Classification automatique des SMS – analyse des caractéristiques langagières de deux groupes d'âge

Mémoire de master
de philologie française
Département des langues
Université de Helsinki
Avril 2018
Julia Poutanen

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto – Fakultet/Sektion – Faculty	Laitos – Institution – Department	
Humanistinen tiedekunta	Kielten osasto	
Tekijä – Författare – Author		
Julia Poutanen		
Työn nimi – Arbetets titel – Title		
Classification automatique des SMS – analyse des caractéristiques langagières de deux groupes d'âge		
Oppiaine – Läroämne – Subject		
Ranskalainen filologia		
Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä– Sidoantal – Number of pages
Pro gradu -tutkielma	Huhtikuu 2018	53 + 1

Tiivistelmä – Referat – Abstract

Tämän tutkielman tarkoituksena on tutkia aikuisten ja nuorten kielenkäytön välisiä eroja ranskankielisessä tekstiviestiaineistossa. Eroja tutkitaan luomalla koneoppimista hyödyntävä automaattinen luokittelija, joka kykenee erottelemaan aikuisten ja nuorten tekstiviestit toisistaan. Työssä tarkastellaan luokittelijan antamia tuloksia ja pyritään selvittämään, miten luokittelijan toimintaa voidaan parantaa kielenkäytöstä saatujen tietojen valossa esimerkiksi tutkielmassa määritellyillä piirteillä (engl. *feature*).

Teoriaosassa käsitellään tekstiviestikielen piirteiden lisäksi iän ja kielenkäytön välistä suhdetta sekä kieliteknologialle ja korpuslingvistiikalle tärkeitä käsitteitä. Menetelmänä käytetystä tilastollisesta luokittelijasta esitellään siihen liittyvä olennainen teoria sekä muita tutkielman kannalta tärkeitä käsitteitä.

Tutkielman aineisto on kerätty Montpellier'ssä, Ranskassa vuonna 2011, ja se koostuu silloiseen tutkimukseen osallistuneiden lähettämistä tekstiviesteistä. Tekstiviestejä on yhteensä 88 000, ja niistä noin 70 000 käytetään tutkielmassa.

Analyysissä keskitytään sekä kielellisiin että teknisiin piirteisiin: tarkastelun kohteina ovat täten sekä malli että aineiston kielelliset piirteet. Tutkimustuloksista selviää, että luokittelija toimii varsin hyvin tekstiviestien erottelussa, mutta tutkielmassa erikseen määritellyt piirteet eivät paranna merkittävästi luokittelijan toimintaa. Piirteistä voidaan kuitenkin tehdä joitakin johtopäätöksiä: tekstiviesteille on tyypillistä keskustelunomainen kielenkäyttö viestin lähettäjän ja vastaanottajan välillä sekä puhekieli. Analysoitujen viestien perusteella voidaan nähdä, että tekstiviestikielen ominaispiirteisiin kuuluvat ääntämistä ja foneettista muotoa heijastavat sanamuodot ja että tekstiviesti muodostanee oman rekisterinsä ranskan kielessä.

Avainsanat – Nyckelord – Keywords

tekstiviestit, koneoppiminen, ikä, luokittelu, ranskan kieli, kieliteknologia, SMS, apprentissage automatique, âge, langue française, linguistique informatique

Säilytyspaikka – Förvaringställe – Where deposited

Keskustakampuksen kirjasto

Muita tietoja – Övriga uppgifter – Additional information

Table des matières

1 Introduction

De nos jours, il est presque impossible de penser une vie sans ordinateurs ou téléphones portables. La diffusion massive des appareils portables et des ordinateurs personnels a rendu possible l'invention de toutes sortes d'applications qui peuvent faire usage de données énormes. Le développement des applications ne serait pas possible avec des systèmes basés sur des règles, parce que les écrire serait coûteux et prendrait beaucoup de temps. Pour cette raison, il n'est pas surprenant que les développeurs aient tendance à se servir de l'apprentissage automatique (angl. *machine learning*), avec lequel les procédures peuvent être automatisées plus facilement [1]. Maintenant que la performance d'un ordinateur ne pose plus de problèmes, l'exécution des tâches utilisant des mégadonnées, le *big data*, est devenue facile. L'automatisation des tâches dont la réalisation serait impossible par des moyens manuels est presque banale aujourd'hui. C'est pour cela qu'il est important d'étudier les façons les plus appropriées de se servir des données, quel que soit l'objectif.

Les applications de l'apprentissage automatique sont nombreuses. L'une d'entre elles se concentre sur la détermination des indicateurs comme l'âge ou le sexe, *author profiling* en anglais. Le principe sous-jacent dans cette application est que dans n'importe quel type de texte, il est possible de trouver des caractéristiques d'après lesquelles subdiviser les auteurs en groupes, par exemple groupe de femmes et celui d'hommes (Soler-Company & Wanner 2014 : 1315). Avec une grande quantité de données, le profilage serait impossible à faire à la main et, comme la bonne réponse varie selon la personne qui les examine, on ne trouverait jamais de consensus, tandis que pour l'ordinateur, il n'existe qu'une réponse. Le profilage est utilisé dans plusieurs domaines : en criminalistique, pour détecter l'auteur d'un message par exemple sur Internet (par exemple dans Chandramouli *et al*. 2009), dans le cadre commercial, pour voir quels produits sont aimés par les clients, et naturellement aussi en linguistique, pour découvrir quels éléments de la langue sont ceux qui distinguent les différents groupes les uns des autres. Ce type de profilage sera l'objet de notre analyse dans ce travail.

Aujourd'hui, une grande partie de la communication entre les gens se fait sur Internet ou est médiée par un appareil, par exemple les SMS, qui, grâce à leurs restrictions et leurs qualités (Fairon, Klein & Paumier 2006 : 55), sont un moyen de communication intéressant du point

1. Dans les systèmes ou applications basés sur des règles, les règles, c'est-à-dire les phases pour exécuter la procédure, sont définies manuellement, contrairement à l'apprentissage automatique, dans lequel les phases ne sont pas définies par l'humain mais intérieurement par le programme. Par contre, l'apprentissage automatique se base sur l'idée de l'expérience, ce qui veut dire que le modèle créé peut apprendre de chaque expérience pour améliorer les expériences à venir.

de vue linguistique, au point qu'on s'inquiète des connaissances de la norme langagière [2]. Les applications développées à l'aide des données SMS sont utiles non seulement pour mieux faire la classification des messages SMS mais aussi pour la classification d'autres messages dans la communication médiée par ordinateur, par exemple les messages sur Facebook ou sur des forums de discussion. Comme la langue utilisée dans les SMS varie beaucoup, la classification n'est pas facile (Fairon, Klein & Paumier 2006 : 44–45), mais c'est justement pour cela qu'il est important de trouver au moins quelques solutions pour la faire, car plus on a d'informations, plus le développement des applications deviendra facile. C'est pour cela que nous avons décidé d'examiner les SMS.

Pour avoir une perspective complète, et comme c'est le principe par exemple dans les humanités digitales, notre étude sera multidisciplinaire, combinant des aspects linguistique, statistique et linguistique informatique. Dans ce mémoire, le but est de trouver des caractéristiques qui permettent de distinguer les deux groupes suivants : les jeunes et les adultes. Nous utiliserons des méthodes d'apprentissage automatique, plus précisément des méthodes probabilistes bayesiennes pour le faire. Pour plus de précision, nous ferons notre classification en déterminant les caractéristiques, ou plutôt les attributs (angl. *features*), qui distinguent le mieux les deux groupes. Par la suite, nous analyserons de quel type elles sont et essayerons de préciser ce qui serait caractéristique du langage SMS dans les deux groupes. Notre deuxième tâche sera d'examiner la manière dont le traitement du langage peut être effectué si les données contiennent du *bruit* (angl. *noise*), des éléments inutiles ou qui font trop varier le résultat. Par ailleurs, nous ferons l'analyse des méthodes utilisées quant à l'efficacité (angl. *efficiency*) et la performance.

Notre travail sera réparti en quatre sections principales. Nous partirons de la description du corpus utilisé. Ensuite, nous aborderons la théorie linguistique de différents points de vue. En premier sera abordé le langage des jeunes et des adultes et la relation entre les SMS et l'âge. Ensuite, nous passerons au traitement des difficultés liées à ce qu'un mot veut dire en traitement automatique et rappellerons quelques termes sur la terminologie de la linguistique de corpus. Nous terminerons le chapitre portant sur la théorie linguistique par les points importants sur le langage SMS. Dans le chapitre 4, nous traiterons de la théorie informatique, c'est-à-dire la méthode que nous utiliserons. Le chapitre 5 sera consacré à l'analyse des modèles que nous testerons pour ensuite voir les résultats que nous avons obtenus au moyen des différents

2. Ce thème est traité par exemple dans David, Jacques & Goncalves, Harmony (2007) « L'écriture électronique, une menace pour la maitrise de la langue ? » *Le français aujourd'hui* 156 : 1. 39–47. et dans Cougnon, Louise-Amélie (2010) « Orthographe et langue dans les SMS. Conclusions à partir de quatre corpus francophones. » *Ela. Études de linguistique appliquée* 160 : 4. 397–410.

modèles. En un mot, nous analyserons les attributs qui ont été les plus distinctifs pour les deux groupes et des messages SMS pour compléter ces points de vue. Nous continuerons par une discussion des résultats et terminerons par une conclusion sur ce qui a été fait.

2 Corpus

Le corpus que nous utiliserons est constitué d'environ 88 000 SMS recueillis en 2011 par une équipe de chercheurs montpellierains, lors du projet *sms4science*, dont l'objectif était d'obtenir des SMS pour rassembler des données de recherche (www1). Tout le corpus est téléchargeable en ligne [3]. Les participants avaient été invités à envoyer leurs messages SMS et à répondre à un questionnaire sociolinguistique, dans lequel on leur avait demandé d'indiquer, en plus des informations personnelles, leurs habitudes langagières et des renseignements sur le téléphone qu'ils utilisent. Dans cette étude, le questionnaire sera seulement utilisé pour obtenir l'âge des participants. Au total, 425 participants ont accepté d'envoyer leurs messages. Parmi ces 425 participants, 422 ont aussi répondu au questionnaire. Les participants ont pu choisir quels messages donner pour l'étude. En moyenne, chaque participant a envoyé 209 messages, mais la médiane est de 28 messages, c'est-à-dire que la moitié du nombre des messages est égale ou inférieure à 28. En déduction, il y a beaucoup de variation dans le nombre des messages que les participants ont envoyés.

Dans notre analyse, nous nous pencherons sur la différence entre les deux groupes : les jeunes et les adultes, c'est-à-dire que nous ne prendrons pas en compte tous les messages SMS du corpus. Nous avons délimité les groupes ainsi : les jeunes incluent les participants âgés de 10 à 21 ans, les adultes ceux de 30 à 55 ans. Nous ferons une classification en essayant de trouver des attributs qui différencient les jeunes et les adultes. Nous avons choisi de comparer ces deux groupes à cause de leur comportement langagier assez différent. Les jeunes sont des inspirateurs de nouvelles formes du langage, tandis que les adultes, qui sont dans la vie professionnelle, sont plutôt conservateurs et essaient de retenir la forme standard (voir partie 3.1 ; Thibault 1997 : 23–25).

Pour la validité de notre recherche, il faut considérer la distribution d'âge des participants. La médiane des âges se situe à 21 ans, de sorte que la distribution est asymétrique vers la droite, ce qui veut dire qu'il y a plus de personnes jeunes que d'adultes et que la courbe n'a pas la forme d'une cloche de la loi normale (Figure 1). Cela résulte du fait que durant la collecte, les participants ont été autosélectionnés au lieu d'être aléatoirement choisis. Panckhurst *et al.* (2013) nous indiquent que c'est l'aspect médiatique qui pousse les jeunes à participer à la collecte des messages. Il peut aussi être constaté que, naturellement, les étudiants à l'université sont les plus proches des chercheurs et ont, selon toute probabilité, été mieux au courant de cette étude. De plus, selon Fairon, Klein & Paumier (2006 : 17–18), ce résultat indique seulement que les jeunes

3. `http://88milsms.huma-num.fr/corpus.html`

utilisent plus les SMS et que la distribution est ainsi plutôt celle des utilisateurs des SMS.

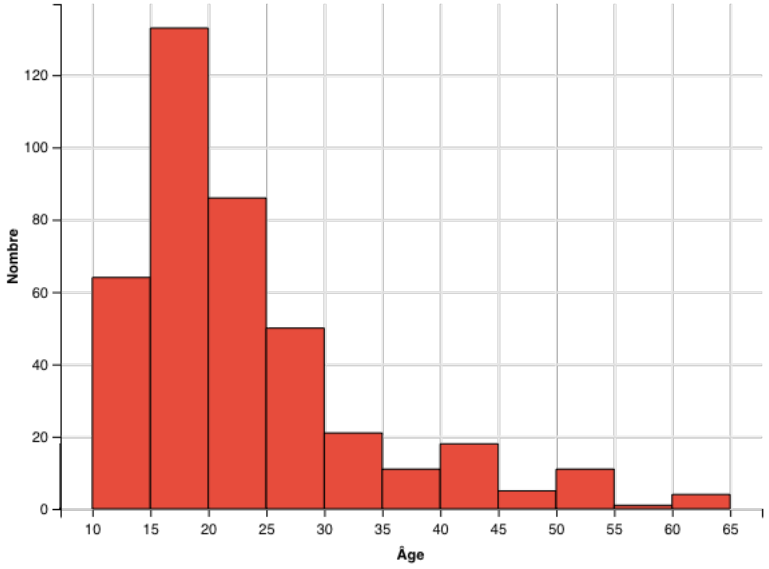


FIGURE 1 – La distribution d'âge des participants dans les données entières.

Au total, les adultes ont envoyé 9995 messages, les jeunes 60 746 messages, ce qui veut dire que le nombre de messages des jeunes est environ six fois plus grand que celui des adultes. Chaque participant adulte a envoyé en moyenne 89 messages, tandis que le chiffre correspondant chez les jeunes est de 248. La longueur d'un message envoyé par un jeune est en moyenne de 13,3 mots, alors que pour les adultes, elle est de 12,5 mots. La distribution asymétrique de l'âge, et par conséquent des messages, ne peut pas être ignorée et c'est pourquoi nous devrons tester notre modèle à la fois avec toutes les données et, d'autre part, avec seulement une partie des messages jeunes, autrement dit avec environ 10 000 messages, pour réduire la différence entre les deux groupes. De cette façon, nous obtenons un corpus équilibré, car les parties sont en équilibre (Manning & Schütze 1999 : 120). Par ailleurs, il est possible de voir l'effet que l'asymétrie peut avoir dans la classification.

3 Cadre théorique

Cette partie résumera les théories pertinentes pour la compréhension de l'aspect linguistique de notre recherche. Nous partirons de l'âge : nous examinerons les manières dont l'âge a une influence sur le langage que l'on utilise et comment cela se voit dans les SMS en particulier. Nous continuerons en expliquant les notions importantes dans le traitement de mots en nous intéressant aux difficultés qui existent en informatique et en linguistique de corpus. La dernière section de ce chapitre sera consacrée au langage SMS. Nous rappellerons les types de procédés utilisés dans les SMS et l'opinion générale des linguistes sur le langage SMS.

3.1 L'impact de l'âge sur le langage

Dans cette section, nous nous demanderons comment l'âge agit sur le langage. L'accent sera particulièrement mis sur la relation entre le langage des adultes et celui des jeunes. Nous traiterons aussi les différentes manières de considérer l'âge. Pour terminer cette section, nous réfléchirons à la manière dont l'âge se voit dans les SMS.

3.1.1 Variable âge

L'âge est une des variables dans le langage des locuteurs. Il est défini comme « l'appartenance à une certaine génération d'usagers de la langue » (Boyer 2001 : 27). Pourtant, l'âge n'est pas une variable simple en sociolinguistique : s'agit-il du changement historique reflété dans le langage ou du changement dans la vie de l'individu (Eckert 1997 : 151) ? Est-il donc possible de dire que la variation relève d'une génération ou s'agit-il du changement normal dans le langage de l'individu en question ?

