes something using its attributes, and for this reason suitable for reviews about products.

After an initial phase of features extraction, many methods can be used to classify the text. The two big families of approaches are: lexicon-based one and machine learning.

For the lexicon-based approach we have talked about the manual one, the dictionary-based and the corpus-based.

For the machine learning approach we have described essentially a part of the supervised ones: decision trees and random forest, Naïve Bayes, support vector machines and logistic regression.

Finally we have analysed some deep learning methods and theories: neural networks, word embeddings, convolutional neural networks, recurrent neural networks and recursive neural network.

CHAPTER 3 - RESOURCES USED AND CREATED

Pre-existing corpus: French Sentiment Corpus

The first corpus used as a baseline is called “French Sentiment Corpus” (FSC) and it has been built originally by (Vincent, et al., 2013). The corpus is in French and it contains 14.000 texts in XML format. The texts are reviews from three websites: Allociné[4] for movie reviews; Amazon[5] for book reviews; TripAdvisor[6] for hotel reviews.

The corpus has also other important metadata such as: review ID, product ID, product name, rating, strength points, product features, and other metadata that are different according to the type of review.

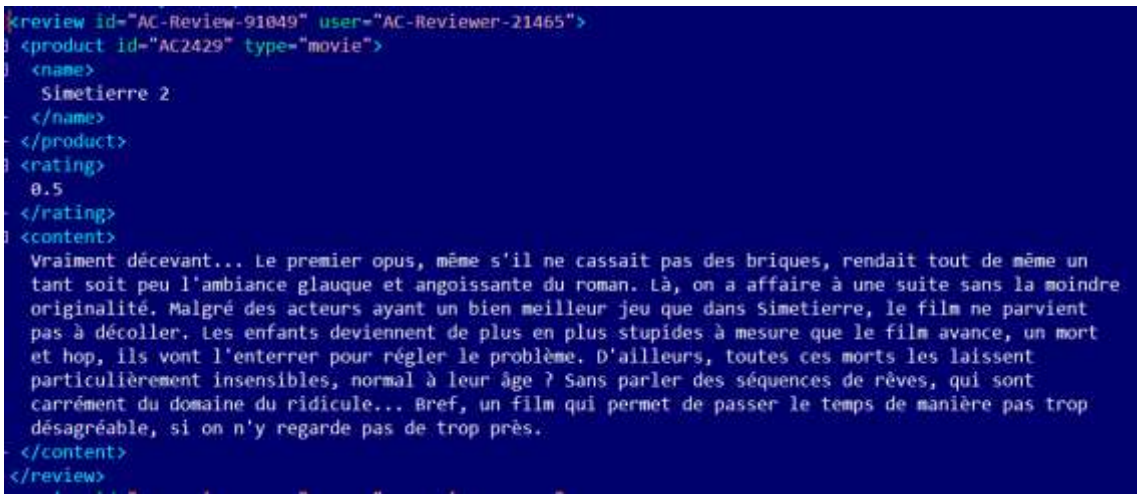

Figure 15 - An excerpt from French Sentiment Corpus

File dimension is about 12.6MB, and it is freely available from the authors upon request. Each review in the corpus is rated in both a nominal and numerical way. Rating goes from 0.5 to 5 and it progressively increases by half a point at a time.

[4] https://www.allocine.fr/
[5] https://www.amazon.fr/
[6] https://www.tripadvisor.fr/

Review type	Reviews count	Words count
Book	3999	318479
Movie	4996	418055
Hotel	5002	560085

Table 1 - Description of the French Sentiment Corpus: reviews types, total number of reviews and words count

A new corpus: New Corpus (NC)

It was interesting to extend the old corpus for multiple reasons:

A. adding more data to analyse could maybe add more variation as a consequence,
B. five years of language evolution and new operating systems with integrated spell checkers could maybe let the analysis free from many misspelled words.

For these reasons, I used a web scraper[7] to collect fresh new data every week. This web scraper can be easily installed as an extension of the browser.
According to the type of web developing technology used by the website and the data type relevant to collect, it is possible to create different sitemaps and traverse the whole website for extraction. Then the data can be exported as CSV or other file types.
In order to establish a sense of continuity, the web scraper was applied on the previous websites of the first corpus: Allociné and Amazon.
In total 4000 reviews were collected and for the 4sake of simplification we took into consideration only 2 types of ranking: the very negative (1) and the very positive (5).

Domain	Words count	Sentence count	Reviews count
Book	160388	11374	2000
Movie	172875	12045	2000

Table 2 - Description of NC Corpus: type of reviews, number of sentences and reviews

Along with this data, other 300 reviews were collected later to analyse the moderate reviews (rating 3). It is possible to find the details in the next section. The choice of the

[7] https://www.webscraper.io/

quantity of reviews is not arbitrary. Other similar works through the years used a similar number of reviews: 500 reviews for electronics products (Hu, et al., 2004), less than 4000 sentences for restaurants (Ganu, et al., 2009), 1000 reviews on various topics from the website Amazon.com (Brody, et al., 2010), 1000 sentences about movies (Thet, et al., 2010). Finally during SemEval 2014 (Pontiki, et al., 2014) the corpus of restaurant and laptop reviews used for the ABSA task featured over 3000 sentences.

Many resources to analyse reviews, but mostly in English and only for some product types.

We need data in order to enhance today's mostly statistical text classification tools with the use of linguistics. Having more data gives us more opportunities to better define and analyse what has been written. Though some annotated data have been produced in challenges as SemEval, resources are still scarce, especially for languages other than English.

Aspect Based Sentiment Analysis (ABSA) aims at "determining the orientation of sentiment expressed on each aspect" (Liu, 2012). ABSA was introduced as a shared task for the first time in SemEval-2014 (Pontiki, et al., 2014). During these years many datasets have been created in English language for laptops, restaurants, hotel (Pontiki, et al., 2015).

In SemEval-2016 (Pontiki, 2016), new and multilingual datasets were provided: restaurant reviews in six languages (English, French, Dutch, Russian, Spanish and Turkish), hotel reviews in Arabic, consumer electronics reviews in three languages (English, Dutch and Chinese), telecom reviews in Turkish and museum reviews in French. Unfortunately, nothing about movie and book domains. As far as I know, there are few works in French for these domains (Hamdan, et al., 2016). As a consequence of this, it is hard to find this kind of data.

Using ABSA on such reviews is a difficult task because the opinion expressed is complex and has various forms. Unlike other kind of reviews which are limited in total amount of usable characters, these reviews are non-predictable in terms of length. In addition to this, they may carry opinions about other products related to the reviewed one which are used as comparison. They can also merge in a same paragraph user's opinion and description of the evaluated product. That's why we created a dataset for ABSA for movies and books that has given as a result the paper (Pecòre, et al., 2018).

Datasets for Aspect-Based Sentiment Analysis: ABSA 2018

This dataset is a sub-dataset of the two corpora NC and FSC. In total it is composed of 1800 reviews, i.e. 4113 sentences for the book domain and 5222 for the movie domain. By taking into account words and line counts of similar projects, it is possible to state that the dataset used in this project is suitable for being used in the context of Aspect-Based Sentiment Analysis tasks. Compared to other corpora in the field, the one used in this project has a higher-than-average number of sentences. Some examples are (Hamdan, Bellot et al. 2016) with 200 books reviews in French, (Alvarez-Lopez, Fernandez-Gavilanes et al. 2017) with 2977 sentences from English books reviews, (Thet, Na et al. 2010) with 1000 sentences from English movies reviews, (Sorgente, Vettigli et al. 2014) with 2648 sentences from movies reviews in Italian.

The corpus deals with books and movies using three types of rating: 1 for extremely negative reviews, 3 for the moderate ones, and 5 for the extremely positive ones. We decided to include 3 stars rating reviews because we noticed that, in such reviews, reviewers do not express a strong opinion: they are therefore induced to justify their balanced opinion to precise what they consider to be the negative and positive aspects of the book or movie.

Dataset statistics

Each part of the dataset is divided per rating and it contains 300 reviews (300 reviews for each domain and rating - M1, M3, M5, B1, B3, B5). The total number of words is 169333, distributed over 9335 sentences.

Corpus	Words count Rating 1	Words count. Rating 3	Words count Rating 5
Movie	31595	42350	22658
Book	29345	20991	22394
Total	60940	63341	45052

Table 3 – Words count per domain and rating

Statistical data show that the number of words of the reviews can vary very easily (some reviews have more than 2000 words). From the table it is evident that there is a great variance on words, a great dispersion of the values and an asymmetrical distribution of data.

	Mean	Variance	Median	Min	MAX
Movie	103.02	139.64	60	1	2094
Book	81.20	118.74	48	12	2052

Table 4 -Statistics of the Corpus used for Aspect-Based Sentiment Analysis

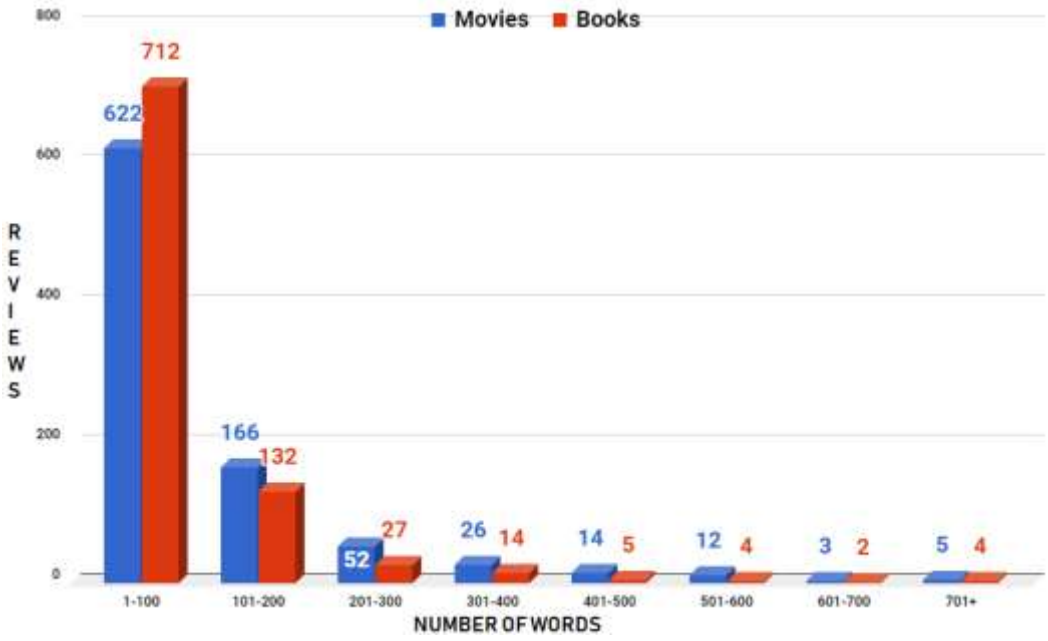

Figure 16 - Histogram showing that most of the reviews are composed by less than 200 words

Most of the reviews contain less than 200 words (Figure 16).

A box plot used to analyse the data (see Figure 17 - Boxplot for Movie reviews and Figure 18 - Boxplot for Book reviews) reveals that, for the movies domain, reviews with more than 254 words are to be considered outliers, and the average amount of words per review is 127.5. (Figure 17)

Figure 17 - Boxplot for Movie reviews

On the other hand, for the books domain, reviews with more than 181 words are to be considered outliers, and the average amount of words per review is 96.5. (Figure 18)

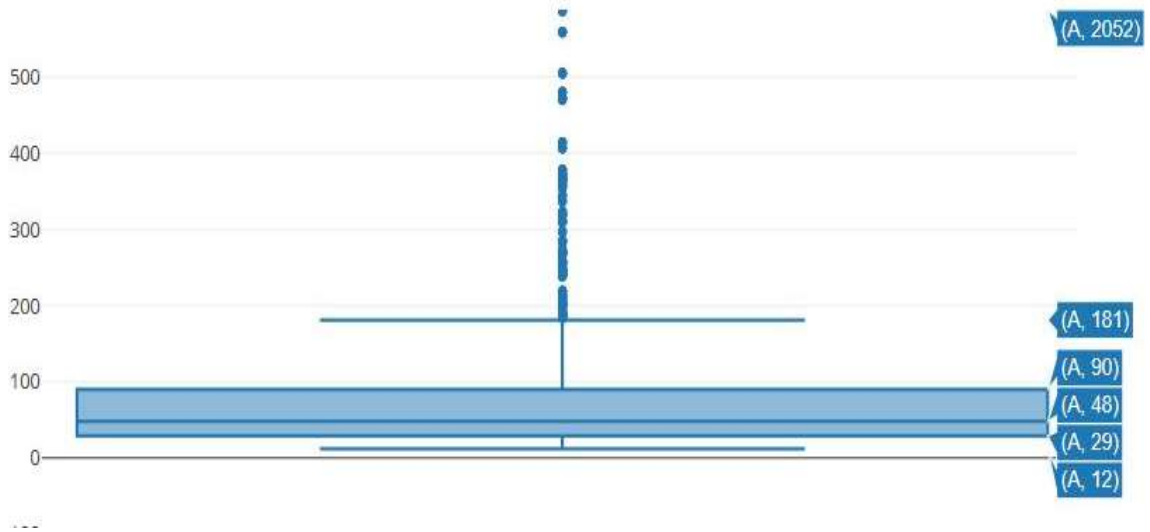

Figure 18 - Boxplot for Book reviews

Thanks to a closer study of the tokens (total number of words) and types (total number of distinct words) in reviews, it is possible to observe that even if movies reviews are generally longer than books reviews, they are less dense. The type/token ratio for book reviews is usually higher than the one for movie reviews. As a consequence of this we can state that probably book reviewers are more creative in language expression than movie reviewers, who tend instead to write longer reviews - especially in a three-star review (M3 has 43289 tokens, compared with negative reviews - 31595 tokens, and positive - 22658 tokens).

Domain	Rating	Token	Type	Ratio
Movie	1	31595	6624	20.9%
	3	43289	8244	19%
	5	22658	5443	24%
Book	1	29345	7492	**25.5%**
	3	21002	4916	**23.4%**
	5	22394	5895	**26.3%**

Table 5 - Token, Type and Ratio Type/Token for each domain and rating

Dataset annotation

Annotation is useful and necessary to evaluate system to study (1) vocabulary used to express a negative or positive opinion and (2) aspects related to a given positive or negative sentiment.

Our annotation tries to be the most general possible concerning the various book or movie involved, regardless of genre. Nevertheless, the annotation scheme has not been specifically conceived for e-books because none of them are present in our dataset and we did not find many big differences between book and e-book reviews.

Annotation scheme and guidelines

Though annotation and classification may be viewed as very precise, it is easy to group classes, depending on the expected use of the corpus.

The annotation scheme is composed of aspects and attributes. For the book domain we have 5 aspects and 19 attributes. The annotation scheme for book domain has been created following GoodReads[8] and LibraryThing[9] review schemas.

Aspects	Attributes
General Feeling	-
Text	*General, Subject, Style, Characters, Pace/Narration, Interest/Accuracy, Translation/Adaptation, Readability*
Illustration	*General, Interest/Accuracy, Graphic quality*
Author	*General, Text Author, Translator, Illustration Author,*
Form	*General, Bookbinding, Typography, Inner structure, Distribution*

Table 6 - Annotation scheme for Book reviews

For the movie one: 7 aspects and 28 attributes. The annotation scheme for movie domain has been created following InternetMovieDataBase[10] review schema.

[8] https://www.goodreads.com/
[9] https://www.librarything.com/
[10] https://www.imdb.com

Aspects	Attributes
General Feeling	-
Direction	*General, Director, Point of view, Direction of actors, Shooting, Sound recording*
Acting	*General, Actor, Stuntman*
Visual	*General, Sets, Costumes, Special FX*
Sound	*General, Music, Sound Effect, Songwriter*
Script	*General, Subject, Plot, Dialogues, Characters, Pace/Narration, Remake/Adaptation/Reboot*
Distribution	*General, Type of Data Storage, Original Version/French Version*

Table 7 - Annotation scheme for Movie reviews

Opinion annotation

Each opinion expression is annotated in three steps.

