Faculté des **Langues**
Université de Strasbourg

Master Technologies des langues
Option Traitement Automatique des Langues

2022-2024

Détection des émotions dans les poèmes professionnels et amateurs

Alexia Artal

Mémoire de Master

Sous la direction de
M. Pablo Ruiz Fabo
Maître de conférences

Remerciements

Ce mémoire marque l'aboutissement de mon projet de master, résultant de mes efforts individuels et nourri par les expériences académiques et humaines acquises durant mon parcours scolaire et universitaire.

Je tiens tout d'abord à exprimer ma profonde gratitude envers l'ensemble de mes professeurs de Master, qui m'ont apporté des connaissances précieuses pour mener à bien ce projet de mémoire. Je tiens à souligner l'implication de Monsieur Pablo Ruiz Fabo, qui a toujours été disponible pour répondre à mes besoins. Ses conseils avisés ont joué un rôle crucial dans l'orientation et la réalisation de ce projet de mémoire.

Je tiens également à remercier Madame Delphine Bernhard pour pour ses divers enseignements, sa pédagogie et ses cours en apprentissage automatique, qui ont été d'une grande aide dans le développement de ce projet. Je remercie également Madame Amalia Todirascu pour ses nombreux cours dispensés durant ma licence en Sciences du Langage, qui ont suscité mon intérêt pour poursuivre un master en Traitement automatique des langues.

Enfin, je souhaite exprimer ma reconnaissance à mes proches pour leur soutien constant tout au long de ce parcours. Leur présence et leurs encouragements précieux ont été déterminants pour mener à bien ce projet.

Table des matières

Table des illustrations

Table des tableaux

Introduction

Les émotions sont une notion difficile à traiter. Nombreuses sont les discussions autour de ce qu'est une émotion. Ainsi, diverses théories sur les émotions ont vu le jour au fil du temps. Les premières théories basées sur des approches physiologiques ont dominé le domaine de la recherche sur les émotions pendant plusieurs années. Par la suite, grâce aux influences des travaux de Darwin, les théories néo-darwiniennes ont émergé se focalisant essentiellement à déterminer les émotions de bases. Bien qu'il existe des différences de point de vue sur la définition de l'émotion et de son origine, il est communément admis que l'émotion résulte de 3 composantes, à savoir : la composante subjective, la composante comportementale et la composante physiologique.

En ce qui concerne la poésie, celle-ci à réussie à se développer tout en étant marginalisé au sein du marché du livre en France. Un certain nombre d'études ont démontré que la poésie n'est pas rentable économiquement en France. L'enjeu principal d'un poète étant de se faire publier, ce mémoire (section 1.2.1) présentera les différents mécanismes et outils afin de se faire connaitre dans le monde de la poésie. Ce mémoire abordera les différences de point de vue sur le terme d'*amateur* et de *professionnel*. Au sein de cette recherche aucun jugement sur les poèmes écrits par des « poètes amateurs » ne sera réalisé. Le terme d'*amateur* sera employé afin de qualifier les poèmes publiés uniquement sur le Web. Tandis que le terme *professionnel* sera quant à lui utilisé pour qualifier les poèmes publiés dans des recueils de poésie.

Plusieurs recherches en traitement automatique des langues (TAL) ont été menées afin de détecter les émotions présentes dans les textes oraux ou écrits. Ainsi, de nombreux lexiques émotionnels, notamment en langue anglaise, existent permettant d'identifier en partie, les émotions dans un texte. Grâce à l'essor de la méthode d'apprentissage par plongements lexicaux, la représentation du sens par ordinateur s'est élargie. Ceci a permis de prendre en compte la sémantique lexicale des mots menant ainsi au développement d'algorithmes plus puissants. Les analyses de sentiments, d'émotions et opinions ont alors exploité cet avancé technologique afin d'accroître les performances de leurs analyses.

Ce mémoire a pour objectif d'identifier les différentes émotions présentes dans des poèmes amateurs et professionnels. L'étude se concentrera sur un échantillon représentatif de poèmes, englobant à la fois des poèmes collectés à partir de plateformes de poèmes amateurs ainsi que des poèmes professionnels, reconnus et publiés dans des anthologies. Différentes techniques relevant du traitement automatique des langues seront appliquées afin de détecter les diverses émotions. De plus, s'il existe des différences significatives liées au lexique émotionnel entre les poèmes amateurs et professionnels, un classifieur pourra être créé afin de classer les divers poèmes.

La structure de ce mémoire est divisée en plusieurs sections clés. La première partie du mémoire est consacrée à la revue de littérature, qui permet d'établir les bases théoriques de l'étude. Elle explore les concepts clés relatifs aux émotions, puis à la poésie et enfin aux technologies du TAL, en s'appuyant sur divers travaux antérieurs réalisés dans ces domaines. Cette revue de

littérature permet de définir les cadres conceptuels et méthodologiques qui guideront l'ensemble de cette recherche. La seconde section vise à décrire la méthodologie adoptée pour développer un classifieur automatique distinguant les poèmes amateurs des poèmes professionnels en se basant sur les termes émotionnels présents. Cette partie détaillera les différentes étapes du processus, depuis la constitution d'un corpus poétique français jusqu'à l'entrainement et l'évaluation de différents modèles de classification, en passant par l'analyse statistiques du corpus constitué.

1. Etat de l'art

1.1. Les émotions

Les discussions autour des émotions sont présentes depuis l'Antiquité. L'éthique aristotélicienne accorde une place privilégiée aux émotions, désignée par *ta pathê* en grec signifiant « affections » ou « passions ». Mais alors, qu'est-ce qu'une émotion ? Cette question est le titre d'un essai écrit par William James en 1884, parut dans la revue *Mind*. Dans cet essai, l'auteur défend l'idée selon laquelle les changements corporels suivent directement la perception d'un fait existant et que le sentiment de ces changements quand ils se produisent est l'émotion (James, 1884).

Dans cette section, les différentes théories de l'émotion seront passées en revue. Ensuite, sera abordé le consensus relatif aux composantes de l'émotion. Enfin, les distinctions entre les divers états affectifs seront présentées.

1.1.1. Théories de l'émotion

1.1.1.1. Théories physiologiques

Les premières théories des émotions traitent essentiellement du rôle de l'activation physiologique dans le déclenchement et le déroulement des processus émotionnels. Les théories de James (1884), Lange (1885) et Cannon (1927) ont été très influentes dans la recherche liée aux émotions. Ces théories, basées sur des approches physiologiques, ont dominé le domaine de la recherche sur les émotions pendant de nombreuses années. C'est lors du premier courant des théories cognitives des émotions qu'elles ont été développées et poursuivies.

On parle classiquement de la théorie de James-Lange. Ces deux auteurs ont défendu à la même époque mais séparément une conception révolutionnaire dite « périphérique » ou « périphéraliste » de l'émotion. Cette théorie suggère que l'émotion est due à une variation au sein du système nerveux périphérique de l'individu. Ces deux auteurs s'accordent sur plusieurs points, notamment sur la séquence émotionnelle qu'ils définissent telle que : un stimulus, des réponses physiologiques, puis la sensation de ces changements périphériques et, finalement, l'émotion. Un autre postulat de cette théorie est que ce sont les perceptions des changements périphériques de l'individu qui constituent en soi l'émotion. Enfin, notons également que selon cette théorie les réponses corporelles sont constituées de changements physiologiques différents selon les émotions que l'individu ressent et qu'une composante nécessaire à l'émotion serait le feedback corporel.

La théorie centrale de Walter Cannon (1927) qui sera poursuivie par Bard constitue une conception de l'émotion fortement opposée à celle de James-Lange sur plusieurs aspects fondamentaux. Alors que la théorie de James-Lange stipule que les différents états émotionnels activent des changements physiologiques différents, Cannon lui, affirme que l'activation

émotionnelle est indifférenciée selon le type d'émotion. Selon cette théorie, les émotions seraient induites par l'excitation du thalamus et le feedback viscéral n'interviendrait pas dans l'induction d'une émotion. Un autre désaccord important entre ces trois auteurs survient dans la possibilité d'un centre du cerveau spécifique qui induirait une émotion. Lange postule l'existence d'un centre vasomoteur, Cannon quant à lui développe des arguments en faveur du rôle joué par le thalamus, et James refuse indéniablement l'idée d'un centre cérébral de l'émotion. Ainsi, la théorie de Cannon stipule que les changements physiologiques ne seraient pas des causes mais des conséquences de l'émotion. La figure 1 illustre les séquences temporelles de l'émotion selon ces deux théories.

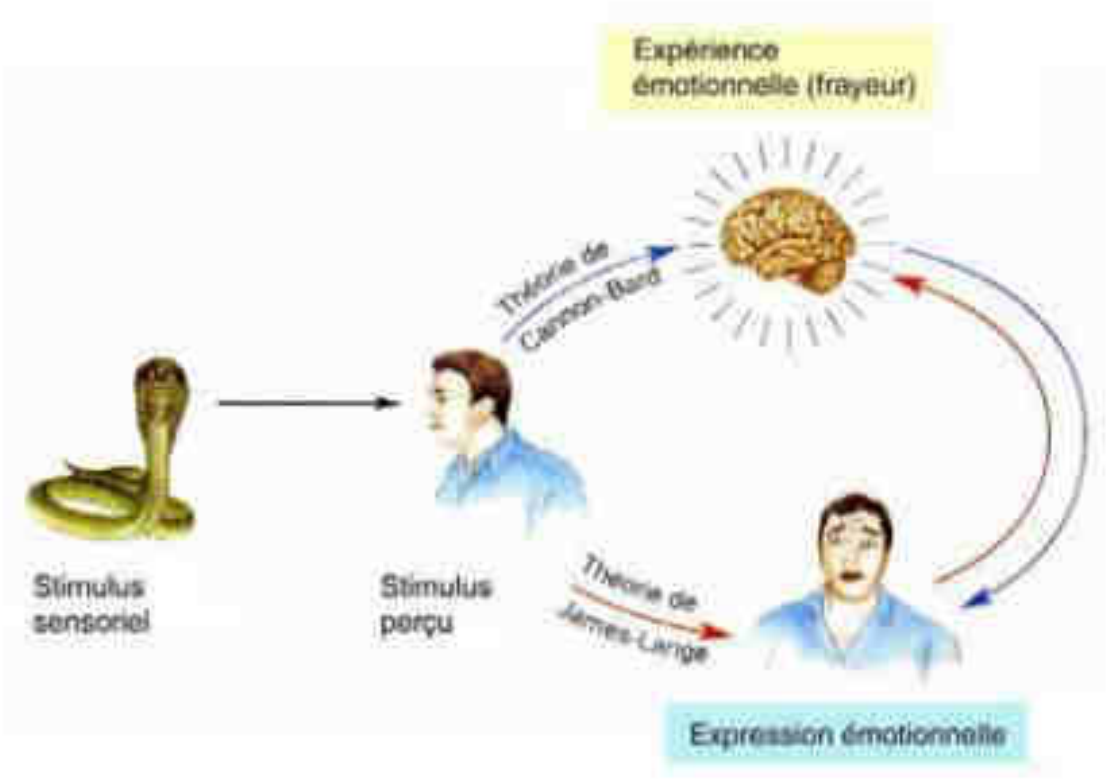

Figure 1 : Schéma de la séquence émotionnelle de James-Lange et Cannon-Bard
Source : Abdat, 2010

Les flèches bleues illustrent la conception de la séquence temporelle de l'émotion selon la théorie de Cannon-Brad. Les flèches rouges représentent le déroulement temporel de l'émotion selon la théorie de James-Lange.

La comparaison des différentes théories permet de mettre en évidence les divergences et les convergences entre les auteurs, tout en soulignant les questions clés qui ont été abordées, telles que le rôle des réponses physiologiques, les perceptions des changements corporels et l'existence d'un centre cérébral spécifique pour l'émotion. Leurs perspectives contrastées ont enrichi la compréhension des mécanismes émotionnels et ont ouvert la voie à de nouvelles perspectives de recherche dans ce domaine.

1.1.1.2. Théories néo-darwiniennes

Les théories suivant une perspective évolutionniste tirent leur origine des travaux darwiniens. Ces théories étudient principalement la fonction communicative des émotions tout en accordant une prédominance aux expressions faciales. L'un des premiers à s'intéresser aux phénomènes émotionnels fut Charles Darwin avec la publication de son ouvrage *L'expression des émotions chez l'homme et l'animal* (1872)[1]. Selon Darwin, les expressions émotionnelles de l'adulte humain sont le reflet de la continuité de systèmes complexes dérivés des autres espèces animales. Ainsi, selon cette perspective, les émotions auraient une qualité primitive adaptative qui serait autant liée à notre passé d'espèce, qu'à notre propre vécu en tant qu'individu.

Les théories néo-darwiniennes se sont essentiellement focalisées à déterminer les différentes émotions de bases en étudiant les expressions faciales liées aux états émotionnels. Le concept d'émotions primaires ou dites de base a été développé dans les années 1970. Les principaux pionniers de ces théories sont les chercheurs Paul Ekman, Carroll Izard et Wallace Friesen. Ainsi, les théories des émotions de base supposent l'existence d'un nombre restreint d'émotions primaires. Ce courant d'inspiration darwinien a permis de dénombrer facilement les émotions et de leur attribuer des propriétés biologiques et fonctionnelles communes. Ce courant repose sur 4 postulats :

1. il existe un nombre limité d'émotions fondamentales ;
2. elle sont universelles dans leur expression ;
3. elles ont chacune une fonction évolutionnaire pour l'espèce humaine ;
4. les émotions complexes résultent de l'essor des interactions humaines et sont le résultat d'une combinaison des émotions de base.

Un désaccord persiste selon les différents partisans de ces théories sur le nombre d'émotions de base. Leur décompte oscille entre 2 et 11. Ainsi, pour Arnold il y aurait 11 émotions primaires : la colère, l'aversion, le courage, l'abattement, le désir, le désespoir, la peur, la haine, l'espoir, l'amour et la tristesse. Selon Ekman, Friesen et Ellsworth, il existerait 6 émotions de base à savoir : la colère, le dégout, la peur, la joie, la tristesse, la surprise. Le tableau 1 permet de rend compte du décompte des émotions primaires selon différents auteurs. A noter que des concepts tels que la détresse, l'émerveillement ou encore l'espérance sont considérés par certains auteurs comme des émotions de base.

[1] Le titre original de cet ouvrage est *The Expression of the Emotions in Man and Animals*

Tableau 1 : Les diverses émotions primaires selon différents auteurs

Auteurs	Emotions primaires
Arnold (1960)	Colère, aversion, courage, abattement, désir, désespoir, peur, haine, espoir, amour, tristesse
Ekman, Friesen, & Ellsworth (1982)	Colère, dégoût, peur, joie, tristesse, surprise
Frijda (1986)	Désir, bonheur, curiosité, surprise, émerveillement, peine
Izard (1971)	Colère, mépris, dégoût, détresse, peur, culpabilité, curiosité, joie, honte, surprise
Mowrer (1960)	Douleur, plaisir
Panksepp (1982)	Espérance, peur, fureur, panique
Plutchik (1980)	Résignation, colère, anticipation, dégoût, joie, peur, honte, surprise
Tomkins (1984)	Colère, curiosité, mépris, dégoût, détresse, peur, joie, honte, surprise

Dans cette lignée, nous retrouvons la théorie de l'amplification de Tomkins. Selon cet auteur, l'affect désignant l'émotion « est le produit d'un ensemble organisé de réponses faciales, musculaires et viscérales suscitées par un programme spécifique inné » (Christophe, 1998). Tomkins propose de faire correspondre à chaque expression faciale une émotion spécifique. Ainsi, cet auteur comptabilise 9 émotions de bases à savoir : la joie, la surprise, la peur, l'intérêt, le mépris, le dégoût, la honte et la colère.

La théorie différentielle des émotions, développée par Izard (1971 ; 1977 ; 1978 ; Izard & Buechler, 1980) stipule que les émotions de base motivent et organisent l'ensemble des comportements afin de servir de fonction adaptative à la vie. De cette manière, les émotions fondamentales sont vues comme un phénomène motivationnel complexe, résultant de trois composantes principales : expressive, subjective et neurophysiologique. Izard dénombre quant à lui 10 émotions fondamentales : la joie, la surprise, la tristesse, la colère, le mépris, l'intérêt, le dégoût, la peur, la honte ainsi que la culpabilité.

Le modèle circumplex développé par Plutchik (1970 ; 1980 ; 1984) dans sa théorie psycho-évolutionniste générale des émotions, détermine 8 émotions de base : la colère, la peur, la joie, la tristesse, l'acceptation, le dégoût, la surprise et l'espérance. Afin d'élaborer une distinction des huit émotions de base, Plutchik identifie huit séquences comportementales différentes ayant toute une fonction adaptive distincte. Ainsi, toutes les autres émotions seraient pour cet auteur des dérivés des émotions fondamentales et constitueraient donc des émotions secondaires. Dans

cette théorie, Plutchik propose de relier ces différentes émotions de base selon trois dimensions à savoir : leur intensité, leur degré de similitude ainsi que leur polarité. La figure 2 représente le modèle du circumplex en trois dimensions de Plutchilk.

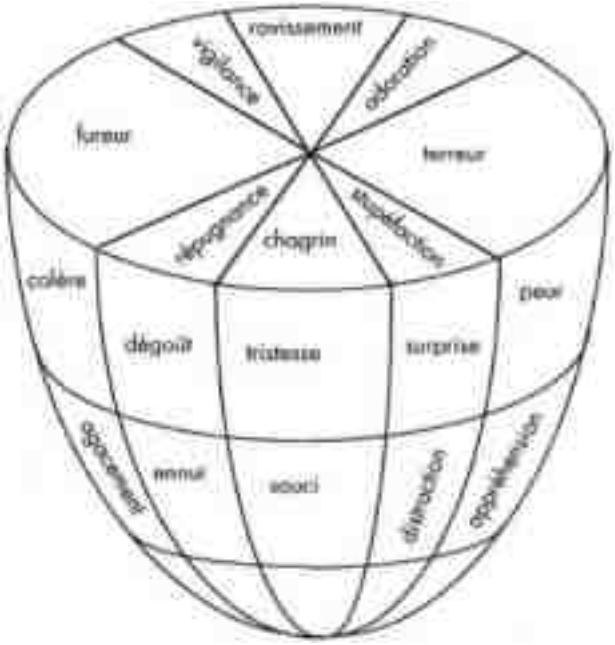

Figure 2 : Le modèle multidimensionnel de l'émotion (Plutchik, 1980)
Source : Christophe, 1998

Ainsi, l'étude de ces différentes théories contribue à enrichir notre compréhension des études des émotions. Ces perspectives théoriques fournissent un cadre conceptuel solide quant à cette recherche. En effet, la typologie des émotions de base d'Ekman, Friensen et Ellsworth servira de base à cette étude. De plus, la possibilité d'intégrer des émotions secondaires sera également envisagée, permettant ainsi d'explorer davantage la complexité du paysage émotionnel. Cette approche théorique diversifiée offre une opportunité de mieux appréhender les nuances et les subtilités des émotions.