Les études concernant la langue se concentrent souvent sur les enfants et les jeunes, laissant de côté les adultes et les personnes âgées, la conception étant que les adultes sont un groupe homogène (Eckert 1997 : 165). En outre, les études reflètent l'idée que le langage à l'âge mûr serait stable, sans variation ni changements (*ibid.*), bien que le langage se développe et change constamment à chaque stade de la vie (*ibid.* : 157). Eckert (*ibid.* : 152) indique que prendre les stades de la vie comme points de départ serait plus logique que l'âge chronologique, car l'âge ne montre pas le stade dans la vie de l'individu. Elle (*ibid.* : 165–166) expose aussi la possibilité de trouver une explication sur la différence de langages dans les réseaux humains. Si les plus âgés sont moins mobiles du point de vue géographique, et que les jeunes le sont plus, cela pourrait se refléter aussi dans leurs modèles linguistiques. Par conséquent, le côté social ne devrait pas être

négligé dans l'étude de la variation liée à l'âge (*ibid.* : 166). Pour nous, les deux points de vue sont un peu problématiques parce que nous connaissons l'âge chronologique des participants, mais pas nécessairement leurs stades de vie. En outre, l'âge chronologique est généralement plus accessible, que ce soit en vue d'une future recherche ou d'une application qui se sert de la classification automatique, mais les réseaux et les informations sur les stades de vie sont moins faciles à acquérir.

Mais quel est le rapport entre le langage des jeunes et des adultes ? Pour les jeunes, il est important à la fois de se libérer des parents et d'« être solidaires » avec leur groupe d'âge, ce qui se voit naturellement dans leur langage (Thibault 1997 : 22). Déjà pour les enfants, ce ne sont pas les adultes qui constituent le modèle primaire pour leur langage mais les autres personnes proches d'eux, comme les frères, les sœurs et les amis (Eckert 1997 : 162). En fait, les jeunes utilisent plus de « formes non standards pour se démarquer de leurs parents » (Thibault 1997 : 25). La jeunesse est la période durant laquelle se développe la langue vernaculaire, et les jeunes la font changer aussi (Eckert 1997 : 163). Quelques caractéristiques du langage jeune sont la troncation, les apocopes, les aphérèses, la verlanisation et l'utilisation de suffixes comme *-os* et *-av(e)*, en outre des emprunts à d'autres langues qui sont très communs (Boyer 2001 : 28). Les trois premiers procédés seront expliqués dans la partie 3.3.1 ; la verlanisation consiste à permuter les syllabes d'un vocable, par exemple *femme* devient *meuf*. Dans un corpus de discussion entre jeunes Français parisiens, Cappeau & Moreno (2017 : 82) ont trouvé de nombreuses fois le mot *trop* dans les cas où *très* serait la bonne forme. Selon eux, « *trop* [...] semble se substituer à *tres* » (*ibid.*). Ce fait est un élément facile à retenir dans nos attributs (voir 4.3.2), les autres faits énumérés ci-dessus demanderaient des analyseurs plus compliqués.

Quant aux adultes, par contre, ils ressentent la pression de parler conformément aux normes standard non seulement au travail mais aussi en famille (Thibault 1997 : 25). C'est pourquoi les gens âgés de 30 à 55 ans sont considérés comme ceux qui maintiennent la norme (*ibid.* : 24). Cela devient clair quand les plus âgés sont considérés : ceux de 60 à 75 ans, en un sens, deviennent jeunes de nouveau en partant à la retraite de leur travail – où le langage normatif atteint son maximum – et en utilisant la langue d'un état antérieur, de leur jeunesse (*ibid.* : 25). Eckert (1997 : 165) signale toutefois qu'il n'y a pas d'études sur ce phénomène et qu'on n'a pas eu de vérifications concernant ce point.

Ce qui est un peu problématique est l'idée que le parler jeune soit innovateur et distant de la langue ordinaire étant donné que les jeunes parlent différemment (Gadet 2017 : 33). La langue n'est jamais stable et ne l'a jamais été. On pourrait dire que le changement se voit dans le parler jeune parce que les jeunes sont ceux qui adoptent l'innovation, mais cela ne veut pas dire que

le langage des jeunes soit une langue autre que le français auquel nous sommes habitués.

3.1.2 Relation SMS et âge

Mais comment l'âge se voit-il dans les SMS ? Il paraît que dans les SMS, le parler jeune et adulte se manifestent comme sous n'importe quelle forme électronique. La relation des SMS et de l'âge n'a pas beaucoup été étudiée. Nous l'examinerons d'un point de vue plus extensif, celui de la langue Internet.

Gruppioni (2011 : 279) montre que même si la cyberlangue (qui, au sens large du terme, peut être comprise comme incluant la langue SMS) est essentiellement associée aux jeunes, elle contient des éléments qui la rendent plus étendue que la langue jeune. La cyberlangue « laisse une large place à la spontanéité, à l'impulsivité », et grâce aux traits qui seront présentés dans la section 3.3.1, comme les émoticônes, nous pouvons en déduire que la cyberlangue, ou le langage SMS, n'est pas réservée uniquement aux jeunes. D'une certaine manière, la cyberlangue peut être vue comme un hyperonyme par rapport à la langue jeune, qui est donc un hyponyme de la cyberlangue.

Schwartz *et al*. (2013 : 9) indiquent que les thèmes dans les statuts Facebook incluent l'école ou les études supérieures parmi les jeunes âgés de 13 à 18 ans et ceux qui ont de 19 à 22 ans ; parmi les adultes, âgés de 30 à 65 ans, c'est le travail qui domine. Cela est confirmé par McKeown & Rosenthal (2011 : 768), qui ajoutent que dans les blogs, les mots *vieux* et *maison* sont communs parmi les adultes. L'utilisation de la cyberlangue est naturellement fréquente parmi les jeunes, car ce sont les jeunes qui adoptent les nouveaux usages langagiers. Les relations humaines deviennent de plus en plus importantes en même temps que les gens vieillissent, ce qui se voit dans le vocabulaire utilisé : les adultes se servent plus du mot *nous*, tandis que les jeunes utilisent plus le mot *je* (*ibid.* ; Gravel *et al*. 2013 : 446). Les mots comme *fils*, *fille*, *mère* et *père* montrent également cela parmi les adultes (Schwartz *et al*. 2013 : 9). Gravel *et al*. (*ibid*.) ajoutent que, dans les tweets étudiés, les jeunes utilisent plus de « modifications stylistiques » [4], qui contiennent des répétitions de caractères dans une séquence ou des séquences écrites entièrement en majuscules.

En résumé, l'âge est naturellement visible dans la langue de chacun, et il s'agit essentiellement du stade de vie que nous vivons à présent. Les gens avec qui nous communiquons ont un effet sur les façons dont nous modifions notre langage, ce qui est courant particulièrement durant la jeunesse. Les SMS n'y font pas exception : ce sont les thèmes de discussion et notre

4. « [...] younger people use more explicit stylistic modifications such as alphabetical lengthening and capitalization of words. »

entourage qui révèlent le plus clairement l'âge que nous avons. En outre, les procédés utilisés dans les SMS, modifications du texte, révèlent s'il s'agit d'une personne jeune ou plus âgée.

3.2 Approche linguistique du corpus

Comme notre étude combine des points de vue de différents domaines, il est important d'aborder des concepts communs à la linguistique et à l'informatique. Nous partirons de la linguistique informatique pour examiner les difficultés qui surgissent quand des corpus de langue écrite sont utilisés pour des applications informatiques. Cette sous-section nous servira lorsque nous développerons notre propre méthode. Même si la méthode finale classifie les messages et pas les mots, nous nous concentrerons sur les mots dans notre analyse, ce qui justifie le choix. Nous finirons cette section en rappelant quelques concepts importants en linguistique de corpus.

3.2.1 Mot en informatique

En informatique et en traitement automatique du langage, le texte est très souvent segmenté en mots, mais ce qu'on considère comme un mot est assez complexe. Dans la segmentation en mots, le texte est divisé en unités, *tokens* (Manning & Schütze 1999 : 124). Un token peut être une séquence de lettres, un chiffre ou un signe de ponctuation, mais tout cela varie selon l'application et le programme. Dans la segmentation en mots, les signes de ponctuation peuvent être omis mais il arrive qu'ils montrent quelque chose sur la structure du texte (*ibid.* : 124–125), c'est pourquoi il faut y réfléchir avant de négliger de les prendre en compte. En ce qui concerne la définition du mot, particulièrement dans le contexte des textes extraits d'Internet, il faut être conscient du fait que les mots peuvent contenir des éléments non conventionnels ou être entièrement composés de signes de ponctuation, comme dans *Micro$oft* ou comme dans une émoticône du type *:-)* (*ibid.* : 125). Dans notre étude, les signes de ponctuation seront pris en compte à cause des émoticônes et de séquences qui pourraient contenir des signes en règle générale inattendus dans la langue écrite.

Normalement, la segmentation est donc faite sur la base des deux espaces autour du token ou d'un signe marquant le commencement d'une nouvelle ligne (Manning & Schütze 1999 : 125), mais que faire avec les signes de ponctuation qui sont souvent attachés à la fin du token ? Le point, par exemple, peut soit marquer une abréviation, soit la fin de la phrase, à cause de quoi les points d'abréviation devraient être pris en compte et ceux de la fin de la phrase séparés du token auquel ils sont attachés. Manning & Schütze (*ibid.* : 135) présentent une procédure selon laquelle la différenciation entre les deux types de points peut se faire pour l'anglais (Tableau 1).

Le tableau indique le signe en question et si le signe indique la fin de la phrase ou non. Nous utiliserons cet algorithme comme point de départ pour le prétraitement du corpus (partie 4.2.2).

Phase	**Signe**	**Fin de la phrase**
1	. ? !	Oui
2	” ”	Non
3a)	Abréviation généralement pas à la fin de la phrase suivie d'un point	Non
3b)	Abréviation non suivie d'une lettre en majuscule, par exemple *etc.*	Non
4	? ! suivi d'une lettre en minuscule	Non
5	Le reste des signes	Oui

TABLEAU 1 – Table des phases d'un algorithme pour la détection de la fin d'une phrase selon Manning & Schütze (1999 : 135).

Il nous reste la question des apostrophes : pour prendre un exemple, *l'ont* peut-il être réparti en deux tokens ou devrait-on le garder comme une unité ? Intuitivement, il y a deux tokens mais comme résultat, nous aurions un token comme *l'* (Manning & Schütze 1999 : 126). Comme dans la langue française les apostrophes sont utilisées très souvent, nous avons décidé de marquer ces deux tokens résultants comme des tokens distincts. Cela nous permettra de prendre en compte les tokens qui sont précédés d'une apostrophe ; par exemple dans *j'ai*, nous pourrons prendre en considération à la fois *j'* et *ai*, au lieu d'obtenir seulement *j'ai*. De plus, Panckhurst (2009 : 43–45) fait une remarque importante concernant le traitement automatique du langage SMS : il y a une infinité de problèmes déjà dans la segmentation des mots, sans parler de la correction automatique et des facteurs typographiques qui influencent la langue des SMS. C'est pour cette raison qu'il n'y a pas de solution simple pour segmenter les mots.

3.2.2 Termes à retenir

La langue suit des régularités, et il a même été proposé que les locuteurs essaient de minimiser l'effort de parler une langue (Manning & Schütze 1999 : 23). Selon la loi de Zipf, il n'y a que quelques mots très fréquents dans la langue et un grand nombre de mots à basse fréquence, ce qui prouve évidemment l'effort minimisé quand les gens se servent des mots très utilisés (*ibid.* : 24). Les listes de fréquence le montrent : les mots à haute fréquence dans un texte sont des mots vides (angl. *stop words*) (*ibid.* : 533–534). Les mots vides sont des mots qui ont une

fonction grammaticale importante et qui sont très communs dans la langue, par exemple *de* ou *elle* (*ibid.* : 20). Dans les applications informatiques, les mots vides sont ceux qui ne sont en général pas pris en compte pour faciliter le fonctionnement du système.

Pour évaluer le fonctionnement d'une application, les notions de *précision* (angl. *precision*) et de *rappel* (angl. *recall*) aident à décrire la fraction des résultats corrects ou pertinents (Manning, Raghavan & Schütze 2008 : 154). Ces notions sont très communes par exemple dans la recherche d'informations. La précision représente le taux de documents que le système a correctement découverts, le rappel celui des documents pertinents dans la collection découverts par le système (www4) :

$$\text{précision} = \frac{\text{nombre de documents pertinents découverts}}{\text{nombre de documents découverts}}$$

$$\text{rappel} = \frac{\text{nombre de documents pertinents découverts}}{\text{nombre de documents pertinents dans la collection}}.$$

Ces deux concepts peuvent encore être illustrés ainsi :

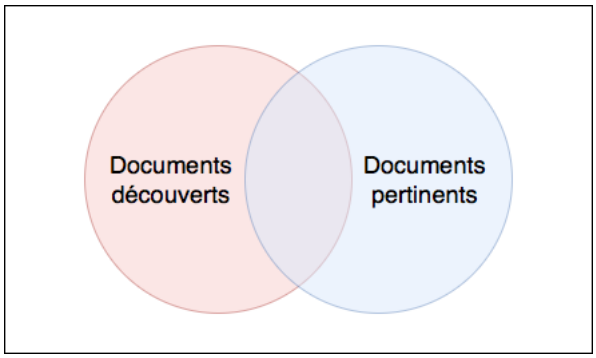


FIGURE 2 – Le diagramme de Venn représentant la précision et le rappel selon Manning & Schütze (1999 : 268).

Dans la figure 2, le cercle à gauche illustre les documents que le système a découverts et le cercle à droite les documents que le système devrait découvrir. L'espace commun aux cercles représente les documents pertinents et découverts. Le cercle à gauche désigne la précision, le cercle à droite le rappel. Nous illustrons ces deux concepts avec un exemple : nous avons une collection de dix œuvres dont trois sont pertinentes pour notre étude. Nous choisissons cinq œuvres de cette collection, y compris les trois œuvres pertinentes. Dans ce cas-là, la précision serait de $\frac{3}{5} = 60\%$ et le rappel de $\frac{3}{3} = 100\%$. Il est important de comprendre qu'il y a toujours un compromis entre les deux concepts : le rappel peut être haut mais aux dépens de la précision, et vice versa (Manning, Raghavan & Schütze 2008 : 156). Par exemple, si le système choisit tous les documents pertinents, le rappel est à 100%, comme dans notre exemple, mais le système ne fonctionne pas très bien parce que la précision reste assez faible.

L'exactitude (angl. *accuracy*) est aussi une représentation de la performance du système et désigne la fraction des bonnes classifications (*ibid*. : 155). L'exactitude est souvent prise comme mesure d'évaluation, mais ne convient pas très bien dans la plupart des cas. C'est pour cela qu'il faut aussi examiner les autres mesures pour avoir une vue d'ensemble plus précise. La dernière mesure que nous inclurons dans cette étude s'appelle score F1. Nous n'entrerons pas en détail dans la théorie du score F1, mais, pour dire la chose très simplement, c'est la moyenne harmonique pondérée de la précision et du rappel (*ibid*. : 156).

3.3 Le langage SMS

La communication médiée par les SMS est limitée par des contraintes technique, économique et situationnelle : le clavier du portable et la longueur limitée du message, le coût de l'envoi du message, l'information transmise et l'interlocuteur (Laporte *et al*. 2011 : 236). Ainsi, il n'est pas surprenant que les gens recourent à des procédés qui facilitent l'écriture d'un message. En général, il est possible de constater que dans les SMS, les phénomènes à la fois de l'oral et de l'écrit sont présents, mais dans une sorte de continuum où les deux peuvent être combinés (Panckhurst 2009 : 39). La langue est utilisée d'une manière créative, qui repose sur la néologie, c'est-à-dire les nouveaux vocables, et sur la néographie, les orthographes créatives. En outre, les SMS se caractérisent par la variation des formes même dans les procédés les plus fréquents (Fairon, Klein & Paumier 2006 : 44), car le scripteur possède une certaine liberté quant à son écriture (Laporte *et al*. 2011 : 237). Dans cette partie, nous décrirons les procédés surtout néographiques qui appartiennent fréquemment à l'usage des SMS. En outre, nous essaierons de donner un aperçu des SMS comme langage.

3.3.1 Procédés dans les SMS

Les procédés langagiers utilisés dans les SMS peuvent être classifiés selon une typologie. Nous présenterons ici deux classifications (Fairon, Klein & Paumier 2006 et Panckhurst 2009) qui incluent presque les mêmes phénomènes avec des dénominations différentes. Nous illustrerons les procédés avec les exemples donnés par les auteurs mentionnés Fairon, Klein & Paumier (2006) et Panckhurst (2009). Nous indiquerons les exemples de Fairon, Klein & Paumier par les lettres *FKP* et ceux de Panckhurst par un *P*.

