N° d'ordre :

N° de série :

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE ECHAHID HAMMA LAKHDAR - EL OUED

FACULTÉ DES SCIENCES EXACTES

Mémoire de Fin D'étude
Présenté pour l'obtention du Diplôme de

MASTER ACADEMIQUE

Domaine Mathématiques et Informatique

Filière : Informatique

Spécialité : Systèmes Distribués et Intelligence Artificielle

Thème

La détection automatique du discours abusif, offensant et obscène dans le dialecte algérien

Présenté par :

MANSOUR Abd Elfettah
TRAD Aissa

Encadrant :

NAGOUDI El Moatez Bellah

Soutenue le : 04-05-2018, Devant le jury :

- M : BOUCHRIT Ammar
- M : BELLILA Khaoula

Année Universitaire : 2017/2018

Remerciements

Nous tenons à remercier du fond du cœur, avant tous, «ALLAH» le tout puissant de mes avoir donné la force, la volonté et le courage de réaliser ce travail.

Nous tenons à remercier chaleureusement et respectivement tout ceux qu'ont contribué de prés et loin à la réalisation de ce projet de fin d'étude et nous tenons remercier notre encadrant Mr. NAGOUDI El Moatez Bellah, Maître a l'Université Echahid hamma lakhdar El-Oued, d'avoir proposé et encadré ce sujet pendant 6 moins. Nous lui exprimons notre profonde gratitude pour nous avoir fait profiter de ses connaissances, mais aussi de ses méthodes de travail, et surtout de sa rigueur scientifique.

Nous remercions les membres de jury pour avoir accepter de juger notre travail.

Notre reconnaissance va aussi à tous ceux qui ont collaboré à notre formation en particulier les enseignants du département d'Informatique, Universitaire Echahide Hamma Lakhdar El-Oued Aussi à nos collègues de la promotion 2017-2018. On remercie également tous ceux qui ont participé de près ou de loin à élaborer ce travail.

Résumé

Les réseaux sociaux représentent aujourd'hui un moyen de communication incontournable. Cependant, ils peuvent également être une source fiable pour diffuser les nouvelles et de la propagande dans les pays arabes, notamment en Algérie. Certains utilisateurs créent actuellement des comptes pour diffuser des contenus abusifs dans des publications écrites en dialecte algérien. Par ailleurs, ce langage abusif est interdit par nos cultures arabes et la détection automatique d'un tel langage est un défi majeur à surmonter. Dans ce travail nous avons utilisé l'apprentissage automatique et profond afin de proposer un système de détection automatique de langage abusif dans le dialecte algérien sur Facebook.

Mots clés : Réseaux sociaux, dialecte algérien, langage abusif, apprentissage automatique, apprentissage profond.

Abstract

Social networks are now an essential means of communication. However, they can also be a reliable source for spreading news and propaganda in Arab countries, especially in Algeria. Some users are currently creating accounts to distribute abusive content in publications written in Algerian dialect. Moreover, this abusive language is forbidden by our Arab cultures and the automatic detection of such language is a major challenge to overcome. In this work we used machine learning and deep learning to propose a system of automatic detection of abusive language in the Algerian dialect on Facebook.

Key words : Social networks, Algerian dialect, Abusive language, Machine learning, deep learning.

ملخص

الشبكات الاجتماعية هي الآن وسيلة أساسية للتواصل الاجتماعي، كما يمكن أن تكون مصدرًا موثوقًا لنشر الأخبار والدعاية في الدول العربية، وخاصة في الجزائر. يقوم بعض المستخدمين حاليًا بإنشاء حسابات لتوزيع المحتوى المسيء في المنشورات المكتوبة باللهجة الجزائرية. علاوة على ذلك، هذه اللغة المسيئة محظورة من قبل ثقافاتنا العربية والكشف التلقائي عن هذه اللغة هو تحد كبير يجب التغلب عليه. في هذا العمل، استخدمنا التعلم الآلي وكذا العميق لاقتراح نظام للكشف التلقائي عن اللغة المسيئة باللهجة الجزائرية على فيسبوك.

الكلمات المفتاحية : الشبكات الاجتماعية، اللهجة الجزائرية، اللغة المسيئة، التعلم الآلي، التعلم العميق.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Les réseaux sociaux sont un outil le plus populaire pour diffuser des pensés et partager des photos et des vidéos, etc. pour cela ils sont devenus une partie essentielle de notre vie quotidienne, bien qu'ils contiennent de nombreux avantages mais aussi ils ne sont pas sans risque, parmi ces risques de contenir des continus obscènes et offensants et dangereux, ils ne sont pas complètement contrôlés, où nous devons les battre.

Les adolescents passent la plupart de leur temps sur les différents sites de réseaux sociaux. Alors qu'ils sont les premières victimes de harcèlement et de contenu abusif ce qui peut affecter négativement leur moral et leur état mental, cependant des sites populaires tels que Twitter, Facebook et Google essaient d'offrir des outils pour filtrer le contenu pour protéger les enfants en particulier et les utilisateurs en général. De nombreux chercheurs rédigent des articles scientifiques en langues étrangères et quelques-uns en langue arabe, mais dans les dialectes et surtout le dialecte algérien il n'existe pas des efforts.

Facebook est l'une des sources les plus populaires pour diffuser des nouvelles et de la propagande dans la région maghrébine et surtout l'Algérie, contrairement à Twitter, qui est largement utilisé dans l'orient arabe. Plusieurs utilisateurs créent maintenant des comptes abusifs pour distribuer du contenu adulte dans les publications arabes, ce qui est interdit par les normes et les cultures arabes. La détection des obscénités et les contenus indécents est très utile pour protéger les enfants et pour déterminer les langages abusifs. Notre travail exploite différents algorithmes d'apprentissage automatique pour détecter les contenus abusifs en arabes avec le dialecte algérien en réalisant notre propre corpus.

Le dialecte algérien est très complexe principalement en raison de sa complexité morphologique et de la disponibilité limitée de logiciels compatibles avec la langue arabe.

Dans notre travail, nous nous concentrons sur la classification du contenu abusif des publications, statuts, commentaires, etc. (Obscène, offensant ou propres) sur Facebook en utilisant cinq algorithmes d'apprentissage automatique et profond, à savoir : Nous avons créé et comparé les résultats des cinq modèles sur notre propre corpus arabe en dialecte algérien : machine à vecteurs de support, arbre de décision, forêt d'arbres décisionnels, naïve bayésienne et finalement avec réseau neuronal convolutif. Notre meilleur classificateur archive un très bon résultat de F-mesure de 0.85.

Chapitre 1

Etat de l'art

1.1 Introduction

Les médias sociaux ce sont des moyens permettant la distribution, la diffusion ou la communication avec le partage des contenus ou des expressions, comme tout un moyen de communication le contenu peut être propre, obscène, cordial, offensif, etc. Les sites de réseaux sociaux sont soumis à une pression croissante pour s'attaquer aux langages abusifs au cours des dernières années.

La détection des obscénités et les contenus indécents est très utile pour protéger les enfants et pour déterminer les discours de haine. Plusieurs sites ont des paramètres de recherches sécurisés comme Google, Youtube, alors que le filtrage des contenus indésirables est très difficile, parce qu'il est impossible de couvrir tous les mots, particulièrement écrit dans les dialectes locaux ou dans certaines cultures, ou l'écriture des nouveaux mots par le remplacement des lettres par des chiffres (ex : O par 0), ou dans l'arabe les lettres sont écrite en latin (ex : ع par 3).

1.2 Le discours de haine

Le terme « discours de haine » doit être compris comme couvrant toutes formes d'expression qui propagent, incitent à, promeuvent ou justifient la haine raciale, la xénophobie ou d'autres formes de haine fondées sur l'intolérance ou l'obscénité. ManuCE (2016).

Qu'est-ce que le discours de haine ?

Certaines personnes se refusent à agir contre le discours de haine parce qu'elles y voient une limite inacceptable à la liberté d'expression. Pour cette raison, elles réservent le terme « discours de haine » à ses formes d'expression les plus graves, par exemple lorsque des menaces immédiates pèsent sur la vie ou la sécurité d'un individu.

Alors que, le discours de haine couvre «toutes les formes d'expression», autrement dit, non seulement le discours, mais aussi les images et les vidéos, ou encore toute autre activité en ligne. La haine en ligne relève par conséquent aussi du discours de haine [1].

La question est maintenant de savoir comment identifier et catégoriser les discours de haine sur Internet en particulier les réseaux sociaux.

Pourquoi faut-il s'attaquer au discours de haine en ligne ?

- **Le discours de haine blesse**
  Les mots blessent, la haine aussi ! Le discours de haine est un problème majeur qui peut constituer une violation des droits de l'homme. Le discours de haine en ligne n'est pas moins grave que sa forme hors ligne, mais il est souvent plus difficile à identifier et à combattre.

- **Les attitudes nourrissent les actions**
  Le discours de haine est dangereux non seulement parce qu'il est préjudiciable en soi, mais aussi parce qu'il peut conduire à des violations des droits de l'homme plus graves, dont la violence physique. S'il ne fait l'objet d'aucun contrôle, le discours de haine en ligne a des répercussions hors ligne, favorisant la montée des tensions raciales et d'autres formes de discrimination et de violence. Le potentiel de la haine à se répandre rapidement dans le monde virtuel aggrave les dommages qu'elle est susceptible de causer.

- **La haine cible les individus et les groupes**
  Le discours de haine peut viser des groupes qui, bien souvent, sont vulnérables à d'autres égards, comme les communautés religieuses,

1. Conseil de l'Europe, Comité des Ministres, Recommandation n^o (97) 20.

les races ou encore les personnes handicapées. Pour autant, les individus eux-mêmes sont de plus en plus ciblés par les discours de haine en ligne. L'impact en est parfois fatal, comme dans le cas du cyberharcèlement, qui a conduit à des suicides dans plusieurs affaires rapportées par les médias[2]. Le discours de haine menace aussi la sécurité et la confiance en soi de quiconque s'identifie aux cibles du discours de haine.

- **Internet est difficilement contrôlable**
  Le discours de haine est davantage toléré en ligne, et y fait l'objet de moins de contrôles. Il est aussi plus facile de harceler, ne serait qu'à cause de la possibilité de se cacher derrière l'anonymat.

1.3 Travaux connexes

Plusieurs travaux similaires ont été réalisés dans le domaine de détection des langages abusifs, mais un peu d'effort qui étudie la relation entre eux, cependant le chevauchement entre ces travaux a produit des différentes descriptions dans les travaux antérieurs.

Durant ce travail, nous avons eu l'occasion d'aborder plusieurs projets liés à notre travail de recherche. Nous présentons dans ce qui suit les différents travaux récents dans ce contexte divisés en travaux qui s'intéressent à l'étude des langues étrangères et d'autres travaux intéressés par l'arabe.

1.3.1 Langues étrangères

Plusieurs travaux ont été réalisés sur les contenus dans les langues étrangères, qui peuvent être éthique, discriminatoire, raciste, etc.

1.3.1.1 Contenus éthiques

Ce qui concerne les contenus éthiques, Davidson et al. (2017) proposent d'étiqueter les tweets en trois catégories : discours de haine, langage offensant ou aucun des deux, en formant un modèle pour différencier ces catégories et analysent ensuite les résultats afin de mieux comprendre comment nous pouvons les distinguer.

2. Conseil de l'Europe, Comité des Ministres, Recommandation n^o (97) 20.

Leur dataset sont les mots et les expressions identifiés par les utilisateurs d'internet comme des discours de haine, recueilli à partir de Hatebase.org[3]. Avec l'API (Application Programming Interface) Twitter[4], ils ont recherché les tweets contenants des termes du lexicon, résultant d'un ensemble de 33.458 utilisateurs de Twitter avec 85.4 million de tweets.

A partir de ce corpus, un échantillon aléatoire de 25k tweets a été pris et les a fait coder manuellement par les travailleurs CrowdFlower (CF)[5], qui ont été demandé à étiqueter chaque tweet comme l'une des trois catégories suivantes : discours haineux, discours offensant mais non haineux ou aucun des deux.

Le score d'accord d'inter-codeur fourni par CF est de 92%, et cela donne un échantillon de 24.802 tweets étiquetés, certains tweets n'ont pas été attribués d'étiquettes car il n'y avait pas de classe. 5% des tweets sont identifiés par la majorité comme discours de haine et 1,3% ont été codé à l'unanimité, ce qui démontre l'imprécision du lexicon Hatebase.

Après avoir transformer tous les lettres en minuscule et normaliser chaque tweet, un ensemble des fonctionnalités (ou bien des caractéristiques) a été construit pour former un classificateur comme les fonctionnalités de bigramme, unigramme et trigramme, les informations sur la structure syntaxique, les scores de sentiment de chaque tweet, des indicateurs binaires et de comptage pour les hashtags, les mentions, les retweets et les URL, ainsi que des fonctionnalités pour le nombre de caractères, de mots et de syllabes dans chaque tweet.

Ce modèle utilise une régression logistique avec régularisation L1 pour réduire les dimensions des données. Une variété de modèles utilisés dans des travaux antérieurs ont été testé : régression logistique, naïve Bayésien, arbre de décision, random forests, et machine à vecteurs de support (SVM) linéaire.

Le teste comporte la validation croisée 5 fois, en tenant 10% de l'échantillon pour l'évaluation pour aider à prévenir le surajustement, et ils trouvent que la régression logistique et la SVM linéaire sont significativement meilleures que les autres modèles, et de décider d'utiliser la

3. Hatebase.org : référentiel en ligne de discours de haine structurés, multilingues et basés sur l'usage.

4. Une API a été mise en place par Twitter, permettant d'interroger sa base de données, de récupérer des informations concernant les tweets et de poster des tweets.

5. www.crowdflower.com, a été changée dernierement en : 03/03/2018 à www.figure-eight.com

régression logistique L2 pour le modèle final, car il permet plus facilement d'examiner les probabilités prédites d'appartenance à une classe et a bien fonctionné dans les articles précédents.

Le modèle le plus performant a une précision globale de 0,91%, un rappel de 0,90% et un score F1 de 0,90%, Cependant que 40% de discours de haine était mal classifié, les tweets avec les probabilités prédites les plus élevées d'être un discours de haine ont tendance à contenir de multiples racial ou homophobe.

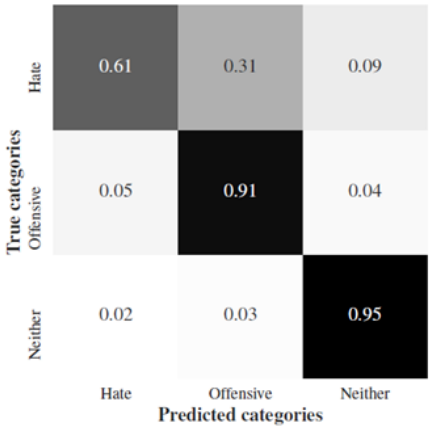


FIG. 1.1: Catégories vraies et prédites, extraite de Davidson et al. (2017)

Les résultats illustrent également comment le discours de haine peut être utilisé de différentes manières. Alors que les tweets racistes et homophobes sont plus susceptibles d'être classés comme discours de haine, mais les tweets sexistes sont généralement classés comme offensants. Les tweets sans mots-clés haineux explicites sont également plus difficiles à classer.

1.3.1.2 Contenus discriminatoires

Pour les contenus discriminatoires, Waseem et al. (2017) ont proposé une solution pour lutter contre la propagation du discours de haine comme les discours racistes ou discriminations sexuelles. Pour cela, ils ont présenté une liste de critères établis dans le mouvement de la théorie de la course critique (CRT)[6], cette liste est utilisée pour annoter un en-

6. CRT est une collection d'activistes et de chercheurs intéressés par l'étude et la transformation de la relation entre la race, le racisme et le pouvoir.

semble de plus de 16K tweets publiquement disponible[7]. Ils ont analysé l'effet des fonctionnalités linguistiques supplémentaires en conjonction avec des n-grammes de caractères pour la détection de la propagande haineuse, ils ont également fourni un dictionnaire basé sur la plupart des étiquettes dans leurs données.