1.1.2. Consensus sur les composantes de l'émotion

Aujourd'hui l'émotion est représentée de façon consensuelle selon trois composantes : la composante subjective, la composante comportementale et la composante physiologique. De nombreux travaux ont ainsi affirmé que l'émotion pouvait être relevée par des traits comportementaux. La composante comportementale de l'émotion renvoie à toutes les manifestations comportementales et expressives d'une émotion. Apparaissent dans ces traits comportementaux la tonalité de la voix, la posture mais également les expressions faciales. Ces dernières ont été énormément étudiées étant donné que le visage représente un canal majeur de la communication émotionnelle.

Toute émotion est accompagnée de réactions physiologiques périphériques. Ces manifestations sont souvent distinguées en fonction de leur provenance. Ainsi, on différencie celles qui sont liées au système nerveux autonome, celles provenant de l'activité cérébrale et celles liées au système endocrinien. Lorsqu'un individu ressent une émotion, son activité neurophysiologique établit une fonction adaptative attribuée à l'émotion. Ceci permet à l'individu de réagir rapidement aux différents stimuli externes. Mais s'il est prouvé qu'un même individu réagit habituellement au même processus, alors il est également certain que deux personnes qui éprouvent une émotion similaire ne réagissent pas physiologiquement de la même façon. La composante physiologique de l'émotion réfère donc à l'ensemble des manifestations physiologiques liées à un événement émotionnel.

Nos émotions seraient le produit de notre subjectivité. Lorsqu'un individu ressent un sentiment, une émotion, il n'est pas toujours capable de la nommer mais peut cependant exprimer si ce qu'il ressent est plutôt agréable ou désagréable. La composante subjective étudiée par Jouvent et Carton (1994) et nommée « affect » par ces auteurs est décrite de la façon suivante : « la subjectivité d'un état psychique élémentaire, inanalysable, vague ou qualifié, pénible ou agréable, qui peut s'exprimer massivement ou sous forme d'une nuance, d'une tonalité » (Pasquier, 2021).

En intégrant ces trois composantes - subjective, comportementale et physiologique - une vision plus complète de ce que représente l'émotion est obtenue. Ces différentes dimensions interagissent et s'influencent mutuellement pour former notre expérience émotionnelle. Comprendre ces composantes est essentiel pour explorer les mécanismes sous-jacents aux émotions, et ainsi fournir une analyse satisfaisante des émotions afin de créer un algorithme capable de les détecter.

1.1.3. Différenciation des divers états affectifs

Que ce soit dans la littérature scientifique ou dans la vie courante, le terme d'*émotion* est parfois interchangé avec celui d'*affect*, de *sentiment*, d'*humeur* ou encore de *tempérament*. En France, le terme d'*émotion* est employé en psychologie et en neurosciences. En revanche, le terme d'*affect* est quant à lui employé presque exclusivement dans le champ de la psychanalyse. Selon toute époque, langue ou champ disciplinaire, le concept d'émotion ne renvoie pas à la même chose. Il existe une importante diversité de point de vue concernant le terme d'*émotion*. L'usage du terme de manière populaire servant couramment d'étiquette recouvrant tous les phénomènes affectifs, ajoute une difficulté définitoire. Toutefois, certains auteurs différencient les divers états affectifs tels que le tempérament, l'humeur ou encore les attitudes.

Scherer (comme décrit dans Petit, 2009) est un des auteurs dont les travaux ont tenté de différencier les divers états affectifs. Afin d'opérer cette distinction, Scherer à utiliser 7 dimensions : la rapidité de changement, l'impact comportemental, la durée, la focalisation sur l'événement, le produit de l'évaluation, la synchronisation de réponse et l'intensité. Pour chacun des types d'affect, l'auteur a repéré la force du trait, ce qui lui a permis de mettre en

relief quelles caractéristiques les différents affects partageaient-ils et leurs poids. Ainsi, l'auteur a distingué les préférences, les attitudes, les humeurs, les dispositions affectives, les postures interpersonnelles, ainsi que les émotions dites « esthétiques » et les émotions dites « utilitaires ».

Les attitudes sont décrites telles que des croyances durables ou des prédispositions vers des objets, personnes ou situation. Les attitudes produisent des états affectifs qui peuvent être décrits par les termes *désirer*, *détester* ou *évaluer*. Ainsi, on retrouve dans la catégorie des attitudes de Scherer les concepts de « sympathique », « affectueux », « haineux » ou encore « désireux ». Les humeurs sont définies par une prédominance persistante de différents sentiments influençant le comportement et par conséquent l'expérience d'un individu. Ainsi, une personne peut être d'une humeur gaie, sombre, allègre, dépressive ou encore indifférente. Les préférences sont caractérisées par Scherer comme « des manifestations affectives évaluatives relativement stables qui portent sur les objets, les situations ou les personnes, qui nous attirent ou que nous rejetons » (Petit, 2009). Le tempérament est une disposition affective qui accompagne généralement l'individu tout au long de sa vie. Dans cette classe Scherer propose d'intégrer l'anxiété, la morosité, la témérité et l'hostilité. Les postures interpersonnelles sont décrites par Scherer comme des styles affectifs que l'individu développe lorsqu'il est en interrelation avec une ou plusieurs personnes. Par exemple, lorsqu'une personne est en relation avec autrui, celui-ci peut adopter une posture interpersonnelle froide, distante, attentionnée, polie, méprisante ou chaleureuse.

Enfin, Scherer postule que l'émotion est multidimensionnelle et qu'elle est composée de 5 dimensions : une composante d'évaluation cognitive des stimuli et des situations, une composante physiologique (les réactions corporelles), une composante d'expression motrice, une composante motivationnelle (la tendance à l'action) et une composante de sentiment subjectif. Ainsi, il définit les émotions comme « des réponses adaptatives au monde, pas simplement des sensations abstraites, comme l'insinuent les théories dimensionnelles » (Ellsworth & Scherer, 2003).

Bernard Rimé différencie divers états affectifs : le tempérament, les troubles émotionnels, l'humeur, les affects et les émotions. Selon cet auteur, le tempérament « correspond aux traits affectifs stables des personnes comme le « névroticisme » (disposition à éprouver des affects négatifs tels l'anxiété, la tristesse, le ressentiment, le pessimisme, l'insatisfaction, etc.) ou l'« extraversion » (Piolat, 2009). Les troubles émotionnels sont de durée variable. On retrouve dans cette classe la dépression, la manie ou encore les troubles anxieux. Les humeurs correspondent selon Rimé, à des états affectifs ressentis comme positifs ou négatifs sans que l'individu ne sache ce qui les a déclenchés ou interrompus. Les humeurs peuvent s'étendre de quelques instants à plusieurs semaines voire plusieurs mois. Les affects sont des manifestations émotionnelles positives ou négatives. Ainsi, l'angoisse, l'espoir, l'exaltation, le découragement, l'excitation joyeuse et la morosité, sont identifiés par l'auteur comme des affects. Enfin, les émotions sont caractérisées par Rimé comme des épisodes frappants et d'une durée relativement courte. Les émotions sont reconnaissables chez l'individu à travers les expressions vocales, faciales, posturales et comportementales.

En conclusion, les diverses études présentées contribuent à clarifier les concepts et à distinguer les différents états affectifs, offrant ainsi une base solide pour une compréhension approfondie des émotions. Ces distinctions permettent de mieux appréhender les caractéristiques et les spécificités des attitudes, des humeurs, des préférences, des postures interpersonnelles et des émotions, tout en soulignant la multidimensionnalité des réponses émotionnelles. Cependant, compte tenu de la complexité inhérente à la distinction des différents états affectifs, cette recherche privilégiera l'acceptation de l'émotion comme un terme englobant tous les états affectifs, sans établir de distinctions spécifiques entre eux. Cette approche permettra de prendre en compte la diversité et l'interconnexion des phénomènes émotionnels, tout en évitant de se perdre dans des catégorisations trop rigides et potentiellement limitantes.

1.2. La poésie

La poésie est un genre littéraire très ancien. Son nom vient du grec *poiêsis*, mot dérivé du verbe *poien* signifiant « créer ». La poésie se présente donc comme un processus de création où le poète est acteur de ce qu'il construit. Cette construction mêle à la fois une recherche de création d'images dans l'imaginaire du lecteur et une recherche d'accord par un rythme et une forme particulière. En effet, ce genre littéraire est marqué par l'oralité et la musicalité apportées par les rimes ou les vers. En France, dès la deuxième moitié du XIXe siècle, la poésie s'est affranchie des règles particulières de rime, de métrique et de scansion pour introduire le vers libre.

Au sein de cette section, la littérature traitant de la place de la poésie en France sera présentée. Par la suite, la figure du poète sera abordée en explorant la dichotomie entre l'amateurisme et le professionnalisme. La perception négative associée au terme *amateur* sera également discutée. Enfin, une partie sera consacrée à l'étude de la relation entre les poètes et le Web.

1.2.1. La place de la poésie et du poète en France

En ce qui concerne la littérature traitant spécifiquement de la poésie en France, les thèses de Christian Vogels, Aude Mouaci, Sébastien Dubois ainsi que Adrien Cassina, constituent de réels apports dans ce domaine. Le travail réalisé par Dubois représente une véritable entreprise visant à établir un panorama de la poésie en France, grâce à un travail d'analyse de statistiques et d'enquêtes.

Nombreux travaux soulignent que la poésie en France est faiblement présente dans l'espace public mais également peu visible dans les médias. Baudelaire à son époque déjà, écrivait à propos de la poésie en France : « […] la France n'est pas poète ; elle éprouve même, pour tout dire, une horreur congéniale de la poésie. Parmi les écrivains qui se servent du vers, ceux qu'elle préférera toujours sont les plus prosaïques. […] Cela vient non-seulement, je crois, de ce que la France a été providentiellement créée pour la recherche du Vrai préférablement à celle du Beau […] » (Baudelaire, 1885). Toutefois, la poésie française a su, dans certains grands

moments de l'histoire française, positionner la figure du poète dans le débat public. Il est vrai, que la parole poétique a trouvé un terreau favorable lors des révolutions et des guerres en France. Citons notamment Hugo et Lamartine, figures emblématiques du poète-prophète dans les révolutions de 1830-1848. La Seconde Guerre mondiale ainsi que la résistance ont permis l'émergence d'une poésie engagée avec notamment Aragon, Char et Eluard.

Aujourd'hui la poésie est un genre littéraire peu rentable. Les théories et statistiques sur la consommation et les pratiques culturelles (Donnat, 2009) démontrent que les produits de haute valeur culturelle sont consommés par un faible nombre de personnes. Le sociologue Sébastien Dubois affirme dans son œuvre *The French Economy Poetry* que la poésie contemporaine française est absente de nombreux points de vente notamment dans les grandes surfaces non spécialisées. On retrouve ce genre littéraire que dans quelques librairies et grandes surfaces culturelles en particulier la FNAC. Selon le Ministère de la culture[2], en 2019 la poésie (ainsi que le théâtre) représentait 0,3% du chiffre d'affaires du marché du livre. Cependant, il existe des institutions indépendantes ou non, qui contribuent au développement et au maintien de la poésie en France. La Commission poésie du Centre National du Livre (CNL) attribue des subventions afin d'inciter les librairies à créer ou maintenir un rayon poésie. Ces librairies ont alors un label LiR. D'autres institutions non publiques soutiennent également la poésie. Le centre international de poésie Marseille, la Maison de la Poésie, ou encore le Printemps des poètes, accordent des subventions et publient des revues et recueils de poésie. Ils organisent également des lectures de poésies ainsi que d'autres événements.

La poésie a réussi à se développer tout en subissant une marginalisation sur le marché du livre. Ceci est le résultat de l'essor de réseaux spécialisés. Nombreuses littératures sociologiques présentent comment ces réseaux relationnels se sont développés afin que la poésie s'inscrive de manière autonome dans un espace socio-économique cohérent. Aujourd'hui, la poésie à ses propres lieux de diffusion, d'éditeurs et d'institutions. La production de la poésie est basée essentiellement sur les réseaux. L'enjeu principal d'un poète en France est d'être publié et par conséquent se construire une réputation. Les études de Dubois démontrent que 61% des poètes sont des éditeurs de poésie, des critiques ou même des professeurs d'université. De plus, 56,5% des éditeurs questionnés par Dubois déclarent que leurs choix éditoriaux sont faits par recommandation ou parce qu'ils ont déjà publié l'auteur. Ainsi, selon divers sociologues, la réputation serait le capital sur lequel repose une collaboration. En effet, Thierry Guichard (2004) métaphorise cette situation comme un « paysage d'archipel », c'est-à-dire qu'un réseau acquiert force et prestige en accueillant des auteurs respectés venant d'autres horizons. Ceci explique pourquoi les acteurs du marché de la poésie française empêchent les amateurs de s'adresser aux éditeurs et critiques. Ainsi, les poètes les plus connus sont ceux qui sont au centre d'un ou plusieurs grands réseaux qu'ils soient institutionnels ou littéraires. Le poète en France doit s'intégrer dans cet espace relationnel presque confidentiel de la poésie en France.

[2] Les données sont accessibles à l'adresse suivante : https://www.culture.gouv.fr/Thematiques/Livre-et-lecture/Actualites/Chiffres-cles-du-secteur-du-livre-2018-2019

Ainsi, ces paragraphes fournissent des informations essentielles sur la situation de la poésie en France, en mettant en évidence les défis contemporains auxquels ce genre fait face tels que les problématiques économiques et les dynamiques des réseaux dans ce domaine.

1.2.2. La figure du poète amateur

1.2.2.1. Amateur versus Professionnel

Plusieurs auteurs ont présenté les poètes amateurs d'un point de vue négatif. Nous venons de l'évoquer, être un poète n'est pas chose aisée, toutefois être un poète qualifié d'amateur le serait encore moins. Adrien Cassina indique : « A la lecture des recherches précédentes, le qualificatif d'amateur s'est présenté comme une ligne de partage, une frontière qui délimiterait entre eux les poètes. Amateur, amateurisme porteraient les germes d'une dépréciation voire d'une disqualification. L'amateur s'opposerait au professionnel. » (Cassina, 2022).

Cassina discute l'opposition poète amateur et poète professionnel. Dans sa thèse l'auteur envisage la dualité amateur versus professionnel en délestant la notion d'amateur des jugements négatifs qui semblent l'environner. Adrien Cassina ne veut pas classer deux sortes de poètes ni en fixer des contours saillants. L'auteur souligne néanmoins : « J'ai pris garde de ne pas succomber à une tentative relativiste où aucune différence n'existerait et où tout se vaudrait. Plus encore, j'ai prêté attention à cette opposition d'amateurs et de professionnels qui existe dans les propos des poètes que j'ai rencontrés. » (Cassina, 2022).

Dans sa thèse, Sébastien Dubois exclut les poètes amateurs. Il prête néanmoins attention aux termes d'*amateur* et de *professionnel* en considérant l'ambivalence du terme *professionnel*. En effet, *professionnel* peut renvoyer à la fois, à l'activité rémunératrice mais également peut permettre de qualifier les compétences individuelles du poète et l'insertion dans des réseaux de celui-ci.

Aude Mouaci quant à elle se focalise exclusivement sur les amateurs et autoédités. Suite à ses enquêtes et observations, elle indique que chercher à délimiter une frontière entre les termes d'amateur et de professionnel est une tâche relativement difficile. En effet, les poètes interrogés par Mouaci ont amplement rejeté le terme d'*amateurisme*. Ainsi, dans son ouvrage, la sociologue ne porte aucune forme de jugement sur la valeur des textes mais déplace simplement son regard sur les pratiques des poètes. Pour ce faire Robert A Stebbins apporte à l'auteur le cadre théorique des loisirs sérieux (serious leisure). A Stebbins distingue les loisirs occasionnels (casual leisure) des loisirs sérieux par une certaine qualité d'engagement. Cette qualité d'engagement qui compte notamment le sérieux, la persévérance, l'éthique ainsi que la sincérité donne à la pratique amateur une dimension épanouissante s'inscrivant dans une durée. Ainsi, ce cadre théorique permet à Mouaci de « penser la poésie comme une activité de loisir réalisée avec sérieux et implication par des hommes et des femmes poètes qui n'en tirent pas pour autant rémunération. » (Cassina, 2022).

En somme, ces études offrent différentes perspectives sur la distinction entre poètes amateurs et poètes professionnels. Ces auteurs mettent en évidence les préjugés associés à l'amateurisme, ou explorent l'ambivalence du terme professionnel, ou encore, abordent l'engagement et l'épanouissement liés à la pratique de la poésie en tant que loisir sérieux.

1.2.2.2. Amateur : un terme dépréciatif

L'autoédition selon Christian Vogels est considérée comme une pratique disqualifiante. L'auteur affirme que : « cette population d'écrivains [les poètes autoédités] n'a pas de références littéraires sur « ce qui se fait dans le milieu » littérairement parlant ni de références sociologiques, historiques, politiques, sur la configuration du milieu et sur « ce qui fait le milieu » (Cassina, 2022). Les propos du chercheur quant aux poètes autoédités, ne s'arrêtent pas là, il indique : « parmi ceux-là [les poètes autoédités], nous trouvons de tout : des auteurs débutants, des naïfs, des malades mentaux, des écrivains sans entregent, mais aussi des écrivains à qui leur, habituel (et puissant), éditeur propose cette solution [l'autoédition] » (Cassina, 2022). Enfin Vogels prétend que ces auteurs sont empreints à un vif désir de publication alors même qu'ils ne connaissent aucunement le milieu littéraire, de diffusion ainsi que le lecteur. Selon l'auteur les poètes autoédités sont « seuls, démunis de tout pouvoir, dans le champ littéraire » (Cassina, 2022).

Le sociologue Sébastien Dubois reconnait que les poètes amateurs sont omniprésents sur Internet. Leurs publications sont facilitées par l'essor de nouveaux outils tels que les plateformes de blogs. L'auteur nomme ces poètes amateurs publiant leurs œuvres quasi exclusivement sur le Web, les « poètes Internet », ceux-ci partageant, selon le chercheur des « […] représentations communes du poète avec les poètes amateurs papier, et pas celles de la poésie sérieuse » (Cassina, 2022).

Néanmoins, avec Hennion, Gomart et Maisonneuve, le concept d'amateur a été reconsidéré pour désigner une personne qui explore un goût par une activité. De plus, le concept de « Pro-Ams » développé par Leadbeater et Miller en 2004 propose d'envisager l'amateur non pas en opposition avec le professionnel, mais de placé ceux-ci sur un continuum. Ils indiquent qu'au cours des dernières décennies, un nouveau type d'amateur est apparu nommé « Pro-Ams » par les auteurs. Ces Pro-Ams seraient des amateurs qui travailleraient selon des normes professionnelles.