Les procédés basés plutôt sur la phonétique sont répartis par Fairon, Klein & Paumier (2006 : 31–32, 33–37) en deux groupes : *phonétisation des caractères*, et *orthographe phonétique*. Selon eux (*ibid*. : 31–32), la phonétisation est considérée comme « un des [éléments] les plus

caractéristiques du langage SMS » et consiste à simplifier le texte en recourant à des graphies qui ont « une valeur phonétique ». Comme exemple, on peut citer des expressions comme (1) et (2), dont la première contient un chiffre et la seconde les lettres *K* et *C* qui représentent respectivement les sons [ka] et [se]. Dans l'orthographe phonétique (*ibid.* : 33–37), le but est d'introduire des formes de la langue parlée. Les procédés sont nombreux dans ce groupe, et nous en mentionnons six. Le premier procédé consiste en la suppression des fins muettes (exemple (3)). Les simplifications suivantes appartiennent aussi aux procédés de l'orthographe phonétique : les digrammes et trigrammes – autrement dit deux ou trois lettres qui représentent un son – sont simplifiés, comme dans l'exemple (4) où *eau* devient *o*, les consonnes doubles sont simplifiées, comme *ff* dans (5), et les semi-voyelles sont simplifiées comme dans (1) *oi* devient *wa*. En plus des simplifications, les lettres *k* et *z* sont utilisées : dans les exemples (2) et (3), *K* représente *ca* et *z* la forme phonétique de la lettre *s*. Le dernier procédé de ce groupe est l'assimilation de consonnes, qui produit un écrasement. Dans l'écrasement, la prononciation et la forme écrite sont altérées de manière qu'il n'y a qu'un son au lieu de deux, comme dans l'exemple (6), [ʒə sɥi] devient [ʃɥi].

1) *7 swaré* [*cette soirée*] FKP
2) *KC* [*casser*] FKP
3) *bizou* [bisous] FKP
4) *ls cado de noel* [les cadeaux de Noël] FKP
5) *cet efé la* [cet effet-là] FKP
6) *chui en forme* [je suis en forme] FKP

Panckhurst (2009 : 41) classifie les procédés phonétiques un peu différemment ; ce sont des *substitutions phonétisées* et des *réductions phonétisées*. Les substitutions peuvent être partielles ou entières : l'orthographe du vocable est complètement modifiée, comme dans l'exemple (1) où *sept* est devenu *7*, ou seulement en partie, comme c'est le cas pour *cadeaux* (4), dont la forme n'a pas complètement changé. Quant aux réductions phonétisées, qui incluent des troncations et des sigles ou acronymes, elles sont traitées dans Fairon, Klein & Paumier (*ibid.* : 41) comme des phénomènes lexicaux. Les troncations – dans lesquelles un segment est abrégé – sont soit des aphérèses (sons initiaux abrégés) (7), soit des apocopes (sons finaux abrégés) (8) (*ibid.*). Les sigles sont des abréviations formées des lettres initiales d'une expression composée de plusieurs vocables. On peut en donner comme exemple le (9). Les acronymes sont des abréviations prononcées comme vocables, comme *SIDA* (*ibid.*).

7) *tain* [putain] FKP
8) *ordi* [ordinateur] FKP
9) *mdr* [mort de rire] FKP

D'un autre côté, les SMS sont fabriqués en utilisant des procédés graphiques. Pour Fairon,

Klein & Paumier (2006 : 32–33, 37–40), il existe deux groupes de procédés liés à l'aspect graphique, à savoir des *rébus*, qui consistent à mêler différents signes comme des chiffres et des lettres, et des *phénomènes graphiques*, tandis que Panckhurst (2009 : 41) en voit quatre : *substitutions graphiques*, *réductions graphiques*, *suppressions ou absences graphiques* et *augmentations et ajouts graphiques*. Partons des substitutions graphiques, qui contiennent des remplacements typographiques du type de l'exemple (10) – dans lequel l'apostrophe est remplacée par une espace –, des icônes, des caractères spéciaux et des rébus (11) ; dans cet exemple, @ représente les oreilles. Les icônes et caractères spéciaux sont traités plus tard comme un procédé à part, suivant la classification de Fairon, Klein & Paumier (*ibid.* : 40).

10) *m en* [m'en] P

11) *de grandes* @ [de grandes oreilles] FKP

Le deuxième groupe se constitue de réductions graphiques qui correspondent aux procédés de l'orthographe phonétique de Fairon, Klein & Paumier, ce qui nous montre que les mêmes procédés peuvent être considérés à la fois comme phonétiques et graphiques. Y sont ajoutés deux procédés : des agglutinations et des abréviations. Dans les agglutinations, les vocables sont agglutinés l'un à l'autre au lieu d'être séparés par une apostrophe, comme dans l'exemple (12). Quant aux abréviations, Fairon, Klein & Paumier (*ibid.* : 39) signalent que l'abréviation est un « phénomène strictement graphique » et qu'elle diffère d'autres types d'abrègements, comme des troncations et des sigles. En effet, dans l'abréviation, le vocable devient plus court mais reste reconnaissable bien que des lettres soient supprimées ; le mot *bonjour* est facilement reconnu grâce aux lettres *bjr* dans l'exemple (13).

12) *jattends* [j'attends] P

13) *bjr* [bonjour] FKP

Les suppressions et les augmentations, les deux derniers groupes de procédés graphiques, sont des phénomènes plutôt typographiques. Par exemple dans (14), la lettre *ç* devient plus simple avec la lettre *c*, et dans (15), il y a une répétition des caractères *u*, *p*, *e* et *r*. Fairon, Klein & Paumier (*ibid.* : 37) appellent ces derniers « graphies à fonction expressive ». Des caractères peuvent aussi être ajoutés, comme dans l'exemple (16) où *z* a été ajouté au début de la séquence. Fairon, Klein & Paumier (*ibid.* : 38) indiquent d'ailleurs qu'il y a des agglutinations supplémentaires qui peuvent apparaître à cause de liaisons des types *je vou**s** entends* ou *mo**n** amie*. Il est donc possible de trouver, par exemple, des expressions comme *mn nange* ou *petit namour*, où on peut voir que ce trait persiste même si la liaison n'y est pas.

14) *ca* [ça] P

15) *suuuuppeeeerrr* [super] P

16) *les zamours* [les amours] FKP

D'autres procédés fréquents sont les *icônes* et les *symboles*. Il s'agit des smileys, émoticônes comme *<3* ou *:D* qui représentent des dessins (Fairon, Klein & Paumier 2006 : 40). Les symboles mathématiques et d'autres types de caractères sont aussi fréquemment utilisés : + pour *et* et = pour l'équivalence. Nous ajoutons à ce groupe de procédés les onomatopées, mentionnées par Panckhurst (2009 : 41), par exemple *bof*. Il faut noter que les procédés présentés peuvent être présents simultanément dans une séquence, ce qui rend évidemment la classification très difficile. Panckhurst (*ibid.* : 42) estime que dans ce type de cas, les procédés peuvent être considérés comme complexes.

Quant à la morphosyntaxe et à la syntaxe, deux phénomènes principaux surgissent : *conversion* et *omission*. Dans la conversion, la classe grammaticale est changée, comme par exemple dans (17), le nom *SMS* est devenu un verbe (Fairon, Klein & Paumier 2006 : 42). L'omission concerne les mots grammaticaux (*ibid.* : 43). L'omission la plus fréquente est celle de *ne* de la négation, mais naturellement cette omission n'est pas fréquente seulement dans les SMS. Dans l'exemple (18), l'omission concerne le sujet. Pour notre modèle, les procédés morphosyntaxiques et syntaxiques sont difficiles à remarquer sans un analyseur syntaxique, si bien que que nous ne pourrons pas en profiter (voir la partie 4.3.2).

17) *sms-moi* [envoie-moi un SMS] FKP

18) *pas eu mon exam* [je n'ai pas eu mon examen] FKP

3.3.2 Langage ou registre ?

À cause de ses caractéristiques spécifiques, le SMS est souvent considéré comme un représentant du langage nouveau (Fairon, Klein & Paumier 2006 : 49). En réalité, la variété dans les formes utilisées montre déjà que l'établissement d'une norme pour le langage SMS serait impossible (*ibid.* : 51) : par exemple, le mot *aujourd'hui* peut s'écrire par *ojrd8*, *aujourd8* ou *ojord8*. Cela indique aussi qu'il n'y a pas de lexique propre au SMS, quoiqu'il existe des dictionnaires qui essaient d'en décrire un (*ibid.* : 52). Souvent ce type de dictionnaires donne l'illusion qu'il y aurait une langue nouvelle à traduire, mais il s'agit seulement de « 'transcrire' [les SMS] en graphies standard » (*ibid.* : 53). D'autre part, il n'est pas rare d'entendre les gens dire que le langage SMS est oral. Pourtant, personne ne l'articule ou ne le parle comme langue maternelle. Il s'agit d'un code écrit qui se sert de procédés de la graphie (*ibid.* : 55). De plus, les phénomènes oraux, comme l'omission de *ne*, sont des procédés de la langue parlée, et non uni-

quement des SMS (*ibid*. : 67). Les traits de la langue parlée sont fréquents dans les SMS à cause du fait que nous envoyons des SMS généralement aux gens que nous connaissons bien. Ainsi, la communication se déroule à un niveau familier, où la langue n'est pas toujours très soignée (*ibid*. : 52). Ces deux caractéristiques – l'utilisation des graphies et le fait d'en utiliser au niveau familier – prouvent que le code SMS n'est pas un langage, un système de communication, mais plutôt une version du français. Si un registre se définit par une variation de la langue utilisée selon la situation (Boyer 1991 : 18), il serait plus approprié de dire que le SMS est un registre ou au moins une variété du registre familier.

4 Méthode

Dans cette partie, nous expliquerons le principe de la méthode à l'aide de laquelle nous classifierons les messages. Comme notre classifieur se base sur un modèle probabiliste, qui s'appelle *bayesien naïf*, nous devrons aussi y faire quelques modifications pour que le modèle marche mieux. Tout cela sera clarifié dans les sections qui suivent.

4.1 Théorie statistique : modèle bayesien naïf

L'apprentissage automatique est le contraire de la programmation à base de règles, dans laquelle les règles sont définies par l'humain (Manning, Raghavan & Schütze 2008 : 255). Dans l'apprentissage automatique, les règles sont apprises automatiquement à partir des données d'entraînement, c'est-à-dire qu'une partie des données utilisées (85-90% des données totales) est tirée pour entraîner le programme ; le reste (10-15%), que nous pourrions nommer données de test, sert à le tester (Manning & Schütze 1999 : 206–207). Les données sont munies de classes, ou de labels, qui signalent la bonne classification (Manning, Raghavan & Schütze 2008 : 256). Dans notre étude, il s'agit de messages qui sont classés « adulte » ou « jeune » selon l'âge de la personne qui les a envoyés. Pour mieux illustrer le processus, « un enseignant surveille et dirige le processus d'apprentissage » suivant l'information fournie par les données d'entraînement (*ibid.*) [notre traduction] [5]. Quand une classification prédéfinie de cette manière est utilisée, on parle d'apprentissage *supervisé*.

Pour classer les données de manière supervisée, il faut compter les probabilités. La probabilité d'un événement est généralement estimée par les fréquences relatives (Ross 2010 : 153). Pour donner un exemple, la probabilité qu'un token apparaisse dans un texte serait le nombre d'occurrences du token dans le même texte divisé par le nombre total de tokens. On peut nommer cet événement A et noter sa probabilité $P(A)$. Quand une probabilité conditionnelle nous intéresse, par exemple l'apparition d'un token à condition qu'avant le token apparaisse tel ou tel token, on le note $P(B|A)$ (*ibid.* : 167–168).

Le théorème de Bayes se base sur ces notions. L'équation du théorème de Bayes se définit ainsi (Manning, Raghavan & Schütze 2008 : 220–221) :

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}.$$

L'idée est de trouver la probabilité de l'événement A à condition que l'événement B se produise. La probabilité de A, *probabilité a priori*, est multipliée par la probabilité conditionnelle de B

5. « [...] a supervisor [...] serves as a teacher directing the learning process. »

sachant A, appelée *vraisemblance*. Cette multiplication est divisée par la probabilité de B. Le résultat $P(A|B)$ de ces calculs s'appelle *probabilité a posteriori* (*ibid.*).

Le classifieur bayesien naïf compte ces probabilités pour chaque token et s'en sert pour déterminer quelle classe est la plus probable pour un token (*ibid.* : 258). Dans notre cas, c'est la probabilité d'être de la classe adulte ou jeune sachant qu'il s'agit du token x : $P(\text{classe}|\text{token } x)$. L'essentiel est de créer un modèle qui puisse être utilisé pour la classification de nouvelles instances, de données que le modèle n'a jamais vues avant (*ibid.* : 257–258).

Normalement, l'efficacité (angl. *efficiency*) du modèle est améliorée par une procédure de sélection d'attributs (angl. *features*) (*ibid.* : 271). Très simplement, on choisit un sous-ensemble de tokens – nommés attributs – qui sont par la suite utilisés dans le modèle au lieu de tous les tokens du corpus. Ces attributs peuvent être choisis parmi les données utilisées ou à l'extérieur des données, par exemple par le moyen des résultats dans les études antérieures. Ainsi, le modèle devient plus simple et le surapprentissage (angl. *overfitting*), le fait de trop compter sur l'information fournie par les données, est évité, car le modèle ne compte pas trop sur les instances dans les données d'entraînement.

Mais d'où la « naïveté » du classifieur vient-elle ? La naïveté signale la supposition que les attributs soient conditionnellement indépendants les uns des autres (Manning & Schütze 1999 : 237). Selon Manning & Schütze (*ibid.*), il en résulte deux choses : premièrement, l'ordre des mots et la structure de la phrase sont laissés de côté, ce qui dans la littérature est appelé un modèle de « sac de mots ». Deuxièmement, la présence d'un token est indépendante de celle d'un autre token. Ces deux faits ne sont naturellement pas du tout vrais dans les langues naturelles, mais malgré ces défaillances, le modèle fonctionne relativement bien : il est rapide et efficace, exact et imperméable au bruit dans les données (Manning, Raghavan & Schütze 2008 : 269–270).

4.2 Méthode utilisée

Cette section sera consacrée à l'examen qui nous permet de voir de quels éléments notre méthode est composée plus concrètement. En premier lieu, nous rappelerons le processus et les étapes qui concernent la classification automatique en nous concentrant sur ce qui est important dans notre recherche. En second lieu, nous traiterons les manières dont le corpus a été modifié pour que notre méthode puisse être appliquée.

4.2.1 Processus

La classification automatique est un processus à multiples étapes. Pour faciliter la lecture, nous expliquerons dans cette section le schéma que suivra notre étude à l'aide d'une illustration (Figure 3).

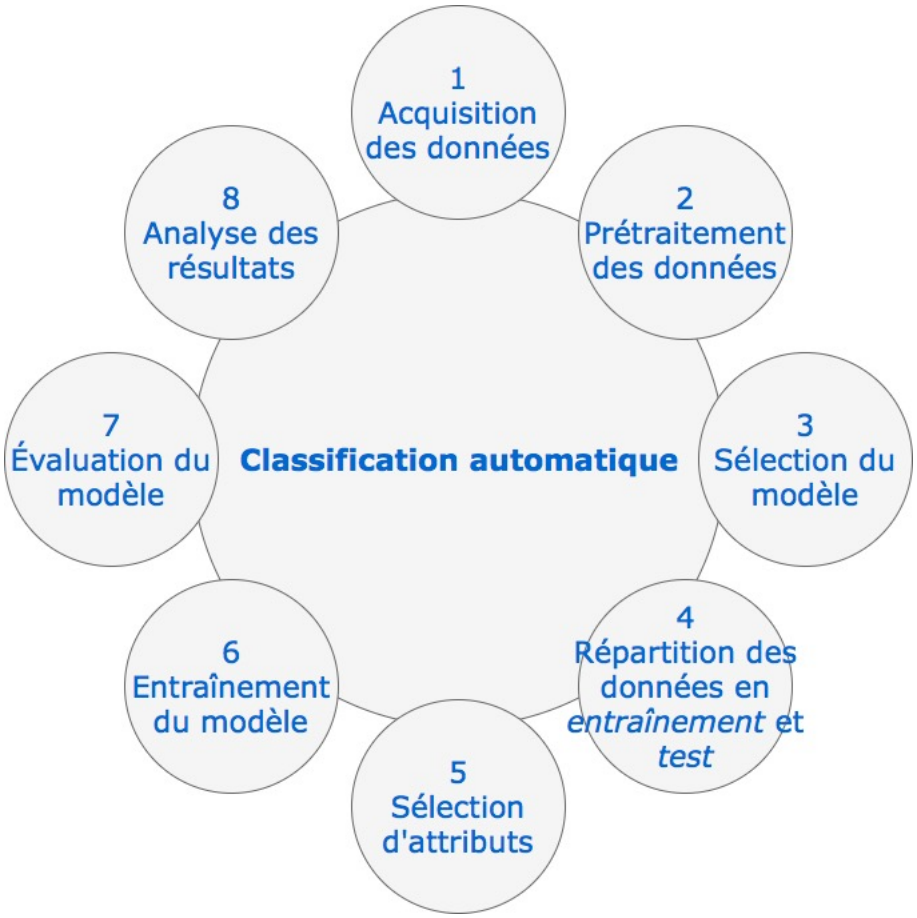


FIGURE 3 – Le processus de la classification automatique.