1. The first step is to select a group of contiguous words that indicate a positive or negative opinion. Opinion is evaluated by an ordinal value: -1 or -2 for a negative sentiment, according to its intensity; 1 or 2 if the opinion is positive.
2. The second step is to detect the entity to which the opinion is reported. This entity is not always expressed, especially if it is the movie or book that is evaluated. When it is expressed, it is most of the time a name or a nominal group. Since including co-reference resolution is beyond the scope of this work, pronouns are not selected as entities. Whenever opinion expression refers to a pronoun, the entity is reported to its previous closest reference. If the entity is detected, a relation is created, which joins opinion expression with entity phrase.
3. In the third step, an aspect and an attribute are chosen in the annotation scheme.

Some examples are:

- **c'est un navet** (eng. *it's a rubbishy movie*) the word navet (eng. *rubbishy movie*) indicates a very negative sentiment (value: -2) and refers to the entity at the same time. The aspect, given by the entity is General Feeling.

- **Le style est très agrèable** (eng. *it's a very pleasant style*), extracted from the book corpus, very pleasant indicates a very positive sentiment (value: 2). The entity is style. The aspect is Text#Style.
- A very negative book review such as **la bobo au style frelaté** (eng. *the bobo with degenerated style*), degenerated refers to a very negative opinion (-2). It can be reported to the entity Style and classified in Text#Style. Because of the reference to the style, one can say that bobo refers to the author; like in *un navet*, *la bobo* expresses in a single word the entity and the opinion of the reviewer.

Previous examples, though being very simple, show how entities, opinion phrases and contexts should be combined to determine the aspect to which they have to be reported. The complexity of these expressions makes it difficult to allocate aspects only to entities, as it is classically done, for example in SemEval 2016 annotation (Apidianaki, 2016).

Entities related to other products

Some phrases indicate a positive or negative sentiment related to another book or movie, most of the time to be compared with the reviewed one.

> ***"Rien à voir avec le seigneur des anneaux, carrément passionnant »***
> ***[eng. (this film has) nothing to do with the Lord of the rings, (that is) downright fascinating]***

the phrase downright fascinating indicates a very positive opinion, but it is applied to another movie: (the Lord of the rings). On the contrary, the full sentence indicates a negative feeling about the movie.

Comparisons inside a review are frequent and can be a problem for automatic opinion detection if it is not possible to distinguish the reviewed product from a comparison with another one; that is why we wanted the possibility for the annotation to report them precisely.

To cope with the problem, the annotation of the entities indicates whether they are or not related to the evaluated product, a product of the same series, another product, etc. So, in the previous example, the very positive phrase "downright fascinating" is reported to "the

Lord of the rings" classified as an entity which refers to another product. The phrase "(this film has) nothing to do with the Lord of the rings" is annotated as a negative opinion, reported to an entity which refers to the evaluated product.

Annotation Process

The annotation has been done by two experts: a native speaker and a non-native speaker. After the choice of the annotation form and the redaction of the guidelines, experiments have been conducted to estimate inter-annotator agreement. The most difficult task was the selection of the phrases related to an opinion, with particular attention to the determination of their scope.

For word selection, Cohen's K (Cohen, 1960) was equal to 0.71, an acceptable result given the difficulty and subjectivity of the task. However, to improve the reliability of the corpus, we decided to perform a cross-reading of the annotations between the two annotators.

The annotation was performed via Glozz software (Widlocher, et al., 2012). Glozz is a multi-purpose text annotation tool, which comes with a full WYSIWYG interface. It makes it possible to create units, defined as contiguous spans of text and relations between them. Annotations may be exported in several file formats and especially as SQL data.

Annotation Results

We annotated 5001 opinion phrases on movies (M1, M3, M5) and 3274 on books (B1, B3, B5). Annotations on negative reviews outnumber annotations on positive reviews, with circa 2899 on M1 and B1 corpora against around 1992 annotations on M5 and B5 corpora.

Figure 19 shows how annotations are distributed between the main classes. Nearly half of annotations are classified as General Feeling in both corpora.

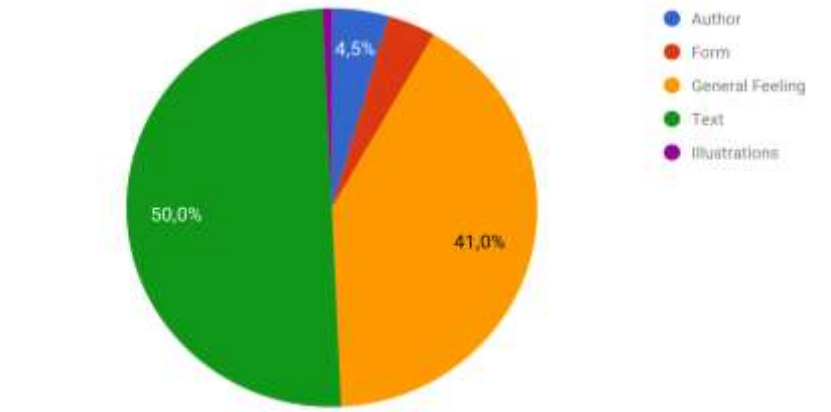

Figure 19 - Annotation distribution for Book reviews

In the book corpus, the most important six classes related to specific aspects are Text#Interest and Accuracy, Text#Pace-Narration, Text#Style, Text#Characters, Text#Readability and Text#Subject. All of them are related to the aspect Text and they collect 44.7% of the annotations not classified as General Feeling. In the very wide variety of the assessed books, the textual aspect represents therefore a very large majority, with great importance given to the interest in the content.

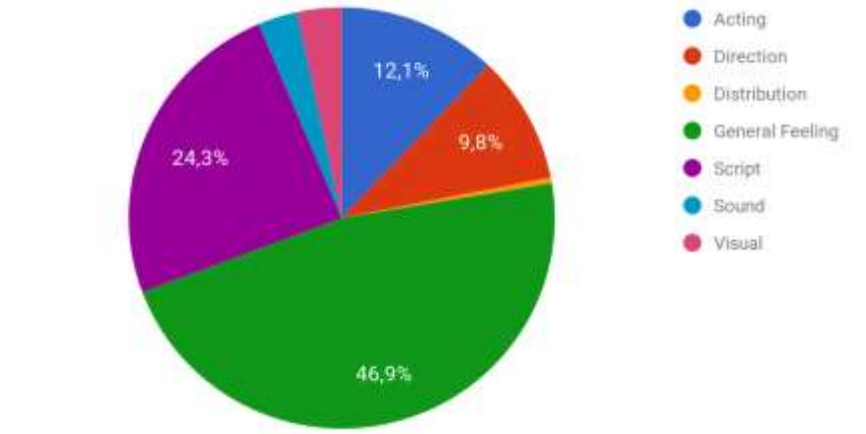

Figure 20 - Annotation distribution for Movie reviews

In the movie corpus (Figure 20), the most important five classes related to specific aspects are Acting#General, Script#Plot, Direction#General, Script#Pace/Narration and Script#General. They comprise 47.6% of the annotations not classified as General Feeling. Apart from Distribution, all the aspects collect a significant number of annotations: the

cinema is a multimodal media, which combines sound and images to tell a story played by actors.

Entities analysis

6392 entities have been selected; 5693 of them are related to the book or movie on which the review was written; this means that around 5% of the opinion expressions are related to another book or movie: a little more than 3.8% for the movie corpus and slightly less than 6.9% for the book corpus. Since most of them are very different from the general opinion of the reviewed book or movie, it is interesting to detect them.

	Relations no.	Aspects no.	Entities no.
Movie	5001	6235	3903
Book	3274	4787	2489

Table 8 - Summary concerning the number and type of annotations: relations, aspects and entities

Differences between corpora and domains

A. There is a difference between the old corpus (from now FSC) and the new corpus (from now NC) in terms of misspelled words. Analysing the misspelled words of a sample of 10.000 words, FSC showed an error rate of 7.2% (ratio between misspelled words and total number of words), whereas NC shows only 3.4%. Even though It is not possible to determine whether the cause is the operating system or social factors, there is a huge difference between the two corpora.
B. Another interesting point: while collecting texts for the corpus, I noticed that collecting negative books reviews was harder than the positive ones, because there was an overwhelmingly higher amount of positive reviews compared with the negative ones. Maybe unlike people reviewing movies, people who don't like the book that they are reading, prefer to close it and not to write a review - unless they have read other work by the author and they want to criticize the author's choices compared to the previous ones. Unfortunately there's no real scientific evidence. We can support the idea that Amazon

prefers to show first the books with the best reviews, then the books with the worst reviews. It is difficult to evaluate due to the huge dimension of Amazon database. For example, if we take a sample of the first 50 random books using some filters\parameters to balance the results, such as: “publication before 2017”, “rising price”, number of reviews more than 20, and no specific average rating, the rating distribution is

   a. for rating 1 - between 0% (no reviews) and 16%
   b. for rating 5 - between 53% and 78%

C. In terms of the linguistic point of view, we can observe that, more than in movies reviews, the book genre constitutes a linguistics sub-domain of books, i.e. user’s style changes based on the genre of the book that is reviewed. Therefore it has been important to collect reviews of different types of books in order to have a more complete annotation.

D. Books’ reviewers seem to make fewer misspellings than movies’ reviewers. Using a sample of 200 reviews/10.000 words, books’ reviewers show an average number of about 500 misspellings, many of them concerning missing capital letters at the beginning of the sentences and foreign names. Movies’ reviewers show, at the same conditions, an average number of about 800 misspellings.

E. It seems that there is a difference between the diversity of words used in movies and books reviews, even when they share the same topic. The difference is to be found in the fact that, in movies reviews, users do an evaluation by just labelling the movies as good or bad (following a description of the several aspects of the product) and by using very simple sentence structures (i.e. “*I’ve seen a movie*” versus “*the writer’s talent lets us perceive*”). In books reviews, instead, users seem to develop a critique that uses a variety of context-specific and refined words, with more complex syntactical features. In the future it would be interesting to investigate this observation too. Let’s take an example, comparing two reviews of the same type, but from the two different sources:

a. from a **movie** review: "*Ben moi j'ai vu le film aujourd'hui à 16h à Torcy (...) Eh ben j'ai pas été *déçut (...) C'est juste un beau film, j'ai aimé les images magnifiques, les décors somptueux, rien que pour ça déjà j'ai trop kiffé. (...)*Pourkoi cette hostilité alors qu'il est bien filmé, *je vois pas *ou il s'est senti trahi??? bon voilà, c'est que mon avis, pas une *kritic mytho. Ey puis y avait la musique *vraimant belle. j'ai passé un bon moment et faut que j'y retourne avec ma mère *parceuq'lle voulait pas y aller a cause des polémiques. Moi je trouve que ça parle *vraiament de ce *ke se passe dans les quartiers.*"
b. from a **book** review: *"L'habileté de l'écrivain nous permet de voir la société dans laquelle nous vivons, avec ses excès individualistes (...). N'importe quel peuple qui baisse les bras et cède à la facilité par lâcheté (...), remords historiques et repentances incessantes finit par être chassé de ses terres. Cela n'a rien de surprenant en soi (...) seulement il y a une sorte de couvercle posé sur l'Islam en France, parce qu'il est le résultat d'une politique d'immigration (...). Ce livre n'est pas un pamphlet de la haine (...) Ceux qui y voient une incitation à la haine sont les mêmes qui, au nom de la paix sociale, acceptent les milliers de viols, vols, trafics et agressions que subissent depuis trop longtemps les Français (es) honnêtes.*"

We can notice that the movie review shows several misspellings (*déçut [for déçu - disappointed], *ou [for où - where], *vraimant — *vraiament [for vraiment - really], *kritic [for critique - critics], *parceuq'lle [for parce qu'elle - because she]), slang (kiffé [love],*Pourkoi [for Pourquoi - why],*je vois pas [for je ne vois pas - I don't see (the reasons why...)], *ke [for que - what]) and a general vocabulary to evaluate a movie and not specifically designed for the topic (un beau film, j'ai aimé les images magnifiques, les décors somptueux (...) y avait la musique vraimant belle [a good film, I loved the wonderful images, the gorgeous sets (...) the music was really good]).

On the other hand in book reviews we find an evaluation about the strengths and weaknesses of the reviewed object (L'habileté de l'écrivain nous permet de voir la société

dans laquelle nous vivons [eng. *the writer's talent lets us perceive the society in which we live*]).

In addition to this, the reviewer usually develops a critique about the topic of the book using more context-related expressions:

- excès individualistes [eng. *individualistic excess*],
- facilité par lâcheté [eng. *facilitated by cowardice*],
- remords historiques [eng. *historical remorses*],
- repentances incessantes [eng. *continuous repentances*],
- *ê*tre chassé de ses terres [eng. *to be chased out of his lands*],
- politique d'immigration [eng. *immigration policy*],
- pamphlet de la haine [eng. *pamphlet of hatred*],
- paix sociale [eng. *social peace*],
- viols, vols, trafics et agressions [eng. *rape, robbery, traffic and assault*])

Lexicons

Pre-existing Lexicons

ValEmo (Syssau, et al., 2005) lexicon

ValEmo is a psychological standard and study before being a lexicon. The two authors and researchers created this lexicon to study the emotional valence of 604 words. The study involved 600 people (484 women and 116 men) aged between 18 and 26 and it was divided into two parts.

In the first part the valence was judged using nominal values: positive, negative and neutral. In the second part, valence was measured using an ordinal scale from -5 to 5.

The list of words implied different parts-of-speech: verbs, nouns, adjectives, adverbs, etc. In addition to this, some words from two more lexicons (Ferrand, et al., 1998), (Ferrand, 2001) were used, creating a mix of abstract and concrete words.

F-POL (Vincze, et al., 2011) lexicon

In 2011 Bestgen & Vincze exploited several lexicons to study co-occurrences in texts. They created a system to expand some already existing lexicons concerning different linguistic aspects: concreteness - F-ABS (Hogenraad, et al., 1981), abstraction - F-IMA, polarity - F-POL (Hogenraad, 1995) arousal (F-ACT), emotion (F-EMO). During the first experiments we used only the F-POL lexicon. This lexicon implied 3.252 words using a 7 scale rating system: from *very unpleasant* (1) to *very pleasant* (7).

SyssauBestgen lexicon (SB)

This emotion lexicon has been created from merging F-POL and ValEmo lexicons. More details in Chapter 4 - Experimentations, in the section called Lexicons correlation.

New lexicons

Lexicon (Lex)

Lex is a handmade opinion lexicon, with more than 2800 words, including: 1500 nouns, 950 adjectives, 400 verbs. Words were chosen according to the presence of a term of endearment or of disparagement. The annotation used a -5/5 scale. The lexicon is free of use under request.

Regression Logistique Lexicon (RLex)

This lexicon is composed by 283 words considered discriminant by logistic regression to classify reviews as positive or negative. The logistic regression has been applied using the *glmnet* package of R. For each word, a score was associated. This score has then been normalized following the same scale for the other lexicons (-5/5).

Different nature of the lexicons and overall coverage

The interesting point of these experiments was testing different types of lexicon corresponding to different linguistic points of view of the same word. In other words it was possible to test whether for certain type of reviews one can distinguish the positive from the negative ones or not, using lexicons that measure different linguistic aspects. This was useful to choose the lexicon that was better designed for opinion analysis.

Table 9 shows that the three lexicons cover two different linguistic aspects (opinion and emotion) having different percentage of coverage with the corpus.

We expected that SB would be the most useful when classifying the reviews due to its coverage rate.

	RLex	Lex	SB
Nature	Statistics	Opinion	Emotion
Words no.	283	2850	3079
Use of lexicon	100%	68%	82.2%
Scale	-5/5		

Table 9 - Number of words per lexicon and percentage of lexicon present in the corpus

Table 10 shows shared words among lexicons. It is normal to observe that there are some words that are shared among lexicons. This is due to the fact that, as it has been said before, a word can have different interpretations and nuances. This is one of the reasons why it is important to experiment different lexicons that study different linguistic aspects, even though the problem we aim to solve is always the same.

	RLex + Lex		RLex + SB		Lex + SB		RLex + Lex + SB		
Shared words among lexicons	44% (RLex)	4.3% (Lex)	61.5% (RLex)	5.6% (SB)	34.6% (Lex)	32.7% (SB)	30% (RLex)	3% (Lex)	2.82% (SB)

Table 10 - Percentage of shared words among lexicons

Software

Misspelled words (Language Tool)

As it has been already pointed out, many words were spelled in the wrong way. Some examples are:

- oeuvre (œuvre),
- drole (drôle),
- extraordinnaire (extraordinaire).

It was important to correct (unfortunately not all) some of them because during the pre-processing of the text, Treetagger (the POS tagger used for this project) did not recognise many words.