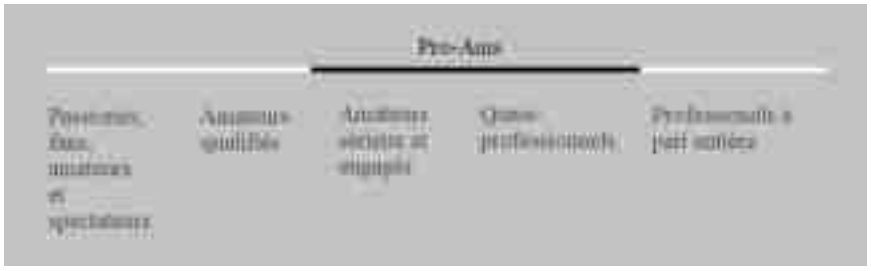

Figure 3 : Schéma du continuum traduit de Leadbeater et Miller (2004)
Source : Leadbeater & Miller 2004

La figure 3 illustre le schéma du continuum développé par Leadbeater et Miller. Les auteurs indiquent qu'au fur et à mesure que l'on se déplace sur ce continuum (de gauche à droite), la quantité de connaissance, de temps et d'investissement dans l'activité considérée augmente. Selon ces auteurs, les Pro-Ams seraient des personnes bien éduquées, engagées et mises en réseaux grâce aux nouvelles technologies.

Ainsi, le qualificatif amateur apparait pour certains comme un terme désignant négativement des personnes exerçant une activité et pour d'autres, un qualificatif qu'il convient d'employé pour désigner des personnes passionnées. Cassina évoque : « L'amateur n'est pas le contraire du professionnel. La médiocrité, l'ignorance ou encore la passivité ne font plus partie des qualificatifs pouvant le définir. » (Cassina, 2022).

Ces perspectives fournissent un aperçu complexe et nuancé de l'image du poète amateur. Au sein de cette étude, les termes d'amateurisme et d'amateur ne seront pas utilisés négativement mais serviront uniquement à référer aux personnes publiant leurs poèmes sur Internet.

1.2.3. Les poète et le Web

Internet a indéniablement permis l'essor des pratiques poétiques. Les auteurs amateurs de poésie, ou encore nommés « poètes Internet » par Sébastien Dubois, ont su exploiter les différents outils qui ont émergé sur le Web. Blogs, sites, en passant par Wattapad, X (anciennement Twitter) ou encore Instagram, nombreux poètes amateurs ont ainsi pu exprimer leurs paroles poétiques sur Internet.

En octobre 1998, soit avant la création de X, Instagram, etc., Aude Mouaci recensait 144 sites francophones consacrés en totalité ou en partie à la poésie. Aujourd'hui de nouveaux paysages et de pratiques en ligne ont vu le jour pour les poètes amateurs. La création de réseaux sociaux a notamment permis l'émergence de ces poètes publiant sur le Web. En 2004, Facebook est lancé mais ne sera qu'accessible en 2006 pour le grand public. S'en suit la création de Wattpad et X en 2006 et Instagram en 2010. Olivier Belin dans son article intitulé *Vers une poésie commune ? Les poètes amateurs de Twitter, Instagram et Wattpad*, nomme les poètes publiant sur le Web les : « twittopoètes » (pour ceux publiant sur X) et les « instapoètes » (ceux publiant sur Instagram). L'auteur indique que : « les twittopoètes et les instapoètes s'emparent des réseaux sociaux comme outils de médiation et de médiatisation » (Belin, 2020).

Les poètes amateurs sont très présents sur le réseau social X. Premièrement, ce réseau est très ouvert au partage en permettant la republication et l'utilisation de mots-clés (hashtags). Ainsi, les poètes qui le veulent peuvent se donner plus de visibilité en utilisant ce système de mots-clés. On peut retrouver notamment : #poesie, #poeme, #getpoed, #poesiefrançaise et bien d'autres. De plus, ce réseau social était caractérisé par une certaine contrainte de brièveté. En effet, les utilisateurs de X devaient respecter dans leurs publications un maximum de 280

caractères, qui avant 2017 était de 140 caractères[3]. Cette contrainte de brièveté du message a permis l'émergence de défi devenu des tendances avec notamment l'écriture de haïku, se prêtant facilement à l'exercice. Les haïkus sont des poèmes caractérisés par une forme très concise à savoir, dix-sept syllabes réparties en trois vers. On remarque également que les poèmes publiés sur X s'accompagnent généralement d'images et de photographies permettant d'illustrer les vers ou la prose de ces poètes.

À la différence de X, Instagram présente non pas des messages contenant des poèmes mais précisément des images de poèmes. En effet, sur Instagram l'image du poème et par conséquent celle du poète également, est différemment travaillée. Les poètes d'Instagram introduisent une nouvelle esthétique du poème en ligne en reproduisant une forme typographique ou un ductus manuscrit telle que ci-dessous.

Figure 4 : Poème du compte Instagram @lesouffledekaruno
Source : compte Instagram @lesouffledekaruno

Instagram utilise le même fonctionnement de mots-clés, permettant aux publications d'être davantage visibles par les utilisateurs. On retrouve des mots-clés tels que : #instapoetry #poetry, #instapoet, #poet et #poetress.

Wattpad créé en 2009, est un réseau social permettant aux utilisateurs inscrits d'éditer et de partager des poèmes, récits, fanfictions ou encore romans, en les rendant accessibles en ligne. Wattpad est également disponible en application mobile, rendant l'accès beaucoup plus facile pour les utilisateurs. C'est également la possibilité d'enregistrer les œuvres pour une lecture hors connexion qui rend ce réseau très intéressant pour les utilisateurs. Wattpad connait un grand succès auprès des écrivains amateurs. Le réseau dispose d'une section poésie où les poètes amateurs peuvent ainsi publier leurs productions poétiques.

[3] Depuis octobre 2023 les utilisateurs de la plateforme ont une limite de 25 000 caractères

Ainsi, les poètes amateurs ont su s'emparer des réseaux et du Web pour partager leurs paroles poétiques au plus grand nombre. Ces différents réseaux permettent un échange entre auteurs et lecteurs. Les systèmes présents sur ces réseaux sociaux tels que les likes, les votes, les partages, les commentaires ou encore les messages privés permettent aux poètes d'ajuster leur style, corriger leurs formes orthographiques et d'améliorer leurs compétences poétiques. Sur Instagram de nombreux poètes à présent connus du grand public ont émergé. Rupi Kaur est une poétesse dont ses débuts se sont passés sur la plateforme Instagram. C'est elle qui a donné l'impulsion à un grand nombre de poètes amateurs de publier sur Instagram. Cette auteur écrit des poèmes courts, incisifs, sur des thématiques fortes et essentiellement féminines, voire féministes. La particularité des poètes publiant sur Instagram comme évoqué précédemment est qu'ils ont souvent une approche esthétique de leur écriture, celle-ci accompagnée d'illustrations ou de photographies. Là encore, c'est Rupi Kaur qui a été la précurseur de cette tendance. Les poètes le savent, les chances d'être lu augmentent si le poème est illustré. Cependant, Rupi Kaur n'a pas échappé à la case autoédition, passant par Amazon pour son recueil Milk and Honey. Quelques mois plus tard, l'auteur signe un contrat avec une maison d'édition. Depuis, plus de 3 millions d'exemplaires se sont écoulés et son recueil a été traduit en 40 langues différentes, dont le français, publié aux éditions Charleston.

X a également permis de faire émerger l'auteur américaine Melissa Broder avec son compte @sosadtoday crée en 2012. Sous forme d'essais poétiques elle y aborde des sujets sensibles tels que la dépression et la santé mentale ou encore la mort. En 2016, l'auteur publie un recueil de chroniques, *So Sad Today*, intitulé d'après son compte X. Elle a également publié divers recueils de poésie comme *Last Sext* (2016) ou encore *Superdoom : Selected Poems* (2021). En 2017, Melissa Broder remporte le prix Pushcart pour son poème « Forgotten Sound », du recueil *Last Sext* (2016).

D'autres plateformes sur le Web ont permis à des auteurs de se faire reconnaitre dans le monde du livre et de l'édition. C'est notamment le cas pour Leav Lang, poète canadienne qui s'est fait connaître en publiant des poèmes d'amour sur la plateforme Tumblr. La maison d'édition Andrews McMeel Publishing publie son recueil *Love & Misadventure*, déjà autopublié précédemment. En un mois, son œuvre s'est vendue à plus de 10 000 exemplaires. Depuis, elle continue de publier des recueils de poésie et des romans.

Ainsi, les plateformes du Web ont permis l'émergence de poètes, aujourd'hui professionnels ne passant plus par l'autoédition. Internet est devenu un outil efficace pour faire reconnaitre son travail et de se construire une réputation dans le monde de la poésie.

En résumé, cette section présente comment Internet et les réseaux sociaux ont ouvert de nouvelles opportunités aux poètes amateurs, leur permettant de partager leurs œuvres, de recevoir des retours et de se faire reconnaître dans le monde de la poésie. Cette partie souligne également l'importance des plateformes en ligne telles que X, Instagram et Wattpad dans la création d'une communauté poétique et la diffusion des paroles poétiques au plus grand nombre. Dans cette recherche les poèmes amateurs seront collectés sur différents sites internet spécialisés dans la publication de poèmes amateurs.

1.3. Le Traitement Automatique des Langues

Le traitement automatique des langues est un domaine de recherche à l'intersection de la linguistique, de l'informatique et de l'intelligence artificielle. Le traitement automatique des langues vise essentiellement à analyser des langues au moyen d'un ordinateur. Le domaine des technologies des langues est devenu un domaine-clé permettant de résoudre des problèmes et répondre aux besoins actuels de notre société. Concernant la poésie, peu d'études ont été réalisées afin de définir quelles sont les variables associées au talent poétique. Les seules études se fondant sur la linguistique computationnelle sont en langue anglaise. La première étude à utiliser la linguistique computationnelle sur un corpus poétique a été celle de Forsyth en 2000. Une autre étude menée par Justine Kao et Dan Jurafsky en 2012 s'est intéressée elle aussi à isoler les caractéristiques significatives des poèmes amateurs et des poèmes professionnels. Enfin, Dalvean à tenter de reprendre ces deux études et de développer des classificateurs afin de pouvoir les placer sur un spectre. Ainsi, le talent poétique peut être identifié par des variables textuelles. Cependant, la détection des émotions par le TAL peut également contribuer à l'identification du talent poétique. L'identification des émotions en traitement automatique des langues repose sur différentes approches. L'approche automatique s'appuie sur des techniques d'apprentissage automatique à partir de données. L'approche à base de règles vise à créer des systèmes qui effectuent une analyse d'émotion basée sur des règles prédéfinies. Enfin, il existe l'approche hybride qui quant à elle combine les approches basées sur des règles et des approches automatiques.

Dans cette section, seront présentées les différentes études réalisées dans le domaine du traitement automatique des langues, se concentrant d'abord sur l'analyse de la poésie, puis sur les émotions.

1.3.1. La poésie en Traitement Automatique des Langues

1.3.1.1. Approches statistiques

La première étude à utiliser la linguistique computationnelle afin de distinguer la poésie amateur de la poésie professionnelle a été celle de Forsyth en 2000 en langue anglaise. Dans cette étude la fréquence d'apparition dans des anthologies a été utilisée, suivant ainsi les traces de Simonton (1989), Matindale (1990) et d'autres. Pour créer son échantillon de référence, Forsyth a sélectionné vingt anthologies publiées entre 1966 et 1997 et a pris les poèmes qui apparaissaient dans plus de cinq d'entre elles. Son échantillon de référence a été constitué de 85 poèmes de 54 auteurs différents. Concernant le groupe de contrôle, 54 poètes ont été sélectionnés en prenant le soin de trouver un poète témoin moins éminent de même sexe et né dans les dix ans suivant chaque poète de l'échantillon de référence. Cela a donné 85 poèmes « obscurs » écrits également par 54 poètes différents. Afin de vérifier si la différence était significative, le chercheur a utilisé le *Little Oxford Dictionnary of Quotations* (Ratcliffe, 1990) afin de compter le nombre de citations de ces auteurs. Grâce au test de Wilcoxon-Mann-Whitney, Forsyth a pu conclure que cette différence était hautement significative. Certains

aspects de la langue des deux sous-ensembles ont ensuite été analysés. Malgré le fait que les poèmes populaires soient en moyenne plus longs que les poèmes obscurs, avec une longueur médiane respective de 155 et 127 mots, cette différence n'a pas été retenue. En effet, grâce au test *U* de Mann-Whitney, le chercheur en a conclu que cette différence n'était pas statistiquement significative. À la suite de différents tests statistiques, Forsyth a relevé un certain nombre de différences significatives entre les deux groupes de poèmes. Premièrement, l'utilisation du test non paramétrique de Wilcoxon-Mann-Whitney a permis de montrer que les poèmes populaires avaient significativement moins de syllabes par mot dans leurs premières lignes. Deuxièmement, le chercheur a démontré grâce au test du khi-deux que les poèmes populaires étaient plus susceptibles de commencer par une ligne initiale composée entièrement de monosyllabes. Une autre différence entre ces deux groupes a été le nombre moyen de lettres par mot dans les poèmes. Le test de Student (ou test-t) non apparié a permis cette différenciation. Le nombre moyen de lettres par mot étant significativement inférieur pour les poèmes populaires que pour les poèmes obscurs. L'étude a également montré que le vocabulaire des poèmes populaires était en moyenne moins riche que celui des poèmes obscurs. Enfin, Forsyth a étudié les différences syntaxiques et en a conclu que les poètes obscurs avaient tendance à utiliser plus de noms singuliers et de verbes au présent que les poètes populaires. Une autre différence syntaxique révélée est que les poèmes populaires présentaient un taux plus élevé de conjonctions de coordination et de pronoms personnels que les poèmes obscurs.

Kao et Jurafsky en 2012 (Kao & Jurafsky, 2012 ; Kao, 2012) analysent en passant par la construction d'un modèle de régression logistique un corpus de poèmes amateurs et de poèmes professionnels. Ils ont pour but de distinguer quelles sont les caractéristiques qui rendent un poème plus attrayant esthétiquement qu'un autre. Leur corpus est composé de 200 poèmes dont 100 poèmes amateurs et 100 poèmes professionnels écrits par 67 poètes différents pour chaque groupe. Les 100 poèmes écrits par des professionnels sont sélectionnés selon le paramètre d'une publication dans une collection de poésie américaine contemporaine, appuyant leur décision sur le fait que les noms de ces poètes se retrouvent présents sur le site de l'Academy of Americain Poets et qu'ils ont, pour la plupart, reçu des prix prestigieux. Cela confirme que ces poètes sont célèbres et reconnus pour leur art. Les chercheurs ont ainsi sélectionné 1 à 3 poèmes de chaque poète proportionnellement au nombre de poèmes que le poète avait dans sa collection. Concernant la longueur des poèmes, un contrôlé a été effectué en supprimant les poèmes qui excédaient 500 mots et en ajoutant un autre poème du même auteur. Ainsi, la longueur des poèmes de cette section variait entre 33 et 371 mots avec une longueur moyenne de 175 mots. La partie réservée de leur corpus aux poèmes amateurs a été construite par le choix aléatoire de 100 poèmes sur le site Amateur Writing[4]. Kao et Jurfasky ont procédé à une correction des erreurs grammaticales et des fautes orthographiques afin de contrôler l'effet des compétences de base du poète. Leur sélection finale de poèmes amateurs comptait entre 21 et 348 mots par poème avec une longueur moyenne de 136 mots.

Différentes caractéristiques ont été calculées à partir de PoetryAnalyser développé par Kaplan en 2006 ou encore du système General Inquirer. Le General Inquirer est un projet né dans les

[4] Le site est disponible à l'adresse suivante : www.amateurwriting.com

années 1960 par Stone et al. en 1966 visant à développer un programme d'analyse objective de données basé sur un dictionnaire composé de 184 catégories sémantiques. Les chercheurs ont sélectionné 16 variables ciblant un des trois domaines à savoir, le style, le sentiment et l'imagerie. Le tableau ci-dessous présente les différentes variables retenues.

Tableau 2 : Variables retenues par Koa et Jurafsky (traduction et adaptation)
Source : Kao & Jurafsky, 2012 p.11

Variables	**Exemples**
Fréquence de mot	–
Ratio type/token	–
Rime parfaite	lourd / velours
Rime oblique	balance / panier
Allitération	serpent qui sifflent
Consonnance	étrange et pénétrant
Assonance	belle rêve
Perspective positive	capable ; ami
Perspective négative	abandon ; ennemi
Emotion positive	bonheur ; amour
Emotion négative	fureur ; chagrin
Bien-être physique	vivant ; manger
Bien-être psychique	calme ; paisible
Mot concret	bateau ; feuille
Mot abstrait	jour ; amour
Généralisation	aucun ; tous

Pour mesurer l'effet de chaque variable sur la probabilité qu'un poème soit écrit par un professionnel ou un amateur, les chercheurs ont construit un modèle de régression logistique sous R. Concernant la sélection du modèle, la méthode descendante a été choisie. Cette méthode commence par construire un modèle en utilisant les 16 variables. Ensuite, le modèle élimine de façon récursive les variables qui ne sont pas significatives dans la variance de données. La méthode de sélection s'arrête dès lors que l'élimination supplémentaire d'une variable entraînerait une perte significative d'information et d'ajustement du modèle. Ainsi, le type de poème (amateur ou professionnel) est prédit selon 8 variables différentes à savoir : le ratio type/token, la fréquence de rimes parfaites, la fréquence d'allitération, les mots de perspectives positives, les mots d'émotions négatifs, les mots concrets, les mots abstraits et les mots exprimant une généralisation.

Les résultats de cette recherche permettent d'indiquer que les poètes amateurs sont significativement plus susceptibles de faire référence à des émotions négatives que les poètes professionnels. De plus, les poèmes professionnels contiennent davantage de mots concrets et moins de rimes parfaites que les poèmes amateurs.

Ces études ont adopté des approches statistiques pour analyser les différences entre la poésie amateur et professionnelle. Leurs résultats offrent une vision des méthodologies potentielles à suivre pour cette recherche. En s'appuyant sur leurs approches méthodologiques, cette étude peut élaborer une méthode similaire pour la distinction des poèmes amateurs et des poèmes professionnels en se concentrant sur les émotions.

1.3.1.2. Méthodes d'apprentissage

Michael Dalvean en 2016 a poursuivi les études de Forsyth. En reprenant le corpus de Forsyth et en utilisant des méthodes d'apprentissage basées sur la régression logistique, il a développé un classifieur. Pour ce faire, le chercheur a utilisé le Linguistic Inquiry and Word Count (LIWC) afin de sélectionner 65 variables linguistiques. Le LIWC est un programme d'analyse créé par Pennebaker et al. (2007). Ce logiciel permet de compter les mots d'un texte en langue anglaise en les attribuant à des catégories qui ont un sens psychologique. Il offre une analyse en pourcentage de plus de 80 dimensions du langage comme les mots fonctionnels, les thèmes, les termes dénotant des émotions positives ou négatives par exemple. Les données ainsi obtenues peuvent être facilement récupérées et transférées dans des fichiers adaptés à une exploitation statistique. Le but de ce classifieur étant de déterminer ce qui distingue des poèmes qui sont hautement représentés dans des anthologies de ceux qui n'y figurent pas ou presque pas.