Il est évident que tout processus de classification automatique commence par l'acquisition des données – s'il n'y a pas de données, il n'y a rien avec quoi faire apprendre les tendances au modèle. Normalement, il faut transformer les données d'une manière ou d'une autre, que cela soit fait pour normaliser les tokens ou pour changer le format des données. Après le prétraitement, l'apprentissage automatique peut commencer. Si le modèle n'a pas encore été choisi, il faut le faire ; c'est-à-dire choisir de quelles théories se servir pour la classification. En fonction du modèle choisi, les données doivent être réparties en données d'entraînement et en données de test (voir 4.1). Les données doivent toujours être représentées sous une forme numérique, et cette transformation est généralement faite en même temps que la répartition. Si le modèle de sac de mots n'est pas employé, il faut définir les attributs qui peuvent éventuellement améliorer le modèle ; dans notre cas, les attributs sont l'objet d'étude qui nous importe le plus. Ensuite, le

modèle peut être entraîné suivant la théorie choisie ; la théorie que nous utiliserons est la bayesienne naïve. Après, le modèle est testé avec les données de test et évalué : quelle est l'exactitude du modèle, c'est-à-dire combien de fois le modèle a prédit correctement la classe du message, et quelles sont les autres mesures d'évaluation ? Finalement, les résultats peuvent être analysés. Dans notre cas, il s'agit, par exemple, d'analyser les attributs qui ont été les plus distinctifs et de se demander pourquoi ; d'analyser les éléments qui distinguent le langage des jeunes et de celui des adultes.

4.2.2 Prétraitement du corpus

Pour faire fonctionner le programme, il faut un peu modifier le corpus. Premièrement, l'extraction des messages des adultes et des jeunes est réalisée parce que nous n'utilisons pas le corpus en entier. Le prétraitement et le code ont été faits par le langage de programmation Python [6]. Deuxièmement, les tags, qui ont été mis au lieu des noms et d'autres informations considérées comme privées, sont supprimés du corpus. Troisièmement, on fait la segmentation en mots (voir 3.2.1). De plus, après la segmentation, chaque token a été mis en minuscule pour que la machine comprenne qu'il s'agit du même token par exemple dans *Demain* et *demain*. Il faut ajouter que normalement, le prétraitement serait continué avec la racinisation, la réduction des formes déclinées. Il ne sera pas possible pour nous d'exécuter une racinisation parce qu'il n'existe pas de racinisateurs qui connaissent bien la langue variée et les formes rarement conformes aux normes dans les SMS.

La segmentation en mots est faite selon les principes qui peuvent être vus dans le tableau 2. En premier sont traités les tokens avec les émoticônes qui peuvent se composer de divers caractères. Nous avons décidé de considérer comme tokens les caractères *: ;=bx* qui précèdent un ou plusieurs des caractères *-)(/dp*|. Ainsi, les émoticônes comme *;-)* ou *=p* sont conservées comme une entité, un token, mais séparés d'autres tokens qui peuvent les précéder ou les suivre. La même règle s'applique aux émoticônes *<3* et ^^.

Suivant le tableau, les tokens avec l'apostrophe ou ceux qui devraient l'avoir sont transformés. Si le token porte une apostrophe, celle-ci est incluse dans la partie du token précédent, laissant le reste des caractères de la séquence comme un deuxième token. Par exemple, *j'ai* devient *j'* et *ai*. Ce principe est conservé même si la séquence n'est pas grammaticalement correcte, comme dans *j'suis*. Si le token présente une espace au lieu de l'apostrophe, l'espace est conservée : ainsi, *j ai* reste comme il est, *j* et *ai*. Les tokens écrits ensemble, sans apostrophe

6. Pour voir le code : `https://github.com/jupouta`, partie *Codes-for-Masters-thesis*.

et espace, sont aussi laissés comme ils sont, par exemple *jen*. Ces règles sont les meilleures dans notre cas pour voir la fréquence d'utilisation de l'apostrophe et de son omission parmi les jeunes et les adultes.

Token	Séparé ou conservé sous forme originale	Exemple	Résultat
Caractères ***: ; = b x*** suivis de ***-) (/ d p*** |	Conservés mais séparés d'autres tokens possibles	*_:-)_* *_=p_*	*_:-)_* *_=p_*
<3 et ^^	Conservés mais séparés d'autres tokens possibles	*_<3_*	*_<3_*
Tokens avec l'apostrophe			
a) Token avec l'apostrophe au milieu	Séparés mais l'apostrophe conservée	*j'ai* *j'suis*	*j'_ai* *j'_suis*
b) Tokens qui au lieu de l'apostrophe ont une espace	Séparés mais techniquement conservés	*j_ai*	*j_ai*
Caractères ***. ! ? : ;*** attachés à un token alphabétique			
a) Caractère n'est pas répété	Séparé (voir le tableau 1 dans la section 3.2.1)	*demain?*	*demain_?*
b) Caractère est répété deux fois ou plus	Conservé mais séparé du token	*demain???* *demain?!*	*demain_???* *demain_?!*

TABLEAU 2 – Tableau du principe de la segmentation en mots. Les tirets bas représentent des espaces qui se produisent dans la segmentation en mots exécutée: soit l'espace se produit entre l'exemple donné et d'autres tokens possibles (_X_), soit à l'intérieur de l'exemple donné (X_1_X_2), où l'exemple se divise en deux tokens.

D'autres principes concernent les signes de ponctuation. Comme nous l'avons déjà constaté dans la partie 3.2.1, le point à la fin de la phrase doit être séparé du token précédent. Nous ajoutons ici une autre règle : si les points sont répétés deux fois ou plus, le groupe de points est considéré comme un token et est séparé du token précédent. Cela s'applique aussi aux caractères *! ? : ;*. De cette façon, la fréquence des caractères répétés peut être observée plus efficacement par notre classifieur. Finalement, si les règles décrites ci-dessus ne s'appliquent pas au token en question, la règle générale est appliquée – à savoir que les tokens sont segmentés selon l'espace entre eux.

4.3 Attributs

Dans la classification que nous essayerons de faire, il est question de la détermination d'indicateurs comme l'âge, auxquels beaucoup de recherches s'intéressent. La classification s'appuie généralement sur des attributs (angl. *features*), qui sont des variables pertinentes qui à la fois décrivent les données et aident à distinguer les groupes. La plus grande partie de cette section sera destinée aux attributs que les recherches antérieures ont utilisées pour déterminer des indicateurs, car ce sont les attributs qui améliorent la classification. Pour cela, nous présenterons

dans cette section les tendances et les types d'attributs qui sont utiles pour notre corpus et pour notre étude. Les attributs que nous utiliserons se trouvent dans l'Annexe et seront expliqués ci-dessous.

4.3.1 Sélection d'attributs

Partons des attributs basés sur les caractères. Dans ce groupe, nous incluons les informations sur des fréquences, comme le nombre de points dans un message, la longueur du message en caractères et en tokens et le nombre des fois qu'une apostrophe est omise ou remplacée par une espace. Soler-Company (2016) étudie la classification des textes d'opinion selon le sexe et constate que les attributs basés sur les caractères sont assez efficaces et qu'ils sont souvent un signe du sexe de l'auteur. Des 20 attributs les plus distinctifs, à peu près 5 sont basés sur les caractères, ce qui montre leur efficacité. Argamon *et al*. (2009) montrent que l'omission des apostrophes distingue les textes de blog des jeunes et ceux des adultes ; autrement dit, l'omission des apostrophes pourrait être utile quand on fait la différence entre les textes écrits par les jeunes et les adultes. Zheng *et al*. (2006 : 390) et McKeown & Rosenthal (2011 : 770–771) ont aussi obtenu de bons résultats : les résultats de la classification se sont améliorés quand en partie les mêmes attributs que les nôtres sont utilisés avec les courriels et les textes de blogs.

Les mots vides constituent le deuxième groupe d'attributs (voir partie 3.2.2). Selon Chandramouli *et al*. (2009), les mots vides sont un des groupes les plus distinctifs pour identifier le sexe de la personne qui a écrit un courriel ; le même est constaté par Argamon *et al*. (2009) pour les textes de blog, mais le meilleur résultat est assuré en utilisant une combinaison des mots vides et des attributs de caractères. D'après Estival *et al*. (2007), supprimer les mots vides des textes utilisés avant la classification produit les meilleurs résultats, au moins pour la prédiction du niveau d'études. En outre, Zheng *et al*. (2006 : 389) parlent du nombre optimal de mots vides. Il paraît que l'exactitude baisse en même temps que le nombre de mots vides augmente et que les messages courts ne permettent pas de faire la distinction entre les auteurs particuliers. Ce fait nous indique qu'il n'est peut-être pas utile de prendre trop de mots vides comme attributs.

D'autres attributs très utilisés sont ceux de type lexical et stylistique, mais la terminologie varie beaucoup. Par conséquent, nous avons décidé d'appeler ce groupe, tout simplement, attributs lexicaux. Il contient des éléments utilisés dans les SMS, comme des émoticônes et des abréviations, mais aussi des vocables qui reflètent les sujets communs aux jeunes et aux adultes (comme l'école et le travail, respectivement). McKeown & Rosenthal (2011 : 771) trouvent que

les attributs lexicaux n'ont pas vraiment d'impact sur l'exactitude mais que la combinaison des attributs lexicaux avec d'autres en a. Cela est confirmé par Argamon *et al*. (2009) et par Zheng *et al*. (2006 : 388). Même si les attributs lexicaux ne paraissent pas aussi performants que par exemple ceux basés sur les caractères, il nous seront utiles pour faire la distinction entre les vocables les plus communs chez les jeunes et les adultes.

Une autre possibilité est de se servir des n-grammes, des *n* caractères ou tokens qui se succèdent dans un texte, par exemple les bigrammes de la phrase *je suis à l'école* sont *je suis*, *suis à* et *à l'école*. Il s'agit d'une approche importante dans la classification automatique. Le fonctionnement des n-grammes se base sur l'idée qu'en connaissant les tokens précédents dans un texte long, il est possible de prédire le token qui suivra (Manning & Schütze 1999 : 192) ; dans l'exemple ci-dessus, il serait assez facile de prédire que c'est le token *l'école* qui suit les tokens *je suis à*. Les systèmes construits ainsi fonctionnent étonnamment bien (*ibid*. : 195). Martell & Tam (2009 : 36–37) classifient les messages en utilisant des n-grammes : les résultats ont été bons quand il a été question de la classification les jeunes versus les adultes d'un des groupes d'âges 30, 40 ou 50 ans. Dickinson, Hu & Miller (2012 : 148) confirment aussi que les n-grammes ont beaucoup de valeur dans la classification des sexes.

Il n'est pas rare d'utiliser des attributs syntaxiques, et plusieurs recherches démontrent l'utilité qui s'ensuit de leur utilisation. Comme notre étude n'est qu'un mémoire de master, nous n'aurons pas accès à un analyseur syntaxique qui reconnaisse les procédés fréquents dans les SMS. De toute façon, l'exactitude du classifieur se détériore considérablement quand un parseur syntaxique développé pour le langage standard est utilisé pour des données bruyantes, ce que sont les SMS (Soler Company & Wanner 2014 : 1318). Cela dit, il faut tenir compte du fait que les attributs syntaxiques pourraient améliorer la performance du modèle et éventuellement résoudre des problèmes.

4.3.2 Attributs choisis

Nous avons formé trois groupes d'attributs en nous basant sur la théorie présentée et sur l'examen du corpus. Les attributs qui se basent sur les caractères constituent le premier groupe. Ce sont des éléments faciles à compter, par exemple la longueur d'un message en tokens et en caractères. Nous supposons que le nombre de tokens dans un message est plus bas pour les jeunes, comme l'indiquent Gravel *et al*. (2013 : 446) pour les tweets, mais que le nombre de caractères ne l'est pas nécessairement : nous avons pris la répétition des caractères comme un attribut, car les jeunes tendent à en répéter plus, au moins dans les tweets, par exemple dans

le mot *bieeeen* (*ibid*.). Les autres attributs dans ce groupe sont le nombre d'accents, des lettres *k* et des consonnes isolées, et également la présence des apostrophes grammaticalement non correctes. En examinant les accents, nous voulons voir si l'un des deux groupes respecte plus les normes en utilisant les signes graphiques nécessaires. La lettre *k* n'est pas parmi les lettres les plus utilisées en français (www2), mais elle est pourtant très fréquente dans les SMS grâce au procédé de phonétisation (voir la section 3.3.1). En ce qui concerne les consonnes isolées, l'apostrophe qui marque la liaison est plusieurs fois remplacée par une espace – il s'agit du remplacement typographique –, par exemple *j en* au lieu de *j'en*, phénomène dont nous essayons de mesurer la fréquence par la quantité de consonnes isolées. Il est aussi possible de trouver des exemples où les consonnes isolées représentent la phonétisation, par exemple dans les cas où *c* représente *c'est*. Un exemple des apostrophes non conventionnelles est *j'suis*, qui devrait naturellement être écrit *je suis* d'après les règles d'orthographe. Le corpus inclut quelques-unes de ces apostrophes, et nous utilisons cette information pour voir si elles sont fréquentes dans les deux groupes.

Le deuxième groupe se compose des attributs qui sont typiquement courants dans les SMS ; nous l'appelons attributs SMS. Premièrement, ce groupe d'attributs contient des vocables liés soit aux jeunes, soit aux adultes ; ce sont des vocables qui reflètent le stade de vie que les jeunes et les adultes vivent. Il s'agit principalement de vocables associés à l'école, au travail ou à la famille, ce qui nous aidera assez facilement à identifier le groupe en question. Deuxièmement, ce groupe inclut des attributs comme le nombre d'émoticônes, d'abréviations et de sigles, de vocables courants dans les SMS ou dans la communication virtuelle et de numéros à l'intérieur d'un mot. Les émoticônes sont très courantes dans le langage SMS (voir la partie 3.3.1). Les abréviations, les sigles et les vocables typiques des SMS peuvent aussi différencier les groupes l'un de l'autre ; nous supposons que les jeunes en utilisent plus. Nous inclurons des abréviations comme *rdv* pour *rendez-vous*, mais il faut dire que ces abréviations ne sont pas spécifiques des SMS (Fairon, Klein & Paumier 2006 : 49). Les vocables typiques des SMS contiennent des mots comme *coucou*, *mdr* pour *mort de rire* ou *asap*. Nous compterons aussi la fréquence des chiffres à l'intérieur d'un token, ce qui nous donnera le taux de phonétisation dans les messages. Signalons que le résultat selon lequel il n'y aurait pas de différences entre les deux groupes serait important car cela nous montrerait que le langage SMS n'est pas réservé qu'à un groupe d'âge mais constitue plutôt un registre.

Les mots vides et les attributs plutôt syntaxiques sont compris dans le dernier groupe. Nous y avons inclus le nombre des mots *je* et *nous*, car dans les études présentées (voir la partie 3.1.2), il a été montré que les jeunes utilisent plus le mot *je*, tandis que les adultes se concentrent plus

sur *nous*. Ce sont des mots vides. En utilisant ces attributs, nous pourrons voir si ce fait est aussi courant dans les SMS. Il paraît aussi que le mot *trop* est en train de remplacer le mot *très* dans le langage des jeunes, et c'est pour cela que la fréquence du mot *trop* sera aussi incluse dans ce groupe.

Quant aux n-grammes, nous utiliserons seulement les bigrammes des tokens.

4.4 Classifieur

Après le prétraitement et la sélection des attributs, il est possible de construire le classifieur. Chaque message a besoin d'un label qui indique de quel groupe le message vient ; dans notre cas, les groupes sont *adulte* et *jeune*. Pour assurer de la vraie proportion de jeunes et d'adultes dans les données d'entraînement et dans celles de test, les messages seront divisés aléatoirement.

Le texte doit aussi être changé en forme numérique : ce sera un tableau où les lignes correspondent à des messages et les colonnes à des tokens ou à des attributs dans les messages (Manning & Schütze 1999 : 296). Dans les colonnes, seul le nombre de tokens comptés ou la forme binaire – si un token est présent dans le message (représenté par le chiffre 1) ou pas (représenté par le chiffre 0) – est donné, ce qui rend possible une représentation numérique. On peut appeler ce processus *vectorisation*. La vectorisation permet aussi d'implémenter d'autres attributs pour l'amélioration du modèle.