In 2018 there are still many problems with correcting misspelled words in French. Many famous software and recent text editors (MS Word, OO Writer, Grammarly, etc.) are not able to recognise and/or correct French misspelled words.

Anyway, during the thesis I joined for a short time the LanguageTool spellchecker group[11] in order to improve their spell checker for French language and use it for my corpus.

LanguageTool uses Hunspell[12] engine and it is the official spell checker of *LibreOffice*, *OpenOffice.org*, *Mozilla Firefox, Thunderbird*, *Google Chrome*. It is also used by proprietary software packages, such as: *macOS*, *Adobe InDesign*, *memoQ*, *Opera*, and *SDL Trados*. It is written in Java and C.

At that time LanguageTool was able to recognise some misspelled words but it was not able to correct them. The problem, according to the developers, was due to the fact that the tool lacked of dictionary files and maintenance. Dictionary files require usually a lot of work for development and maintenance. I decided to give my contribution for French language.

There are many steps to follow during the creation and the maintenance of the files in order to activate suggestions in the spell checker.

Without entering the details of the java code itself (that is by now available in the original forum of the tool), these are in general the big steps to follow:

- to find different versions of a word
- to find possible misspelled variations
- to transform raw text in a dictionary file - a binary file that the tool should be able to read
- finally, to create some XML rules to recognise the part of the text to be considered as wrong - especially when there are compound words and expressions.

[11] https://languagetool.org/dev
[12] http://hunspell.github.io/

This being a time-consuming operation, I couldn't participate actively to the maintenance, but I modified the source file (with the creators' consensus) to correct many words of the corpus. In particular I took care of the corrections of some fundamental words for classification purposes, such as *chef d'oeuvre (*in its variants **chefdoeuvre*, **chef-d-oeuvre*, **chef d oeuvre*, **chef doeuvre*, etc.), some verbal group (such as **jem, *jaime, *jaim, *jkiff, *jador, *jadore)*, some accents and diacritical signs (**plutot, *maitre*).

Other words such as people names, place names or particular words were not directly corrected by LanguageTool even though it detected them as misspelled.

Speaking of numbers: for FrenchSentiment Corpus, LanguageTool corrected 48.000 words over 423.000.

Linguistic dataset: special words, patterns and verb tenses

Linguistic analysis of words and patterns allows to understand user's stances and enhance the shallow parser in its syntactic analysis.

The analysis concerns conjunctions such as:

- plutôt (eng. *rather*),
- mais (eng. *but*),
- bien que (eng. *although*),
- and conditional moods and tenses.

Plutôt (eng. rather)

- It has been encountered more than 200 times and more in negative reviews compared with the positive ones: (N=74%, P=27%);
- It is generally used to intensify the expression found just after the conjunction:
    - "Il est plutôt bien écrit" / "Il est plutôt terre à terre" / "Il est plutôt facile à lire" / "Je suis plutôt bon public"; (eng. *Is is rather written well / It is rather down-to-earth / It is rather easy to read / I am rather good public*)

- Negative – to make a suggestion (expressed by imperative tense) or express a regret (conditional past tense or Je vais + plutôt + infinitive) and make a comparison with another product of the same type, or somehow linked to the object of the review:
    - "allez plutôt voir le dernier Disney" / "achetez plutôt le livre" / "J'aurais plutôt lu le roman" / "Je vais plutôt lire". (eng. *Rather watch the last Disney movie / rather buy the book / I would rather have read the novel / I am rather going to read...)*
    - This comparison serves to expose the flaws of the reviewed product;
- Negative - In a non-defining relative clause, such as "Les acteurs, plutôt fantoches" (eng. *The actors, (I'd) rather (say) puppets)*, it expresses a negative opinion on the word just before the conjunction;
- Negative/Positive - it expresses two qualities of the same product. In this case, the first is limited in its intensity, while the second defines, in a positive or negative way, the reviewed product: "plutôt chiante mais utile" / "plutôt mignon, mais néanmoins un peu facile" (eng. *Rather boring but useful / rather cute, but nevertheless a little bit easy)*;
- Neutral – It introduces a correction or insists on the preferred choice of one term as opposed to another: "un film ou plutôt un reportage"/ "un film ou plutôt un documentaire" (eng. *A movie or rather a report / movie or rather a documentary)*;

Bien que (eng. although)

- It has been encountered more than 1500 times in the corpus
- Main problem: the system doesn't distinguish the relative clause from the conjunction:
    - "on voit bien qu'il y a une correction" / "Bien que l'idée de départ pouvait être sympa, elle n'est pas originale" (eng. *We see very well that there is a correction / although the initial idea could be nice, it is not original*)
- Regularities have been searched in order to solve the problem:

- “Bien que” can be found at the beginning of a sentence, or in a subordinate clause after a comma
- In a subordinate clause it is introduced by c’est + adjective: “c’est lourd bien qu’il (…) ” (eng. although (...) *it’s also heavy*)
- Like “plutôt” it expresses two qualities of the same product. In this case, the first is limited in its intensity, while the second defines in a positive or negative way the reviewed product. Usually in this case its structure is noun + bien que + adjective:
  - "on essaie après de nous faire croire que l'homme bien que viril et misogyne au possible sait de temps à autre être romantique" / "Le récit bien que dramatique est teinté d’un humour". (eng. *they try later to persuade us that the man although virile and as misogynous as possible knows how to be romantic from time to time / the narrative, although dramatic, is coloured with humour.*)

Mais (eng. but)

- It has been encountered more than 2300 times in the corpus
- The word “mais” carries within itself a tricky problem. Depending on its position in the sentence, but also on the context, its meaning can change. For this reason it is short-sighted to merely analyze it as we did with the others.
- Based on its position we can infer that:
  - Between two adjectives (it has been encountered 60% of times), just as plutôt and bien que, it expresses two qualities of the same product: “Il est simple, mais intense” (eng. it is simple, but intense);
  - At the beginning of exclamative or interrogative clauses (it has been encountered 5% of times) it is used to highlight impatience or surprise: “Mais je vous en prie !” / “Mais vous l’avez vu ?” / “Mais où ?” / “Mais Pourquoi ?” / “Mais qu’est-ce que ça ?” (eng. *But… Please! / Did you see it?! / But where? / but why? / but what is that?*).

- Based on the context we can infer that mais can be similar to other conjunctions:
    - Mais aussi / mais également (it has been encountered 5% of times)= Et - "On en ressort vidé et un brin fataliste, mais également plein d'espoir"; (eng. *but also = and - We resulted as empty as well as a little fatalist, but also hopeful*)
    - Mais bien (it has been encountered 2% of times) = plutôt - It introduces a correction or insists on the preferred choice of one term as opposed to another with the exception of mais bien + past participle tense: "mais bien écrit"/ "mais bien joué"; (eng. *but well = rather - written rather well / played rather well*)
    - Mais bon (it has been encountered 4% of times) = expression of discontent - The reviewer is judging the product with discontent, she/he is not convinced about the quality of the product: "A la base, j'avoue ne pas être très fan de la Saga Camping. Mais bon, j'ai essayé de faire abstraction de ce fait et j'ai tenté de regarder le film"; (eng. Well, - *First, I admit not to be a very good fan of the Camping Saga. But well, I tried to disregard this fact and I tried to watch the movie*)
    - Mais Surtout / Mais Comme (it has been encountered 5% of times) = Surtout / Comme: "non seulement on aura un cinéma nul, mais surtout les spectateurs se contenteront d'aller voir des films américains". (eng. *But Especially / How = Especially / How: Not only we will have a worthless cinema, but especially the spectators will settle for going watching American movies*)
- In a negative context:
    - Mais alors (it has been encountered 3% of times): "L'humour est vraiment très passable mais alors le pire les messages d'une finesse pachydermique, teinté d'homophobie et de sexisme" (eng. *Then: "the humor is really just passable but (if that's true) then the worst are the messages of an elephantine sharpness, tainted with homophobia and with sexism)*

- Mais après (it has been encountered 1% of times): "le réalisateur s'est donné aussi un rôle pas le plus intéressant mais après tout qu'importe vu qu'aucuns des personnages ne l'est !!!" (eng. *But after: The director gave himself a role that's not the most interesting but after all what matters, seen that none of the characters is interesting at all!!!)*
- Mais autant (it has been encountered 1% of times): "On peut y trouver ce que l'on veut. Mais autant lire un mauvais horoscope dans ce cas. Très déçu." (eng. *But as far as: We can find on it whatever we want. But that's as far as we can just read a bad horoscope in this case. Very disappointed)*
- Mais certainement pas (it has been encountered 1% of times): "Je ne sais pas qui gagne sur ce coup-là, mais certainement pas ceux qui ont payé leurs places!" (Eng. But *certainly not: I do not know who wins in that case, but certainly not those who paid their places!)*
- Mais comment…: "Jamie Dornan et Dakota Johnson, que j'aime beaucoup en plus, font ce qu'ils peuvent, mais comment être à l'aise sur un tournage pour interpréter cela après si peu de temps où l'on connaît son partenaire? le résultat est forcément décevant." (eng. *But how?: Jamie Dornan and Dakota Johnson - that moreover I like very much - make what they can, but how can they feel at ease on a shooting after such little time knowing their partner? The result is necessarily disappointing.)*
- Mais franchement / honnêtement (it has been encountered 2% of times): "Alors, comme dans toute comédie, on arrive toujours à trouver une situation, un gag, ou une réplique un peu drôle, mais franchement devoir se taper cet empilement de débilités pendant une heure trente pour rire ou sourire une fois ou deux, c'est quand même vachement maigre."' (eng. *But frankly/honestly: So, as in any comedy, we always manage to find a situation, a gag, or a reply (that's) a bit funny, but frankly having to fight against this pile of shallowness for*

one hour and thirty, to laugh or smile maybe once or twice, it definitely too little of a compensation .)

Conditional mood and tense

As the word "mood" suggests, a mood is a way of using a verb to show the attitude of the speaker toward what he is saying. The *indicative mood* expresses facts as well as the conditional mood is "*used to express a proposition whose validity is dependent on some condition, possibly counterfactual. It thus refers to a distinct verb form that expresses a hypothetical state of affairs, or an uncertain event, that is contingent on another set of circumstances*". (Wikipedia: conditional mood[13]).

In many reviews it is easy to find expressions of discontent introduced by a conditional, such as: j'aurais voulu plutôt… *(*eng. *I would have preferred*), j'aimerais que… *(*eng. *I would love*).

For this reason and to add some rules to the shallow parser, I analysed conditional mood and in particular the differences between topics, polarities, and finally, I took a closer look to the use of the first pronoun plus the verb, guessing that it could have been interesting to know how people refer to themselves while reviewing a product with the help of this mood.

In general we can observe that conditional mood is used in both topics and ratings, and people tend to use more conditional in negative reviews (Figure 21).

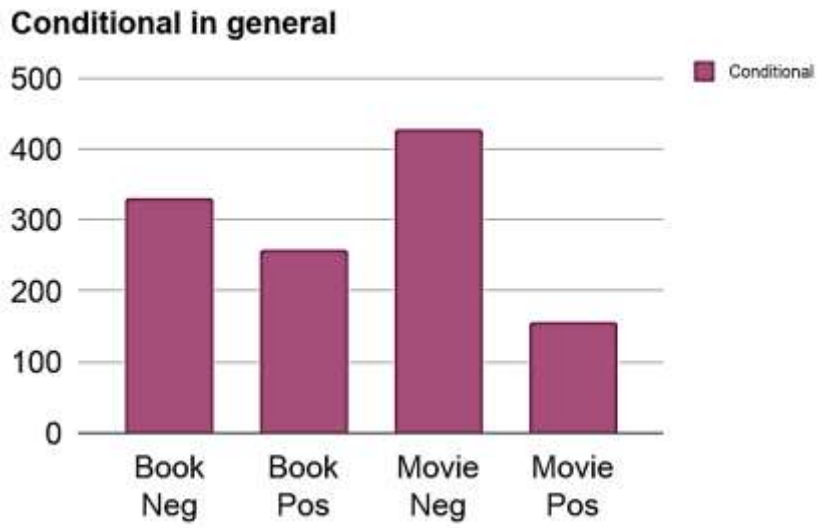

Figure 21 - Use of the conditional mood for both domains

[13] https://en.wikipedia.org/wiki/Conditional_mood

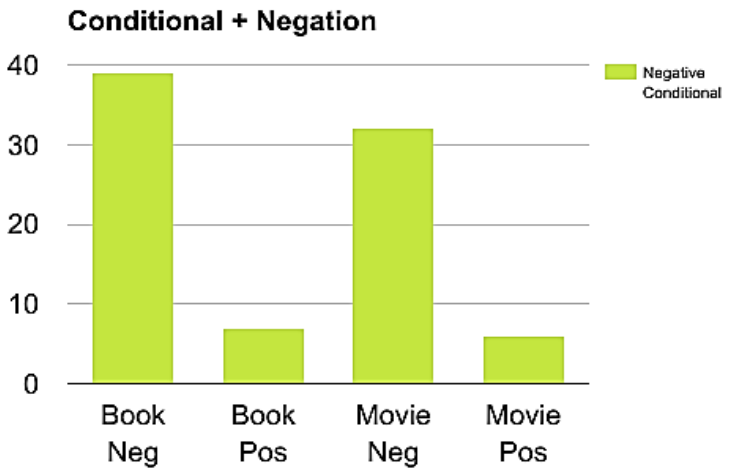

Figure 22 - Use of Negative Conditionals

Taking a closer look, we observe that (Figure 22) it is easier to find many conditional verbs with negations in negative reviews than in positive reviews. This is due to the fact that it is a mood used to express dissatisfaction in reviews.

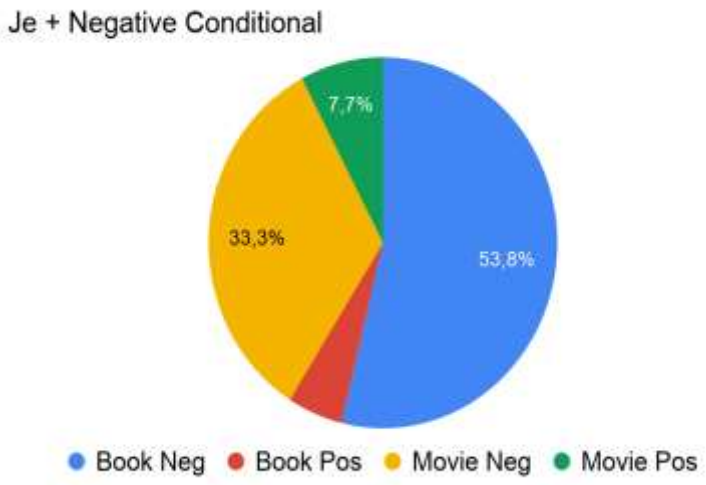

Figure 23 - Distribution per domain and type of review where it appears a first person plus a verb in negative conditional

Eventually, when analysing more closely the conditional plus negation (Figure 23), it seems that when people want to bring themselves and their thoughts in reviews, they really prefer to use conditional plus negation to express their ideas.

Summary of chapter 3

In this chapter we present the resources that have been used and created. In primis, the corpora: French Sentiment Corpus and NC corpus. French Sentiment Corpus was a pre-existing corpus of French reviews where the first experimentations have been carried out.

New Corpus (NC) corpus, a new corpus created during this thesis, where three types of rating have been selected (1-3-5) and they are developed in more than 9000 sentences.

This corpus was also used to create a new dataset for Movie and Book Aspect Based Sentiment Analysis (ABSA) in French, something that was missing in NLP community. The chapter describes the annotation process, some examples, and some thoughts about the difference between domains.

Another section is dedicated to the lexicon used: ValEmo - an emotion lexicon, F-Pol – a polarity lexicon.

Finally, we present other two resources created during this thesis: Lex – an opinion lexicon and RLEX – a lexicon composed by words considered discriminant by logistic regression.

After a section where we discuss about how and why we have modified a spellchecker - called LanguageTool - in order to correct some important words of the corpus. We present a linguistic dataset and a study of some linguistic patterns ("plutôt", "bien que", "mais") that have been introduced in a classification system in the form of rules to take into consideration verb moods and tenses and, precisely, these linguistic patterns.