Initialement une recherche manuelle a été effectuée dans le corpus sur les insultes et les termes relatifs aux minorités religieuses, sexuelles, de genre et ethniques, cette ensemble du corpus est recueilli en utilisant l'API de Twitter.

Ensuite, ils proposent 11 critères pour identifier les discours de haine. Les critères sont partiellement dérivés en annulant les privilèges observés dans McIntosh (2003).

L'ensemble de données a été annotés, à l'aide d'un annotateur extérieur (une femme de 25 ans étudiant les études de genre et une féministe non activiste), afin d'atténuer le biais d'annotateurs.

Par la suite ils ont étudié la propagation du discours de haine selon trois critères : la répartition démographique, la répartition lexicale et la répartition géographique, et les résultats soient :

Dans la répartition démographique et à partir de nom du profile, la distribution du discours sexuel est répandue chez les hommes. Quant à la répartition lexicale, les données ont été normalisées en supprimant les mots d'arrêt, à l'exception de non (not en anglais). En sélectionnant les dix mots ayant l'occurrence la plus fréquente, ainsi, les termes fréquemment utilisés dans chaque classe diffèrent de manière significative.

Les termes les plus fréquents pour le racisme sont nécessaires à la discussion de l'Islam, alors que la discussion aux problèmes des femmes n'exige pas l'utilisation de la plupart des termes qui surviennent le plus fréquemment.

Finalement, la répartition géographique, ils trouvent que l'utilisation de la localisation en tant que fonctionnalité a un impact négatif sur le score atteint de F-mesure. Pour identifier l'origine géographique d'un tweet, il est nécessaire de ne pas prendre en considération que les tags fournis par Twitter, étant donné que seulement 2% des utilisateurs de Twitter divulguent leur localisation Abbas (2015).

L'évaluation de l'influence de différentes fonctionnalités sur la

7. Contenu sexiste envoyé par 613 utilisateurs, 1972 contenu raciste envoyé par 9 utilisateurs et 11559 ni sexsiste ni raciste envoyé par 614 utilisateurs

prédiction dans une tâche de classification a été faite avec un classificateur de régression logistique et une validation croisée 10 fois pour tester l'influence de diverses fonctionnalités sur les performances de prédiction et pour quantifier leur expressivité.

Après la collection des monogrammes, des bigrammes, des trigrammes et des quadrigrammes de chaque tweet et de chaque description d'utilisateur, ils additionnent les coefficients du modèle pour chaque fonctionnalité sur les 10 fois de validation croisée cela permet une estimation plus robuste.

Feature (sexism)	Feature (racism)
'xist'	'sl'
'sexi'	'sla'
'ka'	'slam'
'sex'	'isla'
'kat'	'l'
'exis'	'a'
'xis'	'isl'
'exi'	'lam'
'xi'	'i'
''	''

TAB. 1.1: n-gramme de caractères

Ils ont trouvé que les fonctionnalités les plus influentes de la régression logistique (tableau) correspondent en grande partie aux termes les plus fréquents du sexisme et du racisme. Comme par exemple différentes longueurs n-grammes du mot «Islam» et «Sexiste».

	n-grams	+genre	+genre+loc	mot n-grams
F1	73.89	73.93	73.62	64.58
Precision	72.87%	72.93%	72.58%	64.39%
Recall	77.75%	77.74%	77.43%	71.93%

TAB. 1.2: F1 en utilisant différents ensembles de fonctionnalités, extraite de Waseem et al. (2017)

Les résultats montrent que l'utilisation de n-grammes de longueurs allant jusqu'à 4 caractères, ainsi que le genre comme une fonctionnalité supplémentaire fournissent les meilleurs résultats.

Dans un autre système qui utilise CNN (Convolutional Neural Network) Réseau neuronal convolutif, introduit par Björn and Kumar

(2017), permettant de classifier chaque tweet à l'une des catégories prédéfinies : racisme, sexisme, à la fois (racisme et sexisme) et non-discours haineux. L'ensemble de données utilisé dans ce système est l'ensemble des tweets (6,655 tweets) en anglais fourni par Waseem and Hovy (2016), annoté manuellement par un annotateur expert (qui possède à la fois une connaissance théorique et appliquée du discours de haine) et trois annotateurs amateurs dans CrowdFlower.

Quatre modèles de réseau neuronal convolutif ont été formés, quadri grammes de caractère, vecteurs de mots basés sur des informations sémantiques construites en utilisant word2vec, des vecteurs de mots générés aléatoirement, et des vecteurs de mots combinés avec des n-grammes de caractères.

Ce système consiste premièrement, à générer l'extraction de fonctionnalités pour tous les mots en utilisant les plongements de mots et les caractères de n-grammes. Les plongements de mots ont été générés de deux façons : à travers word2vec Mikolov et al. (2013a,b), où les vecteurs de mots sont générés en fonction du contexte et à travers des vecteurs aléatoires où tous les mots des corpus sont initialisés avec des valeurs aléatoires.

Il existe deux types de plongements : les modèles à sacs de mots continus (CBOW), où le modèle prédit le mot courant à partir d'une fenêtre de mots contextuels environnant et les modèles de skip-gram dans lesquels les mots de contexte sont prédits en utilisant le mot courant.

Les couches de ce réseau utilisées sont : une couche de regroupement (en anglais pooling layer) qui convertit chaque tweet en un vecteur de longueur fixe, capturant l'information de l'ensemble du tweet, une couche de max-pooling capture ensuite les facteurs sémantiques latents les plus importants des tweets, et de côté sortie, une couche softmax calcule les distributions de probabilité de classe pour chaque tweet et affecte les classes de discours haineux en fonction des valeurs de probabilité.

System setup	Precision	Recall	F1-score
Random vectors	0.8668	0.6726	0.7563
word2vec	0.8566	0.7214	0.7829
Character n-grams	0.8557	0.7011	0.7695
word2vec + character n-grams	0.8661	0.7042	0.7738
Logistic Regression with character n-grams (Waseem and Hovy, 2016)	0.7287	0.7775	0.7389

TAB. 1.3: La performance des différents systèmes en utilisant CNN, extraite de Björn and Kumar (2017)

Les résultats de ce système de classement en utilisant CNN (10-fold cross validation) pour les quatre modèles sont comparés au modèle de régression logistique (LogReg) utilisé par Waseem and Hovy (2016). Où le modèle word2vec sans caractères de n-grammes a obtenu les meilleurs résultats de tous les modèles comparés, avec les valeurs de précision, de rappel et F-score de 85,66%, 72,14% et 78,29%, respectivement. D'autre part, les vecteurs de mots aléatoires a atteint des valeurs de précision, de rappel et de score F de 86,68%, 67,26% et 75,63% respectivement, marquant une amélioration très importante de la précision par rapport au modèle LogReg, alors que tous les modèles CNN ont surpassé de manière convaincante la régression logistique en termes de précision et de score F1, tandis que le modèle LogReg a obtenu un meilleur rappel que tous les modèles de réseau neuronal.

1.4 Langue arabe classique

Contrairement aux travaux précédents sur l'obscénité et la détection offensante du langage pour différentes langues, comme l'anglais et l'allemand, les travaux antérieurs sont très limités de cette tâche pour l'arabe. En outre, l'arabe pose des défis intéressants principalement en raison de sa complexité et leurs variations lexicales, ce qui suit est quelques travaux antérieurs qui ont étudié la détection de l'obscénité dans la langue arabe.

1.4.1 Contenus éthiques

Des autres systèmes développés, en utilisant apprentissage profond pour détecter les discours haineux, obscènes et dangereux dans

les données Twitter, dont les données consiste à créer une liste multi dialectale de discours haineux, obscènes et dangereux avec la collection d'un grand ensemble de données de discours haineux, obscènes et dangereux de Twitter, ce qui permet d'étudier les fonctionnalités au niveau de l'utilisateur.

Cette liste contient 68 verbes qui pourraient être utilisés littéralement ou métaphoriquement pour indiquer un abus physique et utilisés dans un ou plusieurs des dialectes suivants : MSA (Modern Standard Arabic), Gulf, égyptien, maghrébin et levantin. La plupart de ces verbes, particulièrement 56, littéralement sont utilisés pour indiquer «wining» dans le sport ou «killing» pour signifier la douleur dans la romance, par ex :«الماتش الجي نضربكم بخمسة اهداف», «le match prochain nous vous battrons par cinq buts», «أقتلك بحناني », «Je vais te tuer avec affection».

Les cinq premiers verbes menaçants physiques dans le corpus annoté, comme indiqué ci-dessus, ne sont pas nécessairement utilisés dans un contexte menaçant. Ceci est également indiqué où moins de 30% des tweets du corpus dangereux annoté indiquent des menaces réelles.

Menace	Fréquence	Pourcentage
slaughter	257	24%
kill	222	21%
hit	159	16%
rape	135	13%
stab	84	8%

Tab. 1.4: Les cinq premiers verbes fréquents dangereux

Ensemble de données	Menaces dangereux	Pourcentage
Tout	10	3%
Dangereux	308	29%

Tab. 1.5: Résultats des annotations «l'ensemble de données» et «l'ensemble dangereux»

Dans ce travail, plusieurs cas des menaces qui ne sont pas considérés comme dangereux, y compris les menaces atténuées incluants également le discours rhétorique, les menaces métaphoriques dans le contexte d'exagérations lors de discussions sportives ou littéraires et les contenus non pertinents (i.e : citations, contenu religieux, etc). Ceux-ci sont moins susceptibles de capturer un discours dangereux.

1.4.2 Comptes d'utilisateurs abusifs

Des comptes abusifs sont créés pour distribuer du contenu adulte dans les tweets arabes, ce qui est interdit par les normes et les cultures arabes.

Les spammeurs exploitent la popularité de Twitter à rester anonymes pour diffuser du contenu malveillant, des mots insultants la pornographie en utilisant : le dialecte, Mots mal orthographiés et de combiner des mots différents en un seul mot pour être indétectable.

Dans ce contexte, Abozinadah et al. (2015) utilisent trois algorithmes basés sur l'apprentissage automatique : naïve bayésienne (NB), machine à vecteurs de support (SVM), arbre de décision (DT) pour détecter les comptes abusifs avec les tweets arabes. Par ailleurs, ils ont comparé les résultats de ces trois classificateurs.

Leurs données sont collectées (à travers l'API de Twitter) comme suit : Ils ont filtrer leur recherche en utilisant cinq mots arabes les plus insultants choisis à partir de insults.net[8], à chaque recherche ils arrivent jusqu'à 200 tweets les plus récents. Le résultat total est de 255 comptes uniques.

Pour chaque compte ils couvrent ces abonnés et ces abonnements, ces profils, les 50 tweets les plus récents qui contiennent des mots arabes ou anglais et des interactions (Likes, hashes, mentions, and links..).

Ils ont étiqueté manuellement 500 comptes qui contients les plus récentes 50 tweets, photos de profil et hashtags. La moitié des comptes sont étiquetés comme des comptes abusifs et les autres comme non abusifs.

l'ensemble de données obtenu : Accounts 350,000, Tweets 1,300,000, Hashes 530,000, Links (URLs) 1,150,000, ensuite, cet ensemble de données est prétraité (normalisation) pour empêcher les spammeurs de contourner les mécanismes de filtrage et de censure et en réduisant l'indexation des mots.

Pour la création des fonctionnalités de détection des comptes abusifs, trois ensembles différents de fonctionnalités ont été définis, à savoir : les fonctionnalités basées sur le profil (nombre de tweets, les abonnés et les abonnements, liens de sites web), les fonctionnalités basées sur le

8. How do I swear in aribic from insults.net. Available : http ://www.insults.net/html/swear/arabic.html

tweet et les fonctionnalités de graphe social (influence de l'utilisateur sur le réseau).

L'évaluation de performance des trois méthodes de classification (NB, SVM et DT-J48) basée sur le moyen de précision (P), rappel (R), F-mesure (F) et accuracy (A) sur les différents ensembles de fonctionnalités et les tweets pour trouver le classificateur approprié qui détecte les comptes abusifs des tweets arabes.

NB a surpassé le SVM et DT-J48 où il atteint une précision de 90%. Alors, que cette évaluation est pour trouver le minimum de tweets et les fonctionnalités permettants d'obtenir les performances les plus élevées du classificateur.

Mubarak et al. (2017), ont centré leur travail sur la classification des utilisateurs du twitter selon l'utilisation des mots obscène ou non. Leur hypothèse est qu'un utilisateur utilise un mot obscène, il utilisera un mot qui n'existe pas à la liste, ce qui permet d'étendre leur liste des mots offensives et cela par :

Création d'un ensemble des mots obscènes (SeedWords SW) à partir de 175 millions tweets en utilisant twitter streaming API avec le filtrage de langue arabe (lang : ar), et de chercher les tweets avec des patterns utilisés dans les communications offensives par exemple : يا **** ابن, يا ولد ****. Les mots après ces patterns sont traités manuellement et évalués d'être obscènes ou non, la liste finale obtenue contient 288 mots et phrases avec 127 hashtag dans les pages pornographiques onlines. Ensuite, en utilisant leur ensemble de 175 millions de tweets (nettoyés et normalisés), ils ont obtenu une liste d'utilisateurs de Twitter (tweeps) qui ont écrit au moins 100 tweets.

La liste des tweeps est divisée en deux groupes, i.e : ceux qui ont écrit des tweets qui n'ont pas inclus un mot obscène de la liste (groupe propre : 166K tweeps), et ceux qui ont utilisé au moins un des mots de la liste (groupe obscène : 23K tweeps).

L'identification des utilisateurs qui utilisent souvent des mots obscènes, puis ils communiquent avec des autres utilisateurs qui n'utilisent jamais les mots, on peut déduire des obscénités supplémentaires.

Avec les tweets des deux groupes, ils ont calculé les nombres unigramme et bigramme dans les deux. Ensuite, ils ont calculé le Log Odds Ratio (LOR) Forman (2008), pour chaque mot unigramme et bigramme

apparu au moins 10 fois.

Les tweets écrits par les tweeps propres sont utilisés comme corpus de fond (background), et les tweets créés par les tweeps obscènes sont utilisés comme un corpus de premier plan (foreground).

Selon eux, les unigrammes et les bigrammes qui produisaient un LOR égal à l'infini, signifie qu'ils n'apparaissaient que dans le corpus obscène mais n'apparaissaient pas dans le corpus propre, et nous les avons ajoutés à notre liste originale.

Le calcul de LOR est donné comme suit :

$$LOR = \log \left[\frac{tp \cdot (pos - tp)}{tp \cdot (pos - tp)} \right]$$

Où : tp et fp sont l'unigrammes et bigrammes pour foreground et background respectivement, pos et neg sont les tweets de foreground et background respectivement.

Dans leur expérimentation, deux évaluations utilisées, intrinsèque et extrinsèque. Pour l'intrinsèque, 100 mots (unigrammes et bigrammes) sélectionnés aléatoirement de la liste générée avec l'équation LOR, ils ont trouvé 59 mots obscènes.

Quant à l'évaluation extrinsèque, 1.100 tweets très discutés accordés à SocialBakers.com et jugés par 3 annotateurs différents d'Egypte pour marquer les tweets comme : obscène, offensant ou propre, les pourcentages de tweets étaient respectivement de 19,1%, 40,3% et 40,6%, où seulement les tweets obscènes considérés par cette évaluation.