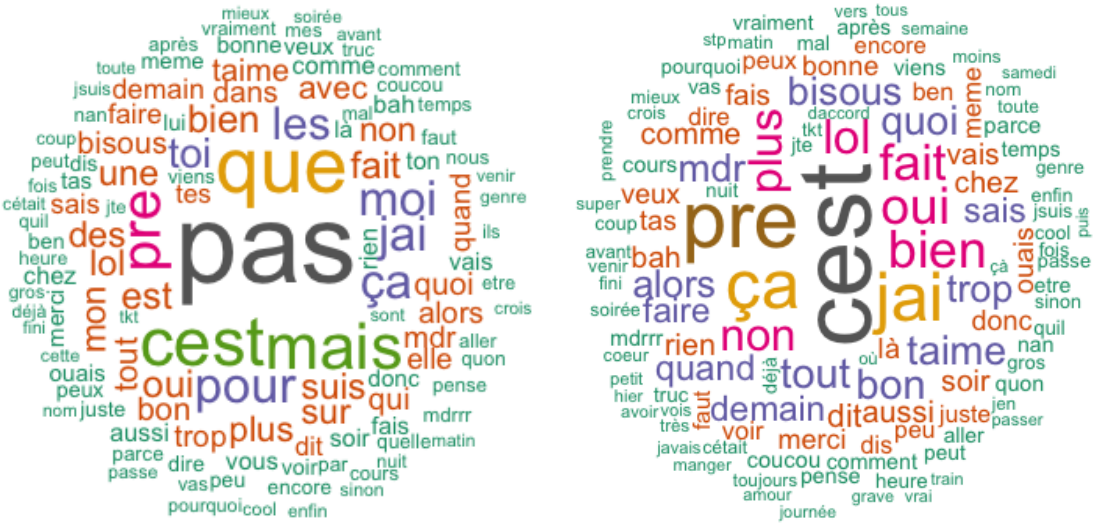
Après la vectorisation, les données sont réparties en données d'entraînement et de test. Pour la suite, le modèle peut être entraîné avec les données d'entraînement. À l'aide de cet entraînement, le modèle est utilisé pour prédire la classe des messages dans les données de test. Durant tout le processus, nous nous servirons de plusieurs bibliothèques de programmation, qui sont des programmes prêts à être utilisés, principalement de *pandas* et de *scikit-learn*. Nous avons besoin de pandas pour manipuler les données sous forme de *data frame*, une sorte de donnée tabulaire qui permet la vectorisation. Avec scikit-learn, nous construirons le classifieur, qui se base sur la théorie déjà traitée dans la partie 4.1.

5 Analyse

Cette section sera consacrée à l'analyse du fonctionnement de l'algorithme utilisé et à l'étude du corpus. Nous commencerons par un aperçu du vocabulaire dans les données en analysant les vocables les plus utilisés. Nous continuerons par l'implémentation du modèle, en examinant ce qui doit se faire pour que le modèle marche le mieux et en étudiant le fonctionnement des versions du modèle. Ensuite, nous passerons à l'analyse des attributs du modèle choisi. Finalement, nous examinerons quelques messages SMS écrits par des jeunes et adultes – pourquoi le modèle les a-t-il classifiés de telle manière et serait-il possible pour l'humain de distinguer les deux groupes ?

5.1 Aperçu du vocabulaire

Pour donner un aperçu du langage et particulièrement du vocabulaire dans notre corpus, nous envisageons des nuages de mots-clés qui sont construits des tokens du corpus. Les nuages de mots-clés dans la figure 4 représentent les tokens les plus utilisés dans les SMS. Le principe dans les nuages de mots-clés est de prendre les tokens les plus utilisés, parmi lesquels les tokens écrits en grands caractères gras sont les plus fréquents, tandis que les tokens en caractères plus petits sont moins utilisés (mais de toute façon, communs dans ce corpus). Comme dans n'importe quel corpus, les tokens les plus utilisés sont presque toujours les mêmes (voir la partie 3.2.2). Sur les deux figures, la première contient les mots vides (angl. *stop words*), et la seconde non. Le corpus n'a pas été proprement segmenté en mots, mais les nombres et les signes de ponctuation n'ont pas été pris en compte, ce qui se voit par exemple dans l'absence d'apostrophes.



(a) Nuage de mots-clés des messages avec les mots vides.

(b) Nuage de mots-clés des messages sans mots vides.

FIGURE 4 – Nuage de mots-clés des messages du corpus.

Il est clair que dans la figure 4a, le token le plus utilisé est *pas*. Il est cependant plus intéressant d'examiner la figure 4b, car sans mots vides, le contenu devient plus visible [7]. Ce qui saute aux yeux est le token *pre*. Les noms et les autres informations privées (adresses, numéros de téléphone, et cetera) ont été marqués par les tags durant la collecte dont le plus utilisé est le tag <PRE_X>, qui désigne des prénoms. C'est pour cela que le token *pre* est si dominant dans notre corpus. Un deuxième point saillant est l'utilisation de mots de salutation comme *bisous* et *coucou*. Ils sont accompagnés par le terme de politesse *merci*. D'autres tokens fréquents sont liés aux questions et aux rendez-vous : *oui*, *ouais* et *non* sont des réponses, tout comme *bon* et *bien*. L'indication la plus claire des rendez-vous se trouve dans les vocables *demain*, *soir* et par exemple *semaine* et *temps*, qui montrent la fonction évidente des SMS comme moyen de communication. Par ailleurs, les abréviations – communes dans la communication médiée par ordinateur – *lol* et *mdr* sont très présentes dans notre corpus. Les caractéristiques énumérées ci-dessus indiquent les fonctions essentielles des SMS : donner des rendez-vous, demander des nouvelles et répondre aux choses demandées en n'oubliant pas la politesse.

Dans les deux nuages, on voit très bien des formes qui renvoient à la deuxième personne du singulier (*taime*, *vas*, *tas*, *jte*), la personne à laquelle le message est adressé, et à la première personne du singulier (*jai*, *mon*, *moi*), la personne qui envoie le message. Ce qui est aussi domi-

7. Le programme utilisé pour produire les nuages de mots-clés n'a pas reconnu tous les mots vides, car le corpus ne contient pas seulement des formes normatives. Bref, il reste quelques mots vides dans 4b.

nant est la construction *c'est*, qui a plusieurs fonctions. C'est pourquoi il n'est pas surprenant qu'elle soit parmi les expressions les plus utilisées, mais il est presque impossible de dire si cette expression est fréquente à cause du registre oral ou non. L'utilisation des pronoms mentionnés ici prouve clairement le registre oral de ce corpus (Cappeau & Moreno 2017 : 76–77 ; Fairon, Klein & Paumier 2006 : 44) ainsi que le fait que la communication se passe entre la personne qui envoie le message et celle qui y répond, comme durant une discussion. Il n'est pas possible de distinguer les mots d'interrogation (*que*, *quoi*, *quand*, *qui*) et les pronoms relatifs, mais il semble que les questions soient fréquentes. Le pronom *ça* est aussi assez présent dans le corpus. *Ça* peut être utilisé dans de nombreuses fonctions, mais comme il est étroitement lié au moment de l'énonciation (Goosse & Grevisse : §692), il est clair que *ça* manifeste la nature conversationnelle du corpus.

Dans l'ensemble, les tokens les plus fréquents sont à peu près les mêmes que les vocables qui figurent normalement dans les listes de fréquence de mots français (www3) : les prépositions comme *pour*, *dans* et *sur* sont assez fréquentes, suivies par les déterminants *les*, *une* et *des*. Les formes des verbes *être*, *avoir* et *faire* sont très présentes aussi (*jai*, *est*, *fait*, *suis*). En plus des caractéristiques énumérées ci-dessus, quelques tokens ont des fréquences un peu divergentes par rapport aux autres listes de fréquence : *pas*, *que*, *alors*, *moi*, *toi* et *trop*. *Pas* et *que* sont les tokens les plus fréquents dans notre corpus, mais cela n'est pas le cas dans la plupart des textes écrits. *Alors*, *moi* et *toi* s'expliqueraient tous par le registre familier et par la fonction essentielle du SMS. *Trop* est peut-être de haute fréquence à cause de la tendance que ce vocable a à remplacer *très* (voir la partie 3.1.1). Les éléments qui paraissent absents dans notre corpus sont les prépositions et les pronoms généralement très utilisés, comme *en* et *se*. Cela s'explique sûrement par le manque de segmentation en mots – car les tokens sont souvent attachés à d'autres tokens au lieu d'une espace ou d'une apostrophe – et par celui d'un analyseur syntaxique, qui pourrait séparer les tokens plus logiquement.

Il est donc possible de dire que le langage conversationnel est présent dans le corpus, de même que le langage qui représente un registre familier. Toutefois, il inclut des tokens qui ne sont pas fréquents dans la majorité des textes, et en effet, ce sont des vocables qui distinguent notre corpus des autres. Par la suite, nous passerons à l'analyse des versions du modèle que nous avons créées.

5.2 Implémentation du modèle

Il n'est pas sans problème d'utiliser toutes les données dans notre cas. Les jeunes sont surreprésentés dans les données, ce qui influence la performance du classifieur. En testant le classifieur avec toutes les données, nous avons perçu que, pour le modèle, il est simplement plus probable de prédire qu'un message a été écrit par un jeune que par un adulte. Cela résulte du plus grand nombre de messages des jeunes. Pour résoudre ce problème, nous utiliserons des données équilibrées : approximativement 10 000 messages des jeunes seront aléatoirement choisis – le nombre de messages des adultes étant à un peu moins de 10 000 – pour assurer un équilibre entre les deux groupes. En outre, la représentativité du modèle sera plus proche de la réalité, car la proportion des messages des jeunes et des adultes sera plus équilibrée.

Comme la loi de Zipf (voir la section 3.2.2) le dit, toutes les données contiennent toujours des vocables qui sont très fréquents, le reste des vocables étant moins fréquents. Pour cette raison, il est courant de supprimer les vocables les plus fréquents du corpus utilisé pour assurer un meilleur fonctionnement du modèle. Nous l'avons aussi fait. Les tokens dont la fréquence est de plus de 7500 occurrences ont été supprimés des données utilisées. Ce sont les suivants : *je*, *pas*, *de*, *tu*, *est*, *et*, *la*, *que*, *le*, *en*, *mais* et *on*.

Le modèle que nous avons construit avec les données équilibrées a plusieurs versions qui utilisent différents attributs choisis. Le tableau 3 contient les résultats obtenus par la classification exécutée. Pour évaluer la performance des modèles, nous donnerons l'exactitude, la précision, le rappel et le score F1 de chaque cas. Comme nous l'avons expliqué dans la section 4.3.2, nous avons certains groupes d'attributs, comme les attributs basés sur les caractères, qui contiennent des types d'informations différents. Par conséquent, nous avons différentes combinaisons d'attributs qui sont par la suite utilisées par le classifieur. Ce qui se voit clairement dans les résultats est que les attributs n'ont pas trop d'influence sur les résultats. L'exactitude, la précision, le rappel et le score F1 sont à peu près les mêmes, mais le modèle sans sac de mots, sans les tokens dans le corpus, nous donne les plus mauvais scores. Les bigrammes fonctionnent un peu mieux mais pas aussi bien que le sac de mots.

Modèle	Exactitude	Précision			Rappel			Score F1		
		Jeune	Adulte	**Moyenne**	Jeune	Adulte	**Moyenne**	Jeune	Adulte	**Moyenne**
Sans attributs	0.822	0.97	0.43	0.89	0.82	0.82	0.82	0.89	0.56	0.84
Tous les attributs	0.814	0.97	0.42	0.89	0.81	0.83	0.81	0.88	0.56	0.84
Caractères	0.813	0.96	0.42	0.89	0.81	0.81	0.81	0.88	0.56	0.83
Attr. SMS	0.824	0.96	0.43	0.89	0.83	0.81	0.82	0.89	0.57	0.84
Attr. syntaxiques	0.820	0.96	0.42	0.89	0.82	0.81	0.82	0.89	0.56	0.84
Caractères + attr. SMS	0.815	0.97	0.41	0.89	0.81	0.83	0.81	0.88	0.55	0.84
Caractères + attr. syntaxiques	0.808	0.97	0.41	0.89	0.80	0.85	0.81	0.88	0.55	0.83
Attr. SMS + attr. syntaxiques	0.822	0.96	0.44	0.89	0.82	0.81	0.82	0.89	0.57	0.84
Sans sac de mots	0.677	0.92	0.25	0.83	0.68	0.66	0.68	0.78	0.37	0.72
Bigrammes	0.742	0.96	0.34	0.87	0.73	0.83	0.74	0.83	0.48	0.78
Bigrammes sans attributs	0.773	0.94	0.35	0.86	0.78	0.70	0.77	0.86	0.47	0.80

TABLEAU 3 – Les résultats des versions utilisées dans la classification.

L'influence des attributs n'est pas grande parce que le modèle de sac de mots utilise déjà les tokens dans les données comme attributs. Le nombre de tokens est d'environ 33 400 [8] et nous avons créé 16 attributs de plus. Les tokens et les attributs créés sont traités équitablement, ce qui aboutit à une situation dans laquelle les attributs créés forment seulement une partie quasi insignifiante dans le modèle. La situation se détériore avec les SMS à cause du volume des tokens qui ne sont pas conformes aux normes. Il paraît aussi que les attributs créés ne fonctionnent pas très bien seuls : l'exactitude et le rappel sont à 0.68, ce qui n'est pas beaucoup mieux que dans le jeu de pile ou face. Une solution aux deux problèmes serait de pondérer les attributs, de donner aux attributs des pondérations pour que le modèle les prenne plus en considération, mais cela n'est pas sans risques : le modèle se base sur les pondérations qui sont comptées automatiquement, et les changer romprait le modèle. Une deuxième solution serait de supprimer plus de mots vides. Nous en avons déjà supprimé quelques-uns, mais il se peut qu'il soit utile d'en supprimer un peu plus, car le nombre d'attributs diminuerait et les tokens qui n'ont pas de rôle dans la différenciation des groupes ne dérangeraient pas le fonctionnement du modèle. La dernière solution serait de supprimer les tokens qui apparaissent rarement. Ces tokens n'ont pas d'influence sur le modèle pour la même raison que les mots vides.

Le grand nombre d'attributs est également problématique à cause de la matrice creuse que tous les attributs du corpus forment. Une matrice creuse est une matrice dont la plupart des éléments sont des zéros. Pour cela, le modèle n'a pas suffisamment d'informations pour déduire

8. Le nombre varie parce que chaque fois les messages dans les données sont différentes.

la classe, car les informations offertes sont trop vagues ou elles ne sont pas distinctives.

Comme cela se voit dans le tableau 4, les attributs seuls ne sont pas très bons en distinguer les deux groupes. La précision est chaque fois très bonne pour les jeunes, mais pour les adultes, elle est presque inexistante. Dans les mesures du rappel, il est néanmoins possible de voir une différence : les attributs basés sur les caractères, les attributs sur les caractères avec les attributs SMS et les attributs sur les caractères avec les attributs syntaxiques paraissent tous avoir une meilleure performance au moins chez les adultes. Ce qui est commun à tous ces modèles est le groupe d'attributs basés sur les caractères ; ces attributs ont probablement une influence à noter sur les modèles. Les scores F1 ne sont pas très bons non plus, mais ils paraissent aussi soutenir notre hypothèse des attributs basés sur les caractères. Le problème avec nos attributs seuls est que le modèle n'a pas beaucoup d'informations à compter et ainsi ne sait pas prédire la classe la plus probable. En déduction, cela nous indique que le modèle sac de mots est essentiel dans la classification des messages SMS.

Modèle	**Exactitude**	**Précision**			**Rappel**			**Score F1**		
		Jeune	Adulte	**Moyenne**	Jeune	Adulte	**Moyenne**	Jeune	Adulte	**Moyenne**
Caractères	0.718	0.91	0.26	0.81	0.75	0.53	0.72	0.82	0.35	0.75
Attr. lexicaux	0.763	0.87	0.19	0.78	0.85	0.22	0.76	0.86	0.20	0.77
Attr. syntaxiques	0.830	0.86	0.16	0.76	0.96	0.05	0.83	0.91	0.07	0.79
Caractères + attr. lexicaux	0.681	0.92	0.26	0.82	0.69	0.65	0.68	0.79	0.38	0.73
Caractères + attr. syntaxiques	0.717	0.90	0.26	0.81	0.75	0.53	0.72	0.82	0.35	0.75
Attr. lexicaux + attr. syntaxiques	0.745	0.87	0.19	0.78	0.82	0.26	0.75	0.85	0.22	0.76

TABLEAU 4 – Les résultats des versions qui utilisent seulement les attributs créés.

Après que nous avons supprimé les attributs peu utilisés – nous avons supprimé les tokens qui se trouvent dans moins de 100 messages –, les résultats ne changent pas en comparant les premières versions aux précédentes. Le nombre d'attributs a baissé jusqu'à approximativement 670. Les chiffres sont un peu plus faibles mais pas dramatiquement. Il semble toujours que les attributs ne sont pas suffisamment distinctifs. Par conséquent, il est utile de supprimer les tokens peu utilisés, spécialement dans le cas où les données ont beaucoup de bruit, mais il faut le faire avec modération.