First experimentation: polarity classification using only a statistical approach

The first attempt was a replication of the work done by (Vincent, et al., 2013). Their goal was to detect the overall opinion of reviews by using the *very positive* (rating 5) and *very negative* (rating 1) reviews to train a classifier. Their approach was purely statistical, based on bag-of-words and representation of documents without the support of a lexicon, missmisusing SVM and logistic regression. The global proportion for their experiments was of 1000/1000 reviews for the hotel domain, 576/858 reviews for the books and 800/800 for movie reviews.

For our first attempt no specific pre-processing of the text was applied. Nevertheless, to avoid over-fitting, asymmetric samples, and related bias, we applied a k-fold validation with k = 10. The result was very near to the one of the two authors but not identical. This was due to some methodological differences: use of different tools for POS tagging, text pre-processing absence, and different part of the corpus used (reviews with rating 0.5-5 for (Vincent, et al., 2013), reviews with rating 1-5 for this thesis).

Global Results	Logistic Regression	SVM
(Vincent, et al., 2013)	0.90	0.88
Thesis	0.86	0.86

Table 11 - a comparison between our overall results and the ones of the baseline for both Logistic Regression and SVM methods (F-Measure)

Review Topic	Logistic Regression (thesis)
Hotel	**0.940**
Movie	0.802
Book	0.821

Table 12 - The Logistic Regression results per review topic (F-Measure)

As it is shown in Table 12, hotel reviews obtained the best results. This is due to the structure and the content of the review itself. In fact, usually a hotel review is more linear and well-structured than a book or movie review. In other words, it is easy to find people

judging their stay and not writing about what they have visited and their emotions concerning their holiday.

Second experimentation: polarity classification using statistical and lexicon approaches

The next step after the replication of the results of (Vincent, et al., 2013) was a content addition. The baseline showed a statistical approach using bag-of-words without opinion lexicon. The first goal of this thesis was to keep on using the bag-of-words method, this time in association with an opinion lexicon, and highlight the differences. Eventually, the results were created adding a score to each word, and dividing the final result by the number of opinion words presented in the text.

One of the most interesting part of this thesis has been the study of how people adopt some words in different contexts. How, and in what measure are words context-dependent? In which measure can we capture some linguistic parameters using linguistics, statistics or both? Can every lexicon fit every type of review in order to detect opinion? I did several tests and used several lexicons to answer these questions.

More specifically, for this part of the thesis our classification was based on two lexical standards (Vincze, et al., 2011), (Syssau, et al., 2005), a lexicon derived from statistical measurements concerning the corpus, and a mix of the three.

Lexicons correlation

In our experiments, the two lexicons described in the resources section were merged. Unfortunately, F-POL (VINCZE, et al., 2011), and ValEmo (Syssau, et al., 2005) don't share the same scale rating system. Indeed, F-POL goes from 1 to 7, whereas ValEmo goes from -5 to 5. In order to merge them and use a scale -5/5, we verified the correlation between the two lexicons (from now the merged lexicon will be called SB) and they shared a R-Pearson linear correlation measure of 0.94.

	SB
Words no.	3079
Use of lexicon	82.2%

Table 13 - Lexicons and corpus coverage

Result and conclusions

	SB	Lex	RLex	RLex + SB	RLex + Lex	Baseline
Hotel	0.826	0.871	0.930	0.905	**0.955**	0.940
Movie	0.638	0.677	0.780	0.792	**0.795**	0.802
Book	0.638	0.642	0.713	0.686	**0.741**	0.821

Table 14 - F1-measure using Logistic Regression and Lexicons

Finally Table 14 shows the results, i.e. the F1-measure using logistic regression. Even if SB has the highest coverage (Table 13) it is not the one that performs best in classification. RLex and Lex are the ones that together seem to perform best, approaching the baseline results. This is probably due to the fact that RLex takes into consideration words that better describe polarity in the corpus, and Lex is focused in measuring opinion and not emotion as SB does.

One example of this reasoning is that there are some words that are emotionally neutral or positive (such as *dormir* - eng. *to sleep*) in hotel review context as well as very negative when they are used in a movie review.

This leads to the conclusion that sometimes the meaning of some words are strictly dependent from the corpus and their context of occurrence and the task itself.

In addition to this, it seems that a more structured and linear review will be easier to classify than a review containing a mix of review and description of the product: it is the case for movie and book reviews.

Conclusions:

- **Lexicon approaches can help in reviews classification.** Even though the baseline has the best results using a Bag-of-Words approach without lexicon, it seems feasible to reach similar results using one. This shows that at present it is possible to play with sequence of words, and not only with statistics, when classifying reviews.

- **One phenomenon analysed, one specific lexicon to describe that phenomenon.** When hypothesis of using a lexicon is taken into consideration, it seems very important to choose the lexicon that describes best the phenomena we want to analyse. That's why even though SB had the best coverage, it performed worse than RLex and Lex in an opinion classification task.
- **Not every lexicon is the right lexicon.** Another important point is the choice of the lexicon domain. From these experiments it seems very important to use both a generic and specific lexicon able to capture both words from intersection of domains and very specific words related to one domain.
- **Single and compound words are both important.** Even when a specific lexicon has been used, some errors highlight that a lexical approach considering only unigrams can be their source.
    - Example: *chef d'oeuvre* (eng. *masterpiece*) is a compound word (*chef*+*de*+*oeuvre*) - when it is considered as three different and separate words, it is obviously easy to lose the valence of the original word.
- **Complex and long review: not every word on them is an opinion.** Actually, it is possible to find some negative terms which are used in a specific context but are just describing something that is not the review of the product itself. In other words, the classification system should consider only the opinion and not the topic.
- **Last but not the least: a text with misspelled words is a text that cannot be well classified.**

This part of the thesis has been described in a paper for the conference TALN-JEP-RECITAL (Pecòre, et al., 2016).

Next challenges

After these first experiments, given the complexity of book and movie reviews, the efforts for this thesis have been focused on studying and analysing these two complex domains.

These objectives were set:

- to correct misspelled words, when it was possible,
- to extend the corpus in order to get new fresh data,
- to study some specific linguistic patterns that can be encountered in reviews,
- finally, taking advantage of the lack of French resources to create a new dataset of French reviews and a dataset of aspects concerning movie and book domains.

Along these activities I analysed some important linguistic patterns (see next section) for the reviews that have been used to feed a classification system, using the output of a chunker called Ritel-NCA (Rosset, et al., 2008). In particular I created some linguistic patterns to be introduced in a shallow parser to detect and classify book reviews. The system will be briefly described in the next section. This part of the thesis has been described in a paper for the conference ECG 2019. (Villaneau, Pecore, Said, Marteau, 2019).

Third experimentation: polarity classification using statistical and lexical approaches, and syntactic information

Linguistic patterns in reviews

As stated before, during the thesis I contributed to the paper (Villaneau, et al., 2018) by retrieving and analysing some linguistic patterns that have been shown and explained in *Linguistic dataset: special words, patterns and verb tenses* section. In that paper, different machine learning techniques with heuristic rules are used to aggregate polarity scores evaluated in chunks provided by a shallow parser. At present, the system is used to analyse the sentiment polarity of single sentences.

Why has it been so important and interesting to study the linguistics behind the reviews?

First of all, the better you formalise and define the human language, the better the system can understand it and classify it. Obviously, it is not possible to constrain a whole language in a set of rules because of its arbitrary nature, rich and changing.

Secondly, sometimes there are some linguistic patterns in a sentence that, if taken individually, seem void, but in reality they drastically change the sentiment of that part of sentence. In this case it is really important to be able to write some rules to insert inside a shallow parser for taking into account these phenomena. This can let us verify the impact that these expressions can have in a classification system based on a shallow parser and some machine learning techniques.

During this part of the thesis the data was submitted to the external system RITEL-NCA as a classification test. This tool has been modified for the task of sentiment analysis. Implemented as an analyzer for the Question-Answering system RITEL, RITEL-NCA uses regular expressions of words and provides outputs with detailed syntactic and semantic information. As RITEL-NCA is designed for Information Extraction and not for Sentiment Analysis, an adaptation for the task was required. This part has been developed by colleague Villaneau, J.. Citing the paper (Villaneau, et al., 2018), the chunker takes into account:

- negations –
    - "*Chunking allows to partially solve difficulties associated with the negation scope, since a specific treatment of the negation is proposed for each form of chunk. (…) Negations are linked with functions which are applied to the score of their scope as an argument*".
- verbal chunks –
    - "*The opinion score of a chunk's head is used as an argument which can be altered by other elements of the chunk: negation or modifiers. Verbal chunks are also used to specify the tense and the mood that are used to calculate the score of the whole sentence. For example, in the phrase je [n'avais pas aimé] [I didn't like], the head aimer [to like] has a positive score (3/5). After application of a negation function. The score value of the chunk becomes negative (-3.9 in the current system). Finally, the employed tense (past perfect) comes to action by reducing the score of the whole clause."*
- nominal chunk –
    - "*The main role of nominal chunks is to enable adjective/name associations to be taken into account. Adjectives such as "mauvais" [bad], "excellent" [excellent], "beau" [beautiful, fine] give to the nominal group a stable*

opinionated orientation, positive or negative. Others, especially size adjectives, increase (as ”grand” in ”grand succès” [great success]), reduce (”succès mitigé” [mitigated success]) or reverse (”faux succès” [false success]) the polarity of the name with whom they are associated. Specifying adjectival functions is difficult, especially in French language where the sense of an adjective can vary, according to its position for instance. Yet we have defined and tested some very frequent adjectives. Also, nominal chunks allow the detection of local negations which involve adjectives such as ”aucun” [no] or prepositional phrases such as ”rien de” [nothing] or ”manque de” [lack of], etc.”

- adjectival and adverbial chunks –
    - *“Adjectival and adverbial chunks have been defined to take into account modifiers and group of modifiers: negatives, intensifiers and diminishers. Each modifier is associated with a function, which is applied to the score of the phrase to whom the modifier is linked. For example, in the phrase ”vraiment trop long” [really too long], really too reinforces the negative opinion generally conveyed by the adjective long when assigned to a book.”.*

$$f_{too}(long) =$$
$$f_{too}(-2) = (x \mapsto 1.3 * x)(-2) = -2.6$$
$$f_{really}(too\ long) = f_{really}(-2.6) =$$
$$(x \mapsto 1.2 * x)(-2.6) = -3.12$$

Figure 24 - Example of a function for the phrase "really too long"

Finally as stated in the paper “*the system assigns scores to the words, expressions and chunks, by using the rules described above. The score of a sentence is obtained by combination of its chunks scores and application of the functions which were defined to take into account negation, mood, tense, etc.*”.

The original goal was to take into account and assign scores to negations, as well as mood and tenses of verbs, determine the opinion orientation of nominal chunks, modifiers, special conjunctions, and common French expressions strictly related to movies and book domains.

The results (Table 15) of these analysis in conjunction with the shallow parser show that using a linguistic pattern for classification can improve results when analysing the whole review.

	Logistic Regression	System (S)	(S) w/o Mood & Tense	(S) w/o Modifiers	(S) w/o "but"
Book	0.821	0.865	0.863	0.862	0.855
Movie	0.802	0.837	0.812	0.837	0.819

Table 15 – Results from the Full System processing linguistic expressions and verbs.

Forth experimentation: ABSA using SVM

Can (very precise) human annotation be the key of a (very) good classification?

The annotation was useful to evaluate a classification system with very efficient data approved by experts. In my tests I used both entities and aspects together, or a combination of them. Aspects has been tested in several combinations and different proportions. In addition to this, I tested more than one classification system: Naïve Bayes, Decision Trees, Support Vector Machines (SVM) and Hierarchical Support Vector Machines (HSVM). The best result came from SVM. Unfortunately I did not have the opportunity to use Deep Neural Network systems because the data in my possession was not enough.

In the following test cases I will show some preliminary tests where I used extremely precise training and test sets that result in a very precise but idealistic classification and very difficult to implement in real life. This is compared to three other preliminary tests where the data is less precise but closer to the reality of things.

For each preliminary test I changed the content of training and testing data, going from very specific data to very general. The SVMs of these tests used RBF kernel, optimized gamma and c parameters via grid-search, and a 5-folds cross-validation. The co-occurrence matrix has been always been built using TF-IDF features selection.

Explanation of the acronyms and table structures used

Acronyms

Before proceeding with the description of the test, please take note of the following acronyms that will be used in this part of the chapter.

The concerned aspects (and their acronyms are):

- GEN_ALL (The General attribute of every aspect),
- GF (General Feeling)
- ACT (Actor and its attributes)
- DIR (Director and its attributes)
- CHAR (Script Characters and its attributes)
- PN (Script Pace/Narration and its attributes)
- PL (Script Plot and its attributes)
- SUB (Subjects and its attributes)
- MUS (Music and its attributes)
- VID (Video and its attributes)

How to read the following tables

Please take note of the following examples (tables) in order to understand the description of the experimentations.

Example #1:

	E1
GEN_ALL	0.66
GF	-
ACT	0.08
DIR	0.25
CHAR	0.20
PN	0.43
PL	0.28
SUB	0.12
MUS	0.11
VID	0.08

Table 16 – Example#1

When there is a numeric value in GEN_ALL, it means that the general attributes of the aspects are evaluated as a distinct class. In the table above, GF will not be present, and each General attribute of each Aspect (ACT, DIR, etc.) will be grouped in GEN_ALL.

Example #2:

	E2
GEN_ALL	+
GF	1
ACT	1
DIR	1
CHAR	1
PN	1
PL	1
SUB	1
MUS	1
VID	1

Table 17 -Example#2

When there is a plus sign (+) without any values in GEN_ALL, it means that the General attributes of each Aspect (ACT, DIR, etc.) have been included in the GF aspect.

Example #3:

	E1
GEN_ALL	0.66
GF	-
ACT	0.08
DIR	0.25
CHAR	0.20
PN	0.43
PL	0.28
SUB	0.12
MUS	0.11
VID	0.08

Table 18 - Example#3

When there is a minus sign (-) the data was not used. In the example above, the GF Aspect was not used for the experiment.

Example #4:

	E1
GEN_ALL	-
GF	-
ACT	0.54
DIR	0.39
CHAR	0.21
PN	0.50
PL	0.36
SUB	0.16
MUS	0.17
VID	0.33

Table 19 - Example#4

When both GF and GEN_ALL Aspects are followed by the minus sign (-), General attributes of each Aspect (ACT, DIR, etc.) are not present in any Aspect.

Preliminary Tests

The first 4 tests were conducted using a small portion of the data (600 documents) in order to observe the limits of the methods and start testing the hypothesis.

For these tests every Aspect is present. In particular, every General attributes of each Aspect (ACT, DIR, etc.) was included in the related Aspect. Documents are present in equal measure for each Aspect, and they have been chosen via stratified sampling with proportion training-test sets of 70%-30% in order to avoid cherry picking imbalance among samples. Macro-average F-score has been used to evaluate the system.

Conclusion of the preliminary tests

	Training Set	Test Set
CASE 1	Entities + Aspects	Entities + Aspects
CASE 2	Entities + Aspects	Aspects
CASE 3	Aspects	Aspects
CASE 4	Aspects	Whole sentences from Corpus

Table 20 - Summary of the preliminary tests

	CASE 1	CASE 2	CASE 3	CASE 4
GEN_ALL	+	+	+	+
GF	1	0.26	0.42	0.40
ACT	1	0.08	0.34	0.17
DIR	1	0.18	0.36	0.40
CHAR	1	0.14	0.32	0.05
PN	1	0.48	0.50	0.32
PL	1	0.20	0.17	0.06
SUB	1	0.00	0.26	0.29
MUS	1	0.00	0.12	0.00
VID	1	0.09	0.26	0.08
F1-AVG	**1**	**0.16**	**0.31**	**0.20**

Table 21 - Results (F1 measures) of the preliminary tests

We see from Table 21, Case 1 is the one giving the best results. At the same time it is the farthest from the reality of things. Actually we have in both training and test sets the entities (for example “the movie”) and the aspect (for example “is a masterpiece”). This structure is repeated for each sentence that has to be classified according to the aspect.

This test is not reliable because:

1. In real reviews it is hard to find every time explicit entities + aspects,
2. Over-fitting: when we feed some documents that don’t have this structure into the system, the F1 measure declines sharply producing the result of Case 2 (F1=0.16),
3. To some extent, we could think that with this repetitive structure we have already classified the data even before launching the classifier which is quite uninteresting.