Plusieurs listes expérimentées à savoir : la liste SW, la liste LOR (unigrammes uniquement), la liste LOR (bigrammes uniquement), les listes combinées LOR (unigrammes uniquement) + SW et les listes combinées LOR (bigrammes uniquement) + SW.

List	P	R	F-mesure
SeedWords (SW)	0.97	0.43	0.59
SW + LOR (unigrams)	0.97	0.44	0.60
SW + LOR (bigrams)	0.89	0.45	0.60
LOR (unigrams)	0.98	0.41	0.58
LOR (bigrams)	0.89	0.44	0.59

Tab. 1.6: Résultats de l'évaluation extrinsèque, extraite de Mubarak et al. (2017)

Les résultats indiquent que la liste SW atteint une grande pré-

cision avec un rappel relativement faible, Quant à la combinaison des listes SW et LOR (unigramme) a produit un rappel légèrement amélioré, tout en maintenant la précision.

Des chercheurs Singh et al. (2016) utilisent l'apprentissage automatique avec neuf fonctionnalités, l'approche a atteint un taux de précision de 91%, l'ensemble de données utilisé est basé sur 6 mots clés liés avec des contenus adultes. Cette approche permet de détecter n'importe quels comptes qui utilisent les mots sexuels, mais elle ne permet pas de détecter les fautes d'orthographe comme «@ss» pour référencer les contenus pornographiques.

Une autre étude Cheng et al. (2015) propose un graphe et un modèle de corrélation collective, le graphe contient deux types de nœuds : les comptes twitter et les hashtags (#) dans chaque tweet. L'approche ignore le contenu du tweet, à l'exception des hashtags et identifie un compte comme un compte de contenu adulte s'il est suivi par des comptes Twitter qui sont intéressés par des comptes des contenus adultes et utilise des hashtags adultes pour que son tweet soit vu par de nombreux utilisateurs.

Des autres recherches ont classifié les comptes abusifs basant sur trois fonctionnalités : profils, graphe social et les fonctionnalités de tweets par l'approche BOW (Bag Of Word). Le classificateur Naïve Bayésien avec 10 tweets et les 100 premières fonctionnalités a atteint un taux de précision de 90%. Cependant, un travail plus récent montre la faiblesse des techniques de catégorisation de texte en utilisant BOW dans le contenu de microblog court Sriram et al. (2010).

Dans ce contexte, Abozinadah and James H. Jones (2017) ont ensuite proposé une approche d'apprentissage statistique qui permet d'analyser les contenus de twitter et de détecter les comptes twitter abusifs en format texte arabe.

Cette approche implique les utilisateurs, quel que soit leur âge, dans les activités sexuelles. En outre, ces comptes pourraient cibler les utilisateurs pour le chantage et l'extorsion.

Selon eux, leur approche a atteint une précision prédictive de 96% et surmonte les imitations de l'approche du sac-de-mot (BOW). Cette approche utilise un algorithme PageRank (PR) pour identifier les mots les plus importants dans le tweet et l'algorithme Semantic Orientation

(SO) pour identifier la cooccurrences de la relation entre eux dans le tweet, en outre les mesures statistiques de base pour trouver le nombre de hashtags, mentions, photos, URLs

Leur ensemble de données utilisé est celui recueilli par Abozinadah et al. (2015). Cet ensemble de données a été collecté en utilisant cinq mots d'insultes arabes dans le moteur de recherche de Twitter. Le résultat de la recherche comprenait 255 comptes Twitter uniques qui utilisaient un ou plusieurs de ces mots d'insultes arabes. Ces comptes ont été utilisés comme des graines pour collecter deux sauts sur chaque compte et inclure les abonnés, les abonnements et les tweets. Alors que, les données recueillies contenaient 350 000 comptes, 1 300 000 tweets, et plus de 300 millions des abonnés et des abonnements.

L'ensemble de données a été analysé manuellement par un choix au hasard de 2500 comptes qui ont tweeté plus de 100 tweets, ces comptes ont été analysés sur la base du contenu du tweet pour étiqueter ceux qui incluent des mots profanes arabes, l'obscénité arabe, les mots insultes arabes, et la signification des tweets arabes. Les analyseurs ont dû détecter cinq tweets qui contenaient des mots abusifs postés sur deux jours différents ou plus pour identifier le compte comme un compte abusif ; sinon, il identifiera soit comme un compte normal ou inconnu. Les comptes normaux contiendraient des tweets sans mots abusifs. Les comptes inconnus étaient les comptes avec des tweets dans d'autres langues que l'arabe, ont des URL, ou des images sans texte, ou ce n'était pas clair pour les analyseurs.

Par la suite, les données a été normalisées par la suppression de tous les mots non arabes, symboles, nombres, les mots vides, la séquence de lettres, sauf le nom d'allah (الله), diacritiques (ِ ّ ْ), espaces blancs et correction les mots mal orthographié.

Leur approche d'apprentissage statistique est basée sur l'algorithme PR, l'algorithme de SO et l'application des mesures statistiques de base sur les contenus du tweets.

L'algorithme Word PageRank a été développé par Larry Page, l'un des fondateurs de Google Page (2001). Cet algorithme a été utilisé dans différents environnements, tels que l'extraction de mots-clés Litvak et al. (2011), l'extraction de phrases Mihalcea and P (2004), et l'influence sur Twitter Wu et al. (2011).

L'algorithme utilisé avec PR est l'algorithme de poids de bord, qui est utilisé pour calculer le PR du tweet en obtenant la moyenne du PR de tous les mots. De plus, le total des PR de tous les tweets de chaque compte est utilisé pour refléter le PR de chaque compte Twitter en fonction du contenu du tweet. Par conséquent, le compte Twitter PR reflétera le type de mots que l'utilisateur utilise dans son tweet et la relation de cooccurrence entre les mots.

L'objectif principal de l'utilisation de cet algorithme est de construire un graphe[9] de mots à partir des tweets et d'avoir une valeur à chaque mot pour refléter l'influence du mot dans le tweet. Les mots sur le tweet auront un score PR, qui est calculé en utilisant deux graphes des mots normalisés des tweets, comptes abusifs et non abusifs.

Le score de PR donne l'impact d'un certain mot est abusif ou non, où chaque mot aura deux scores, un du graphe des comptes abusifs et un du graphe des comptes non abusifs, alors que le score du mot est la soustraction des deux scores.

$$Score(x) = (WPR(Non_{Abisive_{wi}})) - (WPR(Abisive_{wi}))$$

Où :
$WPR(Non_{Abisive_{wi}})$ est le PageRank avec le résultat du poids de bord du graphe non abusif pour le mot i, $i \in I$ *tous les mots du graphe*,

$WPR(Abisive_{wi})$ est le PageRank avec le résultat de poids de bord du graphique abusif du même mot i. Le score de mot définit la polarité du mot comme étant proche d'un mot abusif ou d'un mot non abusif.

Polarité du mot $P(x)$:

$$P(x) = \begin{cases} Abusive\ word, & x < 0 \\ NonAbusive\ word & x \geq 0 \end{cases}$$

La polarité de chaque tweet $P(t)$ sera la moyenne des scores de polarité des mots.

9. Un graphe G = (V, E), où V est l'ensemble de noeuds pour chaque mot unique. E est l'arête entre les nuds qui seront pondérés en fonction de la relation de cooccurrence entre chaque mot dans le tweet (représentée par l'épaisseur des flèches). Le poids reflétera la faiblesse ou la force entre les deux nuds.

$$t = \frac{\sum_{i \in n} P(x)_i}{n}$$

Où :

$$P(t) = \begin{cases} Abusive\ tweet, & t < 0 \\ NonAbusive\ tweet & t \geq 0 \end{cases}$$

La polarité du compte Twitter P (a) sera la moyenne des scores de polarité des tweets.

$$a = \frac{\sum_{i \in n} P(t)_i}{n}$$

Où :

$$P(a) = \begin{cases} Abusive\ account, & a < 0 \\ NonAbusive\ account & a \geq 0 \end{cases}$$

L'orientation sémantique (SO) d'un mot trouve la proximité du mot à un mot positif ou à un mot négatif, elle est calculée avec la méthode PMI-IR (Point-wise Mutual information and Information retrieval) P.D (2002), pour identifier les mots les plus positifs et les mots les plus négatifs dans l'ensemble de données de Twitter.

Dans leur recherche, ils ont identifié les mots les plus positifs et les mots les plus négatifs dans l'ensemble de données de Twitter.

Après le calcul de la fréquence de chaque mot dans l'ensemble de données, le mot «الله» Dieu était le mot le plus fréquent dans le compte légitime, et le mot «سكس» Sexe était le mot le plus fréquent dans le compte abusif. Par conséquent, le mot «الله» Dieu est utilisé pour refléter le mot positif, et le mot «سكس» Sexe pour refléter le mot négatif.

Cette approche d'apprentissage statistique proposée, basée sur deux étapes pour chaque compte Twitter afin de construire l'ensemble de données :

- Le Calcul de PageRank et l'orientation sémantique de mot, pour chaque tweet.
- Et le calcul des mesures statistiques de base pour chaque compte comprenant la moyenne, le minimum, le maximum, l'écart-type et

le total de chaque tweet, pour refléter le comportement du compte Twitter.

L'approche proposée est évaluée et comparée avec l'approche BOW, l'évaluation a fait avec un petit nombre de fonctionnalités qui ajoutent des bruits au classificateur, et minimise leur nombre pour réduire le temps de calcul et améliorer la précision.

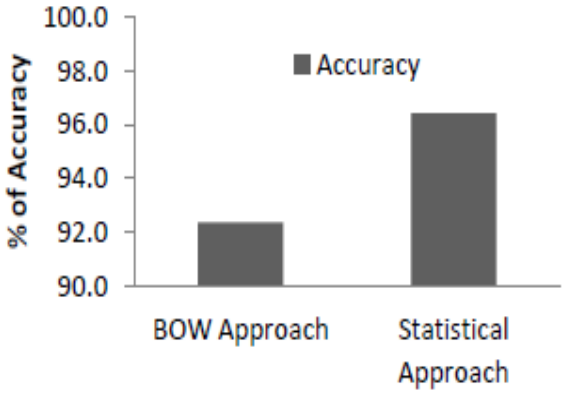


FIG. 1.2: Comparaison des résultats entre les deux approches Bow et Statistique

Le taux de précision atteint 96% pour l'approche d'apprentissage statistique alors que le taux de précision atteint environ 93% pour l'approche BOW, reflétant l'amélioration de l'approche proposée par rapport à l'approche BOW.

1.5 Typologie sur le discours de haine

Le langage abusif est un terme très complexe et très vaste qui contient plusieurs sous termes, qui nécessite l'étude de différentes relations entre eux.

Suite au travail sur le discours de haine, la cyber intimidation et la violence en ligne, une typologie a été proposée par Waseem et al. (2017), qui capture les similarités et les différences entre les sous-tâches et ses implications pour l'annotation des données et la construction de fonctionnalités.

Une variété d'étiquettes utilisées dans les travaux antérieurs, Van-Hee et al. (2015) identifient les remarques de discriminations (raciste,

sexiste) comme un ensemble des insultes, tandis que Nobata et al. (2016) classifient les remarques similaires comme «discours de haine» ou «langage péjoratif». Waseem and Hovy (2016) ne considèrent que le discours de haine sans tenir compte d'un chevauchement potentiel avec un langage autrement offensant, tandis que Davidson et al. (2017) a distingué le discours de haine du langage généralement vulgaire.

Les différents termes atteinte, haine, offense, insolente, etc. existent dans plusieurs domaines (médias sociaux, nouvelles, différents sites, etc.). Alors que l'absence de consensus a entraîné des contradictions dans les différents concepts.

Pour aider à rassembler ces concepts et pour éliminer la contradiction, Waseem et al. (2017) proposent une typologie en deux facteurs primaires :

1. Direct / général : vers un individu spécifique (entité) ou un groupe général,
2. Implicite / explicite : distinction linguistique, sémiotique (les signes et les symboles et leurs signification). Un grand travail peut être synthétisé dans cette typologie qui considère si :
   (a) Vers une cible spécifique : vers un individu, une entité ou un ensemble d'individus, une certaine ethnicité ou orientation sexuelle, par contre dans les travaux antérieurs cette distinction n'est pas prise en compte.
   (b) Le degré de l'explicite : (un langage abusif est explicite ou implicite) c'est la distinction linguistique et sémiotique entre la dénotation, le sens littéral d'un terme ou d'un symbole, et la connotation (les associations socioculturelles), Une langue abusive explicite est celle qui est sans ambiguïté dans sa capacité à être abusive, cependant un langage abusif implicite est ce qui n'implique pas immédiatement un abus.

Ici, la vraie nature est souvent masquée par l'utilisation de termes ambigus et d'autres moyens, rendant généralement plus difficile à détecter à la fois par les annotateurs et les approches d'apprentissage automatique.

Cette typologie a plusieurs implications pour les travaux futurs dans ce domaine.

1. Elle permet d'encourager les chercheurs travaillant sur ces sous-tâches à apprendre des progrès réalisés dans d'autres domaines, les sous-tâches sont distinctes sur les mêmes problèmes en même temps, le discours haineux peut être amélioré en interagissant avec la cyber intimidation, et vice versa, car une grande partie de toutes les sous-tâches consiste à identifier les abus ciblés.

2. Mettre en évidence les distinctions importantes dans les sous-tâches qui ont été ignorées jusqu'ici.

3. Un examen plus attentif de la relation entre les lignes directrices d'annotation et l'intérêt, tandis que le type d'annotation et même le choix des annotateurs devraient être motivés par la nature de l'abus.

4. Encourager les chercheurs à déterminer les fonctionnalités les plus appropriées pour chaque sous-tâche.

5. Il est important de souligner que tous les abus ne sont pas égaux, à la fois en termes d'effets et de détection.

En outre, cette typologie peut rassembler les résultats dans ces différents domaines et de clarifier les aspects clés de la détection du langage abusive. Il y a des distinctions analytiques importantes qui ont été largement négligées dans les travaux antérieurs et on les reconnaît ainsi que leurs implications.

1.6 Discussion

Dans les travaux antérieurs, nous constatons qu'il y a des bons résultats dans les langues étrangères telles que l'anglais, car il existe plusieurs efforts et des outils qui facilitent la continuité de ces travaux dans ce domaine, alors que dans l'arabe classique il n'existe qu'un peu d'efforts mais aucun travail sur les différents dialectes. Cependant, certains travaux ont classé le langage abusif comme abusif ou propre et des autres en obscène, offensant ou propre. Alors qu'il y a un travail qui

s'intéresse à la création d'une topologie pour unifier les définitions du langage abusif, discours direct ou indirect, vers un individu ou vers un groupe.

1.7 Conclusion

Dans ce chapitre, nous avons discuté de la définition du discours de haine, de ses différentes formes et de l'importance de le combattre, puis nous avons abordé un certain nombre des travaux connexes, en arabe et en langues étrangères, classées socialement et éthiquement.

Dans le prochain chapitre, nous discuterons de la particularité de la langue arabe et du dialecte algérien.

Chapitre 2

Particularité de la langue arabe

2.1 Introduction

La langue arabe appartient à la famille des langues sémitiques, et plus précisément au rameau méridional de ces langues, avec un nombre de locuteurs estimé entre 315 millions [1] et 375 millions de personnes au sein du monde arabe et de la diaspora arabe[2], elle est utilisée comme vecteur de transmission religieux pour tous les croyants musulmans au nombre de 1 milliard et demi à travers les cinq continents du globe.