Les bigrammes sont souvent utilisés pour préserver l'ordre des mots et l'information qui se trouve dans les deux tokens successifs. Dans notre cas, les bigrammes ne fonctionnent pas mal, mais pas bien non plus, l'exactitude étant à 0.77. Il résulte de ce fait que les messages

finalement utilisés par le modèle ne représentent plus les messages originaux, spécialement dans leur ordre de mots. Comme nous l'avons dit à propos du prétraitement dans la partie 4.2.2, les tags ont été supprimés, en plus des mots vides, et d'autres modifications ont été effectuées. Si les SMS avaient un minimum de modifications, les bigrammes pourraient mieux fonctionner, car l'information nécessaire serait présérvée.

À la fin, nous devons choisir le modèle dont nous nous servirons dans notre analyse. Le modèle que nous considérons comme le plus performant est celui de tous les attributs avec le sac de mots, car il a donné les meilleurs résultats d'une façon générale. Dans ce modèle, nous avons supprimé les tokens qui sont dans moins de cinq messages, parce que le nombre d'attributs se réduit suffisamment, mais la performance du modèle n'est pas détériorée. Par conséquent, il reste approximativement 6300 attributs. Cette version du modèle sera celle que nous utiliserons dans le reste de l'analyse.

5.3 Analyse des attributs

Dans cette section, nous examinerons les attributs de plus près. Nous analyserons les attributs des deux modèles : ceux du modèle avec tous les attributs et ceux du modèle sans attributs. Par cette comparaison, nous pourrons voir l'effet de nos attributs et des tokens dans les données. De plus, il sera aussi important d'étudier les différences et les similitudes entre les deux groupes en question.

Nous avons quatre listes (voir le tableau 5) qui illustrent les attributs des deux groupes : jeunes et adultes. Les deux premières listes contiennent les attributs du modèle sac de mots avec nos attributs, les deux dernières ceux du modèle sans nos attributs. Les chiffres sont des pondérations que le modèle a données aux attributs et les pondérations sont des logarithmes naturels.

Jeunes		Adultes		Jeunes		Adultes	
Pondération	Attribut	Pondération	Attribut	Pondération	Attribut	Pondération	Attribut
-1.72581	longueur car	-1.69540	longueur car	-3.79767	.	-3.45006	...
-2.20632	longueur mess	-2.20690	longueur mess	-3.80150	,	-3.52788	.
-3.15563	taux diacritique	-2.90375	taux diacritique	-3.86888	?	-4.06299	!
-3.85952	taux *je*	-3.86408	taux répétition	-3.86991	!	-4.06832	,
-4.18929	taux émoticône	-4.02622	...	-4.09575	j’	-4.23593	a
-4.31925	.	-4.08208	.	-4.20147	a	-4.28947	à
-4.32259	,	-4.32789	taux *je*	-4.25430	c’	-4.33638	pour
-4.35654	?	-4.39158	taux consonne	-4.34660	t’	-4.47881	?
-4.36148	!	-4.49369	taux sms	-4.48558	à	-4.52413	bisous
-4.58438	j’	-4.63236	,	-4.63484	ai	-4.54007	! ! !
-4.71526	a	-4.63902	!	-4.65151	un	-4.61681	? ! ?
-4.75341	c’	-4.77715	a	-4.74918	ça	-4.62961	un
-4.78698	taux répétition	-4.83969	à	-4.85193	...	-4.74622	c’
-4.82408	taux sms	-4.90205	pour	-4.89963	moi	-4.80447	j’
-4.86506	taux *k*	-4.94016	taux *k*	-4.91264	pour	-4.83276	il
-4.89928	t’	-5.01323	?	-4.94219	me	-4.89034	les
-5.02004	à	-5.08577	bisous	-4.95425	les	-4.90874	ce
-5.18075	un	-5.09101	! ! !	-4.96187	il	-4.97929	des
-5.18188	ai	-5.17192	? ! ?	-4.96801	:)	-4.98095	ai
-5.18639	taux consonne	-5.19148	un	-5.02988	l’	-5.00964	te
-5.28125	ça	-5.30448	c’	-5.06316	si	-5.03215	moi
-5.39438	moi	-5.34157	il	-5.11526	te	-5.05877	toi
-5.41265	...	-5.35243	j’	-5.12241	toi	-5.09168	suis
-5.48927	me	-5.44379	ce	-5.13870	va	-5.10667	bien
-5.49542	il	-5.46646	les	-5.15341	ce	-5.11614	me

TABLEAU 5 – Les pondérations des 25 attributs dans les deux modèles choisis : le modèle avec tous les attributs à gauche et celui sans nos attributs à droite.

5.3.1 Modèle avec les attributs

Voyons d’abord le modèle avec nos attributs. Les attributs qui se basent sur les caractères sont en haut de la liste parce que les chiffres liés à eux sont toujours grands, spécialement dans l’attribut *longueur des caractères*. Il est un peu problématique que les trois premiers attributs soient les mêmes pour les deux groupes, car distinguer les groupes ne fonctionne pas dans ce cas-là. Plusieurs tokens et attributs sont les mêmes dans les deux listes, ce qui n’est pas surprenant à cause de la loi de Zipf. Les tokens sont aussi très communs en français, par exemple *ça*, *un* et *il*. Pourtant, il y a des différences entre les groupes. Les émoticônes sont en haut de la liste jeune, mais pas du tout dans celle des adultes. Il est ainsi probable que les jeunes s’en servent plus fréquemment que les adultes. Les vocables qui font référence à la première personne du singulier *je* paraissent être plus communs chez les jeunes : l’attribut *je* et les tokens *j’*, *ai*, *moi* et

me. Cette observation pourrait soutenir le point de vue que les jeunes parlent plus d'eux-mêmes, mais il faut se souvenir que la fonction des SMS est de partager ses propres nouvelles avec les destinataires, à cause de quoi l'attribut *je* se trouve aussi dans la liste adulte.

L'attribut *répétition* des caractères est assez haut dans la liste des adultes. Il se trouve deux répétitions : ... et *!!!*. Nous voyons clairement une différence dans l'usage de la répétition des jeunes et des adultes, et il semble que les adultes y recourent plus. Ensuite, le mot *bisous* figure seulement dans la liste des adultes même si c'est une formule de salutation assez courante dans les SMS. Pour les adultes, il peut être plus important d'avoir une salutation claire dans leurs messages, tandis que les jeunes peuvent envoyer des SMS sans expressions de salutation. C'est probablement aussi à cause du vocable *bisous* que l'attribut *SMS* est assez haut dans la liste des adultes, car *bisous* est un des vocables de l'attribut *SMS*. Il est aussi un peu surprenant que les consonnes seules soient en haut de la liste des adultes. Les consonnes seules devraient représenter le taux des substitutions typographiques dans les messages, qui sont un des phénomènes très communs dans les SMS. Il se peut ainsi que les adultes aient assimilé ce procédé SMS dans leurs pratiques scripturales.

Il y a des attributs qui manquent dans la liste. Beaucoup d'entre eux font partie des attributs SMS. Les attributs *jeune* et *adulte* sont parmi eux et ces attributs contiennent des vocables comme *école*, *cours*, *travail* et *famille*. Même s'il est probablement vrai que les thèmes dans les messages des jeunes et des adultes sont liés à l'école et au travail, il est possible que cela ne soit pas visible dans les messages SMS – c'est-à-dire qu'ils n'en parlent pas explicitement ou avec les vocables que nous avions choisis. Quant à l'attribut *abréviation*, les abréviations sont probablement plus utilisées dans les textes écrits en général, et c'est pour cela qu'elles ne sont pas signifiantes pour le modèle. Pour estimer les chiffres à l'intérieur d'un token, autrement dit la phonétisation, nous avons l'attribut *numéro*. Il est un peu surprenant qu'il ne soit pas dans la liste, car ce procédé est commun d'après la littérature.

Les attributs syntaxiques, *nous* et *trop*, ne figurent pas non plus dans la liste. Comme nous l'avons déjà dit, *nous* manque probablement parce que les gens écrivent sur eux-mêmes, sur ce qui leur est arrivé, même les adultes. L'attribut *trop* devrait nous dire si le mot *trop* s'est substitué à *très*, mais il est possible que, en 2011, l'année de collecte des données, cette tendance n'ait pas encore été si dominante. Le dernier attribut qui manque est *apostrophe*, qui désigne le taux d'apostrophes grammaticalement incorrectes, comme *j'suis*. Il n'est pas impossible d'en trouver dans le corpus, mais la tendance à s'en servir ne semble pas être si fréquente.

5.3.2 Modèle sans les attributs

Passons maintenant au deuxième modèle, celui sans nos attributs. Les deux listes, celle des jeunes et celle des adultes, sont assez similaires. Elles incluent toujours beaucoup de mots vides, mais il se trouve aussi quelques différences entre les jeunes et les adultes. L'émoticône *:)* figure seulement dans la liste jeune, ce qui explique aussi pourquoi l'attribut *émoticône* est si important sur le côté jeune. Le vocable *bisous* est toujours seulement dans la liste adulte. On peut aussi constater l'importance de l'attribut *répétition* chez les adultes parce que les deux répétitions figurent toujours dans la liste adulte, mais seulement une d'elles (...) figure sur le côté jeune.

Mais ce qui paraît être similaire sont les vocables liés au scripteur et au destinataire du message. Chacun des deux groupes inclut les tokens *j'*, *ai*, *moi*, *me*, *te* et *toi* dans leurs listes ; les adultes ont même le token *suis* et les jeunes *t'*. Nous pouvons dire sans grand risque d'erreur que, dans les SMS, ce ne sont pas seulement les jeunes qui parlent d'eux-mêmes mais que les adultes et les jeunes le font dans la même proportion.

5.3.3 Résumé

Les différences n'ont pas été grandes entre les deux groupes, mais il va sans dire qu'il y en avait quelques-unes. Il semble aussi que les adultes aient assimilé dans leur langage au moins quelques procédés présentés ci-dessus. Pour autant, nous devons noter que nos attributs n'ont pas été les plus réussis pour révéler les éventuelles différences. Sur les 15 attributs, six ne se sont même pas trouvés parmi les plus attributs les plus distinctifs. Ces attributs ont été les mêmes pour les deux groupes, ce qui provoque une situation où les différences ne sont pas signifiantes. De plus, les attributs en haut de la liste ont été les mêmes, à cause de quoi la distinction entre les groupes ne fonctionne pas très bien.

5.4 Analyse des messages

Dans cette section, nous prendrons aléatoirement dix messages de chaque groupe et les analyserons plus en détail en faisant attention aux procédés SMS et à la performance du modèle. Nous donnerons toujours le message original et le message modifié, mais il faut se souvenir que même le message original a été modifié : la minusculation, la suppression des tags et la segmentation en mots selon les principes décrits dans la section 4.2.2 ont été exécutées. Le modèle se sert des messages modifiés pour mieux fonctionner, mais les informations de nos attributs sont rassemblées soit à partir des messages originaux, soit à partir des messages modifiés. Sous chaque message, les probabilités pour les deux groupes (jeune ou adulte) sont indiquées, sui-

vies par l'indication de si la classification a été correcte. Nous partirons des messages jeunes et continuerons par les messages adultes.

5.4.1 Messages jeunes

1) Original : *j' ai eu pareil et je pensais aussi m' etre loupee ... mais si ça se trouve c' est pour tte la classe ... je sais pas trop*

   Modifié : *j' ai eu pareil pensais aussi m' etre loupee ... si ça se trouve c' pour tte classe ... sais trop*

   Probabilité jeune : **99,85%** Probabilité adulte : 0,15%

   Prédiction : CORRECTE

Le modèle a bien prédit ce SMS, avec beaucoup de certitude. Dans le message, il y a deux répétitions de ..., qui sont plus communs parmi les adultes, mais il n'est pas impossible d'en trouver dans la langue des jeunes ; le token ... est également sur la liste jeune. Le message inclut quatre tokens de l'attribut *je* : *j'*, *je*, *m'* et *je*. On trouve aussi le mot *ça* qui a été un des attributs chez les jeunes. Le token *tte* représente le mot *toute*, qui est une abréviation plutôt graphique. Il y a 196 occurrences de *tte* dans les données, dont 193 dans les messages jeunes. En outre, le message contient une omission très courante : celle de *ne* de la négation. Ainsi, le message inclut assez d'éléments qui incitent le modèle à penser qu'il s'agit d'un message jeune.

2) Original : *ok poupée*

   Modifié : *ok poupée*

   Probabilité jeune : 27,55% Probabilité adulte : **72,45%**

   Prédiction : INCORRECTE

Ce message est très court, et il n'est pas surprenant que le modèle n'ait pas bien réussi dans la classification. Cela est visible dans les probabilités qui ne sont pas aussi fortes que par exemple dans le message (1). Pour l'être humain qui fait de la classification, il ne serait pas non plus très facile de dire si le SMS a été écrit par un jeune ou un adulte parce qu'il n'y a pas beaucoup de contexte ni de tokens pour le découvrir.

3) Original : *on rentre on est fatigué . bonne soirée ;)*

   Modifié : *rentre fatigué . bonne soirée ;)*

   Probabilité jeune : **70,51%** Probabilité adulte : 29,49%

   Prédiction : CORRECTE

Le modèle a choisi la bonne classe pour ce SMS même s'il a peut-être hésité un peu. Il est probable que l'émoticône *;)* a été décisive dans ce cas, car les autres tokens sont des vocables

assez fréquents et normaux – sauf que les phrases sont juxtaposées –, mais les émoticônes comme *;)* sont plus fréquentes parmi les messages jeunes. Par conséquent, classifier ce message serait assez difficile pour quiconque parce qu'il n'y a pas d'autres éléments qui fassent référence à un groupe.

4) Original : *mdrrrrrrrr . you' re welcome !*

   Modifié : *mdrrrrrrrr . you' re welcome !*

   Probabilité jeune : 14,45% Probabilité adulte : **85,55%**

   Prédiction : INCORRECTE

Le modèle a été sûr de la classe, mais la classification n'a pas réussi, car le message n'a pas été écrit par un adulte. Il y a deux raisons à cela : le message contient des tokens en anglais, qui ne sont pas du tout courants dans les données, et, après les tokens anglais, il ne reste que trois tokens peu informatifs : *mdrrrrrrrr*, . et *!*, dont le premier est le sigle *mdr* de l'expression *mort de rire* et qui comprend aussi une répétition du caractère *r*. Comme nous l'avons déjà constaté, la répétition est plus importante chez les adultes, ce qui a probablement mené le modèle vers la classe adulte. Pour une personne censée classifier des messages, il serait peut-être plus facile de découvrir qu'il s'agit d'un jeune, car le sigle SMS *mdr* est présent dans le message.

5) Original : *toute façon je sais pas à quelle heure je vais arrivé !*

   Modifié : *toute façon sais à quelle heure vais arrivé !*

   Probabilité jeune : 27,46% Probabilité adulte : **72,54%**

   Prédiction : INCORRECTE

Ce SMS contient des vocables fréquents et complètement normaux, à cause de quoi le modèle n'a pas eu une bonne performance. Les tokens sont aussi correctement écrits, sans procédés SMS qui changeraient l'orthographe du vocable, même si le dernier n'est pas grammaticalement correct. En outre, le message comprend deux omissions, un procédé syntaxique courant dans les SMS, car il manque *de* à l'expression *de toute façon* et *ne* à la négation. Tous les tokens ici sont tels que leurs pondérations sont grosso modo similaires, sauf peut-être celle de l'attribut *je* sur le côté jeune, ce qui doit être la raison pour l'hésitation du modèle.

6) Original : *mais non pas parce que t' as oublié ton téléphone ! parce que tu répond jamais et t' en à rien à foutre et que tu fait aucun effort !*

   Modifié : *non parce t' as oublié ton téléphone ! parce répond jamais t' à rien à foutre fait aucun effort !*

   Probabilité jeune : **98,48%** Probabilité adulte : 1,52%

   Prédiction : CORRECTE

Ce message contient des fautes d'orthographe et des formes grammaticalement incorrectes. Premièrement, les verbes *répondre* (*répond*), *avoir* (*à*) et *faire* (*fait*) sont mal conjugués : selon la norme du français écrit, on devrait écrire *tu réponds*, *tu fais* et *tu en as*. La dernière faute est également un exemple de la phonétisation, car la forme verbale *as* est simplifiée en *à*, mais nous ne pouvons pas éliminer la possibilité qu'il s'agisse seulement d'une faute d'orthographe. Deuxièmement, on trouve trois omissions de *ne*. Les bonnes formes seraient *tu ne réponds jamais*, *tu n'en as rien* et *tu ne fais aucun effort*. De plus, nous voyons l'expression du registre familier *n'en avoir rien à foutre*. À cause des fautes, une personne lisant ce message aurait probablement tendance à le classifier comme jeune. En tout cas, le modèle a bien fonctionné, car la classe a été bonne, avec un haut niveau de certitude. Cela résulte probablement en grande partie de l'utilisation de *t'*, qui était parmi les attributs les plus jeunes.