The most interesting cases are Case 3 and Case 4. Even though the result is not excellent, the data used in training and test sets is similar to the one that it is possible to find in an review. Both cases show also food for thought:

1. Classification don’t reach perfect results, even if we use annotated and approved by humans sentences. This can be due to the fact that:
    I. The data is not enough
    II. There are too many words in common among aspects
2. This is however a good start in order to improve the results.

Speaking of the classified aspects, we notice that the results are lower in the area of the scripts aspects, music and video aspects. This could make us suspect that we need a way to handle separately these aspects or to distinguish them in some way during the classification.

Fifth experimentation: more data, SVM and hierarchical SVM, and human annotations

These experiments were run to classify aspects but also for testing the limits of a classification in a SVM environment using high accurate selected features. These tests have been done in samples of more than 3000 documents with the usual 70%-30% distribution of training-testing data. The goal of these tests is to choose the best distribution of sets, trying to maximize the final F1-measure avoiding the exclusion of aspects. For this part, the samples are not balanced: that is why I used a micro-average F1-measure to evaluate the system.

Case 5: Classification using the whole aspect attributes set, exception made of General Feeling one

In this test I tried to understand the impact of the class General Feeling (from now GF) on the classification of the other aspects. GF doesn't identify itself as a specific class. It has been usually encountered when the movie was judged good or bad using general concepts such as "that's a good [Aspect] **movie** [entity]". GF is a problematic aspect because it has many words in common with the other aspects, due to the generic content of it.

In this test I expressly did not remove the general attribute of each aspect (GEN_ALL) in order to judge only the absence of the class General Feeling. Nonetheless I was aware that GEN_ALL could be of interference even if it is more specific than GF.

Unfortunately this was the case. More than 2/3 of the test set in fact have been classified as GEN_ALL.

Case 5	F1
GEN_ALL	0.66
GF	-
ACT	0.08
DIR	0.25
CHAR	0.20
PN	0.43
PL	0.28
SUB	0.12
MUS	0.11
VID	0.08
F1-AVG	**0.24**

Table 22 - Case 5: whole aspects set exception made of GF aspect

What if we use very precise Aspects, without any general Aspects?

Case 6: SVMs with only specific attributes

When I decided to exclude the two more general classes (GF and GEN_ALL), I expected to improve the results by taking into account only the specific words related to other aspects.

Even though the improvements are visible, three classes have been highly misclassified. In particular the aspects CHAR, and MUS were absorbed by the class ACT – 51 out of 61 were in fact classified as ACT instead of CHAR, and 23 out of 43 were classified as ACT instead of MUS.

From the analysis of the errors, I discovered that many words were in common, such as: “incredible”, “original”, “love”, “interesting”, etc. I suppose then that is the reason of misclassification: it is possible to imagine that the same test using chunks could represent better these classes and therefore improve the results.

The class SUB has been also misclassified as Director (14 out of 31) and Actor (9 out of 31). In this case, too, many words were found in common, such as: “charismatic”, “intelligent”,”interesting”, etc.

Case 6	F1
GEN_ALL	-
GF	-
ACT	0.54
DIR	0.39
CHAR	0.21
PN	0.50
PL	0.36
SUB	0.16
MUS	0.17
VID	0.33
F1-AVG	**0.31**

Table 23 - Case 6: Aspects without general concepts (GF and GEN_ALL)

Case 7 and 8: All inclusive - every aspect and their general attribute

In the previous tests I excluded the very general classes (GF and GEN_ALL). After these experiments I included in every Aspect the general attributes from GEN_ALL (Case 7), increasing the samples. Did the GEN_ALL attributes in each Aspect resist to the greedy GF class?

In Case 7 we have the same problem of Case 5: the most general class absorbs some of the other classes, where words are in common. This is the case of the class CHAR (36 out of 61 misclassified as GF) and SUB (31 out of 37 misclassified as GF).

For this reason, in Case 8 I decided to replicate the Case 5, this time with more data. We can notice some improvements especially in ACT, DIR and CHAR aspects. However classes such as SUB and MUS suffered of a great misclassification due to ACT and DIR classes.

	Case 7 F1	Case 8 F1
GEN_ALL	+	+
GF	0.74	-
ACT	0.42	0.55
DIR	0.29	0.43
CHAR	0.13	0.24
PN	0.51	0.50
PL	0.29	0.38
SUB	0.09	0.16
MUS	0.22	0.16
VID	0.39	0.25
F1-AVG	**0.34**	**0.33**

Table 24 - Case 7 and 8: F1-Measures of the classificataion system inclulding every aspect with and without GF

Generally speaking, from Case 5 to Case 8 results have improved, which is an incentive to continue using every aspect but changing the way they are presented to the system.

Case 9 and 9_{bis}: to group in hyper classes and proceed with a two-step classification

In these two final tests I tried to group the classes as hyper classes. In particular, I joined ACT and DIR as Human hyper class, CHAR, PN, PL and SUB as Script hyper class, MUS and VID as Media hyper class.

In Case 9 obviously the F1 measures are higher but we have lost the specific details of each class. Remark however that even when the classification is reduced to a few classes or to a binary problem, the results are not so high (see Media hyper class F1-score). This is a proof of the difficulty of the task.

Case 9	F1	
GEN_ALL	+	+
ACT	Human	0.70
DIR		
CHAR	Script	0.63
PN		
PL		
SUB		
MUS	Media	0.40
VID		

Table 25 – F1-measures of the classification system using SVM and hyper classes of Aspects

In Case 9_{bis} I used SmartSVM[14] (van den Burg, et al., 2017), a Python package for H-SVM. A Hierarchical Support Vector Machine (H-SVM) for multiclass classification is a decision tree with a SVM at each node. At the root node of the decision tree, all classes are available for prediction. The number of classes available for prediction keeps decreasing as we descend the tree.

[14] https://smartsvm.readthedocs.io/en/latest/#

SmartSVM is an adaptive hierarchical classifier which constructs a classification hierarchy based on the Henze-Penrose estimates of the Bayes error (Berisha, et al., 2015), (Berisha, et al., 2016) between each pair of classes. Unfortunately also in this case classes such as SUB and MUS do not seem to benefit from this type of classification.

Case 9	F1	Case 9bis	F1
GEN_ALL	+	GEN_ALL	+
Human	0.70	ACT	0.40
		DIR	0.41
Script	0.63	CHAR	0.31
		PN	0.49
		PL	0.31
		SUB	0.15
Media	0.40	MUS	0.15
		VID	0.32
F1-AVG	**0.57**	**0.32**	

Table 26 - Case9 and Case9bis: Using HSVM from hyper classes to specific ones

Concerning the domain of books, a similar task has been done during the writing of this thesis using our dataset and many techniques such as: KNNs, Random Forest, SVM, etc. The best performance (F1-AVG 0.64) has been obtained, in this case too, using SVM. For more details, please refer to (Villaneau, et al., 2018)

Summary of chapter 4

This chapter is written to give an idea of the experimentations of the thesis.

As an introduction of this summary and without entering immediately the details we can say that during the thesis we have tested several approaches: polarity classification using only statistics, and polarity classification using statistical and lexical approaches; then, we added linguistic pattern features in a classification system taking into account the syntactical information, and finally we used the very precise data in our possession (dataset in ABSA), to classify aspects via SVM.

The first experimentation was a replication of another work (Vincent, et al., 2013)*: they used logistic regression and SVM to classify movie, book and hotel reviews without using a lexicon.*

Our first contribution was an extension of the first experimentation adding the four lexicons described in chapter 3 (ValEmo, F-POL, Lex, RLex) and classifying reviews with the help of these lexicons via logistic regression or SVM. Unfortunately, the obtained results did not overcome the replication for movie and book reviews. They showed anyway that it is possible to reach good results using a lexicon oriented to the task. During that experiment we noticed that hotel reviews obtained the best result. For this reason we decided to focus only on the worst results: the ones of movie and book reviews, considered as the most difficult to classify.

The errors were caused by misspelled words, a lexicon too general to be used for both domains and a classification unable to distinguish between the opinion part of a review and just a neutral description of the product.

In order to solve the last problem, we decided to annotate data and we created a dataset for ABSA for movie and book reviews.

Along the work of annotation, we created some linguistic resources concerning particular linguistic patterns and mood/tenses of verbs. The linguistic resources have been used to feed a logistic regression system that used the syntactic outputs of a shallow parser called RITEL-NCA with the information about polarity. This experiment was important to test whether linguistic features could improve the overall results or not, and it was the case, although not much.

Then, we have presented a new set of experiments using our ABSA dataset. The goal of these final experiments was to test whether realistic and very accurate data could improve the

system. For this reason we launched several tests using SVM and H-SVM maximizing the parameters and playing with the combination of data in our possession. The results showed that, even with very accurate data, the presence of words that can occur in different contexts can lead to misclassification in the case of ABSA.

CHAPTER 5 – PRELIMINARY PSYCHOLINGUISTIC STUDY: VALENCE, AROUSAL AND DOMINANCE APPLIED TO OPINION ANALYSIS

Sixth experiment: ABSA classification using SVM and psychological lexicon metrics

The context of the experiment

During the last semester of the Ph.D. I had the chance to initiate a collaboration with *National Research Council Canada* (**NRC**) in particular the *Digital Technologies* department and *Text Analytics* team. This part of the thesis and the related experience have been funded by « *Bourse de Mobilité de l'Ecole des Docteurs de l'Université Bretagne Loire* » and « *Projet Numérique de l'Université Bretagne Sud* » grants.

At CNRC I'm collaborating as Independent Researcher in a parallel research project that aims to use NLP tools to study gender bias in literature by analysing emotions and specific words in contemporary novels. This experience granted me a wider vision of the problems and tools that the domain of this thesis can afford. In addition of this, it inspired me to use a specific lexicon describing polarity-arousal-dominance of a person towards another person or situation or object. In other words, the lexicon that I will describe later has been used – in the context of the thesis - to study the psychological attitude of the reviewers towards the reviewed product.

To do this, I first analysed the difference of VAD levels between the two domains, book and movie by using this lexicon. Then I made a comparison between book and movie reviewers towards different aspects of a product. Finally, I decided to use each VAD numeric value in the previous classification system to improve the system by using more details about the words.

This final chapter is an introduction of a future work and it will briefly describe the background and the reasons behind the chosen approach and the Valence-Arousal-Dominance (VAD) lexicon from C-NRC – i.e. the tool used to describe the psychological attitude of reviewers.

Background

The choice of the words we use is very important. We should have the habit of weighing the words every day because it is easy to notice that they are not the same. For this reason, as we already said, it is quite impossible to find two words which are exactly the same. The reasons behind this can be found in linguistics and in (Saussure, 1916). Saussure defined the word - "la parole" – as an entity totally under free will: there is no innate law that states that the concept behind what we call "chair" should be identified as (the sequence of letters) "chair". The choice of how identifying something is totally arbitrary. Saussure in his *Course in General Linguistics* tries to describe our confused and unstructured thoughts as a structured system. He defines the combination of the signifier and the signified as arbitrary but also - once finally chosen this combination - as a reflection of a whole people mind: "la Langue".

Some works (Ervin, 1964), (Luna, et al., 2008), (Grosjean, 2010) highlight that the use of different words and different structures of languages can affect the way people think. In some cases research found that the choice of our words define exactly how we feel and if we suffer from any type of distress (Mehl, et al., 2017).

Many works and authors have tried to analyse the smallest and powerful meaningful part of our way of expression: words. According to literature it seems that there are three dimensions which are independent from the language and the culture: the Valence (or Polarity), the Arousal, and the Dominance (Russell, 1991).

We will shortly explain these three dimensions (from now VAD or PAD) giving some examples:

1. Valence measures the positive-negative/pleasant-unpleasant level of an object, event, and situation. The term "*refers to the direction of behavioural activation associated with emotion, either toward (appetitive motivation, pleasant emotion) or away from (aversive motivation, unpleasant emotion) a stimulus.*" (Lane, et al., 1999)
2. Arousal is orthogonal to valence and it is the emotional activation level (active-passive/excited-calm)
3. Dominance measures the level of the perception of control on a stimulus (dominant-submissive)

According to (Osgood, et al., 1957), (Russell, 2003), (Mohammad, 2018) we can compare two words using the VAD values:

- The word “banquet” is more positive than “funeral”
- The word “tense” shows a higher level of arousal than the word “depressed”
- The word “president” shows more dominance than the word “idiot”

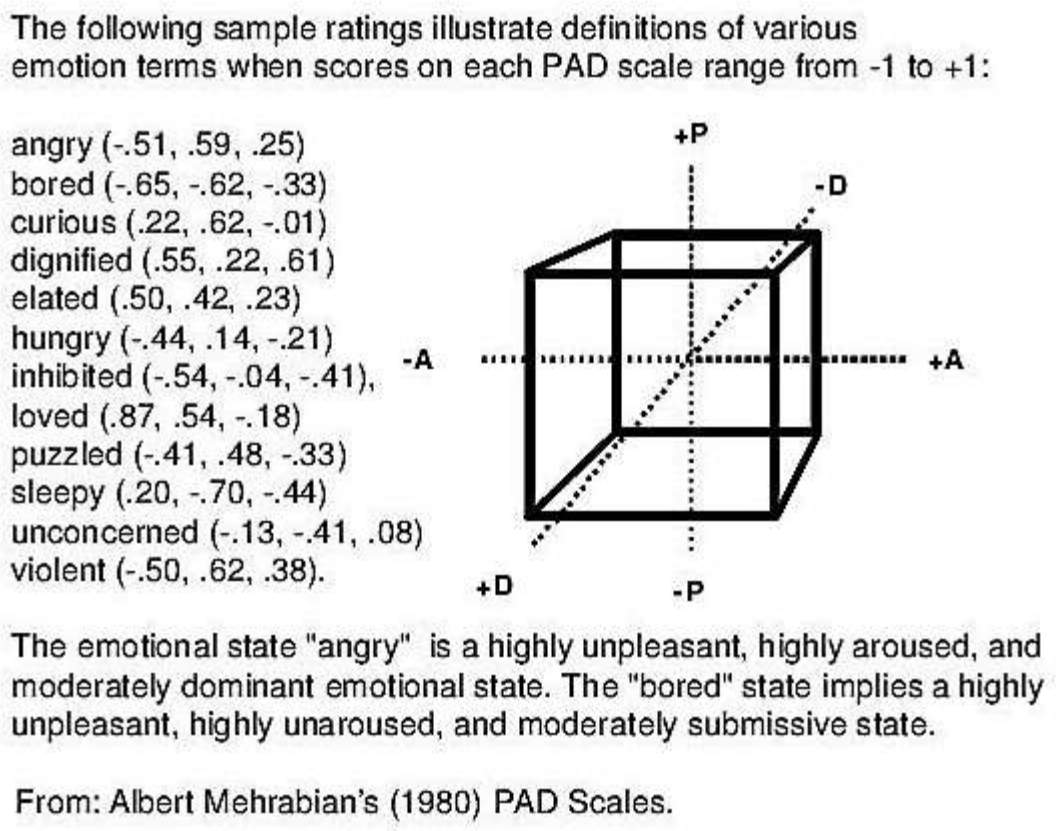

Figure 25 – Polarity (or Valence)-Arousal-Dominance Scale figure by Albert Mehrabian (Mehrabian, 1980)

This kind of information can be very useful to picture how humans react in a certain situation through words. Some examples are in (Graziotin, et al., 2015), (Graziotin, et al., 2015), (Khan, et al., 2011) where words are analysed and the VAD is used to understand how a good productivity is related with a high level of valence (are they happy?) and high dominance (are they in control enough of the situation?).

The most known (and in English) VAD lexicons are:

- Affective Norms of English Words (ANEW) (Bradley, et al., 1999). This lexicon measures VAD for over 1000 words using a 9-point rating scale. The final rating is given by the average of the annotators' ratings.
- (Warriner, et al., 2013) created a lexicon of over 13000 words using a similar scheme.

VAD lexicons can be found also for these languages:

- French: (Vincze, et al., 2011)
- Spanish: (Redondo, et al., 2007)
- German: (LH Vo, et al., 2009)
- Dutch: (Moors, et al., 2013)

For my experiments I used the Valence-Arousal-Dominance lexicon by NRC (from now NRC VAD lexicon[15], (Mohammad, 2018)) the same that I used during the development of the project in collaboration with C-NRC.