Le fait que la langue arabe est la langue du coran elle s'est étendue au-delà du golfe arabo-persique, atteignant l'Afrique du nord et l'Asie mineur. De plus, l'expansion territoriale de l'empire musulman a fait de l'arabe une langue de culture et de sciences. Par ailleurs, la diversité des populations arabes et de leurs cultures ont fait émerger différentes variantes de l'arabe allant de l'arabe classique utilisé dans le coran, à l'arabe standard moderne (ASM).

2.2 Système d'écriture de l'arabe

Comme mentionné précédemment, l'arabe est classé sous le groupe des langues sémitiques qui s'écrit de droite à gauche. Son système graphique se compose d'un alphabet arabe et des voyelles.

1. Languages with at least 50 million first-language speakers ż [archive], sur Ethnologue (consulté le : 29 avril 2018).
2. «Arabic», University of Birmingham.

2.2.1 Alphabet

L'alphabet arabe comporte vingt-huit lettres (si l'on exclut la hamza, qui se comporte soit comme une lettre à part entière soit comme un diacritique).

De nombreuses lettres sont similaires, cela résulte directement du fait que l'écriture est cursive : les formes possibles des lettres s'en sont trouvées diminuées. Pour distinguer les différents sons notés par une même lettre, on utilise des points placés sur ou sous la lettre. Il ne dispose, en réalité, que d'une quinzaine de caractères pour les noter.

Ces lettres peuvent être rangées selon l'ordre traditionnel des alphabets sémitiques (أ ، ب ، ج ، د ، هـ ، و ، ز ، ح ، ط ، ي ، ك ، ل ، م ، ن ، س ، ع ، ف ، ص ، ق ، ر ، ش ، ت ، ث ، خ ، ذ ، ض ، ظ ، غ).

Mais dès les toutes premières époques est apparu un ordre mnémonique, dans lequel des regroupements rapprochent les lettres dont les formes sont semblables. Voici l'ordre traditionnel de l'alphabet arabe, présenté d'une manière qui en fait apparaître les regroupements entre lettres de graphisme semblable (أ، ب، ت، ث، ج، ح، خ، د، ذ، ر، ز، س، ش، ص، ض، ط، ظ، ع، غ، ف، ق، ك، ل، م، ن، هـ، و، ي).

2.2.2 Voyellation

L'arabe possède deux types de signes de notation des voyelles : les voyelles brèves, qui sont notées au moyen de signes diacritiques secondaires et les lettres d'allongement de la voyelle correspondante :

- Les voyelles brèves : (ـِ ـُ ـَ) fatha, damma et kassra, se rajoutent sur les lettres, et ce seulement pour lever des ambiguïtés (rarement) ou dans les ouvrages didactiques ou religieux. Elles servent à préciser la prononciation et le sens d'un mot lorsque le contexte ne suffit pas, par exemple : كَتَبَ (kataba, il a écrit), كُتُب (kutub, livres).

  Il existe d'autres diacritiques dont les plus courants comme l'indication de l'absence de voyelle (سكون - sukun) et la gémination des consonnes (شدة - shadda).

  Les signes (ـٍ ـٌ ـً) indiquent le tanwn, c'est à dire la présence d'une marque *nūn* (ن) à la fin du mot (il s'agit de la marque de l'indéfinition des noms) : prononcez [*un*], [*an*], [*in*].

La graphie (آ) Correspond à une hamza suivie d'une voyelle brève (a) et d'un allongement de cette voyelle : [$\bar{a}$].

- Les voyelles longues : ou lettres d'allongement, (ي / و / ا) appartiennent à l'alphabet et sont incluses dans le corps du mot. Elles se prononcent 2 fois plus longtemps que les voyelles brèves, au nombre de trois :
  - La voyelle brève (ـَ) + la lettre (ا) ou (ى) la voyelle phonétique longue (ـَا / ـَى) prononcée ā; par exemple : حَجّ «pèlerinage» par opposition à حَاج «pèlerin».
  - La voyelle brève (ـُ) + la lettre (و) la voyelle phonétique longue ـُو prononcée ū ; par exemple : قُل «dis !» par opposition à يقُول «il dit».
  - La voyelle brève (ـِ) + la lettre (ي) la voyelle phonétique longue ـِي prononcée ī ; par exemple : طِبّ «médecine» par opposition à طِيب «bonté».

2.3 Etude morphologique de la langue arabe

La morphologie consiste en l'étude de la structure morphémique des mots. Un mot peut être décomposé en unités morphologiques, c'est-à-dire en unités de sens appelées des morphèmes Gombert et al. (2000) .

2.3.1 Morphologie agglutinante

L'agglutination est le rattachement des clitiques aux mots dans un ordre bien précis. Les clitiques sont des morphèmes qui peuvent être réalisés comme des éléments autonomes et possèdent des fonctions syntaxiques indépendantes (telles que l'inversion, la définition, la conjonction ou la préposition). Ils sont invariants, optionnels et ne changent pas la signification de base du mot auquel ils se rattachent.

On distingue deux types de clitiques : des proclitiques qui se rattachent au début du mot et des enclitiques situés à la fin de ce dernier.

Les proclitiques sont répartis dans plusieurs catégories selon leurs fonctions grammaticales comme suite :

catégorie	proclitique
particule	أ interrogative (- est-ce que)
conjonction	و (wa - et) ف (fa - puis, alors)
préposition	ب (bi - par, avec) ك (ka - comme ل (li - pour, à)
particule de futur	س (sa -particule de futur)
particule de négation	لا (lA - ne pas)
particule de négation	ما (mA)
déterminant	ال (Al - le, la, les)

Tab. 2.1: Liste de proclitiques

Les enclitiques présentent les pronoms suffixes qui s'attachent toujours à la fin du mot graphique, leur liste est constituée des éléments suivants : نِي ي نَا كَ كِ كُمَا كُمْ كُنَّ هُ هَا هُمَا هُمْ هُنَّ هِ هِمَا هِمْ هِنَّ . Un mot graphique ne contient qu'un seul enclitique à la fois. Ils s'attachent aux verbes comme étant un complément-objet et aux noms et aux prépositions comme un complément du nom ou complément d'objet indirect. Leurs utilisations est régie par certaines restrictions.

personne	genre	nombre	enclitique
1	masculin	singulier	ـي i ـي niy
	féminin	pluriel	نا nA
2	masculin	singulier	ـك ka
		duel	ـكما kumA
		pluriel	ـكم kum
	féminin	singulier	ـكي ki
		duel	ـكما kumA
		pluriel	ـكنّ kunna
3	masculin	singulier	ـه hu
		duel	ـهما humA
		pluriel	ـهم hum
	féminin	singulier	ـها hA
		duel	ـهما humA
		pluriel	ـهنّ hunna

Tab. 2.2: Liste d'enclitiques

Les clitiques ne sont pas toujours compatibles avec un mot donné, leur compatibilité dépend de la catégorie grammaticale du mot, aussi, l'agglutination avec l'absence de diacritiques présente le problème de

l'ambiguïté qui permet de réduire le nombre de découpages possibles d'un mot.

2.3.2 Morphologie flexionnelle

La flexion est l'ensemble des modifications subies d'un mot à l'aide d'affixes pour dénoter les traits grammaticaux voulus. Les affixes possèdent trois types : les préfixes, qui se situent avant le radical, les suffixes, qui se situent après le radical et les circonfixes qui l'entourent.

2.3.2.1 Traits morphologiques du verbe

- l'aspect : l'arabe distingue trois aspects différents. L'accompli (الماضي) utilisé quand l'action est accomplie. L'inaccompli (المضارع) indique que l'action est en train de se réaliser, sans être achevée. L'impératif (الأمر) indique l'injonction.
- le mode : les modes définis en arabe sont : L'indicatif (المرفوع) dans une proposition principale. Le subjonctif (المنصوب) dans une proposition subordonnée. L'apocopé (المجزوم) exprime la négation, l'interdiction ou le conditionnel.
- la personne, le genre et le nombre du sujet : l'arabe distingue trois personnes, celle qui parle (المتكلم), celle à qui l'on parle (المخاطب) et celle de qui l'on parle (الغائب). Deux genres, le masculin (المذكر) et le féminin (المؤنث) . Et trois valeurs pour le nombre le singulier (المفرد), le duel (المثنى) et le pluriel (الجمع).

2.3.2.2 Traits morphologiques du nom

Elle se base sur :

- L'état : un nom peut être défini (معرف) à l'aide d'un article ou indéfini (نكرة), l'état indéfini est marqué par un diacritique double.
- Le cas : l'accusatif (منصوب), le nominatif (مرفوع) et le génitif (مجرور).
- Le genre et le nombre : comme les verbes, les noms arabes possèdent deux genres et trois nombres.

2.3.3 Morphologie dérivationnelle

La langue arabe est une langue de dérivation (الاشتقاق Ishtiqâq). Les verbes sont dérivés d'une racine (جذر jdr) qui est composée de trois, quatre ou cinq lettres. À partir de ces lettres, on construit un verbe de la forme I الفعل ٱلمجرَّد. Il est possible d'augmenter cette forme I de quelques lettres pour obtenir des verbes dérivés/augmentés فعل مّزيد . Le sens de ces derniers demeure proche de celui de la racine. Le schème (وزن wzn), appelé aussi gabarit, définit le format du radical.

Reprenons l'exemple de la racine كتب k t b, en remplaçant les lettres C de schème CACaC par les lettres correspondantes de la racine, donne naissance au mot K kAtab «correspondre avec». Nous décrivons ci-dessous quelques-unes des dérivations communes en arabe HABASH (2010).

- Le participe actif (اسم الفاعل) et le participe passif (اسم المفعول) ont chacun un schème unique pour chaque forme de verbe. Par exemple, les participes actifs et passifs de la forme I sont CACiC et maC-Cuw3, respectivement : كاتب kAtib «writer» et مكتوب maktuwb «written».

- Les schèmes maCCaC et maCCiC sont utilisés pour indiquer les noms de lieux et de temps (أسماء المكان والزمان), Par exemple : مكتب maktab «bureau» de katab «écrire» et مجلس majlis «conseil» de جلس jalas «asseoir».

- Il existe plusieurs schèmes nominaux qui désignent les instruments (إسم الآلة) utilisés pour le verbe dont ils dérivent. Par exemple, miC-CAC est utilisé pour dériver مفتاح miftAH «clé» de فتح fataH «ouvrir» et منشار minšAr «scie» de نشر našar «scier».

- Le schème CuCayC parmi les autres est utilisé pour dériver la forme diminutive (اسم التصغير) d'un autre nom. Par exemple, شجيرة šujayrah «arbuste» est le diminutif de شجرة šajaraŕh «arbre».

- Le Ya de Nisba (ياء النسبة) est un suffixe de dérivation (يّ iy) qui fait correspondre les noms aux adjectifs qui leur sont liés. Par exemple, أردن Âurdun˜ «Jordanie» → Âurdun˜iy˜, سياسة siyAsa¯h «politique» → سياسيّ siyAsiy˜ «politicien» et ملك malik «roi» → ملكيّ

malakiy˜ «royal». Les deux exemples illustrent comment la dérivation de Ya de Nisba peut supprimer des suffixes (comme la terminaison féminine).

- La contrepartie dénombrable d'un nom collectif est appelée son nom d'unité (اسم الوحدة). Il est souvent dérivé d'un suffixe singulier féminin, par exemple, تمر tamr «dattes» (collectif) تمرة tamrah «date» (singulier). Dans certains cas le passage du singulier au pluriel ne repose pas sur les affixes mais sur les schèmes. Le pluriel bâti sur un schème est appelé pluriel brisé (جمع التكسير) jma Altksyr.

2.4 Le dialecte algérien francophone

L'arabe dialectal est une langue arabe utilisé dans les communications quotidiennes, généralement appelée āmmiyya عاميّة «langue commune» ou dārija دارِجة «langue courante». Elle est définie selon Al-Toma (1969) comme étant «la langue courante des activités quotidiennes, elle est généralement parlé, bien qu'elle soit parfois écrite. Elle varie non seulement d'un territoire arabe à un autre, mais aussi d'une région à une autre au sein du même territoire».

Historiquement, les années de l'occupation française ont laissé leur empreinte sur des générations entières d'Algériens notamment par l'enseignement, même si l'élite algérienne était quasiment inexistante à l'époque coloniale.

La société algérienne est donc bien une société bilingue où les deux langues l'arabe et le français sont constamment utilisées et ce bilinguisme a été imposé par les circonstances de l'histoire. Le multilinguisme en Algérie s'organise autour de quatre langues présentes sur le marché linguistique. Il se compose essentiellement de l'arabe algérien (qui est lui-même divisé en plusieurs variétés régionales), du berbère et de l'arabe classique ou conventionnel (pour l'usage de l'officialité). A tout cela s'ajoute la langue française (première langue étrangère du pays) Medane (2015).

D'un point de vue sociolinguistique, le langage quotidien (algérien) ou le dialecte algérien, connaît une association avec d'autres langues notamment le français, il accepte en son sein des mots et structures grammaticalement tirées de la langue française, tant dit qu'il est

la langue de la majorité des Algériens.

Le vocabulaire de l'arabe algérien est en grande partie issu de l'arabe classique. Toutefois les mots d'origine non-arabe sont nombreux, surtout les mots d'origine française qui deviennent de plus en plus nombreux dans le parler algérien quotidien, surtout dans les grandes villes, par exemple : «lycée», «salon», «épicerie», «normal», etc. L'utilisation de ces mots est réalisée sans aucune adaptation de la phonologie. Ceci crée une situation linguistique et sociolinguistique assez complexe.

Parmi les caractéristiques très répandues dans le dialecte algérien, à savoir l'emprunt qui est fortement présent dans le dialecte algérien. Par ailleurs, l'emprunt est le reflet de l'influence des autres langues sur les dialectes, où nous trouvons beaucoup de mots issus des différentes langues telle que le français et le turc.

Des suffixes empruntés à d'autres langues sont utilisé pour exprimer certains termes comme le suffixe turque جي jiy qui indique la profession, par exemple قهواجي qahwaAjiy «celui qui tient un café».

Aussi, l'impact phonologique des emprunts sur les dialectes, comme c'est le cas pour le mot طابلة TaAbla «table», فرملي Farmliy «infirmier».

En outre, dans certains cas, la phonologie ou la morphologie sous-jacente se traduit par une écriture d'assimilation phonologique régulière, par exemple, طوموبيل Tuwmuwbiyl 'voiture' est aussi écrite طونوبيل Tuwnuwbiyl.

Donnons quelques exemples d'emprunts de mots, d'origines françaises dans le dialecte algérien [3] :

Mots	**Traduction**	**Translittération**
تريسيتي	Électricité	Triciti
دجورنان	Journal	Djornène
كارتي	Quartier	Karti
بوليسيا	Police	Boulissiya
فنيان	Fainéant	Feniénne
كوري	Ecurie	Koûri
لامبة	Lampe	Lamba
تيليفون	Téléphone	Tilifoune
بلاصة	place	Plaça
راسة	race	Rassa

TAB. 2.3: Sens de quelques mots empruntés utilisés dans le dialecte algérien

3. https://fr.wikipedia.org/wiki/Arabe_alg%C3%A9rien#Mots_alg%C3%A9riens_d'origine_fran% . Consulté le : 25/03/2018.

2.5 Conclusion

Ce chapitre résume les particularités de la langue arabe en termes de construction et de composition, ce qui la rend plus complexe et difficile à traiter automatiquement, aussi le dialecte algérien et ses origines.

Dans le prochain chapitre, on va voir quelques études théoriques sur l'apprentissage automatique et profond, ses algorithmes et le plongement lexical.