Les variables indépendantes sélectionnées par Dalvean ont été des variables dérivées de LIWC. En plus de ces variables, Dalvean a sélectionné la variable dépendante de l'analyse, qui était une variable binaire prenant la forme de 1 pour les poèmes fréquemment présents dans les anthologies et de 0 pour les poèmes peu fréquemment présents dans les anthologies. En ce qui concerne l'apprentissage automatique basé sur la régression logistique, Dalvean a opté pour la méthode de validation croisée à 10 plis. En utilisant pas à pas la régression logistique, l'algorithme a procédé à une sélection itérative en ajoutant ou soustrayant une ou plusieurs variables au modèle afin de trouver le meilleur ajustement. Le meilleur modèle dans une partie donnée étant celui obtenant la plus grande précision de classification. Ainsi, les meilleures variables identifiées à travers ces 10 parties lui ont permis de constituer son modèle final. Le classifieur de Dalvean a permis d'identifier 6 variables importantes dans la classification à savoir, les mots exprimant l'inclusion tels que « et » ou encore « avec », les mots fonctionnels, les mots exprimant une causalité, les mots exprimant un loisir, les mots de tristesse et les mots d'émotion positive. A noter que les deux dernières variables étaient les seules à ne pas présenter des coefficients positifs. Ce résultat coïncide avec l'observation antérieure de Kao et Jurafsky (Kao & Jurafsky, 2012 ; Kao, 2012) et Dalvean (Dalvean, 2015) selon laquelle les poètes professionnels ne sont pas enclins à utiliser des termes émotionnels, mais préfèrent plutôt décrire des situations qui éveillent des réponses émotionnelles chez le lecteur. Ainsi, le classifieur développé par Dalvean a obtenu une exactitude de 69%.

Dalvean en 2015 a tenté de présenter ce qui distinguait un poème bien écrit d'un poème moins bien écrit. Pour ce faire, le chercheur a réutilisé les données de Kao et Jurafsky. Dalvean dans cette étude va élargir l'analyse faite par ces deux auteurs de deux façons. Premièrement, il va

utiliser un nombre de variables supérieures. En effet, 68 variables linguistiques tirées du LIWC vont être utilisées. De plus, Dalvean ajoute 32 variables psycholinguistiques provenant des normes de mots (PYMC) développées par Paivio Yuille et Madison (1968) et leurs extensions par Clarke et Paivio (2004). La seconde nouveauté de son étude réside dans la construction d'un classifieur par apprentissage automatique. Ainsi, cette méthode permet la représentation d'un classement de poème sur un spectre plutôt amateur ou plutôt professionnel. Dalvean explique dans son œuvre *Ranking Contemporary Americain Poem* qu'il n'a pas voulu utiliser l'équation logistique de l'analyse de Kao et Jurafsky. Il justifie ce choix par le fait qu'il soit possible que l'équation soit sur-ajustée. Ainsi, le modèle peut simplement apprendre le bruit de l'échantillon et donc ne pas être vraiment généralisable. Si ce modèle a en effet un problème de surajustement alors il devient difficile de l'utiliser afin de classer des poèmes amateurs et des poèmes de professionnels.

Le processus de modélisation de Dalvean s'est déroulé comme suit : division de l'échantillon en un échantillon d'apprentissage (n=100) et en un échantillon de test (n=100). Ainsi, 50 poèmes ont été choisis aléatoirement parmi les poèmes amateurs et également pour les poèmes professionnels. Ensuite, l'étape suivante du processus a consisté à la procédure pas à pas en utilisant les 100 variables sélectionnées. Le chercheur a ainsi construit 5 modèles, les deux premiers construits sur les variables linguistiques du LIWC et les trois derniers sur les variables PYMC. L'auteur a ensuite créé un modèle d'ensemble en combinant les cinq modèles individuels. Ce modèle d'ensemble s'est révélé être le plus précis. En intégrant les résultats des cinq modèles, Dalvean a obtenu un algorithme capable de classer les poèmes comme professionnels ou amateurs avec une exactitude globale de 80 %.

Les études menées par Michal Dalvean offrent une idée de la méthodologie à suivre dans cette recherche, en particulier en ce qui concerne l'élaboration d'un classificateur permettant de distinguer les poèmes amateurs des poèmes professionnels. En utilisant des méthodes d'apprentissage basées sur la régression logistique et en se basant sur des variables linguistiques et psycholinguistiques, Dalvean a développé des modèles capables de classifier avec précision ces deux types de poèmes. Cette approche méthodologique constitue une piste prometteuse pour la présente recherche, offrant une base solide pour l'analyse et la classification des poèmes selon leur niveau de professionnalisme.

1.3.2. Le traitement automatique des émotions

1.3.2.1. Apprentissage automatique

L'apprentissage automatique est un ensemble de techniques de l'intelligence artificielle. L'objectif est de concevoir un algorithme capable d'absorber différentes caractéristiques à partir d'un grand nombre de données. Le modèle ainsi obtenu permet d'appliquer directement des connaissances acquises lors de la phase d'apprentissage afin d'étiqueter un nouveau contenu donné. Les étiquettes peuvent être de deux types, les étiquettes quantitatives servant à répondre à des problèmes de régression et des étiquettes qualitatives permettant de résoudre des

problèmes de classification. Il existe différents types d'apprentissage, à savoir : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage semi-supervisé. L'apprentissage supervisé consiste à produire un modèle basé sur des bases de données préalablement étiquetées. Ce type d'apprentissage permet de réaliser une analyse statistique afin de prédire des résultats sur un ensemble de données fournies en entrée. Plusieurs techniques sont employées pour ce type d'apprentissage.

La classification bayésienne est une méthode de classification statistique. Celle-ci se base sur le théorème de Bayes et est dite naïve à cause de sa forte indépendance. Elle permet de prédire la valeur des paramètres par des probabilités (Lin, 2002). Le support Vector Machines (SVM) est une classe d'algorithmes d'apprentissage permettant la prédiction d'une variable qualitative binaire ou de variable quantitative pour la classification multiclasses (Evgeniou & Pontil, 2001). D'autres algorithmes d'apprentissage existent tels que les réseaux neuronaux. Ceux-ci sont généralement organisés en diverses couches contenant des neurones qui sont reliés. Une autre technique employée est les forêts d'arbres décisionnels (Random Forest) qui constituent un ensemble de classificateurs modélisant les résultats sous formes de branches pour produire des arbres de décision. Le processus se base sur une sélection aléatoire d'un sous-ensemble d'échantillons et des variables d'apprentissage (Breiman, 2001).

Les recherches informatiques en sémantique lexicale se sont développées grâce à l'essor des techniques de plongements lexicaux. Le but de ces représentations, souvent appelées word embeddings en anglais, est de capturer la sémantique lexicale sous forme de vecteurs numériques pour mesurer la similarité des mots. Les représentations vectorielles pour la sémantique existent depuis longtemps. Les modèles plus anciens, appelés modèles vectoriels, utilisent des techniques telles que Term Frequency (TF) ou Term Frequency-Inverse Document Frequency (TF-IDF), expliquées ci-dessous. En revanche, les plongements lexicaux font généralement référence aux nouveaux modèles basés sur des réseaux de neurones et des modèles probabilistes.

Parmi les différentes méthodes, on retrouve la méthode de représentation par sac de mots (Bag of Words). C'est une méthode qui représente un corpus sous forme de vecteurs et qui ignore l'ordre des mots dans un document. Celle-ci permet de représenter un texte par ses mots isolés. Ainsi, cette méthode compte les occurrences de chaque terme présent dans un document et permet de comparer la similarité de ce document avec d'autres documents (Jurafsky & Martin, 2009, p. 653).

D'autres méthodes existent telles que les méthodes de pondération comme le TF et le TF-IDF. Le TF permet de mesurer la fréquence d'apparition d'un terme dans un document. Afin d'éviter un effet de taille, il est souvent admis de diviser la fréquence du terme par la longueur totale du document (Bouillot et al., 2014). Une autre technique est celle du TF-IDF qui est souvent utilisée avec les modèles de sac de mots. Cette technique consiste à donner à chaque mot d'un document un poids pour mesurer non pas sa fréquence d'apparition dans un document mais sa pertinence, définie par sa capacité à distinguer le document par rapport au reste du corpus (Bafna et al., 2016). Le nombre total d'occurrences d'un terme dans un document sera donc

remplacé par un score. Cette technique permet la prise en compte des mots qui apparaissent relativement souvent dans un document. Ainsi, par exemple les déterminants et les pronoms auront moins de poids et seront par conséquent, moins pertinents pour différencier des documents.

Enfin, il existe des méthodes qui apprennent des représentations sur la base de tâches auto-supervisées, telles que la prédiction d'un mot en contexte. Word2Vec est un ensemble de modèles de réseaux neuronaux conçus pour créer des représentations distribuées des mots. Ces modèles, développés par Tomas Mikolov et une équipe de recherche chez Google (Mikolov et al., 2013), permettent de représenter les mots sous forme de vecteurs multidimensionnels. Word2Vec comprend deux modèles principaux : le modèle Continuous Bag of Words (CBOW) et le modèle Skip-Gram. Le modèle CBOW prédit un mot cible en se basant sur son contexte environnant, tandis que le modèle Skip-Gram prédit les mots de contexte à partir d'un mot cible. Ces structures sont illustrées à la figure 4.

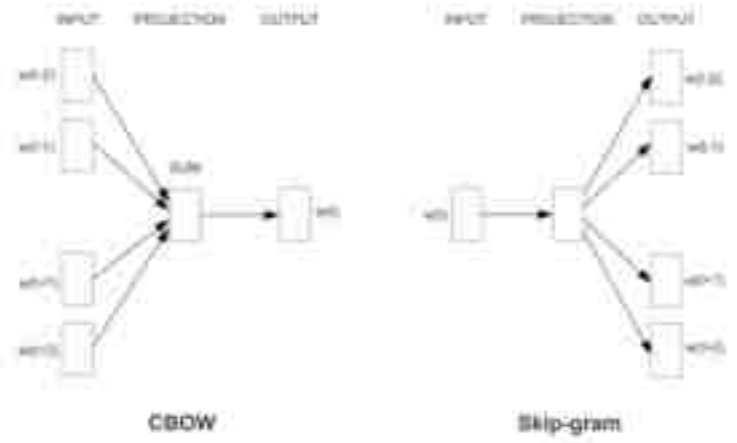

Figure 5 : Structures du modèle CBOW et du modèle Skip-Gram
Source : Ferdi, 2021

Les différentes méthodes d'apprentissage présentées dans cette section offrent une perspective sur les approches envisageables pour la conduite de cette recherche.

1.3.2.2. Les lexiques émotionnels

Plusieurs lexiques de sentiments, d'émotions et d'opinions existent. En ce qui concerne la langue anglaise, celle-ci a vu naître des lexiques émotionnels développés par le Conseil national de recherches Canada (CNRC). Le lexique NRC Emotion Lexicon[5] est un lexique contenant plus de 14 000 entrées créées par Saif M. Mohammad et Peter Turney (Mohammad & Turney, 2013). Ce lexique contient un mot par ligne suivi d'une étiquette d'émotion ou sentiment puis d'un score 0 ou 1. Les étiquettes présentes sont pour le sentiment : négatif ou positif, et pour

[5] Le lexique est récupérable à l'adresse suivante : http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

les émotions : la colère, l'anticipation, le dégoût, la joie, la peur, la tristesse, la surprise ainsi que la confiance. Le score d'association de 0 ou 1 indique si le terme est porteur ou non de l'émotion. Depuis août 2022 un lexique pour la langue française est disponible lors du téléchargement. Ce lexique a été traduit automatiquement par Google Translate.

Un autre lexique développé par Saif M. Mohammad est le lexique NRC Valence, Arousal, and Dominance[6] (VAD) (Mohammad, 2018). Il contient plus de 20 000 mots anglais, chacun associé à un score de valence, d'excitation et de dominance. Pour un terme donné et une dimension spécifique (V, A, ou D), les scores vont de 0 (le plus faible) à 1 (le plus élevé). La valence mesure la positivité ou la négativité du terme, l'excitation évalue le niveau d'activation émotionnelle, et la dominance reflète le degré de contrôle exercé sur un stimulus. Ce lexique est également disponible en français, traduit via Google Translate lors du téléchargement.

Enfin, il existe également le lexique NRC Emotion Intensity Lexicon[7] (NRC-EIL) également développé par Saif Mohammad (Mohammad, 2022). Ce lexique contient près de 10 000 entrées associées à des scores d'intensité pour huit émotions de base identifiées par Plutchilk. Les huit émotions sont : la colère, l'anticipation, le dégout, la joie, la peur, la tristesse, la surprise et la confiance. Les scores vont de 0 à 1 où 0 signifie que le mot transmet une faible quantité d'émotions et où 1 indique que le mot transmet la plus grande quantité d'émotion. Ce lexique est également disponible en langue française lors du téléchargement. La traduction a également été faite par Google Translate.

En France, le projet Emotaix a permis de développer un lexique émotionnel[8] (Piolat & Bannour, 2009). Ce projet est né d'une nécessité à disposer d'un outil français tel que le LIWC pour l'analyse de texte en français. Emotaix est un dictionnaire organisé comportant 2 014 référents. Ce dictionnaire est piloté sous le logiciel Tropes[9] et permet l'identification, la catégorisation ainsi que la comptabilisation automatique du lexique émotionnel dans des corpus oraux ou écrits. Les chercheurs ont choisi 3 dimensions pour le décompte de leur lexique à savoir : la valence (positive ou négative), l'usage (le sens propre ou le sens figuré) ainsi que la nature (les types de contenu regroupé par catégories sémantiques). Les référents présents dans Emotaix sont organisés en 2 x 28 catégories sémantique de base opposées par une valence positive et valence négative comme Dégout versus Désir, par exemple. Ont été ajoutées 3 catégories de base dont la valence n'est pas spécifiée. Ces 28 catégories sont regroupées en 9 catégories qui elles-mêmes sont rassemblées en 3 supra-catégories : Malveillance, Mal-être, Anxiété *versus* Bienveillance, Bien-être, Sang-froid. La figure 6 illustre l'organisation du lexique Emotaix.

[6] Le lexique est récupérable à l'adresse suivante : http://saifmohammad.com/WebPages/nrc-vad.html
[7] Le lexique est récupérable à l'adresse suivante : http://saifmohammad.com/WebPages/AffectIntensity.htm
[8] Le lexique Emotaix est récupérable à l'adresse suivante : https://centrepsycle-amu.fr/outils-recherche/
[9] Le logiciel est disponible à l'adresse suivante : https://www.tropes.fr/

Figure 6 : Roue de l'organisation du lexique Emotaix
Source : Centre PsyClé

Le lexique FEEL[10] (Abdaoui et al., 2017) est un lexique français créé par le laboratoire d'informatique de robotique et de microélectronique de Montpellier (LIRMM). Ce lexique contient plus de 14 000 mots exprimant des émotions ou sentiments. Il a été développé en suivant les deux polarités (négative et positive) et les six émotions de base identifiées par Paul Ekman à savoir : la joie, la tristesse, la surprise, le dégout, la peur, le dégoût et la colère. Il a été créé en élargissant et en traduisant automatiquement le lexique émotionnel NRC Emotion Lexicon (Mohammad & Turney, 2013). Un processus de validation manuelle par une traductrice professionnelle humaine a cependant été mis en place.

Toujours en France, il existe le lexique LIDILEM (Grutschus et al., 2014) incorporé dans la plateforme Emobase[11] créé dans le cadre du projet franco-allemand ANR-DFG EMOLEX (2009-2013). Ce lexique contient des noms, adjectifs et verbes regroupés selon 9 champs sémantiques de l'émotion : le respect, la joie, l'anticipation, la colère, la déception, la jalousie, le mépris, la tristesse ainsi que la surprise. Les termes contenus dans le lexique sont analysés selon diverses dimensions telles que l'intensité, l'aspect, la causalité, la manifestation, le contrôle, la verbalisation, la polarité, et l'expérienciation.

Enfin, la ressource lexicale Polarimot[12] (Gala & Brun, 2012) contient des informations sur la polarité intrinsèque du lexique en français. Ce lexique a été construit en partie, semi-automatiquement, à partir des ressources de Polymots (Gala & Rey, 2008). Il contient près de

[10] Le lexique est disponible à l'adresse suivante : http://advanse.lirmm.fr/feel.php
[11] La plateforme est disponible à l'adresse suivante : http://phraseotext.univ-grenoble-alpes.fr/emoBase/
[12] Cette ressource lexicale est disponible à l'adresse suivante : https://polarimots.lis-lab.fr/

7 500 mots associés à leur polarité. Pour chaque polarité est associé un poids correspondant au taux de fiabilité des annotations.

Ainsi, les différents lexiques émotionnels présentés ci-dessus constituent des ressources précieuses pour la détection des émotions dans le cadre de cette recherche. Ceux-ci pourront être développés et amplifiés dans cette recherche.

Objectif de cette recherche

Pour répondre à la question de recherche concernant la création d'un classifieur capable de distinguer les poèmes amateurs des poèmes professionnels en fonction des émotions qu'ils véhiculent, cette section propose un résumé de l'état de l'art. Elle aborde d'abord les définitions et les interprétations des termes clés de l'étude, à savoir *amateur*, *professionnel* et *émotion*. Ensuite, elle résume brièvement la méthodologie employée pour développer un classificateur automatique, conçu pour différencier les poèmes amateurs des poèmes professionnels en analysant les mots émotionnels qu'ils contiennent.

A la lecture des différentes recherches menées en psychologie et psychanalyse le terme d'émotion semble encore être en débat, comme détaillé dans 1.1.3. Ainsi, la recherche suivante appliquera l'acception suivante pour le terme d'*émotion* : « états émotionnels », incluant alors les humeurs, les sentiments, le tempérament ou encore les attitudes. Le terme d'*amateur* semble quant à lui, porter les germes d'une dépréciation. Lors de cette recherche le terme d'amateur sera employé pour qualifier les poèmes écrits par des hommes et des femmes les publiant sur Internet. Le terme *professionnel* sera quant à lui appliqué pour exprimer les poèmes écrits par des auteurs reconnus, ayant eu un prix dans leur discipline tel que le prix Robert Ganzo de Poésie ou encore le prix Mallarmé.

En ce qui concerne la méthodologie adoptée pour ce projet, plusieurs étapes ont été définies et seront détaillées dans les sections prochaines. La première étape consiste en la création d'un corpus composé de poèmes amateurs et de poèmes professionnels (section 2.1). Des lexiques de mots émotionnels préexistants seront adaptés et utilisés pour l'analyse (voir section 2.2). Des analyses statistiques seront menées sur le corpus afin de rendre compte des données (section 2.3). Par la suite, des classificateurs automatiques disponibles dans la bibliothèque Python Scikit-learn seront entraînés et testés sur ces données (sections 2.4.1, 2.4.2, et 2.4.3). Une recherche d'hyperparamètres sera effectuée, et une visualisation des caractéristiques les plus influentes pour les modèles les plus performants sera réalisée (sections 2.4.4 à 2.4.6).

2. Création d'un outil capable de distinguer poèmes amateurs et poèmes professionnels

Cette section décrit les étapes pour développer un classifieur automatique distinguant les poèmes amateurs des poèmes professionnels en fonction des émotions exprimées. Un corpus équilibré de poèmes francophones a été constitué, et des lexiques émotionnels ont été sélectionnés pour construire un dictionnaire émotionnel détaillé, intégrant plusieurs variables pertinentes. Des analyses statistiques et descriptives approfondies du corpus ont été réalisées pour explorer en détail les données et d'en donner une meilleure compréhension. Enfin, seront présentés le processus de choix des algorithmes de classification, leur entraînement, les tests effectués, ainsi que l'optimisation des hyperparamètres, leurs performances et leurs caractéristiques les plus influentes.