The NRC VAD lexicon is one of the largest manually annotated lexicons and it is composed of the union of several lexicons. The annotated terms denotate / connotate emotions and are commonly used in English. The lexicon is composed of 20,000 words and it has been translated in many languages (French included) by NRC team. The words concern common nouns, adjectives, adverbs and verbs. As we stated before, the domains of Valence, Arousal and Dominance are independent from the language and the culture that is why I decided to use this lexicon for the thesis. In NRC VAD lexicon, each word can assume a value from 0.0 (lowest level of the property of interest) to 1.0 (highest level of the property of interest).

- Example:
  - The word "calm" has Valence = 0.875, Arousal = 0.100, Dominance = 0.282

[15] http://saifmohammad.com/WebPages/nrc-vad.html

- The word “homicide” has Valence = 0.010, Arousal = 0.973, Dominance = 0.518

The NRC VAD lexicon has been annotated as a crowdsourcing task (Figure Eight[16]) by using the Best-Worst Scaling (BWS) method, already known in mathematical psychology and psychophysics (Louviere, 1991), (Louviere, et al., 2015). The idea behind BWS is that the annotator is presented with four words (X_1, X2, X3, and X4) and asked for the word with the highest valence/arousal/dominance level and the word with the lowest valence/arousal/dominance level. Once we know that - for example X_1 has the highest VAD level and X_4 has the lowest ones - we can deduct that:

- $X_1 > X_2$; $X_1 > X_3$; $X_1 > X_4$
- $X_2 > X_4$
- $X_3 > X_4$
- ***X_2 ??? X_3***

In this case, using only the two extreme values (X_1 and X_4), we have automatically solved 5 pairs of values out of the 6 possible. For more details about NRC VAD lexicon, please refer to (Mohammad, 2018).

Research Questions

As it is stated in (Mohammad, 2018), “*VAD lexicon has a broad range of applications in Computational Linguistics, Psychology, Digital Humanities, Computational Social Sciences:*

- *study how people use words to convey emotions.*
- *study how different genders and personality traits impact how we view the world around us.*
- *study how emotions are conveyed through literature, stories, and characters.*

[16] https://www.figure-eight.com

- *obtain features for machine learning systems in sentiment, emotion, and other affect-related tasks and to create emotion-aware word embeddings and emotion-aware sentence representations.*"

As a consequence of this, the questions to which I would like to answer here are:

- Is it possible to identify the domain and polarity of the review via VAD by studying how emotions are conveyed through the review itself? Do people perceive the act of watching a movie or reading a book in a different way?
- Is it possible to identify the type of entity reviewed via VAD by studying the emotional relation between reviewer and entity reviewed?

I will try to answer to these questions and to evaluate the VAD inside the classification framework used in the previous chapter. The idea behind this preliminary chapter is that the VAD could be used as another feature to improve sentiment analysis systems (this is also supported by the VAD authors' declaration of the possible applications of the VAD).

For example we could use the Valence as a parameter to measure the level of positivity/negativity of a sentence due to the presence of positive/negative words. In addition to this, the VAD could help us identifying how the reviewer is influenced by the (aspect of the) product. Recalling the previous Saussure's idea, we can refer to the concept behind "actor", identifying it as "actor" (V=0.653, A=0.561, D=0.629) or as "puppet" (V=0.459, A=0.334, D=0.298); the concept behind the words movie director can be called "movie director" (V=0.740, A=0.570,D=0.691) or "master" (V=0.694, A= 0.490, D=0.849). The reviewer can call the written text in different ways: "novel" (V=0.702, A=0.352, D=0.439), "publication" (V=0.549, A=0.470, D=0.600), "book" (V=0.802, A=0.210, D=0.606), and sometimes people name a book without literary refinement a "magazine article" (V=0.463, A=0.447, D=0.429).

According to this, it seems that the choice of the reviewer's words is not arbitrary: an "actor" is more positive than a "puppet" (V=0.653 vs V=0.459), it is perceived as more active (A=0.561) than a "puppet" (A=334). Finally the puppet (D=0.298) seems to be perceived as more submissive than an actor (D=0.629).

Can we define how people perceive the act of watching a movie and reading a book using the Arousal and Dominance lexicon? For these experiments we will not use the Valence that could be used in a polarity detection task instead.

Experiment A: Overall Valence Arousal Dominance value per domain

In this experiment I analysed each domain by measuring the overall numerical value of the VAD using the whole corpus (book and movie). NRC VAD lexicon shares 72% of words with Book corpus and 70% with Movie corpus.

The measurement has been done in the simplest way possible: sum of each VAD value divided by the number of words with VAD. For this experiment I took 2000 sentences from the annotated corpus and for each domain, in total 4000 sentences. I took the same quantity of sentences in order to measure VAD on – nearly – the same number of words.

	Book	Movie
Valence	0.58	0.52
Arousal	0.51	0.48
Dominance	0.55	0.50

Table 27 - Average value of valence arousal and dominance for each domain using the whole corpus and obtained summing the numerical value of valence-arousal-dominance of each word and dividing by the number of words.

Experiment A - Results

As we see, it seems that the overall value by domain does not capture any particular difference between the two domains. Maybe because the analysis is too general. For this reason, I decided to proceed analysing more closely each single domain and each aspect.

Experiment B1: Valence Arousal Dominance for Aspect

For this experiment I investigated a difference among the aspects for each domain. The aspects for the movie domain are always the same of the previous experiment (page 89). For the book domain instead I regrouped the aspects using this scheme:

- Text (and Illustration Authors)
- Characters
- Form (Bookbinding, Typography, Inner structure, Distribution)

- Interest (Subject, Accuracy, Readability, Translation/Adaptation)
- Narration (Quality, Pace)
- Style

For each domain I measured the average VAD level for each aspect like experiment A: sum of each valence-arousal-dominance value divided by the number of words. The experiments, this time, have been performed on the whole annotated dataset.

Example:

For the sentence S (referring to the Actor aspect) = “a fantastic interpretation”,

- it is composed by “a” + “fantastic” (V=0.969, A=0.696, D=0.831) + “interpretation” (V=0.677, A=0.594, D=0.620)
  - so that the overall result of this sentence S is (V=1.646, A=1.24, D=1.451) / 2 (i.e. the number of words with VAD))
  - **Final result: V=0.82, A=0.64, D=0.72**

Experiment B1 – Results

For the movie domain (Table 28) we can notice that in general the level of VAD are higher when the aspects concern people such as actors and directors, and lower when they refer to the script. Unexpectedly, the video aspect however has both arousal and dominance levels equal to the director one. We can suspect then that the reviewer is more active when talking about what (s)he sees and when judging other people.

	Valence average	Arousal average	Dominance average
Actor	**0.81**	**0.73**	**0.77**
Director	0.73	*0.66*	*0.70*
Characters	0.59	0.54	0.56
Pace/Narration	0.61	0.61	0.59
Plot	0.68	0.65	0.65
Subject	*0.54*	*0.52*	*0.52*
Music	0.67	0.57	0.62
Video	0.78	*0.66*	*0.70*

Table 28 - [Movie] level of valence-arousal-dominance per aspect

For the book domain (Table 29) we can already notice that in general the values are indeed lower than in the movie one. This is due, I suppose, to the fact that the book corpus has fewer words than the movie one. However, I believe that this will not have any consequences on the comparison among aspects. Again, the VAD levels prevail for one aspect: the style. (Examples of style sentences: “facile à lire” [eng. *easy to read*],” se lit très vite”[eng. *you can read the book very quickly*], ”un livre mordant et pertinent” [eng. a pungent and relevant book, ”un livre mal écrit” [eng. *a poorly written book*]) Moreover, the dominance level for this aspect is twice the Characters aspect (Examples of characters sentences: “des personnages un peu trop stéréotypés” [eng. *characters a little too stereotyped*], “les personnages ont perdu toute leur saveur et leur caractère, ils sont tristes, sans relie” [eng. *the characters have lost all their flavour and character, they are sad, without relief*], “le personnage ne me fascine plus comme à ses débuts” [eng. *the character does not fascinate me anymore as in its beginnings*]). The lowest levels of VAD are registered for the Form aspect (Examples of the Form sentences: “ne donne pas envie de l'ouvrir” [eng. *it doesn’t encourage to open it*], “imprimé à l’envers” [eng. *printed upside down*]).

	Valence average	Arousal average	Dominance average
Authors	0.38	0.34	0.35
Characters	0.29	0.26	0.28
Form	*0.22*	*0.21*	*0.22*
Interest	0.51	0.45	0.49
Narration	0.43	0.38	0.39
Style	**0.58**	**0.49**	**0.54**

Table 29 - [Book] level of valence-arousal-dominance per aspect

Essentially after these experiments I was not satisfied enough by the results: I did not see any strong difference among aspects that could represent a key for a change in the classification system results. I suppose that this could be due to the fact that the VAD lexicon is calibrated for general common use words. It means that maybe a VAD lexicon by domain could improve (and describe better) these results. Also for this reason I decided to study the words that are unique for each aspect, assuming that they can be important because of their uniqueness.

Experiment B2: Valence Arousal Dominance for hapax for Aspect

In order to have a better picture of the situation, I decided to analyse the aspects using the words that are unique among aspects – i.e. hapax, assuming that they should be very important to characterize that aspect. In order to be objective and considering the little dimension of the corpus, the hapax have been selected among adjectives, adverbs, common nouns. For now, no further pre-selection criteria were used, except that the frequency was equal to one. I am aware of the possible critical points and I will discuss a possible solution in the Conclusions section.

Experiment B2 – Results

	Valence Frequency =1 average	Arousal Frequency =1 average	Dominance Frequency =1 average
Actor	*0.40*	**0.41**	**0.40**
Director	*0.39*	0.38	0.38
Characters	0.38	0.35	0.36
Pace/Narration	0.29	*0.29*	*0.29*
Plot	0.35	0.36	0.36
Subject	0.38	0.38	0.39
Music	**0.41**	0.38	0.39
Video	*0.28*	0.39	0.28

Table 30 – [Movie] level of valence-arousal-dominance of words per aspect with frequency = 1

As we can see from the (Table 30), except for the valence level, the other two are always highest for the actor aspect. If we exclude the lowest results of Pace/Narration aspect, the margin between the results for the actor class and the other aspects is not so neat.

	Valence Frequency =1 average	Arousal Frequency =1 average	Dominance Frequency =1 average
Authors	**0.26**	**0.24**	**0.25**
Characters	0.05	0.13	0.12
Form	0.17	0.16	0.17
Interest	0.21	0.19	0.21
Narration	0.18	0.17	0.17
Style	0.20	0.17	0.18

Table 31 -[Book] level of valence-arousal-dominance of words per aspect with frequency = 1

For book reviews instead (Table 31), Style is no longer the highest for VAD levels but Authors. This could lead us to the conclusion that every time there is a relation between reviewer and another human being (in the person of the author for the book domain, and in the person of the actor/director for the movies), the VAD levels – especially arousal and dominance – are higher than the other aspects. This is not a surprise: at the beginning of the chapter we saw that this lexicon has been also used to investigate the psychological relation between two people via specific words that bring with them specific levels of VAD.

Finally the question is: can we use the VAD values as features for a classification system?

Experiment C: Valence Arousal Dominance values as SVM features

Finally, I decided to apply the VAD values to the previous classification system (Case 8 - Table 24) using them as features.

First I put in the system only the numerical values of the VAD without any other features (i.e. – words). In other words, I used the numeric value from the VAD of each word in the reviews but I did not introduce the words of the reviews. The results were really low and, referring to the test we did before it was expected that in this way the VAD alone can not define the aspects.

Then I decided to merge together the VAD values with the previous classification features. As a reminder, for the movies: the features of the previous classification system were the aspects of the movie with its general attribute. (Table 24)

Unfortunately results concerning the Movie domain (Table 33) have not always benefited from this approach and in certain cases they made the classification worse, i.e. we have worst results for Actor, Pace/Narration, Plot, Subject, Music. Table 35 shows that for Book domain VAD can improve a little the results even when the sample is smaller than the one seen for the Movies.

Experiment C – Results

Case 8	Precision	Recall	F1-score	Support
Actor	0.42	0.67	**0.55**	226
Director	0.41	0.46	*0.43*	183
Characters	0.34	0.21	0.24	61
Pace/Narration	0.57	0.40	*0.50*	84
Plot	0.41	0.23	0.38	102
Subject	0.17	0.05	0.16	39
Music	0.25	0.09	0.16	53
Video	0.45	0.41	*0.25*	61
Avg / total	0.40	0.42	0.39	809

Table 32 – [Movie] Classification System using SVM and the words for each aspect as features.

	Precision	Recall	F1-score	Support
Actor	0.42	0.64	0.51	226
Director	0.41	0.49	*0.45*	183
Characters	0.32	0.20	0.24	61
Pace/Narration	0.53	0.37	0.44	84
Plot	0.37	0.25	0.29	102
Subject	0.21	0.08	0.11	39
Music	0.24	0.08	0.11	53
Video	0.41	0.34	*0.37*	61
Avg / total	0.39	0.41	0.39	809

Table 33 – [Movie_VAD] Classification System using SVM and the words for each aspect + Valence Arousal Dominance values as features

	Precision	Recall	F1-Score	Support
Authors	0.53	0.42	*0.47*	38
Characters	0.47	0.27	0.35	33
Form	0.50	0.17	0.25	12
Interest	0.56	0.69	**0.62**	237
Narration	0.34	0.22	0.27	89
Style	0.49	0.52	*0.50*	111
Avg / total	0.50	0.52	0.50	520

Table 34 –[Book] Classification System using SVM and the words for each aspect as features.

	Precision	Recall	F1-Score	Support
Authors	0.46	0.42	*0.44*	38
Characters	0.41	0.21	0.28	33
Form	0.67	0.17	0.27	12
Interest	0.57	0.70	**0.63**	237
Narration	0.31	0.20	0.24	89
Style	0.50	0.51	*0.51*	111
Avg / total	0.49	0.51	0.49	520

Table 35 – [Book_VAD] Classification System using SVM and the words for each aspect + Valence Arousal Dominance values as features

Conclusions

At present, probably due to the preliminary character of this study, the VAD lexicon seems not be of much help for the system. This could be explained by several reasons: lack of data, lack of words inside VAD that are expressly conceived for the domain, lack of further pre-processing on hapax that could give the right weight to some key words.

In particular I, maybe naively, chose to take into consideration the hapax in this way to A) speed up the process, and B) because at present I do not have enough words to truly represent specific words for the domain. In other words, some hapax found such as

- “protagoniste” [eng. Protagonist] (Aspect=Character),
- “maitre de la 7ème art” [eng. Seventh art Master] (Aspect=Director)

seem to be more tailored to some specific aspects than other words – always hapax in our corpus – such as:

- “(musique) apocalyptique” [eng. Apocalyptic (music)]

- “pathetique” [eng. Pathetic] (Aspect=Subject)

As a matter of facts with a larger corpus we should be able to find “apocalyptic image” and “pathetic actor” that may reveal the non-specificity character of these words.

One solution to create a better system may be to enlarge the corpus while storing the new hapax up to a taxonomy and deleting the words shared between aspects.

Another solution to preserve specific words may be using some language corpus management and query system such as Sketch Engine[17] to discover the co-occurrences of two words when one word is an entity such as “actor”, “director”, etc.

Another critical point of these experiments has been the lack of syntactic references. To better explain this I will give you an example.

In VAC lexicon we have both “master” and “puppet” words. They have obviously their specific VAD values. In a sentence such as “the actor is a puppet” or “this director is a master of direction”, the sum of the VAD values may give us a good result compared to the reality of things. Things can turn difficult in case of sentences such as “(the director is a) master of puppets” or “(the) absolute 7th art Master”. Here we will not be able to capture the implicit opinions about the actors (which are only puppets) and the ability of the director (that has to handle bad actors) in the first sentence; in the second one we will not appreciate again the movie director considered like a master but also like the absolute master of the movie industry.

In addition to this, some words which are very important for these type of reviews are still missing (for example the word “navet” [eng. Turnip]).

Anyway the work is only at the initial stage: I hope for the future to find a way to weigh specifically some words that seem important for each aspect, to add more words from other lexicons. In particular I would like to add (Vincze, et al., 2011) lexicons that measure not only the VAD levels – that we see that can be of help for describing the human related type aspects, but also abstraction and concreteness levels which I expect to be of help for the media aspects where the imaginary is very strong. In addition to this, it could be interesting to use the Valence values with the polarity values expressed in our annotated data in order to distinguish positive and negative reviews.