Chapitre 3

Apprentissage automatique

3.1 Introduction

Le traitement automatique du langage naturel (TALN) est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle. Il vise à créer des outils de traitement de la langue naturelle pour diverses applications. Il ne doit pas être confondu avec la linguistique informatique, qui vise à comprendre les langues au moyen d'outils informatiques. Le TALN est sorti des laboratoires de recherche pour être progressivement mis en œuvre dans des applications informatiques nécessitant l'intégration du langage humain à la machine.

3.2 Apprentissage automatique

L'intelligence artificielle (IA) initié en 1956 par John McCarthy[1], elle implique des machines capables d'effectuer des tâches caractéristiques de l'intelligence humaineS. and P. (2003). Bien que ce soit plutôt général, cela inclut des choses comme la planification, la compréhension du langage, la reconnaissance des objets et des sons, l'apprentissage et la résolution de problèmes. Nous pouvons mettre l'IA en deux catégories, générale et étroite.

Les algorithmes d'apprentissage peuvent se catégoriser selon le mode d'apprentissage qu'ils emploient : apprentissage supervisé, apprentissage non supervisé, apprentissage par renforcement, etc. On trouvera, figure 3.1 [2], une représentation graphique des différents modes

1. John McCarthy né le 4 septembre 1927 à Boston, mort le 24 octobre 2011 est le principal pionnier de l'intelligence artificielle avec Marvin Lee Minsky.

2. https ://recherche.orange.com/learning-zoo/

d'apprentissage.

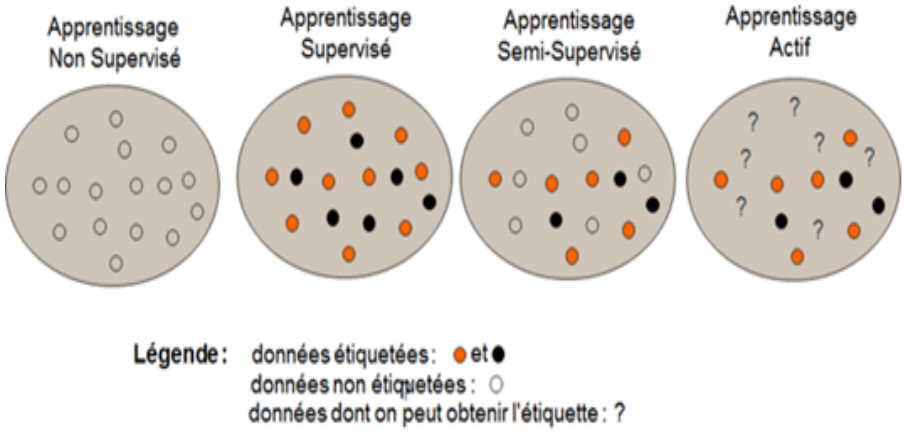


FIG. 3.1: Les différents modes d'apprentissage

Arthur Samuel[3] a inventé la phrase pas trop longtemps après IA, en 1959, la définissant comme « la capacité d'apprendre sans être explicitement programmée ». Vous voyez, vous pouvez obtenir l'IA sans utiliser l'apprentissage automatique, mais cela nécessiterait de construire des millions de lignes de codes avec des règles complexes et des arbres de décision.

Ainsi, au lieu de coder en dur des routines logicielles avec des instructions spécifiques pour accomplir une tâche particulière, l'apprentissage automatique est un moyen de « former » un algorithme pour qu'il puisse apprendre. La « formation » implique d'alimenter d'énormes quantités de données et de permettre à l'algorithme de s'ajuster et de s'améliorer. Comme est décrite dans la figure 3.2[4].

3. Arthur Lee Samuel né le 5 décembre 1901, mort le 29 juillet 1990 était un pionnier américain du jeu sur ordinateur, de l'intelligence artificielle et de l'apprentissage automatique.

4. https ://www.xenonstack.com/blog/data-science/log-analytics-with-deep-learning-and-machine-learning

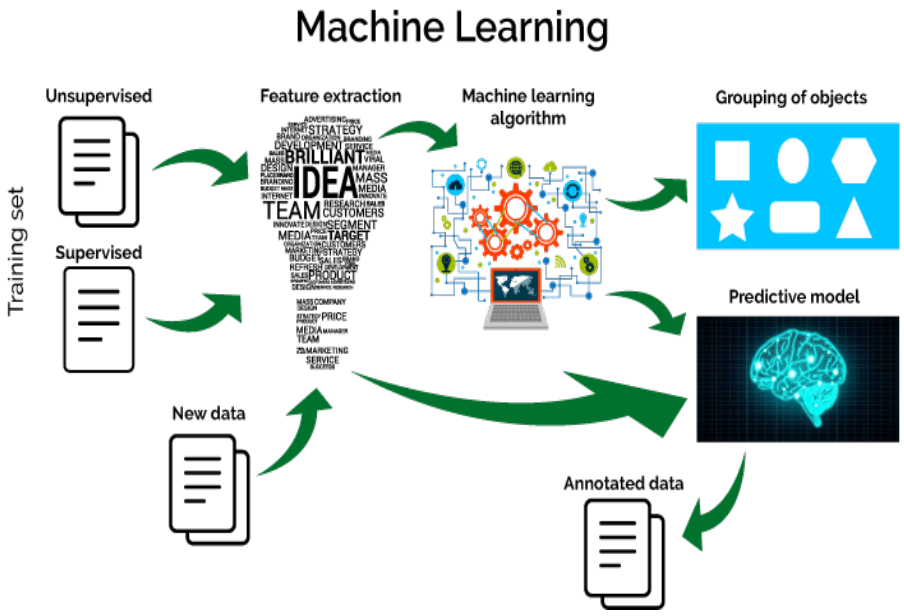


FIG. 3.2: Architecture générale d'apprentissage automatique

3.3 Apprentissage profond

L'apprentissage profond est l'un des nombreuses approches de l'apprentissage automatique. D'autres approches comprennent l'apprentissage de l'arbre de décision, la programmation logique inductive, l'apprentissage par renforcement, les réseaux de neurones, etc.

L'apprentissage profond a été inspiré par la structure et la fonction du cerveau, à savoir l'interconnexion de nombreux neurones. Les réseaux neuronaux artificiels (RNA) sont des algorithmes qui imitent la structure biologique du cerveau [5].

Dans les RNA, il y a des « neurones » qui ont des couches discrètes et des connexions avec d'autres « neurones ». Chaque couche sélectionne une caractéristique spécifique à apprendre, telle que des courbes / bords dans la reconnaissance d'image. C'est cette stratification qui donne son nom à l'apprentissage profond, la profondeur est crée en utilisant plusieurs couches par opposition à une seule couche.

5. extraite de : https ://www.quora.com/What-is-the-difference-between-deep-learning-and-usual-machine-learning. Consulté le : 22/03/2018.

3.4 Plongement lexical

Le plongement lexical des mots (en anglais word embedding) est une manière de représenter des mots comme des vecteurs, typiquement dans un espace de quelques centaines de dimensions. Le vecteur représentant chaque mot est appris via un algorithme itératif, à partir d'une grande quantité de texte, il a été massivement exploités ces dernières années, notamment pour l'apprentissage d'une représentation de mots Mikolov et al. (2013a,b,c).

La représentation dans un espace vectoriel des mots a permis de développer des méthodes qui prennent en compte la sémantique de ces mots. On trouvera dans Mikolov et al. (2013b) un modèle qui construit des représentations continues des mots en se basant sur leur contexte, en prenant en compte la fenêtre d'occurrence d'un mot pour construire une représentation permettant de déduire des relations sémantiques entre différents termes.

3.4.1 Les modèles de Word2vec

Word2vec est un modèle prédictif particulièrement efficace sur le plan informatique pour l'apprentissage des plongées de mots à partir de texte brut. Il se présente sous deux formes : le modèle du sac continu de mots (CBOW) et le modèle de Skip-Gram Mikolov et al. (2013a). Algorithmiquement, ces modèles sont similaires, sauf que CBOW prédit les mots cibles par exemple «tapis» à partir des mots de contexte source «le chat s'assoit sur le», tandis que le skip-gram fait l'inverse et prédit les mots de contexte source de la cible mots. Cette inversion peut sembler un choix arbitraire, mais statistiquement elle a pour effet que le CBOW lisse une grande partie de l'information distributive (en traitant un contexte entier comme une observation). Pour la plupart, cela s'avère être une chose utile pour les jeux de données plus petits. Cependant, skip-gram traite chaque paire contexte-cible comme une nouvelle observation, ce qui a tendance à être meilleur lorsque nous avons des ensembles de données plus volumineux.

Dans ce qui suit, nous allons présenter les deux architectures CBOW et Skip-gram illustrées par la figure 3.3 Mikolov et al. (2013a).

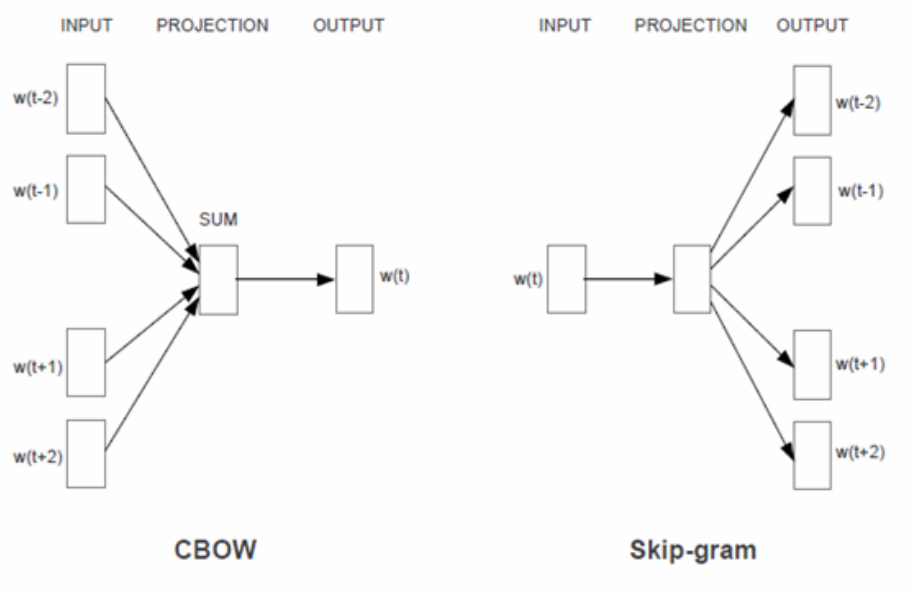


FIG. 3.3: Architectures CBOW et Skip-gram du modèle word2vec.

3.4.1.1 Le modèle CBOW

La première architecture permet de prédire un mot w_i en fonction de son contexte, qui correspond aux mots précédents et aux mots suivants. Dans cette architecture, la couche de projection est partagée par tous les mots : tous les mots sont projetés dans la même position.

Ici, l'apprentissage des word embeddings consiste à prédire un mot en fonction de son contexte. Ceci est fait en calculant la somme des word embeddings du contexte, puis en appliquant sur le vecteur résultant un classifieur log-linéaire pour prédire le mot cible. Enfin, le modèle compare sa prédiction avec la réalité et corrige la représentation vectorielle du mot par rétro-propagation du gradient. Ce modèle cherche à maximiser l'équation suivante :

$$S = \frac{1}{I} \sum_{i=1}^{I} \log p(m_i | m_{i-n}, ..., m_{i-1}, m_{i+1}..., m_{i+n})$$

Où I : correspond au nombre de mots dans le corpus, le modèle reçoit une fenêtre de n mots autour du mot cible m_i. Ce modèle est nommé CBOW pour sac-de-mots continus ou Continuous Bag-of-Words pour deux raisons. D'une part, l'appellation sac-de-mots indique que l'ordre des mots du contexte n'a pas d'influence sur la projection. D'autre part,

l'objectif souligne l'utilisation d'une représentation continue (word embeddings) des mots en contexte, ce qui diffère des modèles sac-de-mots standards.

3.4.1.2 Le modèle Skip-gram

L'architecture skip-gram est l'inverse de CBOW. Le mot cible mi est maintenant utilisé comme entrée, et les mots de contexte sont utilisés en sortie. Elle consiste à prédire, pour un mot donné, le contexte dont il est issu. Dans cette architecture le mot en entrée est projeté dans la couche cachée puis dans la couche de sortie pour produire un vecteur. Ce vecteur est ensuite comparé à chacun des mots du contexte et le réseau se corrige par rétro-propagation du gradient. De cette manière, la représentation vectorielle du mot d'entrée va être proche de celles des mots qui produisent le même contexte que lui. Cette architecture maximise l'équation suivante :

$$S = \frac{1}{I} \sum_{i=1}^{I} \sum_{-n \leq j \leq n, \neq 0} \log p(m_i | m_{i+j})$$

L'architecture Skip-gram avec échantillons négatifs (Negative Sampling) cherche à représenter chaque mot $m_i \in \mathrm{V}_I$ et chaque contexte $c \in \mathrm{V}_C$ par des vecteurs de d-dimensions ($\vec{m}_i$, $\vec{c}$), afin d'avoir des représentations vectorielles similaires pour les mots similaires. Ceci est fait d'une part, en maximisant le produit scalaire $\vec{m}_i . \vec{c}$ pour les paires (m_i,c) qui apparaissent dans le corpus D; d'autre part, en minimisant les exemples négatifs (m_i,c_{Neg}), qui ne sont pas nécessairement observés dans D. Les exemples négatifs sont créés en corrompant stochastiquement les paires (m_i,c) observées dans D, d'où le nom d'échantillons négatifs. En outre, le contexte ne se limite pas au contexte immédiat, et les instances d'apprentissage peuvent être créées en ignorant un nombre de mots dans son contexte, par exemple, m_{i-4}, m_{i-3}, m_{i+3}, m_{i+4}, d'où le nom de Skip-gram. Un exemple de ces deux architectures est présenté dans la figure 3.4 extraite de : Janod et al. (2015) :

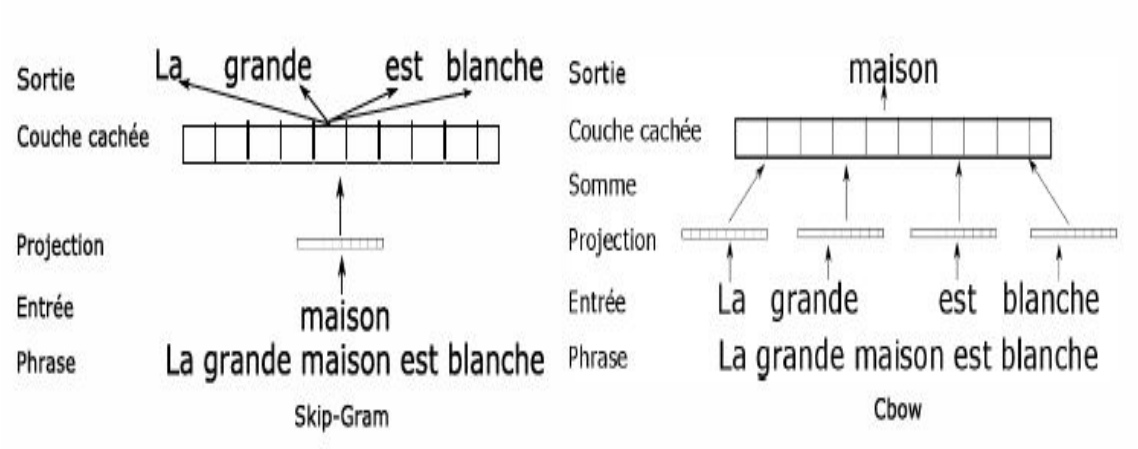


FIG. 3.4: Exemples des architectures CBOW et Skip-gram de Word2vec.