2.1. Création d'un corpus poétique français

Pour la constitution du corpus, il a été nécessaire de créer deux sous-corpus distincts : l'un pour les poèmes écrits par des professionnels et l'autre pour ceux rédigés par des amateurs.

Les critères utilisés pour distinguer les deux groupes de poètes sont précis et durement définis. Pour le sous-corpus des professionnels, seuls les poètes ayant reçu un prix de poésie francophone ont été retenus. La liste des poètes récompensés ainsi que les détails des prix qu'ils ont reçus sont disponibles sur le site Printemps des poètes[13]. Cette distinction garantit la crédibilité et la reconnaissance officielle des poètes sélectionnés. Ainsi, on retrouve des auteurs ayant été récompensés par des distinctions prestigieuses telles que le Grand prix de poésie de l'Académie Française, le prix Heredia, le prix Max Jacob, le prix Apollinaire et le prix Mallarmé. La liste complète des auteurs et des prix reçus est disponible en annexe 1.

A l'opposé, le groupe des amateurs est composé de poèmes publiés sur des forums de poésie en ligne, où l'écriture, parfois moins encadrée par des critères académiques ou professionnels, peut comporter des erreurs grammaticales. Ainsi, 3 forums ayant une large communauté active ont été retenus pour cette sélection : Jepoemes.com[14], mobie.oniris.be[15], et poemes-AZ.com[16].

La création du sous-corpus des poètes professionnels s'est déroulée en plusieurs étapes planifiées pour garantir la qualité et la représentativité des données recueillies. Tout d'abord, 50 poètes ayant tous reçu un prix de poésie francophone ont été identifiés de manière aléatoire à partir de la liste disponible citée ci-dessus. Pour chaque poète, un recueil disponible à la Bibliothèque Nationale Universitaire (BNU) de Strasbourg a été sélectionné. Dans chaque recueil, 5 poèmes ont été choisis de manière aléatoire, tout en respectant une limite d'environ 30 à 400 mots environ par poème pour éviter une hétérogénéité excessive du corpus. Les poèmes sélectionnés ont ensuite été photographiés et les images enregistrées. La phase suivante

[13] Le site est disponible à l'adresse suivante : https://www.printempsdespoetes.com/
[14] JePoemes.com est disponible à l'adresse suivante : https://www.jepoemes.com/
[15] mobile.oniris.be est disponible à l'adresse suivante : http://mobile.oniris.be/
[16] poemes-AZ.com est disponible à l'adresse suivante : https://www.poemes-az.com/

a impliqué l'océrisation (OCR), opérée par le moteur OCR de Azure AI Vision[17]. Cette technologie a permis de convertir les images en texte numérique. Toutefois, des erreurs de reconnaissance de caractères peuvent survenir durant ce processus. Ainsi, une relecture attentive a permis de corriger les erreurs grammaticales introduites par l'OCR. Par la suite, le bruit présent dans les textes a été éliminé (exemple Figure 7). Les poèmes ainsi nettoyés ont été enregistrés au format texte brut pour faciliter leur analyse ultérieure. Enfin, un tableau de métadonnées (voir annexe 2) a été rempli pour chaque poème, incluant des informations telles que le nom de l'auteur, le titre du poème, le titre du recueil, la date de publication, l'éditeur, et d'autres données pertinentes.

La création du sous-corpus des poètes amateurs a suivi un processus similaire, bien que les sources et certains aspects des procédures diffèrent légèrement. Tout d'abord, trois forums de poésie en ligne, précédemment mentionnés, ont été identifiés. Ces forums ont été choisis pour leur popularité, leur activité récente et la diversité des poèmes qu'ils contiennent. À partir de ces forums, 50 poètes ont été sélectionnés de manière aléatoire, garantissant une diversité représentative des poètes amateurs en ligne. Pour chaque poète, 5 poèmes ont été choisis aléatoirement, en respectant la même limite d'environ 30 à 400 mots par poème que pour les professionnels. Les pages web contenant les poèmes sélectionnés ont été enregistrées en format HTML afin de conserver une trace des textes originaux. Cette étape est essentielle pour garantir l'authenticité des données et permettre une vérification ultérieure si nécessaire. Le texte des poèmes a ensuite été copié et une suppression du bruit a été effectuée. Les erreurs grammaticales ont été corrigées afin de garantir une comparaison pertinente entre les deux groupes de poèmes. Les poèmes nettoyés ont été enregistrés au format texte brut, et le fichier de métadonnées a été complété pour chaque poème, incluant des informations similaires à celles des poèmes professionnels et des métadonnées spécifiques aux amateurs. Ces dernières comprennent notamment la date de téléchargement, l'indication de corrections grammaticales éventuelles et la source du poème.

Le corpus ainsi constitué comprend un total de 500 poèmes pour 100 poètes différents, répartis entre les deux sous-corpus. Chacun compte 250 poèmes, soit 50 poètes offrant ainsi une base équilibrée propice à des analyses comparatives.

[17] Version 3.2

Figure 7 : Exemples de suppression de bruit

2.2. Création d'un dictionnaire émotionnel

L'identification de deux dictionnaires émotionnels a été entreprise dans le cadre de cette recherche, notamment le lexique VAD de Mohammad (Mohammad, 2018) en français et le lexique FEEL de Abdaoui et al. (2017). Ces deux lexiques contiennent des mots émotionnels sous leur forme lemmatisée (forme canonique du mot).

Le lexique VAD, contient 15 555 entrées, chaque entrée comprend des scores selon 3 caractéristiques décrites ci-dessous, allant de 0 (le plus faible) à 1 (le plus élevé) pour chaque terme émotionnel. La première caractéristique est la valence qui désigne la polarité émotionnelle du mot. La seconde caractéristique est l'excitation (arousal en anglais) qui correspond à l'activation émotionnelle associée au terme. Enfin, la dernière caractéristique concerne la dominance. Cette dernière correspond au contrôle perçu sur un stimulus. Un extrait du dictionnaire est disponible ci-dessous.

Tableau 3 : Extrait du dictionnaire VAD

Mot émotionnel	V	A	D
abject	0.354	0.590	0.417
agressivité	0.077	0.913	0.648
attaquer	0.276	0.837	0.642
attachant	0.663	0.434	0.604
alcool de contrebande	0.633	0.389	0.491

D'autre part, le lexique FEEL, dérivé de la traduction et de l'extension d'un dictionnaire appelé NRC Emotion Lexicon (Mohammad, Turney, 2013), contient 14 126 entrées et présente des scores de polarité (négative ou positive) ainsi que des scores associés aux six émotions de base de Paul Ekman, à savoir : la joie, la peur, la tristesse, la colère, la surprise et le dégoût. Au sein du lexique FEEL, un score de 0 signifie l'absence de cette émotion, tandis que 1 indique sa présence.

Tableau 4 : Extrait du lexique FEEL

id	word	polarity	joy	fear	sadness	anger	surprise	disgust
1652	abandonner	negative	0	1	1	1	0	1
1887	agacement	negative	0	0	1	1	0	1
1931	aimable	positive	0	0	0	0	0	0
2219	applaudissement	positive	0	0	0	0	1	0
2495	au pouvoir	positive	0	0	0	0	0	0

Une fois les dictionnaires identifiés, un script Python a été élaboré pour effectuer plusieurs étapes. Tout d'abord, la colonne de valence a été supprimée du dictionnaire VAD, puisque cette information était déjà présente dans le lexique FEEL sous la catégorie polarity. Concernant le lexique FEEL, la colonne de polarité a été normalisée. Les polarités initialement représentées par les chaînes de caractères "positive" et "negative" ont été converties en format numérique binaire : 0 pour négative et 1 pour positive. Cette conversion vise à simplifier le traitement ultérieur des données et à obtenir un dictionnaire final normalisé.

Ensuite, les entrées lexicales comprenant plus d'un token telles que « fabriquer de tout pièce » ont été supprimées des deux dictionnaires, car selon la méthodologie adoptée, elles ne se révélaient pas pertinentes. Cette suppression a conduit à la suppression de 2 225 entrées lexicales comprenant plus d'un token du dictionnaire VAD et de 2 148 entrées du lexique FEEL. Par la suite, les dictionnaires ont été fusionnés en conservant uniquement les entrées disponibles dans les deux. Cette fusion a donné lieu à un total de 8 349 entrées, ce qui a engendré une perte totale de 16 959 entrées après les étapes de suppression et de fusion.

Pour élargir le dictionnaire final, une technique basée sur les plongements lexicaux a été employée. Un modèle de Word2Vec[18] a été utilisé pour rechercher deux synonymes par entrée. Cet algorithme rend capable la transformation des mots en vecteurs numériques, permettant ainsi de capturer les relations sémantiques entre les mots. Lors de la recherche de synonymes, les mots figurant dans une liste française de mots vides intégrée par défaut dans la bibliothèque Spacy ont été exclus, afin de maintenir un contrôle strict sur les mots ajoutés dans le dictionnaire. Un seuil minimal de similarité de 0.5 a été défini afin de ne considérer que les synonymes hautement similaires. Les synonymes dont la similarité dépassait ce seuil, ne correspondants pas aux mots vides et n'étant pas déjà inclus dans le dictionnaire ont été ajoutés au dictionnaire final, en conservant les scores d'excitation, de dominance, de polarité et des six

[18] Modèle fr_wiki_no_phrase_no_postag_500_cbow_cut10.bin disponible à l'adresse : https://fauconnier.github.io/

émotions de base du mot d'origine. La fonction *expand_dictionary_with_synonyms()* en figure 8 rend compte de ces étapes. Au total, 2 501 nouvelles entrées ont été ajoutées, portant le dictionnaire final (extrait Tableau 5) à 10 850 mots émotionnels.

Figure 8 : Fonction *expand_dictionary_with_synonyms()*

Grâce à cette recherche, des synonymes tels que : *délaisser* pour le mot émotionnel *abandonner*, *punition* pour *châtiment*, *complot* pour le mot émotionnel *conspiration*, *innocenter* pour *disculper*, *fortuné* pour *riche* ont été trouvés.

Tableau 5 : Extrait du dictionnaire émotionnel final

Mot émotionnel	Excitation	Dominance	Polarité	Joie	Peur	Tristesse	Colère	Surprise	Dégoût
honorable	0.462	0.919	1	0	0	0	0	0	0
intellectuel	0.455	0.845	1	0	0	0	0	0	0
frustrer	0.809	0.243	0	0	0	1	1	0	0
hideux	0.750	0.328	0	0	1	1	0	0	1
immortalité	0.635	0.861	1	0	0	0	0	0	0
injustice	0.817	0.336	0	0	0	1	1	0	0
gifler	0.804	0.518	0	0	0	0	1	1	0
kidnappeur	0.923	0.535	0	0	1	1	1	1	0

2.3. Analyse du corpus

Le corpus initial, comprenant un total de 86 935 tokens, se divise en 34 853 tokens pour le groupe des professionnels et 52 082 tokens pour le groupe des amateurs. En ce qui concerne le nombre de mots après suppression de la ponctuation, des symboles et des numéros, le groupe

des poèmes professionnels comprend 26 738 mots, tandis que le groupe de poèmes amateurs en contient 39 523, totalisant ainsi 66 261 mots.

La bibliothèque SciPy de Python a été utilisée pour réaliser des tests statistiques visant à évaluer la distribution des données. Le test de Shapiro-Wilk[19] a été utilisé sur le nombre de mots pour vérifier la normalité des données, et ainsi déterminer l'approche statistique utilisée. Les résultats ont indiqué une non-normalité des données pour les deux sous-corpus, avec des valeurs p inférieures au seuil alpha communément admis de 0,05. Par conséquent, le test *U* de Mann-Whitney[20], un test non paramétrique, a été utilisé pour comparer la tendance centrale des deux groupes en termes de nombre de mots. Les résultats ont révélé une différence significative entre les deux groupes avec une valeur $p \approx 1.67e\text{-}17$.

Pour remédier à cette disparité, le corpus a été reconstruit en modifiant le sous-corpus des amateurs pour le rendre homogène avec celui des professionnels. Cette décision a été motivée par la facilité d'accès aux données pour ce dernier groupe. Les nouvelles statistiques de la version 2 du corpus ont montré une meilleure homogénéité (voir Tableau 6) avec un total de 75 823 tokens, dont 36 342 pour les poèmes professionnels et 39 481 pour les amateurs. En ce qui concerne le nombre de mots, le sous-corpus des poèmes professionnels comprend 27 812 mots et celui des amateurs en contient 29 573, totalisant 57 385 mots.

Les tests statistiques ont été répétés sur la version 2 du corpus. Les résultats ont montré que les données ne sont toujours pas normalement distribuées, mais que la version 2 du corpus présente une homogénéité significativement améliorée en termes de nombre de mots entre les deux groupes (voir Figure 9). Ainsi, grâce à l'utilisation du test non paramétrique de Mann-Whitney il a été conclu qu'il n'y a pas de différence significative en termes de tendance centrale sur le nombre de mots entre les groupes amateur et professionnel (valeur $p \approx 0.1$).

Tableau 6 : Statistiques comparatives des corpus

Version	**v1**		**v2**	
Type	**Nb tokens**	**Nb mots**	**Nb tokens**	**Nb mots**
Amateur	52 082	39 523	39 481	29 573
Professionnel	34 853	26 738	36 342	27 812
Total	86 935	66 261	75 823	57 385

[19] La documentation du test est disponible à l'adresse suivante : https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html

[20] La documentation du test est disponible à l'adresse suivante : https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html#scipy.stats.mannwhitneyu

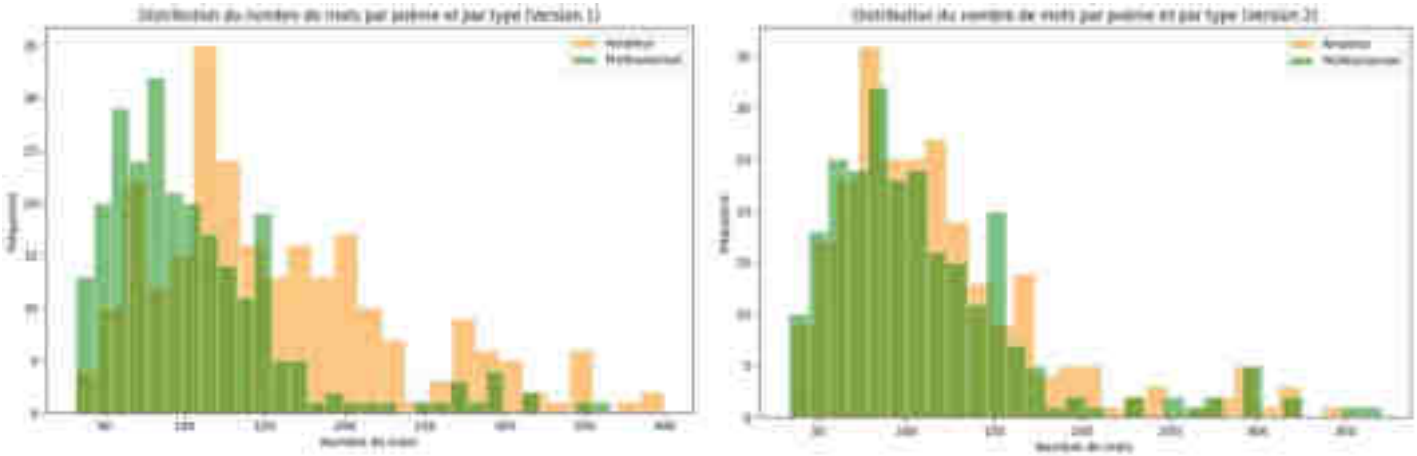

Figure 9 : Distributions du nombre de mots par poème et par type V1 et V2

D'autres analyses supplémentaires ont été menées sur la version 2 du corpus. Premièrement, la moyenne du nombre de tokens a été calculée pour chaque groupe de poèmes. Pour les amateurs, cette moyenne s'élève à environ 158 tokens, tandis que pour les professionnels cette moyenne s'élève à environ 145 tokens. La moyenne totale pour l'ensemble du corpus est estimée à 152 tokens par poème. En ce qui concerne le nombre de mots, la moyenne pour les amateurs est d'environ 118 mots, et pour les professionnels, elle est de 111 mots. La moyenne totale de mots par poème dans l'ensemble du corpus est donc d'environ 115. Le poème le plus long en termes de nombre de mots est le numéro 95, classé comme professionnel, avec 370 mots. En revanche, le poème le plus court est le numéro 110, également classé comme professionnel, avec seulement 33 mots.

Une analyse des mots émotionnels présents dans le corpus a également été menée, en utilisant le dictionnaire émotionnel nettoyé, fusionné et agrandi. Le total de mots émotionnels identifiés pour les poèmes professionnels est de 8 088, tandis que pour les poèmes amateurs, il est de 8 357. En calculant la moyenne du nombre de mots émotionnels par poème pour chaque groupe, nous obtenons les résultats suivants : la moyenne pour les poèmes écrits par des professionnels est de 32.35 mots émotionnels par poème, et pour les poèmes amateurs, elle est de 33.43 mots émotionnels par poème. Le poème ayant le plus grand nombre de mots émotionnels est le numéro 95 (également le plus grand nombre de mots), classé comme professionnel, avec 113 mots émotionnels sur un total de 370 mots. À l'opposé, le poème ayant le moins de mots émotionnels est le numéro 213, appartenant au groupe des professionnels, avec seulement 5 mots émotionnels sur un total de 50 mots.

La polarité des mots émotionnels dans les poèmes des deux groupes, amateurs et professionnels, a également été examinée (Figure 10). Les résultats montrent qu'il n'y a pas de différence significative de polarité entre ces deux catégories. En effet, parmi les mots émotionnels identifiés dans les poèmes professionnels, 67,7 % des mots ont une polarité positive et 32,3 % une polarité négative. De même, parmi les mots émotionnels des poèmes amateurs, 68,3 % ont une polarité positive et 31,7 % ont une polarité négative. Ainsi, il est intéressant de noter que, quel que soit le type de poème (amateur ou professionnel), l'utilisation de mots émotionnels à

polarité positive est nettement plus fréquente dans les données que l'utilisation de mots émotionnels à polarité négative.