[17] https://www.sketchengine.eu/

SUMMARY OF CHAPTER 5

In this chapter we resume the discussion between emotion and sentiment that we had addressed in the introduction. Thanks to a collaboration with the Canadian National Research Council, I had the opportunity to test a lexicon that is linguistically and culturally independent. This lexicon is called NRC VAD lexicon: Valence, Arousal, Dominance lexicon.

It is used to detect levels of these three parameters between a person and another person, or situation. It can be used to understand how people react to a stimulus.

In the case of the thesis it was used to detect different levels of the properties of interest between the two studied domains, and to answer the question that had already been asked during chapter 3: movie reviewers are different from the book ones? For the moment it seems that there aren't any substantial differences.

Another research question was: is it possible to evaluate the reviewer-aspect relationship? In order to answer this question we tested the VAD levels by aspect.

For movies, VAD levels are higher when the aspect involves the presence of a person (actor or director).

For books, style and interest aspects have the highest levels of VAD. Then we had tested words with frequency equals to 1 for each aspect. For movies we did not remark many differences compared to the previous test, for books the highest level of VAD has been remarked for the authors – we could think that there is a relation between reviewer and aspect describing something human.

In addition to this, the lexicon was used without success as a feature in the SVM classification system.

In this chapter we have shown a preliminary test and the work will be improved in the future by trying to check if all the words very relevant for the classification are well represented, by adding words that - in a lexicon which is translated from English - do not necessarily find exact matches, such as "bouquin" (in eng. "tiny book"), "bouquin de gare" (in eng. "tiny station book") or "navet (in eng. "turnip").

Finally, it would also be interesting to test the polarity using the Valence parameter of the lexicon.

CHAPTER 6 - CONCLUSIONS

We have discussed the importance of Sentiment and Emotion analysis not only for the research community but also for people's and enterprises' everyday life. We know that it is hard to define the boundaries between sentiment and emotion and that it is important in different contexts.

These facets of sentiment analysis have taken different names through the years in the research community: sentiment analysis, opinion mining, subjectivity analysis, etc. This shows us the complexity of the task, but also the interest in analysing better this problem and the will to find a solution. Moreover, today's technology and big data give us the chance to carry out more and more experiments. Unfortunately, this is true especially for English, but not for French, that lacks of some resources.

Scrolling through the pages of this thesis we have discovered the vastness and richness of this field, both for the work done and the types of approach: lexicon-based, corpus based, machine learning and deep learning.

During these three years we have worked for the French community, developing many resources that are now freely available for the future research: lexicons, datasets and a whole corpus used for Aspect-Based Sentiment Analysis. We have carried out many experimentations and benefited from the complexity of Sentiment Analysis.

First, we used only a statistical approach for polarity classification for hotel, movie and book reviews, and we obtained very good results. This is true especially for the hotel domain, thanks to its linearity.

Afterwards, we decided to focus on the two most difficult domains: movies and books. In general, we have observed that for polarity detection, we can obtain a good result using statistics, linguistics and syntax. That was the case when we proposed a logistic regression classifier with the use of our lexicons and syntactic information from a shallow parser. However, we experienced some problems in separating and distinguishing, especially in long reviews, the opinion part from the description part. Unfortunately, some words were shared between both parts.

For this reason we decided to isolate the opinion part and we labelled it as an attribute of a product.

In NLP community this is called Aspect-Based Sentiment Analysis (ABSA). For French it seems that a resource describing movie and book using ABSA was not readily available. We did an extensive effort in creating this resource. We annotated book and movie reviews, retrieving aspects and attributes from each review. We knew that it was something useful for the research community: in recent years competitions such as SemEval have been using this kind of data to evaluate new classification systems.

Unluckily, we did not have enough data to test the potentiality of deep learning systems. We decided to use several traditional machine learning methods to classify aspects in our reviews: Naïve Bayes, decision trees, logistic regression, hierarchical SVM. The best approach was obtained using SVM.
The final conclusion is that a simple SVM system or a combination of other machine learning systems can not reach as optimal results as the ones we had during the phase of polarity detection.
This is true also when we use very precise annotations and we evaluate only sentences with just one type of aspect. Although we have normalized the samples for each aspect, the most general classes absorbed always the specific ones. The more classes are involved, the more specific is the system and the result is an increased number of misclassified examples.

The use of decision trees + SVM in a hierarchical classification system was not enough to have a high F1-score.
We should consider for the future finding a way to better weigh some words that are specific for the aspect, and a way to handle a system that can be fed using this kind of words.

During the last months of this thesis, thanks to a collaboration between research centres, I started a psychological analysis of the reviews and their aspects via a lexicon that measures precisely this aspect. It is a preliminary study and at present we know only that there is a difference among some aspects for each domain. More has to be done in order to highlight

some specific words and explain the psychological reasons behind the choice of these words to judge an aspect. I find this analysis interesting and innovative because some recent papers have underlined how the use of language is directly connected to the emotional state of the person choosing and pronouncing certain words. (Mehl, et al., 2017) have shown that it is possible to identify some markers (most of them are adverbs and pronouns) that are connected to the stress level of someone. In other words, the choice of our words – but not every type of part-of-speech - can unveil how we are experiencing some situation: from the joy to have read a very good book, to the stress of writing a thesis.

Finally, we hope that this thesis has given its contribution to the NLP community by describing the issues concerning Sentiment Analysis, Emotion Analysis and ABSA. In addition to this, we have created many resources (corpora, lexicons, etc.) for French and we have tested them in different classification systems. The results have confirmed that the task is really complex, especially for long types of review, which are not always taken into consideration - even in international competitions.

For the future, I think that it could be of great interest to continue the work presented in chapter 5, optimising the use of the psychological lexicon – especially for the dominance and arousal parameters. Moreover, it could be interesting to insert new dimensions such as abstraction/concreteness to evaluate aspects evoking imagination as the Media aspects.

We hope that our whole work and the last chapter can open the way to other works analysing which emotion lies behind the words we use every day.

BIBLIOGRAPHY

Alvarez-Lopez, T., et al. 2017. A book reviews dataset for aspect-based sentiment analysis. *In 8th Language Technology Conference.* 2017.

Apidianaki, Marianna. 2016. Vector-space models for ppdb paraphrase ranking in context. *In Proceedings of EMNLP. .* 2016, pp. 2028–2034.

Baccianella, S., Esuli, A. et Sebastiani, F. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *In LREC.* 2010.

Bellegarda, J. 2010. Emotion analysis using latent affective folding and embedding. *In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.* 2010.

Berisha, V. et Hero, A.O. 2015. Empirical Non-Parametric Estimation of the Fisher Information (2015). *IEEE Signal Processing Letters .* 2015, Vol. 22, 7, pp. 988 - 992 .

Berisha, V., et al. 2016. Empirically estimable classification bounds based on a nonparametric divergence measure. *IEEE Transactions on Signal Processing.* 2016, Vol. 64, 3, pp. 580 - 591 .

Bespalov, Dmitriy, et al. 2011. Sentiment classification based on supervised latent n-gram analysis. *In: Proceedings of the 20th ACM international conference on Information and knowledge management.* 2011, pp. 375-382.

Bhayani, Go A. R. et Huang, L. 2009. Twitter Sentiment. *Stanford Digital Library Technologies Project .* 2009.

Bolton, Ruth N., et al. 2013. Understanding Generation Y and their use of social media: a review and research agenda. 2013, Vol. 24, 3, pp. 245-267.

Bradley, Margaret M. et Lang, Peter J. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. *Technical report C-1, the center for research in psychophysiology.* 1999.

Breck, Eric et Cardie, Claire. 2017. Opinion Mining and Sentiment Analysis. [auteur du livre] Ruslan Mitkov. [éd.] Ruslan Mitkov. *The Oxford Handbook of Computational Linguistics.* 2017.

Breiman, L., et al. 1984. *Classification and Regression Trees.* s.l. : CRC Press, 1984.

Breiman, Leo. 2001. Random forests. *Machine learning.* 2001, Vol. 45, 1, pp. 5-32.

Brody, S. et Elhadad, N. 2010. An unsupervised aspect-sentiment model for online reviews. *In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT 10.* 2010.

Chen, CL Philip et Zhang, Chun-Yang. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences.* 2014, Vol. 275, pp. 314-347.

Chen, P., et al. 2017. Recurrent attention network on memory for aspect sentiment analysis. *In EMNLP 2017.* 2017, pp. 463–472.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960, Vol. 20, 1.

Collomb, A. 2008. Mining Opinions in Comparative Sentences. *In Proceedings of the 22nd International Conference on Computational Linguistics - COLING08.* 2008, pp. 241–248.

Cortes, Corinna et Vapnik, Vladimir. 1995. Support-vector networks. *Machine learning.* 1995, Vol. 20, 3, pp. 273-297.

Das, S. et Chen, M. 2001. Yahoo! for amazon: Extracting market sentiment from stock message boards. *In Proceedings of the Asia-Pacific Finance Association Annual Conference (APFA).* 2001.

Das, Sanjiv R. et Chen, Mike Y. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management science.* 2007, Vol. 53, 9, pp. 1375-1388.

Deerwester, S., et al. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science.* 1990, Vol. 41, 6, pp. 391-407.

Dong, Li et al., et. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. *In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* 2014, pp. 49-54.

Dreyfus, S. 1973. The computational solution of optimal control problems with time lag. *IEEE Transactions on Automatic Control.* 1973, Vol. 18, 4.

Ervin, S. 1964. An analysis of the interaction of language, topic, and listener. [éd.] In John Gumperz and Dell Hymes (eds.). *The Ethnography of Communication, special issue of American Anthropologist,.* 1964, Vol. 66, 2, pp. 86-102.

Esuli, A. et Sebastiani, F. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *LREC.* 2006.

Esuli, Andrea et Sebastiani, Fabrizio. 2006. Determining term subjectivity and term orientation for opinion mining. *In: 11th Conference of the European Chapter of the Association for Computational Linguistics.* 2006.

Fahrni, Angela et Klenner, Manfred. 2008. Old wine or warm beer: Target-specific sentiment analysis of adjectives. *In: Proc. of the Symposium on Affective Language in Human and Machine,.* 2008, pp. 60-63.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database.* s.l. : Bradford Books, 1998.

Ferrand, Ludovic et Alario, François-Xavier. 1998. Normes d'associations verbales pour 366 noms d'objets concrets. *L'Année psychologique.* 1998, Vol. 98, 4, pp. 659-709.

Ferrand, Ludovic. 2001. Normes d'associations verbales pour 260 mots «abstraits». *L'Année psychologique.* 2001, Vol. 101, 4, pp. 683-721.

Forman, George. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research.* 2003, pp. 1289-1305.

Francisco, Virginia et Gervas, Pablo. 2006. Exploring the compositionality of emotions in text: Word emotions, sentence emotions and automated tagging. *AAAI-06 workshop on computational aesthetics: Artificial intelligence approaches to beauty and happiness.* 2006.

Fukushima, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics.* 1980, Vol. 36, 4, pp. 193-202.

Ganu, G., Elhadad, N. et Marian, A. 2009. Beyond the stars: Improving rating predictions using review text content. . *In WebDB.* 2009.

Ghosh, Aniruddha, et al. 2015. SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. 2015.

Goodfellow, I., Bengio, Y. et Courville, A. 2016. *Deep learning.* Cambridge : MIT press, 2016.

Graziotin, D., Wang, X. et Abrahamsson, P. 2015. Do feelings matter? On the correlation of affects and the self-assessed productivity in software engineering. *Journal of Software: Evolution and Process.* 2015, Vol. 27, 7, pp. 467-487.

—. 2015. Understanding the affect of developers: theoretical background and guidelines for psychoempirical software engineering. *In Proceedings of the 7th International Workshop on Social Software Engineering - SSE 2015.* 2015.

Grosjean, F. 2010. Personality, thinking and dreaming, and emotions in bilinguals. *Bilingual: Life and Reality. Cambridge, Mass: Harvard University Press.* 2010, p. Chapter 11.

Grossberg. 1973. Contour enhancement short-term memory and constancies in reverberating neural networks. *Studies in Applied Mathematics.* 1973, Vol. 52, pp. 213-257.

Guan, Ziyu, et al. 2016. Weakly-Supervised Deep Learning for Customer Review Sentiment Classification. *In: IJCAI.* 2016, pp. 3719-3725.

Guyon, Isabelle et Elisseeff, André. 2003. An introduction to variable and feature selection. *Journal of machine learning research.* 2003, pp. 1157-1182.

Hamdan, H., Bellot, P. et Bechet, F. 2016. Sentiment analysis in scholarly book reviews. 2016.

Hand, D. J. et Yu, K. 2001. Idiot's Bayes—not so stupid after all? *International statistical review.* 2001, Vol. 69, 3, pp. 385-398.

Harrington, P. 2012. *Machine Learning in Action.* s.l. : Manning Ed., 2012.

Hatzivassiloglou, Vasileios et Mckeown, Kathleen R. 1997. Predicting the semantic orientation of adjectives. *In: Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics.* 1997, pp. 174-181.

Hickey, R. 2005. *Legacies of colonial English: Studies in transported dialects.* s.l. : Cambridge University Press, 2005.

Hogenraad, R., Bestgen, Y., Nysten, J.L. 1995. Terrorist Rhetoric : Texture and Architecture. [auteur du livre] E. Nissan et K.M. Schmidt. *Information knowledge, intellect.* 1995, pp. 48-59.

Hogenraad, Robert et Orianne, Etienne. 1981. Valences d'imagerie de 1.130 noms de la langue française parlée. *Psychologica Belgica.* 1981.

Hopfield, J.J. 1982. Neural networks and physical systems with emergent collective computational abilities. *In Proceedings of the National Academy of Sciences of the USA.* 1982, Vol. 79, 8, pp. 2554–2558.

Hovy, Eduard H. 2015. What are Sentiment, Affect, and Emotion? [auteur du livre] N. Gala et al. [éd.] N. Gala et al. *Language Production, Cognition, and the Lexicon.* Switzerland : Springer International Publishing, 2015, Vol. Text, Speech and Language Technology 48, pp. 13-24.

Hu, Minqing et Liu, Bing. 2004. Mining and summarizing customer reviews. *n: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* 2004, pp. 168-177.

—. 2004. Mining opinion features in customer reviews. *In: AAAI.* 2004, pp. 755-760.

Huettner, A. et Subasic, P. 2000. Fuzzy typing for document management. *ACL 2000 Companion Volume, Tutorial Abstracts and Demonstration.* 2000, pp. 26-27.

Jindal, Nitin et Liu, Bing. 2006. Identifying comparative sentences in text documents. *In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2006, pp. 244-251.

Kanayama, Hiroshi et Nasukawa, Tetsuya. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. *In Proceedings of EMNLP'06.* 2006, pp. 355–363.

Khan, I.A, Brinkman, W.P. et Hierons, R.M. 2011. Do moods affect programmers' debug performance? *Cognition, Technology & Work.* 2011, Vol. 13, 4, pp. 245–258.

Kim, S.-M et Hovy, E. 2005. Automatic detection of opinion bearing words and sentences. *n Companion Volume to the Proceeding of the International Joint Conference on Natural Language Processing.* 2005.

Kim, S.-M. et Hovy, E. 2004. Determining the sentiment of opinions. *In COLING. Association for Computational Linguistics.* 2004.

Labutov, Igor et Lipson, Hod. 2013. Re-embedding words. *In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.* 2013, pp. 489-493.

Lafferty, John, Mccallum, Andrew et Pereira, Fernando CN. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Lane, R. D., Chua, P. M.-L. et Dolan, R. J. 1999. Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. *Neuropsychologia.* 1999, Vol. 37, 9, pp. 989–997.

Lark, Joseph, Morin, Emmanuel et Saldarriaga, Sebastián Peña. 2015. CANÉPHORE: un corpus français pour la fouille d'opinion ciblée. *TALN 2015.* 2015.

LeCun, Y., et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE.* 1998, Vol. 86, 11, pp. 2278-2324.

LH Vo, M et al., et. 2009. The berlin affective word list reloaded (bawl-r). *Behavior research methods.* 2009, pp. 534–538.

Linnainmaa, Seppo. 1976. Taylor expansion of the accumulated rounding error. *BIT Numerical Mathematics.* 1976, Vol. 16, 2, pp. 146-160.

—. 1970. *The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. Master's Thesis.* Univ. Helsinki : s.n., 1970.

Liu, B. 2012. Sentiment Analysis and opinion mining. *Synth. Lect. Human Lang. Technol.* 2012.