Ces modèles sont capables de capturer des régularités sémantiques et syntaxiques Mikolov et al. (2013c). En effet, la distance qui sépare la projection de deux mots peut représenter une relation complexe telle que la notion de singulier-pluriel ou masculin-féminin Mikolov et al. (2013a) comme le montre la figure 3.5 extraite de : Janod et al. (2015) :

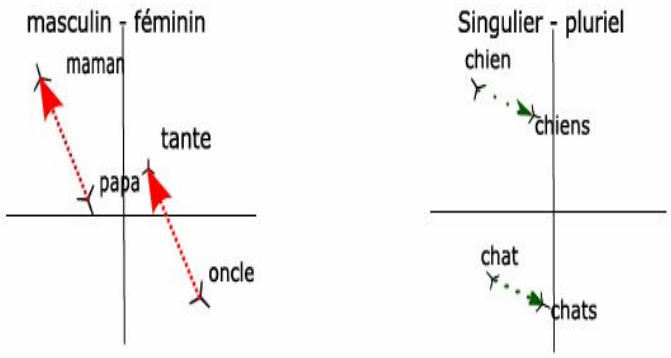


FIG. 3.5: Exemples de relations de mots dans l'espace Word2vec.

3.4.2 Avantages de Plongement lexical

Les outils utilisés pour le traitement automatique du langage existaient avant l'arrivée du plongement lexical, par exemple : les dictionnaires, les ontologies, etc. Néanmoins, un certain nombre d'obstacles est apparu face à ces derniers, on trouve notamment le problème de la dimensionnalité (fléau de la dimension). Un dictionnaire contient, en effet, plusieurs dizaines de milliers de mots différents, la plupart de ces mots

peuvent prendre plusieurs formes (au pluriel, au féminin ou avec différentes conjugaisons pour les verbes), le stockage est donc malaisé. Le plongement lexical permet de remplacer l'ensemble de ces mots par un ensemble de réels, ce qui est bien moins coûteux. En outre, pour plus de performance dans les résultats, ces outils demandent une analyse sémantique, qui n'est pas possible avec les techniques précédant. Ainsi, le plongement lexical est encore aujourd'hui le centre de plusieurs sujets de recherches, comme : la traduction, l'analyse de sentiments, reconnaissance vocale, dialogue homme-machine en langage naturel, l'aide à la création de ressources lexicales, identification automatique de termes similaires, etc.

3.5 Modèles d'apprentissage

3.5.1 Machine à vecteurs de support

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais support vector machine, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination. Les SVMs sont une généralisation des classifieurs linéaires. Schölkopf and Smola (2002).

3.5.2 Arbre de décision

L'arbre de décision est un algorithme d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont situées aux extrémités des branches (les « feuilles » de l'arbre), et sont atteints en fonction de décisions prises à chaque étape. L'arbre de décision est un outil utilisé dans des domaines variés tels que la sécurité, la fouille de données, la médecine, etc. Il a l'avantage d'être lisible et rapide à exécuter. Il s'agit de plus d'une représentation calculable automatiquement par des algorithmes d'apprentissage supervisé.Quinlan (1983).

3.5.3 Forêt d'arbres décisionnels

Les forêts d'arbres décisionnels ou forêts aléatoires (en l'anglais random forest classifier), elles font partie des techniques d'apprentissage

automatique. Cet algorithme combine les concepts de sous-espaces aléatoires et de bagging. L'algorithme des forêts d'arbres décisionnels effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents. Nisbet et al. (2009).

3.5.4 Naïf Bayésien

Le classifieur naïf bayésien est un type de classification bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses, appartenant à la famille des classifieurs linéaires. Il suppose que l'existence d'une caractéristique pour une classe est indépendante de l'existence d'autres caractéristiquesZhang (2004).

L'avantage du classifieur bayésien naïf est qu'il requiert relativement peu de données d'entraînement pour estimer les paramètres nécessaires à la classification, à savoir moyennes et variances des différentes variables. En effet, l'hypothèse d'indépendance des variables permet de se contenter de la variance de chacune d'entre elles pour chaque classe, sans avoir à calculer de matrice de covariance.

3.5.5 Réseau neuronal convolutif

En apprentissage automatique, un réseau de neurones convolutifs ou réseau de neurones à convolution (en anglais CNN ou ConvNet pour Convolutional Neural Networks) est un type de réseau de neurones artificiels acycliques (feed-forward), dans lequel le motif de connexion entre les neurones est inspiré par le cortex visuel des animaux. Les neurones de cette région du cerveau sont arrangés de sorte qu'ils correspondent à des régions qui se chevauchent lors du pavage du champ visuel. Leur fonctionnement est inspiré par les processus biologiques, ils consistent en un empilage multicouche de perceptrons, dont le but est de prétraiter de petites quantités d'informations. Les réseaux neuronaux convolutifs ont de larges applications dans la reconnaissance d'image et vidéo, les systèmes de recommandation et le traitement du langage naturel.Matusugu and Mori (2013).

3.6 Evaluation de systèmes

L'évaluation des classificateurs est très importante pour optimiser leurs capacités de prédiction, ils sont généralement évalués par la F-mesure qui est basée sur la précision et le rappel, ces mesures peuvent être définies par :

$$Precision = \frac{TP}{TP + FP}$$

Nombre d'exemples positifs correctement classés divisé par le nombre total d'exemples positifs prédits.

$$Rappel = \frac{TP}{TP + FN}$$

Nombre d'exemples positifs correctement classés divisé par le nombre total d'exemples positifs Où :

	Classe 1 prédite	**Classe 2 prédite**
Classe 1 actuelle	TP	FN
Classe 2 actuelle	FN	TN

- Classe 1 : Positive
- Classe 2 : Negative
- TP : True Positive
- TN : True Negative
- FP : False Positive
- FN : False Negative

Alors que, F-mesure est la moyenne harmonique de la précision et du rappel :

$$F = 2 \cdot \frac{Precision \cdot Rappel}{Precision + Rappel}$$

3.7 Conclusion

Dans ce chapitre, on a vu un peu de théories sur l'apprentissage automatique, quelques algorithmes (automatique et profond) et le plongement lexical notamment les modèles Skip-Gram, CBOW.

Dans le dernier chapitre, nous allons créer notre propre système, où nous allons collecter et étiqueter un ensemble de données à partir des réseaux sociaux, puis appliquer un ensemble d'algorithmes d'apprentissage automatique et les évaluer et essayer de les améliorer en utilisant l'apprentissage profond.

Chapitre 4

Approche proposée

4.1 Introduction

La détection automatique du langage abusif est une tâche difficile mais importante pour les médias sociaux en ligne. Notre travail est basé sur l'exploitation de plusieurs méthodes de l'apprentissage automatique, consistant à effectuer une classification sur un langage abusif, pour détecter les langues obscènes et offensants. Pour cela nous avons construit notre propre corpus de dialecte algérien, ce corpus contient plus de 80 000 publications de Facebook.

4.2 Création de la liste initiale

Nous avons créé manuellement une liste initiale de 525 mots[1] abusifs du dialecte Algérien collectée pendant de trois semaines, avec la participation de plusieurs personnes de différentes régions de l'Algérie, la figure suivante montre quelques mots de la liste initiale :

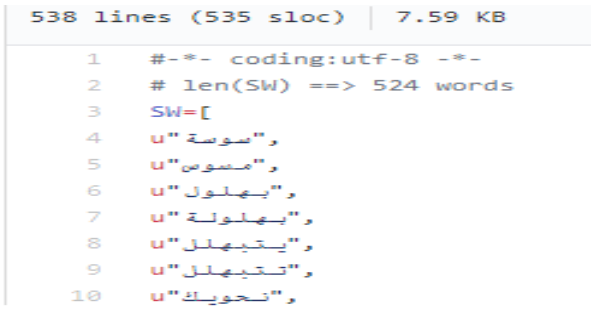


FIG. 4.1: Quelques mots de la liste initiale

1. https ://github.com/javaAgain/NLP-Dialecte-Algrienne/blob/master/initgit.py

A titre d'exemple, la table 4.1 représente 30 mots les plus fréquents dans notre corpus, y compris plusieurs mots en dialecte Algérien.

mot	fréq	mot	fréq	mot	fréq
اليوم	2525	الجزائري	2298	رايح	2046
بلاد	2496	بصح	2294	وسلم	2043
نظن	2444	قلي	2284	حيوان	2041
الشيتة	2439	مليون	2282	محمد	1995
أستغفر	2420	تافه	2261	حنا	1981
تصدي	2378	المنشور	2250	*حبة	1968
وين	2371	جام	2161	خير	1948
الدولة	2370	ناس	2142	دولة	1925
حملة	2332	جاهل	2116	تقول	1922
*يك	2318	كبير	2085	خرطي	1918

Tab. 4.1: Les 30 mots les plus fréquents dans le corpus

4.3 Collection de données

Après la création des comptes développeurs sur les plateformes de Twitter[2] et Facebook[3], on a développé un outil avec le langage de programmation Python, qui permet d'interroger l'API[4] de Facebook pour récupérer des postes et des commentaires recents des 30 pages algériennes les plus connues[5], Les APIs fournis par Facebook sont très limitées par rapport à twitter, seul le contenu des publications sont disponible et autorisé. Pour Twitter on a recuperé plus de 250000 tweets et retweets avec plusieurs attributs [id, user.id, user.location, user.genre, etc.][6], malheureusement Twitter est peu utilisé dans notre pays. Plus de ça, la plupart des tweets proviennent des pays du Golfe..

À titre d'illustration, nous donnons un exemple d'interrogation de Graph API dans la figure 4.2.

2. https ://developer.twitter.com
3. http ://developers.facebook.com
4. https ://graph.facebook.com/v2.12/
5. https ://github.com/javaAgain/NLP-Dialecte-Algrienne/blob/master/initgit.py
6. https ://github.com/javaAgain/NLP-Dialecte-Algrienne/blob/master/initgit.py

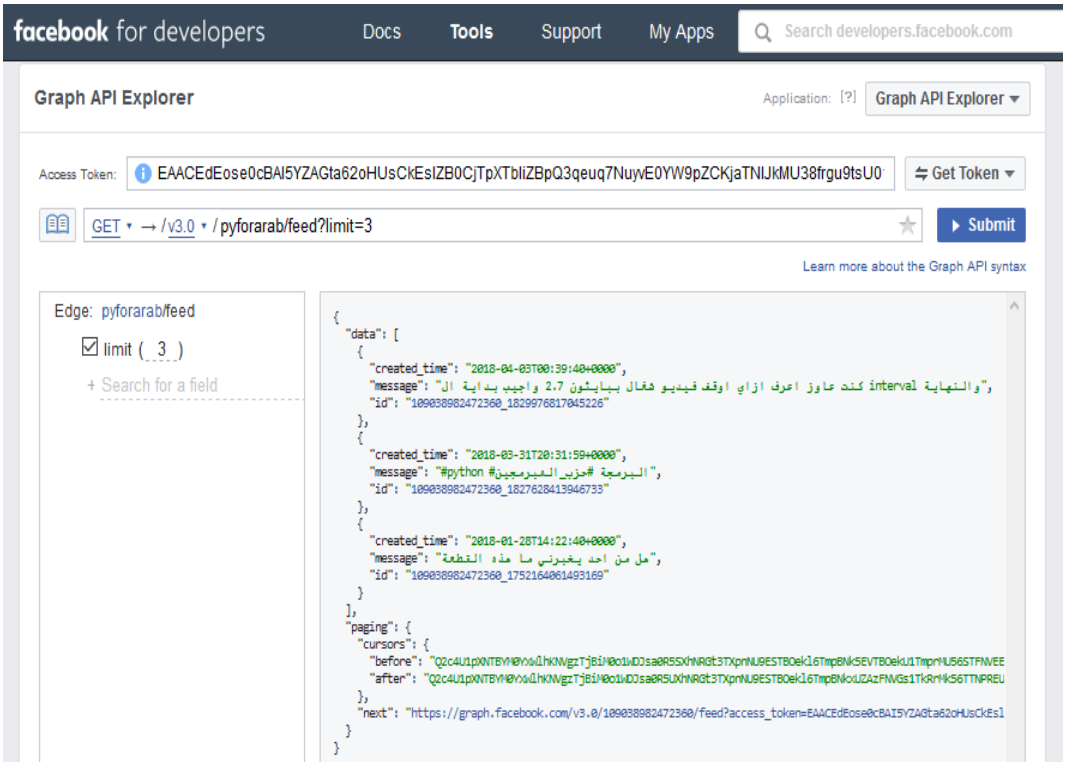


FIG. 4.2: Graph API de Facebook

Ensuite, la table suivante donne quelques informations sur la collection de nos données.

Source	Taille de données	Filtrage
Facebook	80 K	Avec (LI)
	80 K	Sans (LI)
Twitter	250 K	Avec (LI)

TAB. 4.2: Informations sur les données collectées

L'ensemble recueilli des commentaires et des postes a été filtrés durant le téléchargement via les 30 pages algériennes en utilisant les expressions régulières[7] qui permettent de garder seulement les textes arabes. Par la suite, on a éliminé les mots vides (ou stop word, en anglais) en utilisant NLTK[8] Library de python, qui sont tellement communs mais sont inutile de l'indexer ou de l'utiliser dans notre recherche.

L'étape d'acquisition de données à ce stade apparaît dans la figure 4.3

7. re.sub(params)
8. Natural Language Toolkit (NLTK) est une boîte-à-outil permettant la création de programmes pour l'analyse de texte.

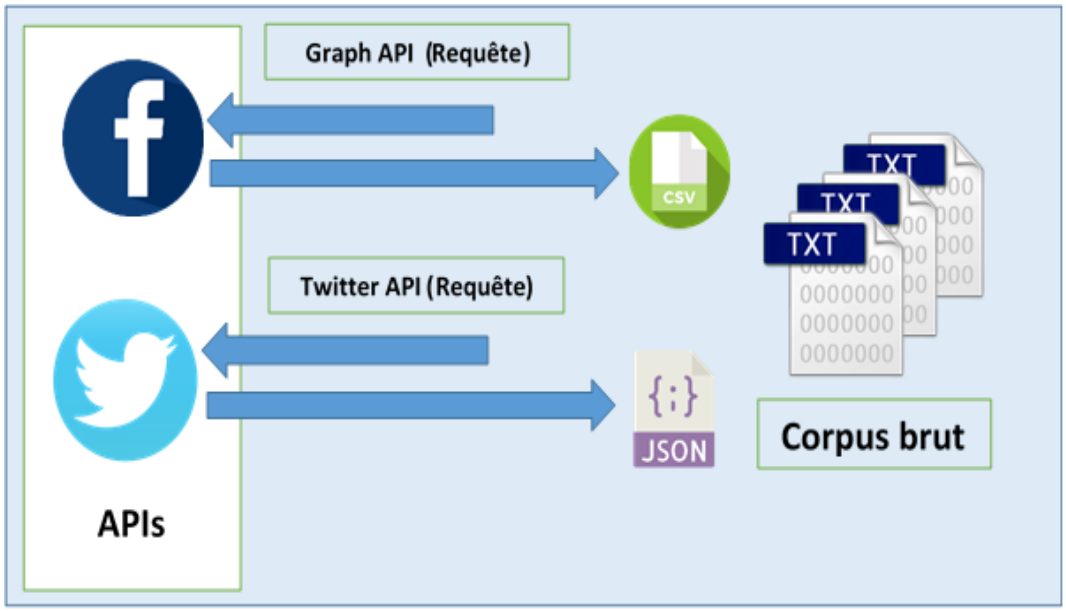


FIG. 4.3: Etape d'acquisition de données

4.4 Exploitation des données

4.4.1 Création d'un word2vec

Afin de générer un modèle word2vec nous avons utilisé la bibliothèque Gensim [9] avec les paramètres suivants :

- Taille de vecteur = 300
- Taille de fenêtre = 5
- sg = 0 (0 pour CBOW, 1 pour skip-gram)

De plus, nous considérons seulement les mots apparaissent plus de deux fois Dans notre modèle word2vec chaque mot est représenté par un vecteur de 300 dimensions de nombres réal entre [-1, 1]. Le calcule de similarité entre deux mots est obtenu par le cosinus de leurs vecteurs. Par la suite, on a obtenu un vocabulaire de taille 108699 mots.