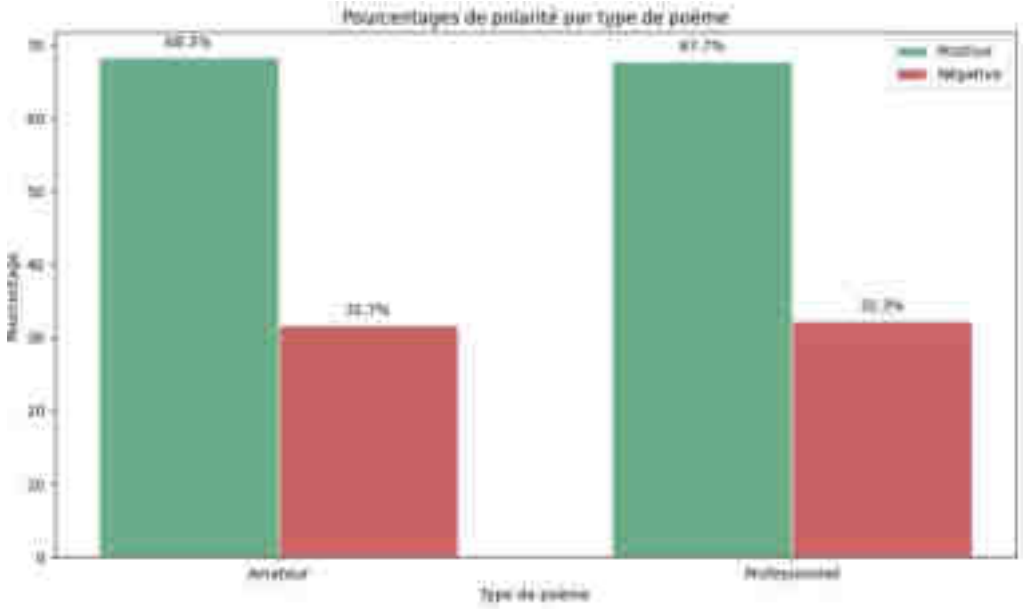

Figure 10 : Répartition des polarités par type de poème

La répartition des six émotions de base selon le type de poèmes a également été examinée (Figure 11). Pour déterminer le pourcentage de chaque émotion dans les deux groupes, le nombre d'occurrences de mots émotionnels associés à chaque émotion spécifique dans les poèmes amateurs et dans les poèmes professionnels a d'abord été compté. Ensuite, le pourcentage a été calculé en divisant le nombre d'occurrences de chaque émotion pour un groupe donné par le nombre total d'occurrences de cette émotion dans les deux groupes combinés.

Ainsi, le pourcentage de joie dans les poèmes amateurs est plus élevé que dans les poèmes professionnels, atteignant respectivement 60,8 % contre 39,2 %. Cela signifie que, parmi tous les mots émotionnels exprimant la joie dans les deux groupes, 60,8 % proviennent des poèmes amateurs et 39,2 % des poèmes professionnels. De même, la proportion de mots émotionnels exprimant la peur est relativement similaire entre les deux groupes, avec une fréquence de 50,7 % pour les poèmes amateurs et de 49,3 % pour les poèmes professionnels. Cela indique que, parmi tous les mots émotionnels liés à la peur dans les deux groupes combinés, 50,7 % proviennent des poèmes amateurs et 49,3 % des poèmes professionnels.

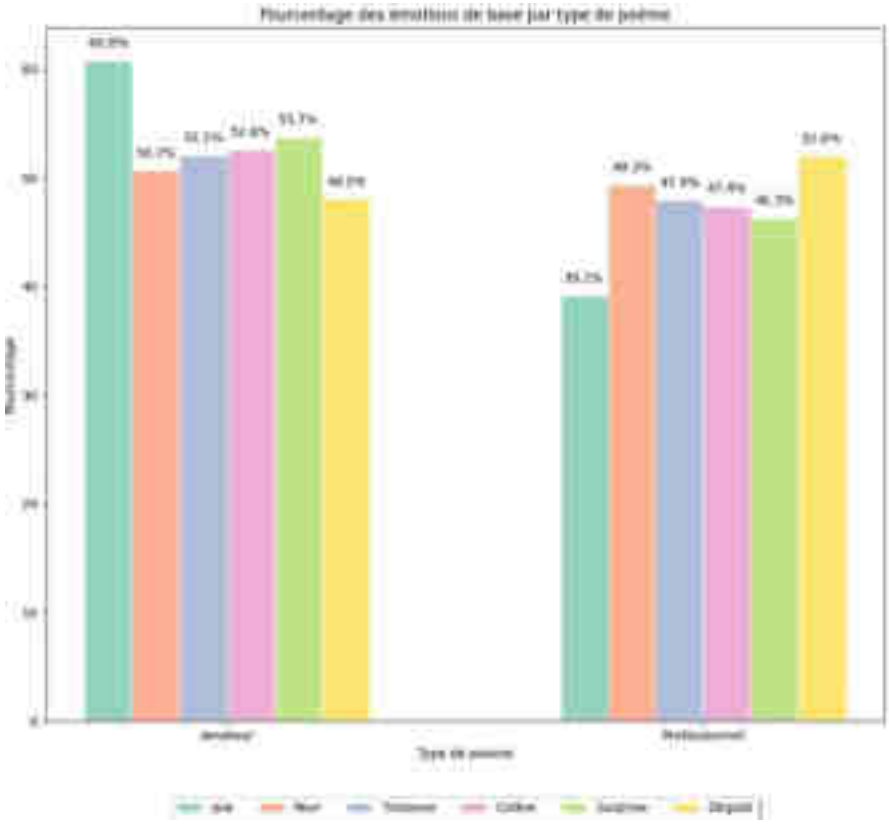

Figure 11 : Répartition des émotions de base par type de poème

Les mots émotionnels évoquant les sentiments de tristesse, de surprise et de colère sont légèrement plus fréquents dans les poèmes amateurs que dans ceux des professionnels. Enfin, l'émotion de dégoût est davantage présente dans les poèmes professionnels que dans ceux des amateurs, avec respectivement 52% contre 48%.

Ainsi, ces analyses offrent une perspective approfondie sur le contenu émotionnel des poèmes présents dans le corpus version 2. Elles révèlent non seulement une meilleure homogénéité dans la distribution des mots au sein des deux sous-corpus, mais aussi des disparités dans la répartition des mots émotionnels et de leurs contenus émotionnels entre les deux groupes de poètes. De ce fait, le corpus constitué permettra la conception d'un algorithme capable de différencier les deux groupes en fonction des émotions véhiculées en leur sein.

2.4. Apprentissage automatique

Le but de ce projet est de développer un classifieur automatique capable de différencier les poèmes amateurs des poèmes professionnels en se basant sur les émotions qu'ils expriment. Pour atteindre cet objectif, des techniques d'apprentissage automatique, qui permettent d'entraîner un modèle à partir des données seront utilisées. En analysant les caractéristiques émotionnelles, le modèle apprend à identifier les schémas émotionnels spécifiques qui distinguent les œuvres amateurs des œuvres professionnelles. Ce processus implique plusieurs étapes, telles que le prétraitement des données, la sélection des caractéristiques pertinentes,

l'entraînement des modèles, l'optimisation des hyperparamètres et l'évaluation de leurs performances. Grâce à cette approche, l'automatisation et l'amélioration de la qualité du classement des poèmes en fonction des émotions véhiculées sont espérées.

2.4.1. Préparation et transformation des données

Dans cette étude, un pipeline de traitement de texte a été développé dans le but d'entrainer un classifieur automatique. Deux transformers personnalisés ont été implémentés pour ce faire. Le premier, nommé *TextLemmatizer*, a été conçu pour tokeniser et lemmatiser les textes en utilisant un modèle disponible dans la bibliothèque Spacy (voir Figure 12). Cette étape de prétraitement avait pour objectif de normaliser le texte et de simplifier les données, facilitant ainsi leur analyse ultérieure.

Figure 12 : Transformer *TextLemmatizer*

Le seconde transformation, le *EmotionalWordScoreCalculator*, a utilisé les mots émotionnels comme fonctionnalités (voir Figure 13). Les scores émotionnels pour chaque texte ont été calculés en multipliant les scores TF-IDF des mots avec les scores émotionnels correspondants du dictionnaire. De plus, le nombre de mots émotionnels dans chaque texte a été intégré comme caractéristiques supplémentaires. Enfin, la faible ou forte présence de chaque mot émotionnel dans chaque poème a été capturée, étendant ainsi les informations disponibles pour une analyse émotionnelle approfondie. Il convient de noter que les données ont été agrégées au niveau du poème en produisant un vecteur par poème, ceci dans le but de faciliter leur traitement et leur analyse dans le cadre de ce projet.

Figure 13 : Vectorisation avec *EmotionalWordScoreCalculator*

Par la suite, le pipeline a été utilisé pour entraîner plusieurs modèles classificateurs, exploitant les caractéristiques émotionnelles extraites pour prédire et catégoriser les poèmes écrits par des amateurs ou des professionnels. Les données ont été séparées en un ensemble d'entraînement représentant 80% du total des données et un ensemble de test représentant 20% du total, soit 400 poèmes pour l'entraînement et 100 pour le test du modèle. Ainsi l'ensemble d'entrainement compte 205 poèmes amateurs et 195 poèmes professionnels. L'ensemble de test comprend donc 45 poèmes amateurs et 55 poèmes professionnels.

2.4.2. Entraînement des modèles de classification

Dans cette section, l'entrainement de divers modèles de classification pour la tâche spécifique de distinction entre poèmes amateurs et professionnels a été entrepris. Une gamme de classifieurs disponibles dans la bibliothèque Scikit-learn[21] a été exploitée, chacun ayant ses propres caractéristiques et hyperparamètres.

[21] Version 1.3.2 (Pedregosa et al., 2011)

Dans cette étude, plusieurs modèles de classification ont été évalués, chacun proposant des approches distinctes pour résoudre le problème de classification. Le *DummyClassifier* a été utilisé en premier, fournissant des prédictions aléatoires ou basées sur des stratégies simples. Ce classifieur sert de référence pour comparer d'autres algorithmes plus complexes. Dans notre implémentation, la stratégie adoptée par ce classifieur a été déterminée par le paramètre *most_frequent*, qui retourne toujours l'étiquette de classe la plus fréquente. Un autre classifieur, *MultinomialNB* a été utilisée. C'est un classifieur naïf bayésien adapté aux données multinomiales, souvent utilisé dans le TAL pour la classification de texte.

D'autres algorithmes de classification ont été utilisés tel que : la classe *DecisionTreeClassifier* qui construit des arbres de décision en divisant récursivement les données en sous-groupes homogènes basés sur les caractéristiques des données, dans le but de créer un modèle prédictif utilisant des règles de décision simples. La Régression Logistique est un modèle de régression utilisé pour la classification binaire et multiclasse, en utilisant une fonction logistique pour estimer la probabilité d'appartenance à chaque classe. Le *RandomForestClassifier* est une méthode d'ensemble qui combine plusieurs arbres de décision en ajustant chacun sur un sous-ensemble aléatoire des données d'entraînement. En moyennant les prédictions de ces arbres, il améliore la qualité prédictive tout en réduisant le risque de surajustement. *L'ExtraTreesClassifier* est une variante de *RandomForest* où les divisions sont aléatoires, ce qui peut améliorer la robustesse et réduire la variance du modèle. *AdaBoostClassifier* est une technique d'ensemble qui commence par entraîner un classifieur sur l'ensemble des données d'origine. Ensuite, il ajuste itérativement des versions supplémentaires de ce classifieur sur les mêmes données, en modifiant les poids des instances mal classées à chaque étape. Cela permet aux classifieurs suivants de se concentrer davantage sur les cas difficiles, améliorant ainsi la performance globale du modèle. Le *GradientBoostingClassifier* construit un modèle de manière séquentielle, ajoutant des modèles successifs pour optimiser des fonctions de perte arbitraires qui sont différentiables. *SVC* (Support Vector Classifier) sépare les classes en maximisant la marge entre les instances les plus proches de chaque classe dans l'espace des caractéristiques. Enfin, *GaussianNB* est basé sur une approche probabiliste qui suppose que chaque classe suit une distribution normale.

Chacun de ces modèles a été testé sur l'ensemble de test pour déterminer lequel offrirait la meilleure performance pour la tâche de classification en question, en utilisant les métriques de précision, de rappel et de score F1 pour évaluer leur efficacité.

2.4.3. Résultats

Cette section est consacrée aux différents résultats obtenus pour certains des meilleurs classifieurs évalués. L'intégralité des rapports de classification de chaque modèle testé peut être consulté en annexe 3 et 3bis. Pour chaque modèle, les mesures de précision, de rappel et de score F1 pour chaque classe (amateur et professionnel) sont rapportées, ainsi que l'exactitude globale du modèle sur l'ensemble de test. Ces métriques permettent d'évaluer la capacité du modèle à prédire avec précision les deux classes ainsi que sa capacité globale à généraliser sur

de nouvelles données. En outre, les métriques de macro-moyenne et micro-moyenne fournissent des moyennes pondérées des métriques de performance pour évaluer le modèle dans son ensemble.

Voici les résultats de quelques-uns des meilleurs modèles testés. Ces scores ont été obtenus en évaluant les modèles sur l'ensemble de test, composé de 100 poèmes (45 amateurs et 55 professionnels). Pour une vue d'ensemble plus claire, les figures 14 et 15 illustrent les performances de chaque modèle testé en termes de score F1 et d'exactitude.

- **Extra Trees** : Ce modèle a obtenu une précision de 0,71 pour les poèmes amateurs et de 0,76 pour les poèmes professionnels, avec une exactitude globale de 0,74. En d'autres termes, 71 % des poèmes classés comme amateurs étaient effectivement écrits par des amateurs, tandis que 76 % des poèmes classés comme professionnels étaient bien attribués à des professionnels, montrant ainsi une meilleure prédiction pour les poèmes professionnels. Le rappel et le score F1 suivent les mêmes valeurs, avec 0,71 pour les amateurs et 0,76 pour les professionnels. En résumé, le modèle Extra Trees offre une performance globalement supérieure pour les poèmes professionnels, comme le démontrent les scores de précision, de rappel et de F1 plus élevés. L'exactitude globale de 0,74 reflète une performance solide sur l'ensemble du jeu de données.

- **Régression Logistique** : Ce modèle a montré une précision de 0,71 pour les poèmes amateurs et de 0,80 pour les poèmes professionnels, atteignant une exactitude de 0,76. Le rappel était de 0,78 pour les amateurs et de 0,75 pour les professionnels, conduisant à un score F1 de 0,74 pour les amateurs et de 0,77 pour les professionnels. Ainsi, le modèle de Régression Logistique, tout comme le modèle Extra Trees, montre une meilleure performance pour les poèmes professionnels que pour les poèmes amateurs, comme l'indiquent les différentes métriques plus élevées pour les poèmes professionnels. L'exactitude globale de 0,76 indique une performance globalement bonne sur l'ensemble du jeu de données.

- **Multinomial NB** : Ce modèle a obtenu une précision de 67% pour les poèmes amateurs et de 86% pour les poèmes professionnels, atteignant une exactitude globale de 75%. Cela signifie que 67% des poèmes classés comme amateurs étaient effectivement des poèmes écrits par des amateurs, tandis que 86% des poèmes classés comme professionnels étaient effectivement des poèmes écrits par des professionnels. Le modèle a correctement classé 75% de tous les poèmes, toutes catégories confondues. Le rappel était de 87% pour les amateurs et de 65% pour les professionnels, indiquant que 87% des poèmes amateurs ont été correctement identifiés, tandis que 65% des poèmes professionnels ont été correctement identifiés par le modèle. Le score F1 était de 76% pour les amateurs et de 74% pour les professionnels. En résumé, le modèle Multinomial NB montre une meilleure performance pour les poèmes professionnels que pour les poèmes amateurs en termes de précision, mais une performance supérieure pour les amateurs en termes de rappel.

- **Random Forest** : Ce modèle a obtenu les meilleures performances, avec une exactitude générale de 0,77. La précision était de 74 % pour les poèmes amateurs et de 80 % pour les poèmes professionnels. Le rappel, qui évalue la proportion de poèmes correctement identifiés dans chaque catégorie, s'établissait à 76 % pour la classe Amateur et à 78 % pour la classe Professionnel, indiquant une légère supériorité dans l'identification des poèmes professionnels. Enfin, le score F1 était de 75 % pour les poèmes amateurs et de 79 % pour les poèmes professionnels, reflétant ainsi une performance globale légèrement meilleure pour les poèmes professionnels.

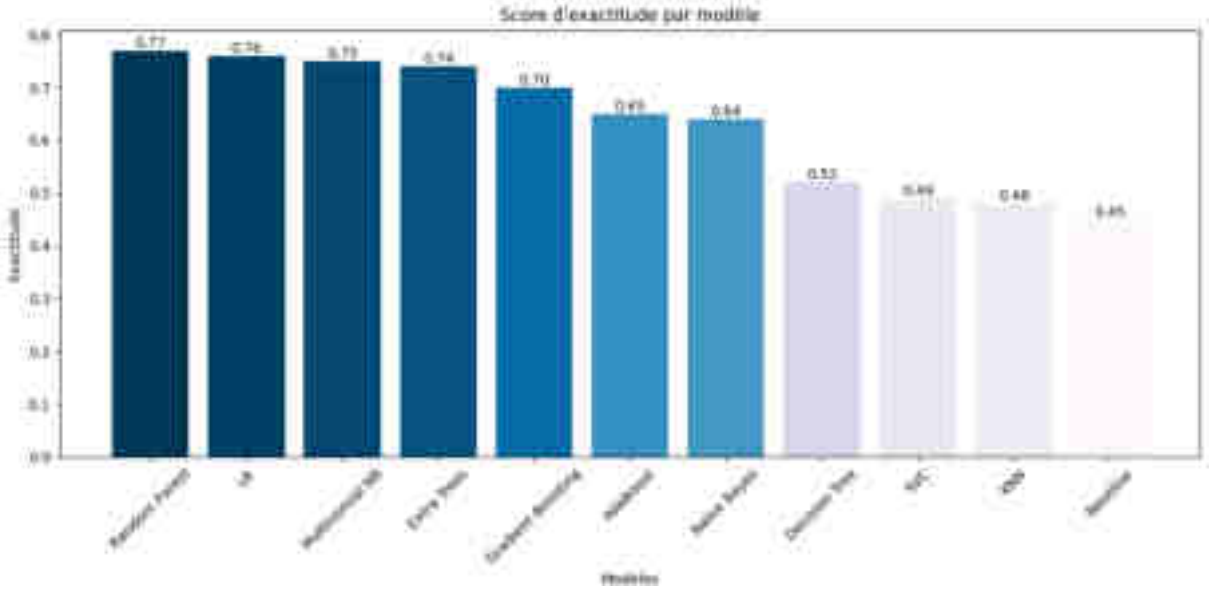

Figure 14 : Score d'exactitude de chaque modèle

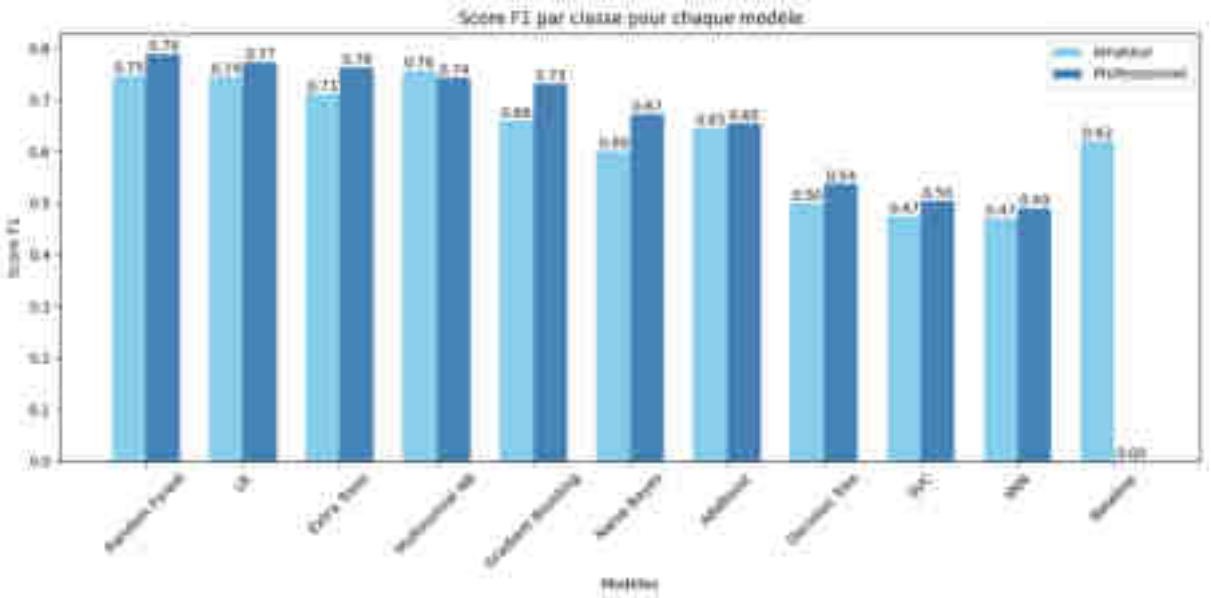

Figure 15 : Score F1 par classe pour chaque modèle

En conclusion, bien que chaque modèle ait ses points forts, les résultats révèlent une tendance générale à mieux prédire les poèmes professionnels que les poèmes amateurs, comme

l'indiquent les scores de précision et de F1 plus élevés pour les poèmes professionnels. Les performances globales montrent que les modèles parviennent à généraliser de manière satisfaisante sur l'ensemble des données de test. Parmi tous les modèles, le Random Forest se démarque par ses performances supérieures, suivi de la régression logistique, du Multinomial NB et du modèle Extra Trees.