—. 2010. Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing.* Chapman & Hall/CRC Machine Learning & Pattern Recognition, second edition, 2010.

Liu, Bing. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies.* 2012, Vol. 5, 1, pp. 1-167.

Liu, H., Lieberman, H. et and Selker, T. 2003. A model of textual affect sensing using real-world knowledge. *In Proceedings of the 8th international conference on Intelligent user interfaces.* 2003, pp. 125-132.

Louviere, J.J, Flynn, T.N. et Marley, A. A. J. 2015. Best-Worst Scaling: Theory, Methods and Applications.. Cambridge University Press, 2015.

Louviere, J.J. 1991. Best-worst scaling: A model for the largest difference judgments. Working Paper. . 1991.

Luna, D., Ringberg, T. et Peracchio, L. 2008. One individual, two identities: Frame switching among biculturals. *Journal of Consumer Research.* 2008, Vol. 35, 2, pp. 279-293.

Maas, Andrew L., et al. 2011. Learning word vectors for sentiment analysis. *In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume.* 2011, p. 14.

Malt Barbara C., Majid Asifa. 2013. How thought is mapped into words. *WIREs Cogn Sci.* 2013, pp. 583-597.

Manyika, J., et al. 2011. Big data: The next frontier for innovation, competition, and productivity. 2011.

Marvin, Minsky et Seymour, A. Papert. 1969. *Perceptrons.* s.l. : MIT Press, 1969.

McDonald, R., et al. 2007. Structured models for fine-to-coarse sentiment analysis. *In Annual Meeting-Association For Computational Linguistics.* 2007, Vol. 45, p. 432.

Mehl, Matthias R., et al. 2017. Natural language indicators of differential gene regulation in the human immune system. *In PNAS 2017.* 2017, Vol. 114, 47, pp. 12554-12559.

Mehrabian, A. 1980. *Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies.* Gunn & Hain, Cambridge, MA. . s.l. : Gunn & Hain, Cambridge, MA. , 1980.

Miao, Q., Li, Q. et Dai, R. 2009. AMAZING: a sentiment mining and retrieval system. *Expert System Application.* 2009, 36, pp. 7192-7198.

Mikolov, Tomas, et al. 2013. Distributed representations of words and phrases and their compositionality. *In: Advances in neural information processing systems.* 2013, pp. 3111-3119.

Miller, G. A., et al. 1990. Introduction to WordNet: An on-line lexical database. *International journal of lexicography.* 1990, Vol. 3, 4, pp. 235-244.

Mitchell, Tom M. 1997. *Machine Learning.* s.l. : WCB/McGraw-Hill, 1997.

Mohammad, S. et Yang, T. 2011. Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. *In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis.* 2011, pp. 70-79.

Mohammad, S. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.* 2018, Vol. 1.

Mohammad, Saif M. 2012. # Emotional tweets. *In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation.* Association for Computational Linguistics, 2012, pp. 246-255.

Mohammad, Saif M. et Bravo-Marquez, Felipe. 2017. WASSA2017 shared task on emotion intensity. *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.* 2017, pp. 34-49.

Mohammad, Saif M. et Turney, Peter D. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. *In: Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text.* 2010, pp. 26-34.

Mohammad, Saif M. et Yang, Tony Wenda. 2011. Tracking sentiment in mail: How genders differ on emotional axes. *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis.* 2011, pp. 70-79.

Mohammad, Saif M. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion measurement.* s.l. : Elsevier, 2016, pp. 201-237.

Mohammad, Saif M., Kiritchenko, Svetlana et Zhu, Xiaodan. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242.* 2013.

Mohammad, Saif, Dunne, Cody et Dorr, Bonnie. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. *In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.* 2009, Vol. 2, pp. 599-608.

Moors, A. et al, et. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 dutch words. 2013.

Moraes, Rodrigo, Valiati, JoãO Francisco et Neto., Wilson P. GaviãO. 2013. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Syst. Appl.* 2013, Vol. 40, 2, pp. 621-633.

Mullen, Tony et Collier, Nigel. 2004. Sentiment analysis using support vector machines with diverse information sources. *In: Proceedings of the 2004 conference on empirical methods in natural language processing.* 2004.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. *In Proceedings of the 10th international workshop on semantic evaluation.* 2016, pp. 1-18.

Narayanan, R., Liu, B. et Choudhary, A. 2009. Sentiment analysis of conditional sentences. *In Proceeding of the 2009 Conference on Empirical Methods in Natural Language Processing.* 2009, pp. 180-189.

Osgood, C.E et Suci, G., Tannenbaum, P. 1957. The measurement of meaning. . University of Illinois Press., 1957.

Pak, Alexander et Paroubek, Patrick. 2010. Twitter as a corpus for sentiment analysis and opinion mining. *LREC.* 2010, pp. 1320-1326.

Pang B. Lee, L. et Vaithyanathan, S. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *EMNLP'02: Proc. Conf. on Empirical Methods in Natural Language Processing.* 2002, pp. 79-86.

Pang, Bo et Lee, Lillian. 2008. Opinion Mining and Sentiment Analysis. *Information Retrieval.* Foundation and Trends, 2008, Vol. 2, 1-2, pp. 1--135.

Pang, Bo, Lee, Lillian et Vaithyanathan, Shivakumar. 2002. Thumbs up?: sentiment classification using machine learning techniques. *In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing.* Association for Computational Linguistics, 2002, Vol. 10, pp. 79-86.

Pecòre, S. et Villaneau, J. 2018. Complex and Precise Movie and Book Annotations in French Language for Aspect Based Sentiment Analysis. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* 2018.

Pecòre, S. et Villaneau, J., Said, F. 2016. Combiner lexique et régression logistique dans la classification d'avis laissés sur le Net : une étude de cas . *TALN-JEP-RECITAL.* 2016.

Pennington, Jeffrey, Socher, Richard et Manning, Christopher. 2014. Glove: Global vectors for word representation. *In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).* 2014, pp. 1532-1543.

Pontiki, M., et al. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. *In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014),.* 2014, pp. 27-35.

Pontiki, M., et al. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. *In Proceedings of the 9th International Workshop on Semantic Evaluation.* 2015, pp. 486-495.

Pontiki, Maria et al., et. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. *In: Proceedings of the 10th international workshop on semantic evaluation.* 2016, pp. 19-30.

Pontiki, Maria, et al. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. *In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016).* 2016, pp. 19-30.

Price, Richard. 1763. *An Essay towards solving a problem in the doctrine of chances.* 1763.

Quinlan, JR. 1986. *Induction of decision trees. Machine Learning.* 1986.

—. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

Rabelo, J.C.B., Prudêncio, R.B.C. et Barros, F.A. 2012. Using link structure to infer opinions in social networks. *IEEE International Conference on Systems, Man, and Cybernetics.* 2012.

Read, Jonathon et Carroll, John. 2009. Weakly supervised techniques for domain-independent sentiment classification. *In: Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion.* 2009, pp. 45-52.

Recupero, Diego Reforgiato et Cambria, Erik. 2014. Eswc'14 challenge on concept-level sentiment analysis. *In: Semantic Web Evaluation Challenge.* 2014, pp. 3-20.

Redondo, J., et al. 2007. The spanish adaptation of anew (affective norms for english words). . *Behavior research methods.* 2007, Vol. 39, 3, pp. 600-605.

Rosenblatt, F. 1957. The Perceptron a perceiving and recognizing automation. *Report Cornell Aeronautical Lab.* 1957.

Rosenthal, S., et al. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. *In Proceedings of the 9th international workshop on semantic evaluation.* 2015, pp. 451-463.

Rosset, S., et al. 2008. The LIMSI participation to the QAst track. *Actes de Working Notes of CLEF.* 2008.

Russell, J.A. 2003. Core affect and the psychological construction of emotion. *Psychological review.* 2003.

—. 1991. Culture and the categorization of emotions. *Psychological Bulletin.* 1991, Vol. 110, 3, pp. 426–450.

Sajous, Franck, Hathout, Nabil et Calderone, Basilio. 2013. Glàff, un gros lexique à tout faire du français. *In: Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013).* 2013, pp. 285--298.

Saussure, Ferdinand de. 1916. *Cours de linguistique générale.* [éd.] C. Bally et A. Sechehaye. Illinois : s.n., 1916.

Schmid, Helmut. 2013. Probabilistic part-ofispeech tagging using decision trees. *In: New methods in language processing.* 2013.

Seni, G. et Elder, J. 2010. *Ensemble methods in Data Mining.* 2010.

Siegle, G. 1994. 1994. http ://www.sci.sdsu.edu/cal/wordlist/origwordlist.html. .

Sorgente, A., Vettigli, G. et Mele, F. 2014. An italian corpus for Aspect-based sentiment analysis of movie reviews. *CLICIT.* 2014.

Strapparava, Carlo et al., et. 2004. Wordnet affect: an affective extension of wordnet. *LREC.* 2004, pp. 1083-1086.

Syssau, Arielle et Font, Noëlle. 2005. Evaluations of the Emotional Characteristics of a Set of 604 words. *Bulletin de psychologie.* 2005, Vol. 3, pp. 361-367.

Thet, T. T., Na, J.-C. et Khoo, C. S. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *J. Inf. Sci.* 2010, Vol. 36, 6.

Turney, Peter D. et Littman, Michael L. 2003. 2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information System, 2003, Vol. 21, 4, pp. 315–346.

Turney, Peter D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. 2002, pp. 417-424.

Turney, Peter. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *In: Proceedings of the 40th annual meeting on association for computational linguistics.* Association for Computational Linguistics, 2002, pp. 417-424.

van den Burg, Gerrit J.J et Hero, Alfred O. 2017. Fast Meta-Learning for Adaptive Hierarchical Classifier Design. *arXiv preprint arXiv:1711.03512.* 2017.

Villaneau, J. Pecòre, S., Said, F., et Marteau, P.F. 2019. Aspect Based Sentiment Analysis for Book reviews: Shallow Parsing Combined with Statistical Methods. ECG 2019.

Villaneau, J., Said, F. et Pecòre, S. 2018. Aspect Detection in Book Reviews: Experimentations. *NL4AI@AI*IA.* 2018, pp. 16-27.

Vincent, Marc et Winterstein, Grégoire. 2013. Building and exploiting a French corpus for sentiment analysis. Construction et exploitation d'un corpus français pour l'analyse de sentiment. *Proceedings of TALN 2013.* 2013, Vol. 2, 76.

Vincze, N. et Bestgen, Y. 2011. Une procédure automatique pour étendre des normes lexicales par l'analyse des cooccurrences dans des textes". *"Ressources linguistiques libres" de la revue Traitement Automatique des Langues.* 2011, Vol. 53, 3.

Vincze, Nadja Et Bestgen, Yves. 2011. Identification de mots germes pour la construction d'un lexique de valence au moyen d'une procédure supervisée. *Actes de TALN11.* 2011, Vol. 1, pp. 223-234.

Vincze, Nadja et Bestgen, Yves. 2011. Identification de mots germes pour la construction d'un lexique de valence au moyen d'une procédure supervisée. *Actes de TALN11.* 2011, pp. 223-234.

Wang, Xingyou, Jiang, Weijie et Luo, Zhiyong. 2016. Combination of convolutional and recurrent neural network for sentiment analysis of short texts. . *In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers.* 2016.

Warriner, A.B., Kuperman, V. et Brysbaert, M. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods. 2013, Vol. 45, 4, pp. 1191–1207.

Werbos, P.J. 1974. *Beyond regression: new tools for prediction and analysis in the behavioral sciences. Ph.D. thesis.* Harward : Harward University, 1974.

Widlocher, Antoine et Mathet, Yann. 2012. The glozz platform: A corpus annotation and mining tool. *In: Proceedings of the 2012 ACM symposium on Document engineering.* 2012, pp. 171-180.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. 2004. Learning subjective language. *Computational linguistics.* MIT Press, 2004, Vol. 30, 3, pp. 277-308.

Wiebe, Janyce et Riloff, Ellen. 2005. Creating subjective and objective sentence classifiers from unannotated texts. *In: International Conference on Intelligent Text Processing and Computational Linguistics.* Springer, 2005, pp. 486-497.

Wiebe, Janyce. 2000. Learning subjective adjectives from corpora. *Aaai/iaai.* 2000, Vol. 20.

Wilson, T., et al. 2005. OpinionFinder: A system for subjectivity analysis. *In Proceedings of hlt/emnlp on interactive demonstrations.* Association for Computational Linguistics., 2005, Vol. 34-35.

Wilson, T., Wiebe, J. et Hoffmann, P. 2005. Recognizing contextual polarity in phrase level sentiment analysis. *In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP).* 2005, pp. 347–354.

Wilson, Theresa et al., et. 2005. OpinionFinder: A system for subjectivity analysis. *n Proceedings of hlt/emnlp on interactive demonstrations .* 2005, pp. 34-35.

Yi, J. Nasukawa, T., Bunescu, R. et Niblack, W. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. *Third IEEE International Conference on Data Mining.* 2003, p. 427.

Yu, H. et Hatzivassiloglou, V. 2003. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. *In Proceedings of the conference on Empirical methods in natural language processing EMNLP 2003.* 2003, pp. 129–136.

Yussopova, N., et al. 2015. A Decision Support Approach based on Sentiment Analysis Combined with Data Mining for Customer Satisfaction Research. *The International Journal on Advances in Intelligent Systems.* 2015.

Zhang, L., Wang, S. et Liu, B. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.* 2018.

Zhao, Zhou et al., et. 2017. Community-Based Question Answering via Asymmetric Multi-Faceted Ranking Network Learning. *In: AAAI. 2017.* 2017, pp. 3532-3539.

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

Titre : Analyse des sentiments et des émotions de commentaires complexes en langue française.

Mots clés : *fouille d'opinion, analyse des sentiments, apprentissage automatique, analyse des émotions, traitement automatique du langage*

Les définitions des mots « sentiment », « opinion » et « émotion » sont toujours très vagues comme l'atteste aussi le dictionnaire qui semble expliquer un mot en utilisant le deux autres. Tout le monde est affecté par les opinions : les entreprises pour vendre les produits, les gens pour les acheter et, plus en général, pour prendre des décisions, les chercheurs en intelligence artificielle pour comprendre la nature de l'être humain. Aujourd'hui on a une quantité d'information disponible jamais vue avant, mais qui résulte peu accessible. Les mégadonnées (en anglais « big data ») ne sont pas organisées, surtout pour certaines langues – dont la difficulté à les exploiter. La recherche française souffre d'une manque de ressources « prêt-à-porter » pour conduire des tests.

Cette thèse a l'objectif d'explorer la nature des sentiments et des émotions, dans le cadre du Traitement Automatique du Langage et des Corpus. Les contributions de cette thèse sont plusieurs : création de nouvelles ressources pour l'analyse du sentiment et de l'émotion, emploi et comparaison de plusieurs techniques d'apprentissage automatique, et plus important, l'étude du problème sous différents points de vue : classification des commentaires en ligne en polarité (positive et négative), Aspect-Based Sentiment Analysis des caractéristiques du produit recensé. Enfin, un étude psycholinguistique, supporté par des approches lexicales et d'apprentissage automatique, sur le rapport entre qui juge et l'objet jugé.

Title : Sentiment and emotion analysis of complex reviews

Keywords: *sentiment analysis, emotion analysis, opinion mining, machine learning, natural language processing*

"Sentiment", "opinion" and "emotion" are words really vaguely defined; not even the dictionary seems to be of any help, being it the first to define each of the three by using the remaining two. And yet, the civilised world is heavily affected by opinions: companies need them to understand how to sell their products; people use them to buy the most fitting product and, more generally, to weigh their decisions; researchers exploit them in Artificial Intelligence studies to understand the nature of the human being. Today we can count on a humongous amount of available information, though it's hard to use it. In fact, the so-called "Big data" are not always structured – especially for certain languages. French research suffers from a lack of readily available resources for tests. In the context of Natural Language Processing, this thesis aims to explore the nature of sentiment and emotion. Some o four contributions to the NLP research Community are: creation of new ressources for sentiment and emotion analysis, tests and comparisons of several machine learning methods to study the problem from different points of view - classification of online reviews using sentiment polarity, classification of product characteristics using Aspect-Based Sentiment Analysis. Finally, a psycholinguistic study - supported by a machine learning and lexical approaches – on the relation between who judges, the reviewer, and the object that has been judged, the product