Les figures suivantes montrent quelques exemples de notre modèle et illustrent les capacités et la cohérence de ce modèle.

9. Nous allons la touché dans section des outils requis.

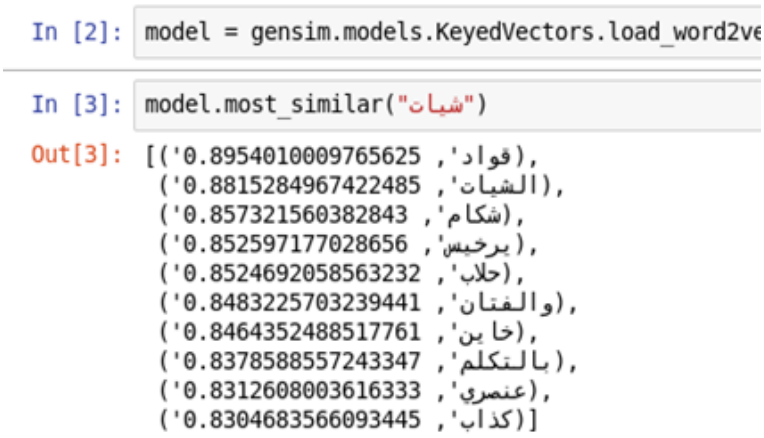


FIG. 4.4: Exemple de dix mots les plus proches d'un mot شيات

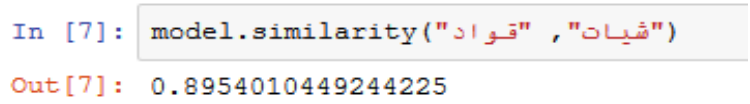


FIG. 4.5: Degré de similarité entre deux mots

FIG. 4.6: Exemple d'un mot hors contexte

FIG. 4.7: Degré de similarité entre deux contextes

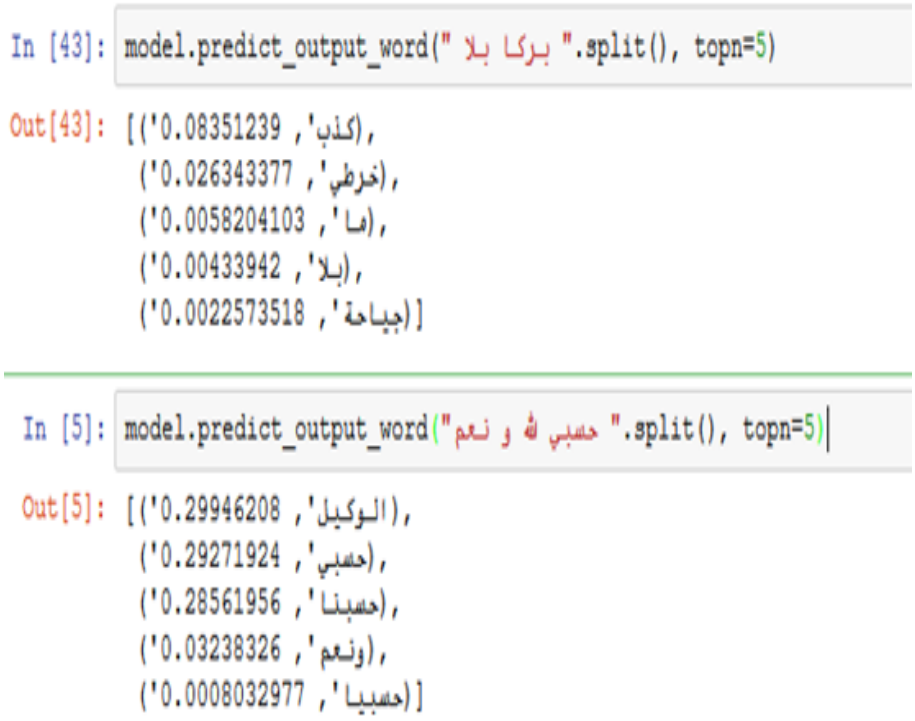


FIG. 4.8: Extraction les mots les plus proches d'un contexte

Dans la figure suivante on a utilisé Tensorboard pour visualiser 500 mots en trois dimensions :

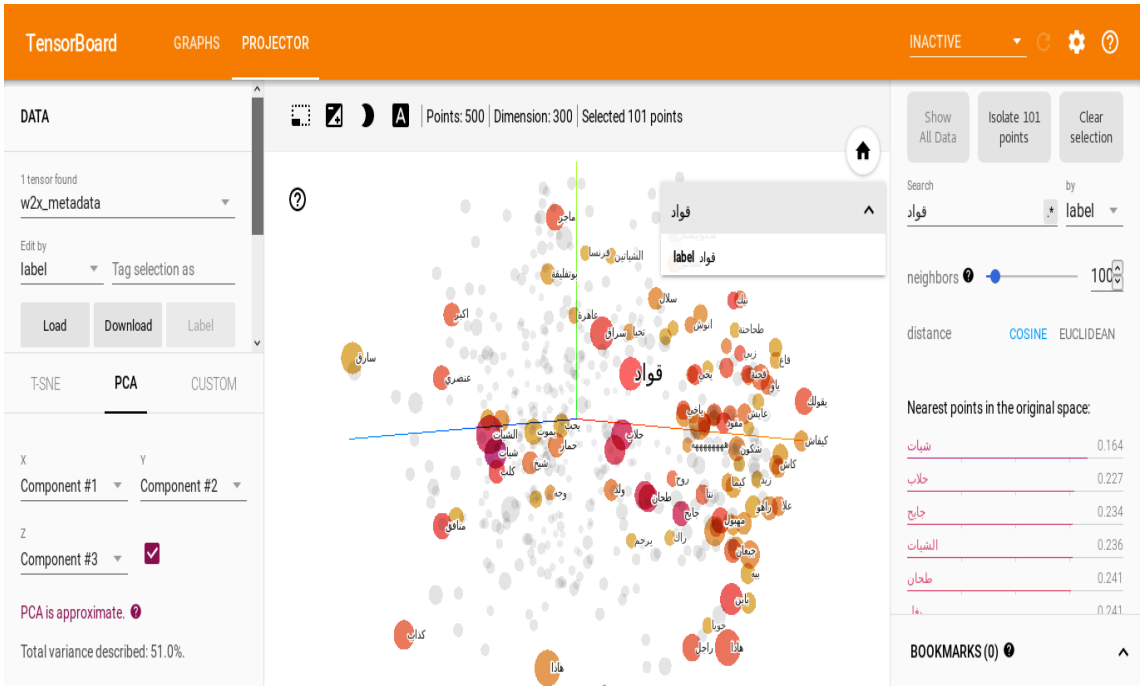


FIG. 4.9: Exemple de similarité des mots à un mot en trois dimensions

4.4.2 Extension de la liste initiale

Nous avons utilisé le modèle pour faire l'extension de la liste initiale (LI) et nous avons obtenu une nouvelle liste qui contient plus de 920 mots abusifs[10] de dialecte Algérien, cette liste est encore nettoyée manuellement en raison de supprimer quelques mots non pertinents.

4.4.3 Etiquetage

Dans notre travail, on a divisé les types de discours en trois catégories :

1. Offensant : Est un discours adressé aux individus, groupes ou entités morales, dans le but de les provoquer et de les harceler, tout ce qui est un comportement violent et haineux.

2. Obscène : Est un discours sale et non éthique qu'il est dirigés ou non dirigés.

10. https ://github.com/javaAgain/NLP-Dialecte-Algrienne/blob/master/data$_w2v.txt$

3. Normal : est un discours qu'il ne soit pas offensants pour les individus ou les groupes, par exemple une publication qui contient des nouvelles ou des éloges ou exprimer une opinion poliment.

Nous avons choisi 2500 publications/commentaires depuis l'ensemble des données collecté en utilisant la liste initiale (LI) et 2500 autres publications/commentaires téléchargées aléatoirement sans l'utilisation de la liste initiale. L'étiquetage est fait manuellement par nous-même, ou chaque publication classifié comme normal, offensant ou obscène. Le nombre total pour chaque catégorie est : normal : 2892, offensant 1497 obscene 611 de l'ensemble de publications.

4.4.4 Extraction des fonctionnalités

Dans l'apprentissage automatique, une fonctionnalité est une propriété mesurable d'un phénomène observé, L'extraction de fonctionnalités discriminantes est une étape fondamentale du processus d'apprentissage préalable à la classification. Les fonctionnalités sont généralement numériques mais il peut s'agir de chaînes de caractères, etc. Dans notre travail l'extraction des fonctionnalités depuis le texte basée sur six propriétés comme suit :

Fonctionnalité	Valeur
Contient au moins un mot abusif	0 ou 1
Contient au moins un mot purement obscène	0 ou 1
Le nombre des mots abusif	$>= 0$
Le nombre des mots dans le texte (publication)	> 0
Le discours est direct ou indirect	0 ou 1
Le degré de positivité	$\sum_{i=1}^{n} neg(mot_i) + \sum_{i=1}^{n} pos(mot_i)$

Tab. 4.3: Table to test captions and labels

- Où, le degré de positivité désine le sentiment de la publiction, pour çala on a utilisé un ensemble de données fournie par Mohammad Salameh [11].

- Contient au moins un mot purement obscène : une liste supplémentaire des mots purement obscènes.

11. http ://saifmohammad.com/WebDocs/lexiconstoreleaseonsclpage/SemEval2016-Arabic-Twitter-Lexicon.zip

- Le discours est direct ou indirect : est-ce que la publication est direct (vers un cible spécifique) ou une publication général, pour cela nous avons créé une liste des mots adressés arabes qui indiquent que le discours est direct, voici quelques mots de la liste : « أنتَ, نت, انت, نتا, انتا , أنتِ, أنتي, نتي, نتيا, أنتم, انتم, نتم,نتوما, نتما, أنتنّ, انتن, إياكَ, ياك, ردبالك, تابه, إياكِ, ردي, بالك, تابهي, إياكم,ردوبالكم, تابهو, إياكنّ»

4.5 Expérimentation

4.5.1 Apprentissage automatique

Afin de bien choisir nos classificateurs nous avons utilisé ce schéma[12] :

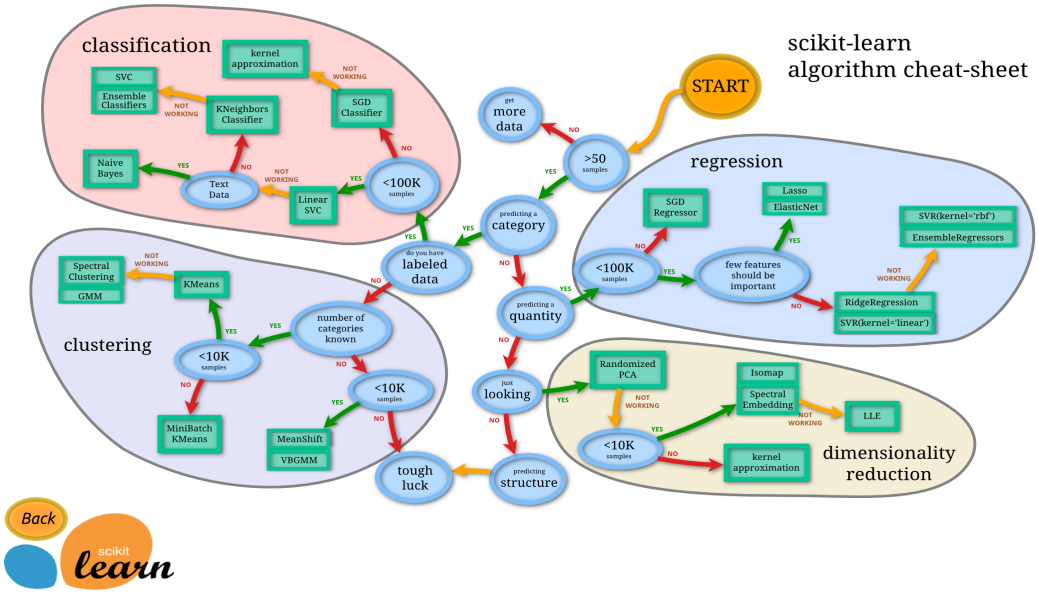


FIG. 4.10: Schéma assistant à choisir l'algorithme approprié

La taille de notre données est 5k échantillons alors en utilisant le schéma précédent nous choisissons quatre classificateurs : machine à vecteurs de support, arbre de décision, forêt d'arbres décisionnels, naïve bayésienne.

L'architecture générale des étapes de l'utilisation de l'apprentissage automatique est illustré dans la figure ci-dessous :

12. Extraite de http ://scikit-learn.org/stable/tutorial/machine$_l earning_m ap/index.html$

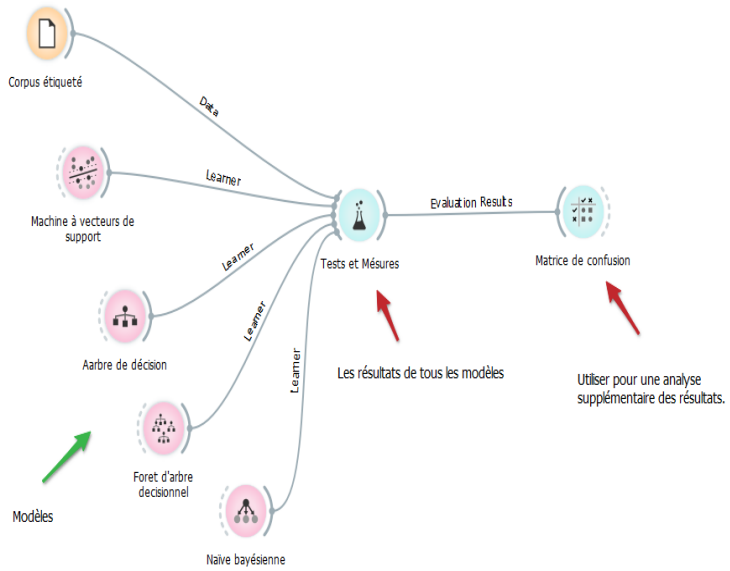


FIG. 4.11: Architecture générale de système

4.5.2 Apprentissage profond

Nous avons ainsi utilisé le réseau neuronal convolutif (CNN) pour essayer d'améliorer les résultats obtenus pour les quatre classificateurs d'apprentissage automatique en utilisant les couches illustrées dans la figure suivante :

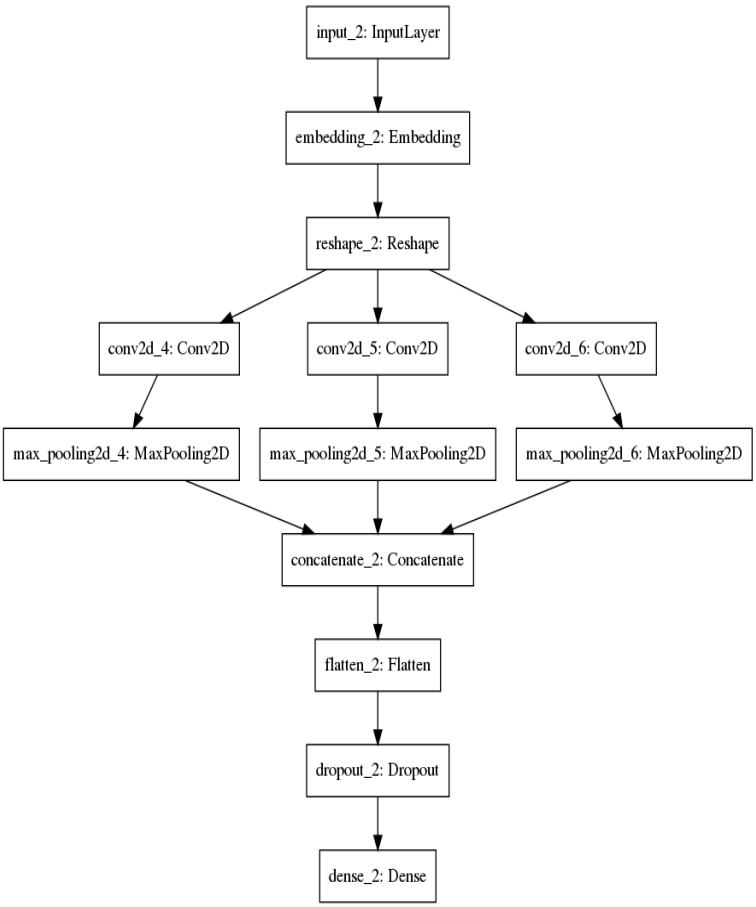


FIG. 4.12: Les couches de notre modèle CNN.