2.4.4. Optimisation des hyperparamètres

Dans cette section, une optimisation des hyperparamètres pour les modèles de classification utilisés a été effectuée. L'objectif était d'identifier les combinaisons d'hyperparamètres qui maximisent les performances de chaque modèle. Pour ce faire, la technique de recherche d'hyperparamètres par validation croisée a été appliquée à l'aide de l'outil GridSearchCV de la bibliothèque Scikit-learn, avec une validation croisée stratifiée configurée à 5 plis.

La recherche d'hyperparamètres par validation croisée consiste à évaluer différentes combinaisons d'hyperparamètres en utilisant la validation croisée sur l'ensemble d'entrainement. Cela permet d'estimer les performances de chaque combinaison d'hyperparamètres de manière robuste, en évitant le surajustement aux données d'entraînement. Les meilleurs hyperparamètres identifiés pour trois des modèles les plus performants sont présentés ci-dessous.

Pour le modèle **Extra Trees**, les meilleurs hyperparamètres sont les suivants :
- profondeur maximale de chaque arbre : *None*
- nombre maximal de caractéristiques à considérer pour diviser un nœud de l'arbre : *log2*
- nombre d'arbres dans la forêt : *200*

Ces paramètres permettent d'optimiser les performances du modèle en trouvant un bon équilibre entre biais et variance. La profondeur maximale définie sur *None* permet aux arbres d'apprendre des relations complexes sans restriction, mais cela peut augmenter le risque de surapprentissage. Le choix de *log2* pour le nombre de caractéristiques à considérer lors de la division d'un nœud signifie que le modèle évaluera un nombre de caractéristiques équivalent au logarithme en base 2 du nombre total de caractéristiques disponibles, ce qui aide à contrôler la complexité des arbres. Enfin, un ensemble de *200* arbres offre une bonne couverture des données, bien que cela puisse augmenter la charge computationnelle.

Pour le modèle **LR**, les meilleurs hyperparamètres sont les suivants :
- C : *1*
- penalty : *l2*
- solver : *liblinear*

Ces paramètres contrôlent divers aspects du modèle. Le premier paramètre correspond à la force de régularisation. Plus la valeur de *C* est faible, plus la régularisation est forte. Dans ce cas, une valeur de *1* indique une régularisation faible, ce qui permet au modèle d'être plus complexe. Le

second paramètre spécifie le type de régularisation à appliquer. Ici, *l2* indique l'utilisation de la régularisation de Ridge/Tikhonov. Cela tend à distribuer la pénalisation sur toutes les caractéristiques et à réduire le risque de sur- ajustement. Le dernier paramètre définit l'algorithme utilisé pour optimiser la fonction de coût. Ici, *liblinear* est un solveur qui utilise une approche de descente de gradient adaptée aux petits jeux de données. Il est particulièrement efficace pour les problèmes de classification binaire.

Pour le modèle **SVC**, les meilleurs hyperparamètres sont les suivants :
- C : *0.1*
- kernel : *linear*

Tout comme le modèle de régression logistique, le modèle SVC utilise le paramètre *C* pour contrôler la force de régularisation. Ici la valeur de *0.1* indique une régularisation forte, ce qui entraîne un modèle plus simple en limitant davantage les coefficients. Le second paramètre spécifie le type de noyau à utiliser dans le modèle SVC. Ici, *linear* indique l'utilisation d'un noyau linéaire, ce qui signifie que le modèle cherche à trouver un hyperplan linéaire qui sépare au mieux les différentes classes (Amateur et Professionnel) dans l'espace des caractéristiques. En utilisant ces paramètres optimisés, le modèle SVC est configuré pour obtenir de bonnes performances lors de la classification tout en maintenant une bonne capacité de généralisation sur de nouvelles données, grâce à une régularisation appropriée et à l'utilisation d'un noyau linéaire pour la séparation des classes.

Ainsi, après avoir déterminé les meilleurs hyperparamètres pour plusieurs modèles, cette optimisation permettra de réentraîner les modèles en utilisant ces paramètres ajustés, dans le but d'améliorer leurs performances globales.

2.4.5. Résultats après optimisation des paramètres

Après le réentraînement des modèles sélectionnées et leurs hyperparamètres configurés, leurs performances finales ont été évaluées sur les données de test. Les résultats des 3 meilleurs modèles sont présentés dans cette section.

À la suite de l'optimisation des hyperparamètres du modèle **LR**, les scores demeurent inchangés : la précision pour la classe Amateur reste à 0,71 et à 0,80 pour la classe Professionnel. Le score F1 est maintenu à 0,74 pour Amateur et à 0,77 pour Professionnel, tandis que l'exactitude globale reste à 0,76. Le tableau complet des performances après cette optimisation est présenté ci-dessous.

Tableau 7 : Rapport de classification modèle LR

		Après optimisation des hyperparamètres			
		Précision	**Rappel**	**Score F1**	**Support**
LR	**Amateur**	0,71	0,78	0,74	45
	Professionnel	0,8	0,75	0,77	55
	Exactitude			0,76	100
	Macro-moyenne	0,76	0,76	0,76	100
	Micro-moyenne	0,76	0,76	0,76	100

Pour ce modèle, la matrice de confusion (voir Figure 16) indique 35 prédictions correctes de la classe Amateur et 10 prédictions incorrectes de la classe Professionnel. De plus, elle montre 14 prédictions incorrectes de la classe Amateur et 41 prédictions correctes de la classe Professionnel.

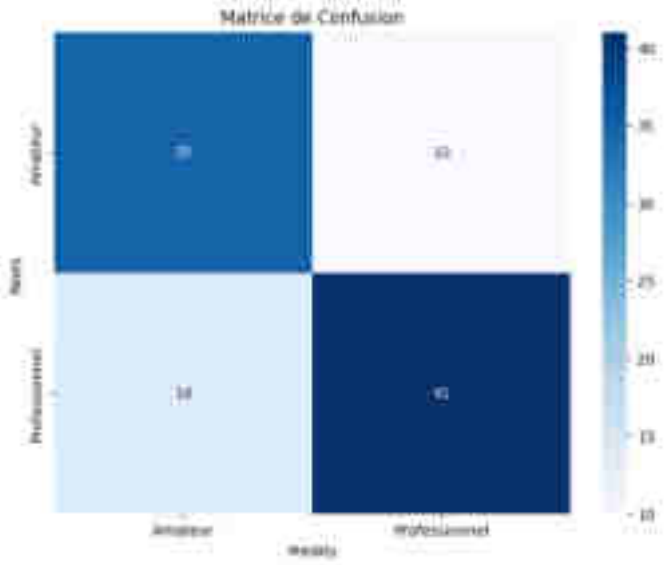

Figure 16 : Matrice de confusion modèle LR

Après l'optimisation des hyperparamètres du modèle **Extra Trees**, des améliorations notables ont été constatées dans les performances pour les deux classes. Avant l'optimisation, la précision pour la classe Amateur était de 0,71, tandis qu'elle est montée à 0,85 après l'optimisation, avec un rappel également amélioré, passant de 0,71 à 0,73. Pour la classe Professionnel, la précision est passée de 0,76 à 0,80, et le rappel a fortement augmenté de 0,76 à 0,89. Ces ajustements ont entraîné une hausse du score F1 pour la classe Amateur, de 0,71 à 0,79, et pour la classe Professionnel, de 0,76 à 0,84 (voir Figure 17). En outre, les mesures macro-moyenne et micro-moyenne ont également montré des progrès, avec les scores F1 combinés augmentant de 0,74 à 0,82. Ces résultats mettent en évidence l'impact positif de l'optimisation des hyperparamètres sur les performances globales du modèle Extra Trees, en améliorant la reconnaissance des instances des différentes classes. Le rapport de classification complet est disponible ci-dessous (Tableau 8).

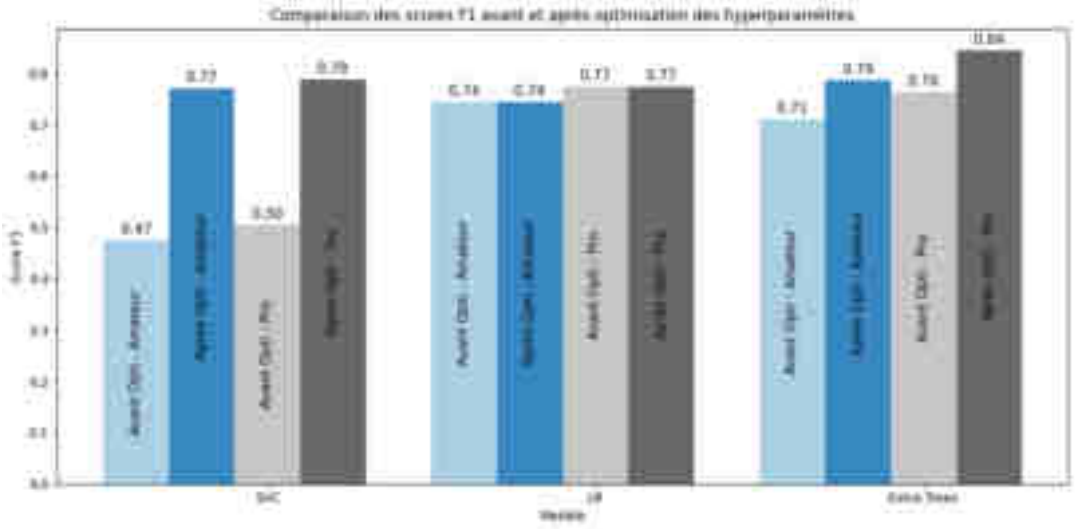

Figure 17 : Comparaison des scores F1 avant et après optimisation des hyperparamètres

Tableau 8 : Rapport de classification modèle Extra Trees

		Après optimisation des hyperparamètres			
		Précision	**Rappel**	**Score F1**	**Support**
Extra Trees	**Amateur**	0,85	0,73	0,79	45
	Professionnel	0,80	0,89	0,84	55
	Exactitude			0,82	100
	Macro-moyenne	0,82	0,81	0,82	100
	Micro-moyenne	0,82	0,82	0,82	100

Concernant ce modèle, la matrice de confusion affiche qu'il y a eu 33 prédictions correctes de la classe Amateur et 12 prédictions incorrectes de la classe Professionnel. De plus, il y a eu 6 prédictions incorrectes de la classe Amateur et 49 prédictions correctes de la classe Professionnel (voir Figure 18).

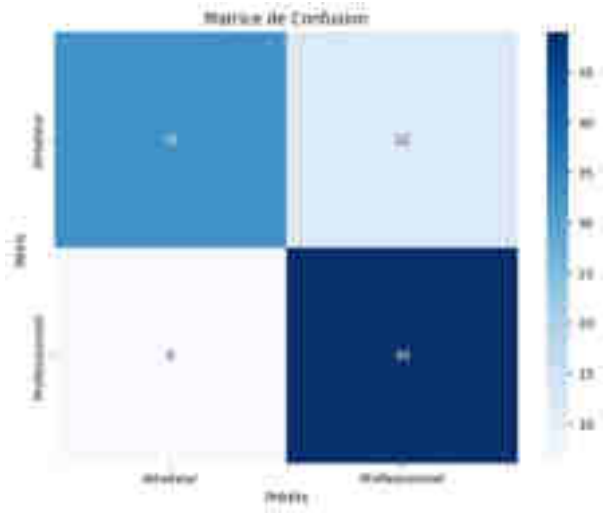

Figure 18 : Matrice de confusion modèle Extra Trees

Après l'optimisation des hyperparamètres du modèle **SVC**, des améliorations marquées ont été constatées dans les performances pour les deux classes, comme le montre le Tableau 9 ci-dessous. Avant l'optimisation, la précision pour la classe Amateur était de seulement 0,44 ; elle a considérablement progressé pour atteindre 0,73. Le rappel pour la classe Amateur a également connu une hausse significative, passant de 0,51 à 0,82, ce qui témoigne d'une meilleure capacité du modèle à identifier correctement les exemples de cette classe. Concernant la classe Professionnel, la précision est passée de 0,54 à 0,84 après l'optimisation, tandis que le rappel a augmenté de 0,47 à 0,75. Ces ajustements ont entraîné une amélioration substantielle du score F1 pour les deux classes, passant de 0,47 à 0,77 (Amateur) et 0,50 à 0,79 (Professionnel) (voir Figure 17). L'exactitude globale du modèle a également croît, atteignant 0,78 contre 0,49 auparavant. Ces résultats illustrent clairement l'impact positif de l'optimisation des hyperparamètres sur les performances du modèle SVC, améliorant sa capacité à classer avec précision les deux classes cibles.

Tableau 9 : Rapport de classification modèle SVC

		Après optimisation des hyperparamètres			
		Précision	**Rappel**	**Score F1**	**Support**
SVC	**Amateur**	0,73	0,82	0,77	45
	Professionnel	0,84	0,75	0,79	55
	Exactitude			0,78	100
	Macro-moyenne	0,78	0,78	0,78	100
	Micro-moyenne	0,79	0,78	0,78	100

Pour ce modèle, la matrice de confusion (Voir Figure 19) montre qu'il y a eu 37 prédictions correctes pour la classe Amateur et 8 prédictions incorrectes classées comme professionnel. De plus, 14 prédictions incorrectes ont été faites pour la classe Amateur, tandis que 41 prédictions étaient correctes pour la classe Professionnel.

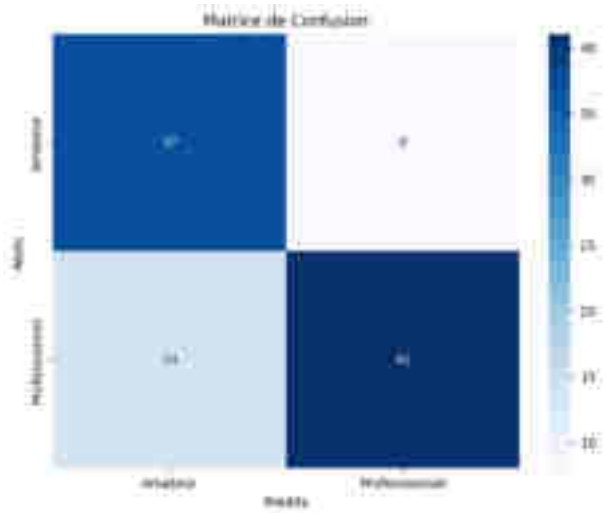

Figure 19 : Matrice de confusion modèle SVC

Ainsi, grâce à l'optimisation des hyperparamètres, le modèle Extra Trees se distingue par les meilleures performances, atteignant une exactitude globale de 82%. En comparaison, le modèle de régression logistique reste stable à 76% d'exactitude. Il est également intéressant de noter que le modèle SVC a considérablement amélioré ses performances après l'optimisation, passant d'une exactitude de 0,49 à 0,78, soit une progression de 0,29 (voir figure ci-dessous).

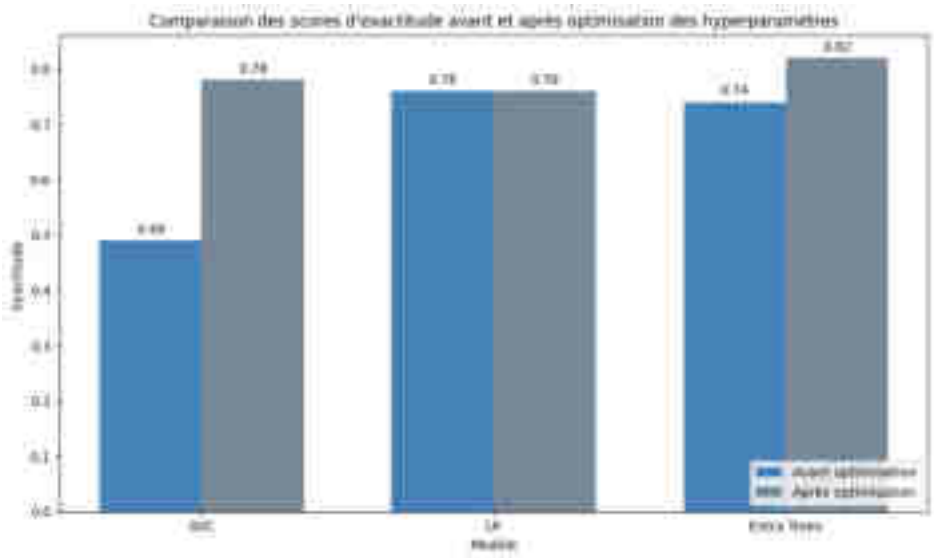

Figure 20 : Comparaison des scores d'exactitude avant et après optimisation des hyperparamètres

2.4.6. Caractéristiques influentes

A la suite de l'entraînement et l'évaluation des modèles, les caractéristiques les plus influentes pour la prédiction des modèles les plus performants (LR, SVC, Extra Trees) ont été identifiées. Les résultats de cette analyse sont détaillés ci-dessous. Pour rappel, les caractéristiques prises en compte incluent la faible ou forte présence de chaque mot émotionnel du dictionnaire précédemment établi, le nombre de mots émotionnels présents dans le poème, ainsi qu'un score par poème pour l'excitation, la dominance, et les six émotions de base (joie, peur, tristesse, colère, surprise et dégoût).