7) Original : *je l' appellerai la semaine prochaine je vais m' allonger*

   Modifié : *l' appellerai semaine prochaine vais m' allonger*

   Probabilité jeune : **57,39%** Probabilité adulte : 42,61%

   Prédiction : CORRECTE

Ce message inclut trois tokens liés à l'attribut *je* : deux fois *je* et *m'*. La juxtaposition des phrases est l'unique chose qui frappe, le reste est grammaticalement correct dans ce SMS. D'une certaine manière, c'est probablement pour cela que le modèle n'a pas été très sûr de la prédiction. Ainsi, le taux de l'attribut *je* paraît être la raison du bon fonctionnement du modèle.

8) Original : *oui mais ca t empeche que tu peux faire qq chose de tes journees !*

   Modifié : *oui ca t empeche peux faire qq chose tes journees !*

   Probabilité jeune : **50,53%** Probabilité adulte : 49,47%

   Prédiction : CORRECTE

De nouveau, le modèle a hésité dans la classification. Dans ce SMS, on remarque des éléments intéressants. Premièrement, au lieu d'écrire *ça*, le participant a écrit *ca*, qui serait classifié comme une suppression graphique. De la même façon, *t empeche*, un exemple de substitution graphique, s'écrirait *t'empêche* dans la langue standard. En outre, l'abréviation *qq*, qui représente *quelque*, est incluse dans la phrase. Malgré les éléments énumérés, la phrase ne contient pas de vocables curieux ou de vocables qui distingueraient bien les deux groupes. L'attribut *consonne*, qui était plus haut dans la liste adulte, a probablement été le seul à mener le modèle vers la classe adulte, même si le résultat s'est avéré bon.

9) Original : *wai et twa ?*

Modifié : *wai twa ?*

Probabilité jeune : **99,49%** Probabilité adulte : 0,51%

Prédiction : CORRECTE

Ce message très court contient des séquences caractéristiques des SMS. Il s'agit de l'orthographe phonétique, et pour être plus précise, de la simplification des semi-voyelles. Les séquences *wai* et *twa* représentent les vocables *ouais* et *toi*. On pourrait aussi dire que l'orthographe du vocable *ouais* est modifiée, autrement dit qu'il s'agit d'une substitution phonétisée. Ce qui surprend un peu est la certitude que le modèle a eue dans la prédiction malgré la brièveté du message. Cette certitude est due aux deux vocables mentionnés : le modèle n'a vu que des cas où *wai* et *twa* étaient dans les messages des jeunes. C'est pour cela que le taux de probabilité est si élevé.

10) Original : *j' ai été voir sur internet le tatoueur dont tu m parlais . je t' aime*

Modifié : *j' ai été voir sur internet tatoueur dont m parlais . t' aime*

Probabilité jeune : **61,3%** Probabilité adulte : 38,7%

Prédiction : CORRECTE

Dans ce SMS, il y a deux occurrences de l'attribut *je*, mais le destinataire du message est mentionné deux fois lui aussi. Les vocables ici font partie du langage assez courant, sauf la consonne isolée *m* qui s'écrirait *me* dans la langue standard. Dans la liste des attributs jeunes, *t'* était parmi les plus utilisés, de la même manière que l'attribut *je*, le token *j'* et l'attribut *consonne*. Toutefois, ces attributs étaient aussi dans la liste adulte, sauf l'attribut *t'*. Pour conclure, il n'a pas été facile de décider de la classe du message, ce qui a pour conséquence que la différence entre les taux de probabilité n'était pas grande.

5.4.2 Messages adultes

11) Original : *fyi . i am likely to be stuck at work late tonight . trials with .*

Modifié : *fyi . i am likely to be stuck at work late tonight . trials with .*

Probabilité jeune : 0,32% Probabilité adulte : **99,68%**

Prédiction : CORRECTE

Ce message a été écrit en anglais [9]. Ce qui surprend le plus est que le modèle ait été si sûr de

9. Il peut sembler que l'anglais soit fréquemment utilisé dans les messages, mais les messages ont été choisis au hasard. Le plus courant est de trouver un message qui contient un ou deux vocables en anglais, mais les messages complètement écrits en anglais sont très rares dans le corpus.

classer le message comme adulte, même si cette classification est correcte. Pour une personne qui parle anglais, le mot *work* pourrait révéler plus facilement qu'il s'agit d'une personne adulte.

12) Original : *ok . bisous bonne route*

Modifié : *ok . bisous bonne route*

Probabilité jeune : 1,96% Probabilité adulte : **98,04%**

Prédiction : CORRECTE

Le mot *bisous* est présent dans le message, et comme nous l'avons vu dans la section 5.3, c'est un vocable significatif pour le groupe adulte. À part ce vocable, il n'y a pas d'autres tokens qui font penser particulièrement aux adultes. Le taux de probabilité pour la classe adulte est très élevé ; nous arrivons à la conclusion que le classifieur a plus facilement de bons résultats qu'un être humain faisant la classification.

13) Original : *ohhhhh ma ... trop contente de te voir ...*

Modifié : *ohhhhh ma ... trop contente te voir ...*

Probabilité jeune : 1,23% Probabilité adulte : **98,77%**

Prédiction : CORRECTE

De nouveau, les éléments dans le message ne sont pas frappants – il n'y a que trois fois des répétitions : deux fois de ..., ce qui est très commun parmi les adultes, et les *hhhhh*. C'est probablement grâce à ces éléments que le modèle a fonctionné avec beaucoup de certitude. De plus, il manque le sujet et le verbe dans le message.

14) Original : *je suis en robe sans collants : mon cul c' est un igloo .*

Modifié : *suis robe sans collants : mon cul c' un igloo .*

Probabilité jeune : **58,08%** Probabilité adulte : 41,92%

Prédiction : INCORRECTE

Voilà un exemple d'hésitation du classifieur. Les probabilités pour les deux classes sont proches l'une de l'autre, ce qui a pour conséquence que le modèle ne sait pas quel groupe choisir. Le message n'inclut pas beaucoup d'éléments surprenants à l'exception du token *cul*, qui appartient au registre familier, et du token *igloo* qui apparaît uniquement dans ce message de notre corpus.

15) Original : *bah oui si tu veux ! ! ! c' est à sète ? ! ? cardinal ? ! ?*

Modifié : *bah oui si veux ! ! ! c' à sète ? ! ? cardinal ? ! ?*

Probabilité jeune : 0% Probabilité adulte : **100%**

Prédiction : CORRECTE

Dans ce SMS, nous trouvons toujours une répétition mais aussi des signes de ponctuation qui sont communs parmi les adultes, à savoir *?!?*. C'est probablement, de nouveau, pour cela que le modèle a été si sûr de la classe du message. Pour l'humain, la classification ne serait sans doute pas si simple parce que le message ne donne pas d'autres signes qui feraient référence aux adultes.

16) Original : *je pense que je dormirai !*

    Formatted : *pense dormirai !*

    Probabilité jeune : **55,5%** Probabilité adulte : 44,5%

    Prédiction : INCORRECTE

Ce message est très court, et c'est pour cela que le modèle l'a classifié de façon erronée. En plus des vocables fréquents dans le langage ordinaire, on trouve deux occurrences de l'attribut *je*. Comme nous l'avons déjà mentionné plusieurs fois, c'est l'un des attributs en haut de la liste jeune. C'est aussi le cas pour le point d'exclamation. Il faut néanmoins se souvenir que ces deux attributs se trouvent aussi dans la liste adulte, mais que les pondérations sont plus petites. En conclusion, le modèle n'a pas su décider de quel groupe il s'agissait, mais ce serait difficile pour quiconque, car le SMS ne donne pas beaucoup de contexte.

17) Original : *ok moi j fini à 17 h j récupéré et on s rejoint à la maison vers 17 h 30 sinon j pense k sera à la maison*

    Modifié : *ok moi j fini à h j récupéré s rejoint à maison vers h sinon j pense k sera à maison*

    Probabilité jeune : 0% Probabilité adulte : **100%**

    Prédiction : CORRECTE

Ce message contient beaucoup de consonnes seules, plus précisément cinq (*h* n'est pas inclus), attribut qui est assez haut dans la liste adulte. Les consonnes *j*, *s* et *k* sans la lettre *e* représentent la prononciation car, prononcé rapidement, l'e caduc ne s'entend pas. Ces consonnes isolées sont ainsi des vocables simplifiés de l'orthographe phonétique. Ce qui surprend est l'illogisme des temps verbaux : les deux premiers *j* s'écriraient-ils *j'ai* ou *je* ? Le premier *j* devrait être au temps présent parce que l'expression du temps *à 17h* fait référence au futur proche ; dans ce cas-là, *fini* devrait aussi s'écrire *finis*. Mais dans le deuxième *j*, la forme verbale est au participe passé, ce qui indiquerait *j'ai*. S'il s'agit du premier cas, tous les *j* représentent la même séquence de sons [ʒə], mais s'il s'agit du dernier cas, les *j* représentent deux séquences : [ʒə] et [ʒɛ].

Mais pourquoi le modèle a-t-il été si sûr de la classe ? L'attribut *consonne* explique en partie la certitude. D'ailleurs, les vocables contiennent plusieurs diacritiques, six au total. L'attribut *diacritique* se trouvait en haut des deux listes. L'attribut *k* a probablement aussi eu un effet sur

le résultat. Le vocable *maison* est inclus dans l'attribut *adulte*, et il est possible que cet attribut ait influencé le taux de probabilité.

18) Original : *cc ma belle j' espère k vou avez passé une bonne soirée et que été content même s' il ore préféré otre chose . dis moi s' il est déçu ou pa j' arrête pa d culpabilise j sui mal bisous bisous*

Modifié : *cc ma belle j' espère k vou avez passé une bonne soirée été content même s' il ore préféré otre chose . dis moi s' il déçu ou pa j' arrête pa d culpabilise j sui mal bisous bisous*

Probabilité jeune : 0% Probabilité adulte : **100%**

Prédiction : CORRECTE

Ce message est assez long et contient beaucoup d'éléments intéressants à la fois dans les attributs et dans les caractéristiques SMS. Partons des attributs. Premièrement, l'attribut *je* est fortement présent dans le message – on en trouve trois occurrences. Deuxièmement, le token *bisous* est un des attributs associés aux adultes, mais aussi un des vocables de l'attribut *SMS*. À la lumière de tout ce que nous venons d'analyser, nous pouvons dire que les deux tokens *bisous* seraient suffisants que le modèle prédise qu'il s'agit d'un SMS écrit par un adulte. Troisièmement, trois consonnes sont isolées, dont *k* est la plus intéressante, car c'est l'un des attributs trouvés dans les listes. Bien que l'attribut *diacritique* ne distingue pas très bien les groupes, le taux de 12 diacritiques a bien pu avoir un effet sur la classification. En résumé, tous ces attributs ont contribué au fait que le taux de probabilité a été complètement en faveur de la classe adulte.

En ce qui concerne les autres caractéristiques, le plus frappant dans ce message est la quantité de segments en orthographe phonétique. Le plus commun est la suppression des fins muettes, dont les séquences sont les suivantes : *vou*, *pa*, *sui* et *été*. Il s'agit ainsi d'omettre des lettres, ce qui peut rendre le message plus bref et plus simple, mais le premier vocable, *vou*, ne respecte pas les règles de liaison. L'autre caractéristique consiste en la simplification des digrammes et des trigrammes : *ore*, *otre*, *culpabilise* et *été*. Dans *ore*, *o* et *e* remplacent *au* et *ait*, respectivement. *Au* a de nouveau été remplacé par *o* dans *otre*, mais dans *culpabilise*, *e* se substitue à *er*. La fin *ait* est remplacé par *é*. En plus des caractéristiques énumérées, nous trouvons des e caduc supprimés dans *k* et *d*, comme c'était le cas dans l'exemple (17). D'autres éléments intéressants sont l'abréviation *cc*, qui désigne le vocable *coucou*, l'omission de *ne* et la juxtaposition des énoncés.

19) Original : *tu te sens comment avec lui ? ! ?*

Modifié : *te sens comment avec lui ? ! ?*

Probabilité jeune : 0,22% Probabilité adulte : **99,78%**

Prédiction : CORRECTE

Le modèle a bien prédit la classe de ce SMS, mais comme le message est assez court, il est surprenant qu'il ait été si sûr de la prédiction. En fait, l'unique élément qui fait penser aux adultes dans ce message est la séquence de signes *?!?*, car elle se trouve dans la liste des attributs adultes, mais pas dans la liste jeune. Les autres vocables ne sont pas exceptionnels dans la langue standard.

20) Original : *envoyé*

Modifié : *envoyé*

Probabilité jeune : 11,18% Probabilité adulte : **88,82%**

Prédiction : CORRECTE

Durant la collecte des SMS, les participants ont reçu un message qui leur indique que le message a bien été envoyé à l'équipe de recherche. Ce SMS est l'exemple d'un tel message. En ce qui concerne le taux de probabilité assez élévé, il est possible que les adultes aient donné plus de messages de ce type à la recherche, ce qui expliquerait la certitude du modèle.

5.5 Résumé

Pour nous servir du corpus peu équilibré, nous avons dû équilibrer les données. Malgré l'équilibre, les premiers résultats nous ont menée à faire d'autres modifications, par exemple supprimer quelques tokens pour réduire le bruit et l'influence de la matrice creuse. Le problème est que nos attributs ne se sont pas avérés bons pour distinguer les deux groupes. Néanmoins, il semble qu'avec les attributs basés sur les caractères nous ayons obtenu les meilleurs résultats même si le modèle choisi fait l'usage de tous nos attributs avec le sac de mots.

Les prédictions marchent relativement bien avec le modèle sac de mots, spécialement quand on examine les probabilités. Sur les 20 messages analysés, 15 ont été correctement prédits, ce qui nous donne une exactitude de 0.75. Il est aussi visible que dans les messages longs, les taux de probabilité sont très hauts et les prédictions correctes. Cela signifie que le modèle est bon – avec une quantité importante d'informations, le modèle marche impeccablement. Si le modèle hésite dans la prédiction, il s'agit de messages dont l'humain non plus ne pourrait pas déterminer le groupe. À cet égard, le modèle fonctionne très bien.

Les pondérations créées par le modèle paraissent bien fonctionner pour la classification des messages. Les attributs les plus distinctifs, d'après l'analyse des attributs et des messages, ont été les émoticônes, par exemple *:)*, les abréviations communes dans les SMS et le mot *ça* pour les jeunes ; la répétition des caractères et le mot *bisous* pour les adultes. Il semble, néanmoins, que les différences ne soient pas si grandes : en haut des listes, il y a des attributs

qui sont exactement les mêmes pour les deux groupes. Ces attributs sont assez inutiles, parce qu'il n'aident pas à distinguer les adultes des jeunes. Il existe également des attributs qui ne se sont pas trouvés parmi les 25 attributs relevés ; leur utilité pour le modèle est ainsi faible.

À l'opposé, les similitudes trouvées sont assez importantes. En plus des vocables très fréquents en français, le langage lié au scripteur et au destinataire du message est commun aux deux groupes. Le langage conversationnel et parlé domine dans le corpus. En outre, dans certains messages analysés, l'écriture imite la prononciation et la forme phonétique dans chacun des deux groupes. Nous avons aussi trouvé des cas où l'écriture ne suit pas les règles d'orthographe, ce qui nous fait nous demander si c'est aussi une caractéristique des SMS ou si cela reflète plutôt des difficultés à écrire en langue standard. D'après ces observations, nous pourrions constater que le registre de SMS est en train de se former.

6 Discussion

Les données utilisées ont été équilibrées, ce qui veut dire que le nombre de SMS jeunes a été réduit au même niveau que celui de SMS adultes. Pour cette raison, la proportion ne représente plus celle de l'échantillon des données. Toutefois, cette proportion peut mieux représenter la réalité des utilisateurs SMS – les messages SMS envoyés par les jeunes et les adultes sont plus en équilibre. Cela signifie que l'algorithme utilisé dans cette étude peut facilement être appliqué dans d'autres études. Malgré tout, très souvent, dans les données utilisées, les adultes, et spécialement les plus âgés, sont en minorité, quelle que soit l'étude.