4.5.3 Résultats avec apprentissage automatique et apprentissage profond

Nous avons divisé notre ensemble de données étiquetées en deux sous-ensembles, ensemble d'entraînement (80%) et ensemble d'évaluation (20%). On a entraîné les classificateurs avec l'ensemble l'entraînement, ensuite nous les testons avec l'ensemble d'évaluation. Les résultats de nos expérimentations sont présentés dans le tableau suivant :

Type d'apprentissage	**Algorithme**	**F-mésure**
apprentissage automatique	Machine à vecteurs de support	0.80
	Arbre de décision	0.76
	Forêt d'arbres décisionnels	0.77
	naïve bayésienne	0.80
apprentissage profond	réseau neuronal convolutif	0.85

TAB. 4.4: Résultats obtenus avec différents types et méthodes d'apprentissage

Plusieurs détails sur la capacité de prédiction de notre système pour chaque algorithme sont exprimés dans le schéma suivant :

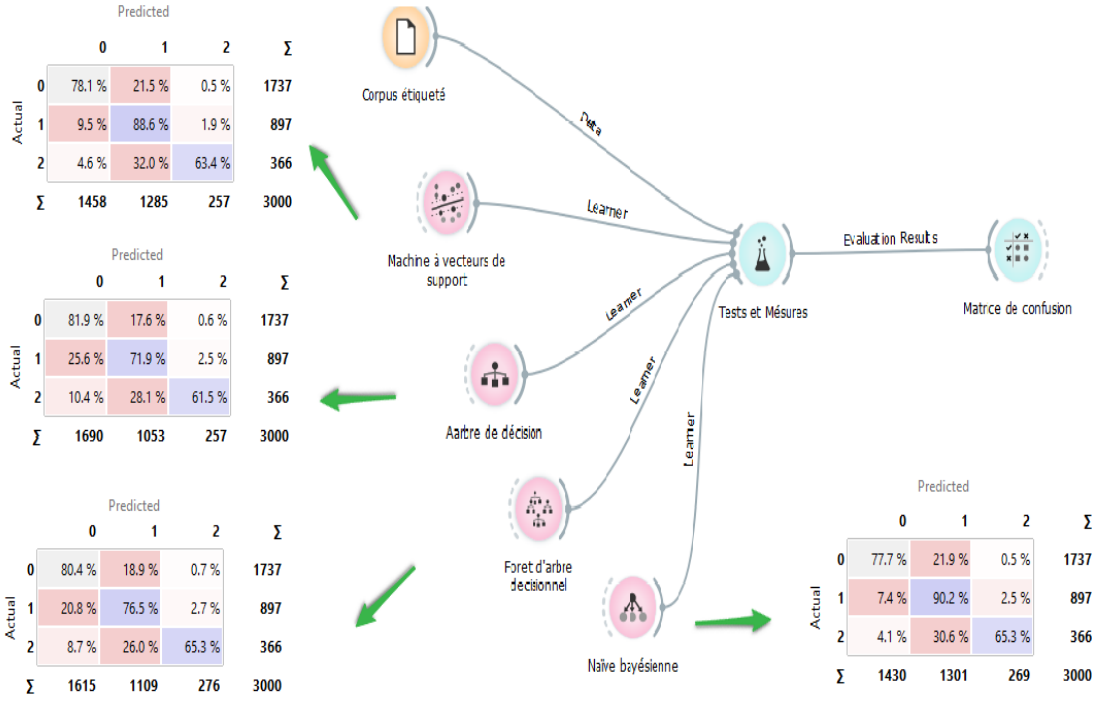


FIG. 4.13: Capacité de prédiction des algorithmes.

4.6 Les outils requis

- Python : est puissant et rapide joue bien avec les autres courses partout est convivial et facile à apprendre est Open source. Python est un langage de programmation open source, puissant et facile à apprendre. Il dispose de structures de données de haut niveau, sa syntaxe est élégante, que son typage est dynamique et qu'il est interprété. Il est idéal pour l'écriture de scripts et le développement rapide d'applications dans nombreuses plateformes et nombreux domaines : développement Web et Internet, Scientifique et Numérique, éducation, GUI de bureau, développement de logiciels, applications commerciales, etc. [13]

13. https ://docs.python.org/fr/3.5/tutorial/

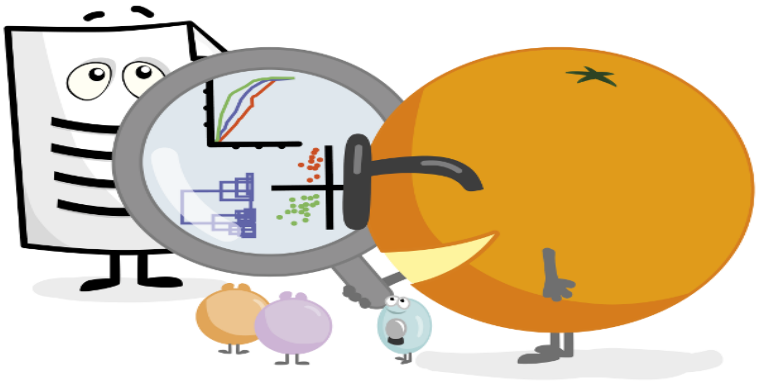
- Gensim : est une bibliothèque libre de Python conçue pour extraire automatiquement des sujets sémantiques des documents, pour traiter les textes numériques bruts et non-structurés. Les algorithmes de gensim, tels que Word2vec, n'ont besoin que d'un corpus de documents en texte brut.[14]

- Scikit-learn : est une bibliothèque Python pour les algorithmes d'apprentissage automatique, elle se concentre sur l'apprentissage de la machine à des non-spécialistes en utilisant un langage général de haut niveau. L'accent est mis sur la facilité d'utilisation, la performance, la documentation et la cohérence de l'API. Il a des dépendances minimales et est attribué sous la licence BSD simplifiée, enchaînant son utilisation dans les milieux académiques et commerciaux. Le code source, les fichiers binaires et la documentation peuvent être téléchargés depuis `http://scikit-learn.sourceforge.net`.

14. https ://radimrehurek.com/gensim/intro.html

- Keras : est une API de réseaux neuronaux de haut niveau, écrite en Python et capable de fonctionner sur TensorFlow, CNTK ou Theano. Il a été développé dans le but de permettre une expérimentation rapide. Pouvoir passer de l'idée au résultat avec le moins de retard possible est la clé pour faire de bonnes recherches.[15]

- Orange3 : est un outil open source d'apprentissage automatique, de visualisation de données et d'analyse de données interactifs avec une grande boîte à outils.[16]



4.7 Conclusion

Dans ce chapitre, nous avons proposé une approche pour détecter les contenus abusifs sur Facebook en dialecte algérien, pour cela on a exploité cinq classificateurs d'apprentissage automatique et profond en utilisant notre propre corpus collecté depuis Facebook. Nous avons

15. https ://keras.io/
16. https ://orange.biolab.si/

trouvé un très bon résultat de F-mesure de 0.85, où le modèle CNN a un meilleur résultat que tous les classificateurs machine à vecteurs de support, arbre de décision, forêt d'arbres décisionnels, naïve bayésienne comparés avec les valeurs de 0.80, 0.76 et 0.77 et 0.80 respectivement.

Conclusion générale et perspectives

Notre travail fournit un grand effort pour essayer de lutter contre la propagation du langage abusif dans le dialecte algérien sur les réseaux sociaux. En outre, notre système permet de classifier le langage abusif comme obscène, offensant ou normal.

Nous avons étudié les caractéristiques de la langue arabe en général et le dialecte algérien en particulier et sa complexité, nous avons créé une liste des mots abusifs du dialecte algérien, cette dernière a été utilisé dans la collection de données à partir Facebook et Twitter. De plus, nous avons utilisé notre modèle word2vec afin d'agrandir de façon automatique cette liste. Ce modèle est créé en utilisant seulement les données recueillies par Facebook car les plus part données de twitter représente que les pays de Golf.

Notre approche exploite différents algorithmes d'apprentissage automatique et profond, qui ont été entrainé sur nos propres données. Ces algorithmes ont été évalués et on a obtenu des bons résultats dans les algorithmes d'apprentissage automatique et d'excellents résultats dans l'algorithme d'apprentissage profond. Nous ne sommes pas au courant des autres travaux existants de mots ou des contenus abusifs en utilisant le dialecte algérien. Notre travail apporte les contributions suivantes :

- Une liste initiale de 525 des mots abusifs du dialecte algérien de différentes régions.
- La liste précédente est élargie jusqu'à 950 mots similaires en utilisant le modèle word2vec entrainé sur l'ensemble de tweets et de commentaires.
- Un ensemble des Id de 50 pages algériennes plus connues et utilisée sur Facebook.

- Un ensemble de plus de 80K commentaires et statuts via Facebook.
- Un modèle word2vec à partir d'un corpus de dialecte Algérien qui contient 108K mots. Ce modèle peut être utilisé dans des travaux connexes.
- Un ensemble de plus de 250K tweets via Twitter API.

Les différentes pistes explorées pendant ce travail nous ont amenées à envisager de nombreuses perspectives. Nous présentons ici celles qui nous paraissent les plus prometteuses.

Nous essayons de généraliser ce travail sur la langue amazighe et d'autres dialectes tels que les dialectes maghrébins dans les pays voisins comme la Tunisie et le Maroc, et aussi d'intégrer d'autre sites des réseaux sociaux qui utilisent d'autres moyens de discours telles que la vidéo et la voix à titre d'exemple YouTube.

Bibliographie

Diana Abbas. What's in a location. talk at twitter flight. 2015.

Ehab A Abozinadah and Jr James H. Jones. A statistical learning approach to detect abusive twitter accounts. *Proceedings of the International Conference on Compute and Data Analysis - ICCDA*, pages 6–13, 2017.

Ehab A. Abozinadah, Alex V. Mbaziira, and James H. Jones Jr. Detection of abusive accounts with arabic tweets. *International Journal of Knowledge Engineering*, 1:113–118, 2015.

Salih J Al-Toma. The problem of diglossia in arabic : A comparative study of classical and iraqi arabic. *Cambridge, Mass : Harvard University Press*, 21, 1969.

Björn and Utpal Kumar. Using convolutional neural networks to classify hate-speech. *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, 2017.

H Cheng, Xing X, Liu X, and Lv Q. An iterative social based classifier for adult account detection on twitter. *IEEE Transactions on Knowledge and Data Engineering*, pages 1045–1056, 2015.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International AAAI Conference on Web and Social Media*, pages 512–515, 2017.

George Forman. Bns feature scaling : an improved representation over tfidf for svm text classification. *In Proceedings of the 17th ACM conference on Information and knowledge management.*, 2008.

J.-E Gombert, Colé P., Valdois S., Goigoux R., Mousty P., and Fayol M. Enseigner la lecture au cycle 2. *Paris : Nathan.*, 2000.

N. HABASH. Introduction to arabic natural language processing. *Morgan Claypool Publishers*, 2010.

Killian Janod, Mohamed Morchid, Richard Dufour, and Georges Linares. Apport de l'information temporelle des contextes pour la représentation vectorielle continue des mots. 2015.

M Litvak, Last M, Aizenman H Gobits I, and Kandel. A language-independent graph-based keyphrase extractor. advances in intelligent web mastering. pages 121–130, 2011.

ManuCE. Connexions - manuel pour la lutte contre le discours de haine en ligne par l'éducation aux droits de l'homme, edition révisée. *Conseil de l'Europe*, 2016.

Masakazu Matusugu and Katsuhiko Mori. Subject independent facial expression recognition with robust face detection using a convolutional neural network. 16:555–559, 2013.

Peggy McIntosh. Understanding prejudice and discrimination. pages 191–196, 2003.

Hadjira Medane. L'interférence comme particularité du «français cassé» en algérie. *TIPA. Travaux interdisciplinaires sur la parole et le langage [En ligne]*, 2015.

R Mihalcea and Tarau P. Textrank : Bringing order into texts. pages 404–411, 2004.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR abs/1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *In Advances in Neural Information Processing Systems 26 (NIPS 2013). Curran Associates, Red Hook, NY, USA*, pages 3111–3119, 2013b.

Tomas Mikolov, YIH W.-T., and ZWEIG G. Linguistic regularities in continuous space word representations. *In HLT-NAACL*, pages 746–751, 2013c.

H Mubarak, Darwish K, and Magdy. Abusive language detection on arabic social media. *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, 2017.

Robert Nisbet, John Elderand, and Gary Miner. *Handbook for Statistical Analysis And Data Mining*. Academic Press, 2009.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. *In Proceedings of the 25th International Conference on World Wide Web*, pages 145–153, 2016.

L Page. Method for node ranking in a linked database, united states patent : 6285999. 2001.

Turney P.D. Thumbs up or thumbs down ?: Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (Stroudsburg, PA, USA, 2002)*, pages 417–424, 2002.

R. Quinlan. *Learning efficient classification procedures, Machine Learning : an artificial intelligence approach*. Morgan Kaufmann, 1983.

Russell S. and Norvig P. *Artificial intelligence- a modern approach, 2ed.* 2003.

Bernhard Schölkopf and Alexander J. Smola. Learning with kernels : Support vector machines, regularization, optimization and beyond. *MIT Press*, 2002.

M Singh, Bansal D, and Sofat S. Behavioral analysis and classification of spammers distributing pornographic content in social media. *Social Network Analysis and Mining*, page 41, 2016.

B Sriram, Fuhry D, Demir E, Ferhatosmanoglu H, and Demirbas M. Short text classification in twitter to improve information filtering. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (New York, NY, USA, 2010)*, pages 841–842, 2010.

Cynthia VanHee, Ben Verhoeven, Els Lefever, Guy De Pauw, Vťeronique Hoste, and Walter Daelemans. Guidelines for the fine-grained analysis of cyberbullying. *Technical report, LT3, Ghent University, Belgium*, 2015.

Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people ? predictive features for hate speech detection on twitter. *In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies. ACL, San Diego, California*, pages 88–93, 2016.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse : A typology of abusive language detection subtasks. pages 78–82, 2017.

S Wu, Hofman J.M, Mason W.A, and Watts D.J. Who says what to whom on twitter. *Proceedings of the 20th International Conference on World Wide Web (New York, NY, USA, 2011)*, pages 705–714, 2011.

Harry Zhang. The optimality of naive bayes. *Conférence FLAIRS2004*, 2004.