De plus, il est un peu problématique de traiter les groupes d'âge : le groupe d'âge est plus étendu chez les adultes, 25 ans, tandis que l'étendue de celui des jeunes est de 10 ans. Comme nous l'avons vu dans la section 3.1, qui traite de la relation entre le langage et l'âge, les groupes d'âge ne sont pas du tout homogènes et ce n'est pas le cas en particulier s'il y a 25 ans entre le plus jeune et le plus âgé du groupe. Mais il faut se souvenir que les SMS n'ont pas été recueillis au niveau national – ceux qui ont voulu contribuer à la collecte des messages ont pu le faire. C'est pour cela que l'ensemble peut être biaisé. De plus, il aurait été utile d'examiner les résultats des questionnaires pour voir les conclusions qu'il est possible de faire à propos de ce corpus.

Les attributs que nous avons créés n'ont pas eu d'effet sur le fonctionnement de l'algorithme à cause du modèle sac de mots qui se sert de tous les tokens dans les données. Pour réduire le nombre de tokens, il est normal de recourir à la racinisation qui consiste à supprimer les affixes dans le token. Dans notre cas, cela aurait été d'une grande importance, mais comme les formes de vocables varient de manière inattendue, il n'a pas été possible d'utiliser des racinisateurs. L'unique solution aurait été de prendre un racinisateur (Martins, Matsubara & Monard 2003 : 229) qui connaisse les formes et les procédés des SMS, mais au moins au moment de rédaction de ce travail, nous n'étions pas consciente de l'existence d'un tel racinisateur. La racinisation pourrait être la solution pour améliorer le fonctionnement de l'algorithme et pour assurer que la classification marche bien pour n'importe quel texte médié par ordinateur. À l'aide de la racinisation, il serait également possible d'améliorer la performance des bigrammes. Les bigrammes n'ont pas très bien fonctionné, probablement à cause de la perte des tokens et par conséquent, de l'ordre des mots original. Cela dit, les futures recherches nous montreront si les bigrammes, ou bien les n-grammes, aideraient dans la classification des SMS.

Le problème du modèle de sac de mots est que l'ordre des mots n'est pas pris en considération : le modèle tient compte seulement des fréquences pour compter des probabilités. Il en

résulte deux choses. Premièrement, la longueur du message, c'est-à-dire le nombre de tokens, a beaucoup d'influence sur la performance, car plus il y a de tokens, plus il y a d'informations à compter. C'est un problème quand le message contient peu de tokens, et comme nous l'a montré la loi de Zipf, la présence d'un token peut beaucoup compter dans la prédiction de la classe, car la plupart des tokens sont des vocables très fréquents. La performance du modèle ne devrait pas dépendre de la longueur du message, mais les messages devraient être traités équitablement (Martins, Matsubara & Monard 2003 : 229). Deuxièmement, la sémantique n'est pas prise en considération dans le modèle. Les vocables qui sont sémantiquement proches entre eux pourraient être groupés dans un attribut, au lieu de les prendre chacun séparément. De cette manière, il ne s'agirait pas seulement de chiffres mais les significations des vocables pourraient également être prises en compte.

Quant au langage des SMS, il apparaît qu'il n'y a pas beaucoup de différences entre les deux groupes. D'une part, il se peut que le SMS soit en train de devenir un registre ou au moins une variété du registre familier. D'autre part, la communication est la fonction la plus importante du SMS, ce qui se reflète naturellement dans le langage aussi. Nous nous demandons si c'est en fait la combinaison du registre familier et de la communication qui produit le langage SMS. En tout cas, les adultes paraissent aussi être plus au courant des procédés des SMS qu'il ne l'a été observé dans les études faites sur ce sujet. De toute façon, il serait utile de comparer des SMS d'aujourd'hui à ceux d'il y a 20 ans parce que cela pourrait nous indiquer des changements dans les habitudes langagières.

Il est important de noter que faire la classification serait quasiment impossible pour l'être humain qui serait, selon toute probabilité, plus faible pour distinguer les groupes, sans parler du travail manuel que cela demanderait. En revanche, le modèle a fonctionné relativement bien selon tous les critères. Il est surtout bon que les probabilités soient souvent fortes, en particulier dans les cas où le message est long, car cela nous prouve que le modèle fonctionne. Si la probabilité était de 50% à chaque fois, le modèle prédirait la classe en jouant à pile ou face.

7 Conclusion

Le but de ce mémoire était de trouver des caractéristiques qui puissent distinguer la langue utilisée par les deux groupes d'âge, les adultes et les jeunes. Pour réaliser cette comparaison, nous nous sommes servi d'un corpus de 88 000 SMS, parmi lesquels les messages des jeunes et des adultes ont été choisis pour faire une classification automatique. La classification a été faite par l'apprentissage automatique supervisé, et pour être plus précise, par un modèle bayesien naïf. En somme, le classifieur a bien fonctionné, et en règle générale, se servir de cette méthode est bien possible.

Les messages SMS contiennent beaucoup d'éléments et de procédés qui les rendent distincts d'autres textes. La raison en est la créativité qui y joue un rôle important. En général, il semble que les jeunes et les adultes ont tous les deux adopté certains procédés SMS. C'est pour cela que notre analyse nous a menée à dire que le SMS pourrait plutôt être considéré comme un hyponyme du registre familier, car le langage utilisé dans les messages n'est pas uniquement réservé aux jeunes ou aux adultes. En général, le langage est conversationnel et imite le parlé. Il apparaît également que l'imitation de la prononciation est l'un des éléments les plus importants dans le SMS, visible par exemple dans la simplification des lettres finales. La plus grande différence se trouve, toutefois, dans les émoticônes, dans les abréviations comme *mdr*, dans la répétition des lettres et dans l'utilisation de quelques mots, comme *bisous* et *ça*. Nous nous demandons si la répétition et ces mots sont seulement des traits de notre corpus, même si *bisous* paraît être un élément important du registre SMS. Ainsi, il serait intéressant d'analyser les réponses des questionnaires plus en détail pour avoir une idée globale des SMS : le corpus représente-t-il les habitudes du langage SMS des parlants français ou francophones ?

Nous avons aussi étudié les méthodes elles-mêmes : que faire si le corpus contient beaucoup de bruit et de variation et quelle est la performance du modèle ? Les solutions pour éliminer l'effet de la variation ne sont pas simples, au moins elles ne s'appliquent pas à toutes les situations. Nous avons vu que supprimer des tokens dont l'occurrence était très restreinte pouvait aider, mais il faut toujours réfléchir à la proportion d'informations qu'on est prêt à perdre. Une autre solution claire serait de développer un racinisateur ou lemmatisateur qui comprenne au moins quelques procédés SMS et puisse en donner les formes en français standard. Cela serait utile, non seulement pour les données de SMS mais également pour d'autres données issues d'Internet.

En outre, les futures recherches pourraient appliquer les résultats que nous avons obtenus à d'autres données SMS, mais aussi aux données des médias sociaux. Les SMS et les médias

sociaux sont un phénomène récent, et c'est pour quoi les résultats de tout genre sont importants quand nous essayons de décrire ce phénomène. Nos résultats indiquent que le langage dans les SMS et celui dans les médias sociaux se ressemblent, ce qui facilitera évidemment le développement d'autres modèles.

Mais quelle approche devrait-on adopter dans la classification automatique des messages SMS ? Dans l'apprentissage supervisé, les solutions ne sont pas simples. L'exactitude à 100% n'est et ne sera pas possible, au moins pas dans un proche avenir. Nous avons choisi le modèle le plus simple au sein de l'apprentissage supervisé à cause de sa simplicité et sa facilité. Naturellement, il existe d'autres méthodes qui pourraient fonctionner même mieux dans la classification des SMS ; de nos jours, il est à la mode de se servir des réseaux de neurones (angl. *neural networks*) et de l'apprentissage profond (angl. *deep learning*). Il serait également important d'utiliser plus d'analyses statistiques pour s'assurer des résultats obtenus. Les futures recherches nous montreront de quelle manière les autres méthodes pourraient améliorer les résultats de la classification automatique des SMS.

Notre étude a touché à la fois le domaine de l'informatique et celui de la linguistique, et les deux aspects ont été présents tout au long de l'étude. Il va sans dire que les études multidisciplinaires seront de plus en plus importantes à l'avenir pour que nous puissions comprendre de manière exhaustive les phénomènes dans le monde. Il est aussi important de noter que notre mémoire s'est concentré sur le français et ses caractéristiques alors que l'anglais est très souvent la langue observée. D'après notre observation, la meilleure disponibilité des données anglaises, entre autres, mène les recherches de prendre l'anglais comme objet d'étude, ce qui n'est évidemment pas souhaitable à long terme.

Bibliographie

Argamon, Shlomo ; Koppel, Moshe ; Pennebaker, James W. & Schler, Jonathan (2009) « Automatically Profiling the Author of an Anonymous Text. » *Communications of the ACM* 52 : 2. 119–123.

Boyer, Henri (1991) *Éléments de sociolinguistique*. Dunod, Paris.

Boyer, Henri (2001) *Introduction à la sociolinguistique*. Dunod, Paris.

Cappeau, Paul & Moreno, Anaïs (2017) « Les tendances grammaticales. » Gadet, Françoise (éd) *Les parlers jeunes dans l'Île-de-France multiculturelle*. Éditions Ophrys, Paris. 73–99.

Chandramouli, R. ; Chen, Xiaoling ; Cheng, Na & Subbalakshmi, K. P. (2009) « Gender Identification from E-mails. » *2009 IEEE Symposium on Computational Intelligence and Data Mining*. 154–158.

Dickinson, Brian ; Hu, Wei & Miller, Zachary (2012) « Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. » *International Journal of Intelligence Science* 2 : 4A. 143–148.

Eckert, Penelope (1997) « Age as a Sociolinguistic Variable. » Coulmas, Florian (éd) *The Handbook of Sociolinguistics*. Blackwell Publishers Ltd, Cornwall. 151–167.

Estival, Dominique ; Gaustad, Tanja ; Hutchinson, Ben ; Pham, Son Bao & Radford, Will (2007) « Author profiling for English emails. » *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*. 263–272.

Fairon, Cédrick ; Klein, Jean René & Paumier, Sébastien (2006) *Le langage SMS. Étude d'un corpus informatisé à partir de l'enquête 'Faites don de vos SMS à la science'*. Presses universitaires de Louvain, Louvain-la-Neuve.

Gadet, Françoise (2017) « Pour étudier les 'parlers jeunes'. » Gadet, Françoise (éd) *Les parlers jeunes dans L'Île-de-France multiculturelle*. Éditions Ophrys, Paris. 27–53.

Goosse, André & Grevisse, Maurice (2011) *Le bon usage*. 15ème édition. Groupe De Boeck s.a., Bruxelles.

Gravel, Rilana ; Meder, Theo ; Nguyen, Dong & Trieschnigg, Dolf (2013) « 'How Old Do You Think I Am ?' : A Study of Language and Age in Twitter. » *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. 439–448.

Gruppioni, Elisa (2011) « La cyberécriture des adolescents : hétérogénéités et permanences de la 'langue jeune'. » Leinard, Fabien & Zlitni, Sami (éds) *La communication électronique : enjeux de langues*. Éditions Lambert-Lucas, Limoges. 277–291.

Laporte, Aurore ; Le Galloudec, Morgane ; Servent, Domitille ; Tran, Thi Mai & Trancart, Marine (2011) « Évolution des pratiques scripturales et caractérisation des troubles du langage écrit : vers une évolution des frontières. » Leinard, Fabien & Zlitni, Sami (éds) *La communication électronique : enjeux de langues*. Éditions Lambert-Lucas, Limoges. 235–247.

Manning, Christopher & Schütze, Hinrich (1999) *Foundations of Statistical Natural Language Processing*. MIT Press, Massachusetts Institute of Technology.

Manning, Christopher D. ; Raghavan, Prabhakar & Schütze, Hinrich (2008) *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

Martell, Craig & Tam, Jenny (2009) « Age Detection in Chat. » *2009 IEEE International Conference on Semantic Computing*. 33–39.

Martins, Claudia ; Matsubara, Edson & Monard, Maria-Carolina (2003). « Reducing the dimensionality of bag-of-words text representation used by learning algorithms. » *Proceedings of the 3rd IASTED International Conference on Artificial Intelligence and Applications*. 228–233.

McKeown, Kathleen & Rosenthal, Sara (2011) « Age Prediction in Blogs : A Study of Style, Content, and Online Behavior in Pre- and PostSocial Media Generations. » *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. 763–772.

Pankhurst, Rachel (2009) « Short Message Service (SMS) : typologie et problématiques futures. » Arnavielle, Teddy (éd) *Polyphonies, pour Michelle Lanvin*. Université Paul-Valéry Montpellier 3, Montpellier. 33–52.

Panckhurst, Rachel ; Détrie, Catherine ; Lopez, Cédric ; Moïse, Claudine ; Roche, Mathieu & Verine, Bertrand (2013) « Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS. » *Épistémè – revue internationale de sciences sociales appliquées* 9. 107–138.

Ross, Sheldon M. (2010) *Introductory Statistics*. Academic Press, Canada.

Schwartz, H. Andrew ; Eichstaedt, Johannes C. ; Kern, Margaret L. ; Dziurzynski, Lukasz ; Ramones, Stephanie M. ; Agrawal, Megha ; Shah, Achal ; Kosinski, Michal ; Stillwell, David ; Seligman, Martin E. P. & Ungar, Lyle H. (2013) « Personality, Gender, and Age in the Language of Social Media : The Open-Vocabulary Approach. » `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073791` (consulté le 10/10/2017)

Soler-Company, Juan (2016) « Use of Discourse and Syntactic Features for Gender Identification. » Pearce, David & Pinto, H. Sofia (éds) *STAIRS 2016 : Proceedings of the 8th European Starting AI Researcher Symposium*. IOS Press, Amsterdam. 215–220.

Soler Company, Juan & Wanner, Leo (2014) « How to Use Less Features and Reach Better Performance in Author Gender Identification. » Calzolari, Nicoletta *et al*. (éds) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association, Reykjavik. 1315–1319.

Thibault, Pierrette (1997) « Âge. » Moreau, Marie-Louise (éd) *Sociolinguistique. Concepts de base*. Pierre Mardaga, Sprimont. 20–26.

www1 = Corpus « 88milSMS. » `http://88milsms.huma-num.fr` (consulté le 9/11/2017)

www2 = Bépo. `http://bepo.fr/wiki/Accueil` (consulté le 6/11/2017)

www3 = Éduscol : Liste des mots les plus fréquents de la langue française. `http://eduscol.education.fr/cid47916/liste-des-mots-classee-par-frequence-decroissante.html` (consulté le 9/11/2017)

www4 = Cours de Recherche d'information textuelle. `http://www-connex.lip6.fr/~gallinar/gallinari/uploads/Teaching/2017-2018-RI.pdf` (consulté le 15/4/2018)

Zheng, Rong ; Li, Jiexun ; Chen, Hsinchun & Huang, Zan (2006) « A Framework for Authorship Identification of Online Messages : Writing-Style Features and Classification Techniques. » *Journal of the American Society for Information Science and Technology* 57 : 3. 378–393.

Annexe : Attributs

Attributs basés sur les caractères

— Répétition de caractères (par ex. *bieeen*) BINAIRE

— Longueur du message en tokens et en caractères FRÉQUENCE

— Nombre d'accents FRÉQUENCE

— Nombre de *k* FRÉQUENCE

— Fréquence des consonnes seules (par ex. *j* qui ne précède pas une apostrophe et/ou une voyelle) FRÉQUENCE

— Apostrophes incorrectes (élision non nécessaire, par ex. *j'sais*) BINAIRE

Attributs SMS

— Nombre d'émoticônes FRÉQUENCE

— Nombre de vocables liés aux jeunes (*école*, *cours*, *jeune*, *ado*, *devoirs*) FRÉQUENCE

— Nombre de vocables liés aux adultes (*travail*, *famille*, *maison*, *vieux*, *fils*, *fille*, *mère*, *père*, *maman*, *papa*) FRÉQUENCE

— Nombre d'abréviations et de sigles (plutôt graphiques) (*pr*, *tt*, *ds*, *qd*, *bcp*, *stp*, *svp*, *pcq*, *ss*, *rdv*) (selon Fairon, Klein & Paumier 2006 : 49) FRÉQUENCE

— Nombre de vocables typiques dans les SMS (de tokens et d'abréviations phonétisées) (*coucou*, *bisous*, *lol*, *mdr*, *ptdr*, *jtm*, *asap*, *jtd*, *jtad*, *tvb*, *tlm*) (les premiers [sauf *coucou* et *bisous*] selon Fairon, Klein & Paumier 2006 : 104, les deux derniers selon Panckhurst 2009 : 41) FRÉQUENCE

— Fréquence des chiffres à l'intérieur d'un token FRÉQUENCE

Mots vides et d'autres attributs syntaxiques

— Nombre de *j'*, *me*, *m'* et *moi* FRÉQUENCE

— Nombre de *nous* FRÉQUENCE

— Nombre de *trop* FRÉQUENCE

N-grammes

— Bigrammes