Pour la régression logistique et le SVC, les caractéristiques les plus influentes sont déterminées à l'aide des coefficients du modèle et des hyperparamètres linéaires. L'attribut *coef_* est utilisé pour ces deux classifieurs, fournissant les coefficients associés à chaque caractéristique. Ces coefficients indiquent l'importance relative de chaque caractéristique pour prédire les classes. Les scores des caractéristiques varient de -1 à 1 :

- Les scores proches de -1 indiquent que la caractéristique contribue fortement à prédire la classe négative (ici, Amateur).
- Les scores proches de 1 indiquent que la caractéristique contribue fortement à prédire la classe positive (Professionnel).

Pour le modèle Extra Trees, la mesure d'importance des caractéristiques est obtenue à l'aide de l'attribut *feature_importances_*. Contrairement à la régression logistique et au SVC, qui

fournissent des scores de caractéristiques indiquant leur contribution à chaque classe, Extra Trees ne permet pas une distinction explicite entre les classes. Les scores de *feature_importances_* vont de 0 à 1 et indiquent simplement l'importance globale de chaque caractéristique pour le modèle, sans spécifier pour quelle classe cette importance est maximale.

En résumé, pour Extra Trees, les caractéristiques les plus influentes pour la prédiction globale sont identifiées, mais la contribution spécifique de chaque caractéristique à chaque classe ne peut pas être discernée.

Régression logistique

Pour prédire la classe Professionnel, les 30 caractéristiques les plus influentes pour ce modèle sont les suivantes (ordre décroissant) : *temps*, *mort*, *nuit*, *mot*, *fille*, *long*, *presque*, *oiseau*, *homme*, *colline*, *chair*, *arbre*, *lourd*, *lire*, *cri*, *pays*, *voix*, *ombre*, *monde*, *femme*, *bouche*, *sou*, *histoire*, *saisir*, *entrer*, *chose*, *porte*, *Polarité*, *vol* et *maigre* (voir Figure 21). Il est notable que, parmi ces caractéristiques, seule la polarité est un score émotionnel, tandis que les autres sont des lemmes associés aux émotions (faible ou forte présence).

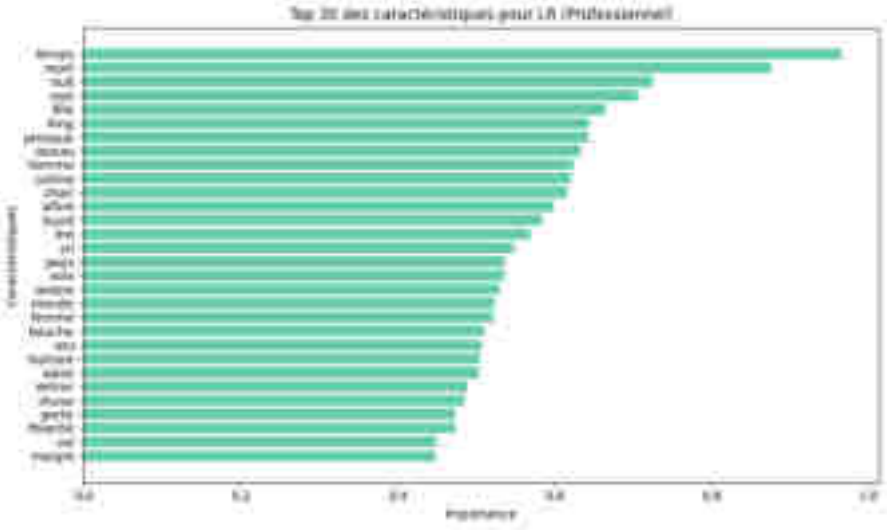

Figure 21 : Caractéristiques les plus influentes modèle LR (Professionnel)

Pour prédire la classe Amateur, les 30 caractéristiques les plus influentes sont : *amour*, *aimer*, *sentir*, *beauté*, *bien*, *pleurer*, *rester*, *poète*, *morne*, *secret*, *fruit*, *rêve*, *mal*, *petit*, *dépasser*, *bon*, *aller*, *vague*, *offrir*, *océan*, *brume*, *amoureux*, *espérer*, *vert*, *rime*, *larme*, *jolie*, *Joie*, *soupir* et *caresse* (voir figure ci-dessous). Il est intéressant de noter qu'un seul score émotionnel, la joie, se distingue parmi ces caractéristiques.

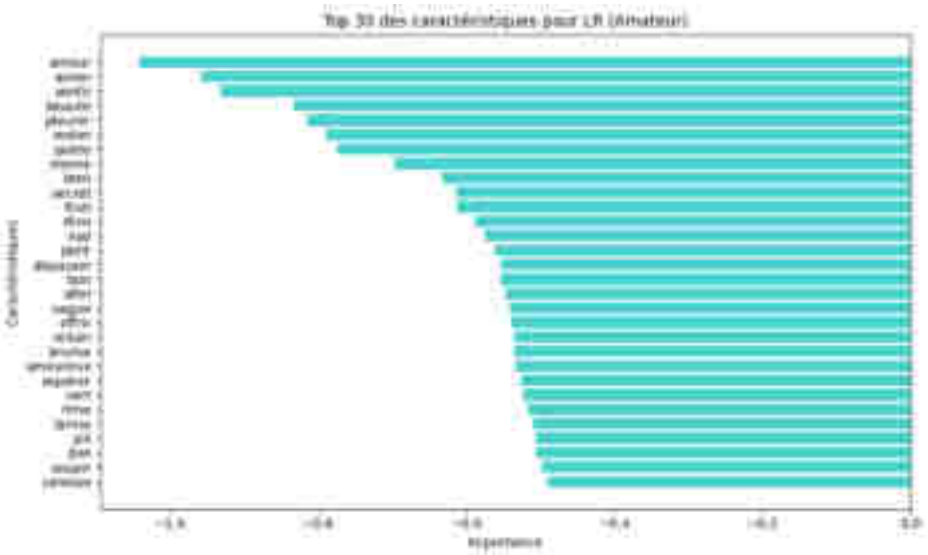

Figure 22 : Caractéristiques les plus influente modèle LR (Amateur)

SVC

Pour prédire la classe Professionnel, les 30 caractéristiques les plus influentes pour ce modèle sont : *temps*, *mort*, *presque*, *homme*, *fille*, *long*, *mot*, *colline*, *lire*, *lourd*, *saisir*, *histoire*, *nuit*, *oiseau*, *chair*, *arbre*, *monde*, *créer*, *cri*, *pays*, *ombre*, *chose*, *poussière*, *maigre*, *vol*, *fureur*, *habiter*, *forêt*, *usure*, *porte* (voir Figure 23). Dans ce modèle, seule la faible ou forte présence de lemmes émotionnels a été prise en compte.

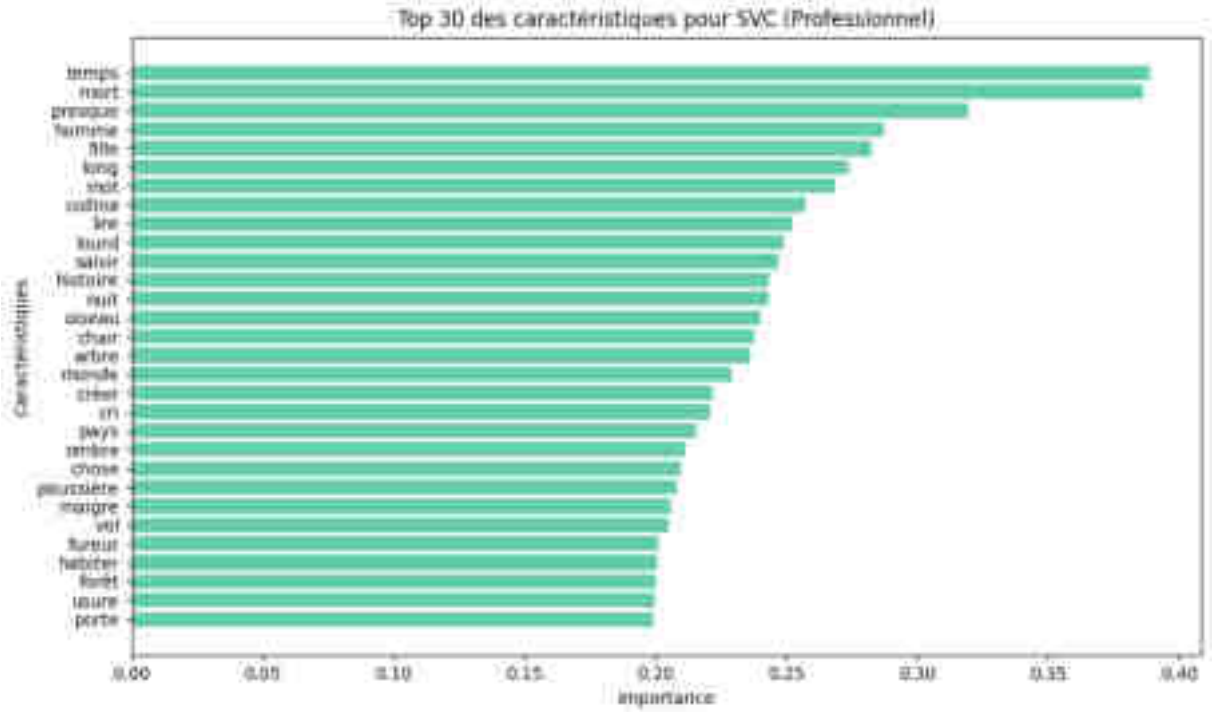

Figure 23 : Caractéristiques les plus influentes modèle SVC (Professionnel)

Pour prédire la seconde classe, le modèle SVC a utilisé les 30 caractéristiques les plus influentes suivantes : *sentir*, *amour*, *aimer*, *beauté*, *rester*, *poète*, *pleurer*, *morne*, *dépasser*, *secret*, *fruit*, *aller*, *soupir*, *océan*, *vert*, *rime*, *espérer*, *brume*, *vague*, *bien*, *amoureux*, *rêve*, *longer*, *poésie*, *voile*, *petit*, *offrir*, *bateau*, *peur*, *bon*. La même observation s'applique que pour la classe Professionnel, seule la faible ou forte présence de lemmes émotionnels a été prise en compte dans ce modèle (voir figure ci-dessous).

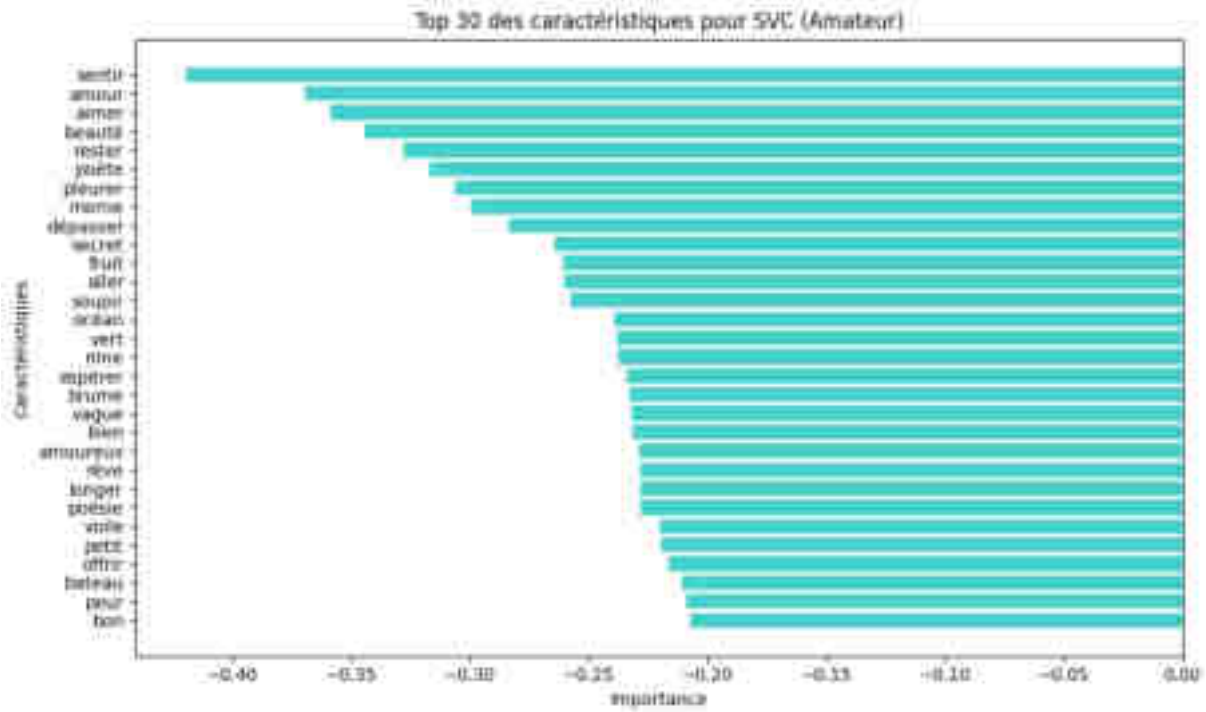

Figure 24 : Caractéristiques les plus influentes modèle SVC (Amateur)

Extra Trees

Comme expliqué précédemment, ce modèle ne peut montrer que les caractéristiques les plus influentes pour sa prédiction, sans spécifier si elles concernent la classe Professionnel ou Amateur[22]. Ainsi, parmi les meilleures caractéristiques, classées par ordre décroissant, on retrouve : *Joie*, *amour*, *Colère*, *nombre de mots émotionnels*, *Peur*, *Tristesse*, *Excitation*, *Dégoût*, *Polarité*, *aimer*, *Surprise*, *Dominance*, *nuit*, *petit*, *temps*, *pleurer*, *bien*, *mort*, *sentier*, *entrer*, *poète*, *lumière*, *voir*, *mot*, *sou*, *aller*, *vie*, *main*, *arbre*, *beauté* (voir Figure ci-dessous).

[22] Des outils d'explicabilité externes à Scikit-learn, tels qu'ELI5 (https://eli5.readthedocs.io/en/latest/), LIME (https://github.com/marcotcr/lime/tree/master) et SHAP (https://github.com/shap/shap), existent. Cependant, ces approches n'ont pas été explorées en raison de contraintes de temps.

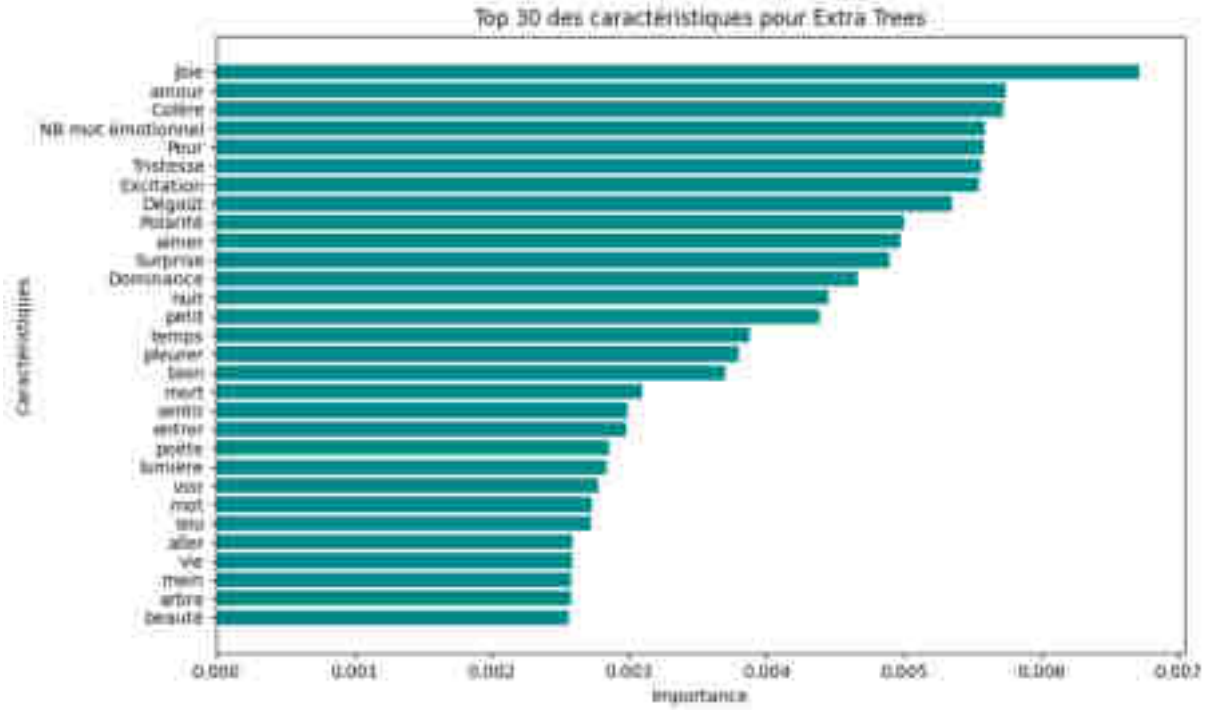

Figure 25 : Caractéristiques les plus influentes modèle Extra Trees

Il est intéressant de noter que tous les scores des émotions de base font partie des caractéristiques les plus influentes, ainsi que le score de Polarité, tous issus du dictionnaire FEEL. Les scores issus du dictionnaire VAD, à savoir l'excitation et la dominance, figurent également parmi ces caractéristiques. Ce qui est également remarquable, c'est que ces scores se retrouvent parmi les 12 premières places des caractéristiques les plus importantes, et que seuls deux lemmes émotionnels, amour (2ème position) et aimer (10ème position), sont inclus dans ce groupe. Ainsi, ce modèle, qui a d'ailleurs atteint les meilleures performances avec une exactitude globale de 0,82, prend en compte tous les « types » de caractéristiques disponibles parmi les plus influentes.

Une comparaison des différentes caractéristiques de chaque modèle a été réalisée. Des nuages de mots ont été créés pour identifier plus clairement les caractéristiques communes aux trois modèles, ainsi que celles partagées par les modèles SVC et LR pour chaque classe.

Figure 26 : Nuage de mots des caractéristiques les plus pertinentes des 3 modèles (deux classes confondues)

Les lemmes *temps*, *mort*, *nuit*, *arbre*, *amour*, *bien*, *aimer*, *beauté*, *pleurer*, *poète*, *petit* et *aller* apparaissent dans les trois modèles. Ces lemmes se rattachent à divers champs lexicaux, notamment la nature (*arbre*), les émotions (*amour*, *aimer*, *pleurer*), la temporalité (*temps*, *mort*, *nuit*), ainsi que l'esthétique représentée par le lemme *beauté*.

Il est également possible de comparer les caractéristiques communes de chaque classe :

Figure 27 : Nuages de mots (classe Professionnel et classe Amateur)

Après une analyse des nuages de mots émotionnels, il apparaît que le champ lexical de la nature est commun aux deux classes, avec les lemmes suivants : *océan*, *brume*, *vague*, *vert*, *fruit*, *arbre*, *oiseau*, *colline*, *monde* et *pays*. Une autre observation qui se dégage est la présence de stéréotypes associés, qui semble manifeste au sein des lemmes de la classe Amateur, mais plus subtile dans la classe Professionnel. Dans la classe Professionnel, des lemmes tels que *mort* et *chair* sont liés à la condition humaine universelle. Ils renvoient à des concepts fondamentaux de l'existence humaine, où la chair symbolise le corps, et la mort représente sa défaillance. Pour la classe Amateur, en revanche, des lemmes tels que *rêve* et *beauté* reflètent davantage une perception stéréotypée de la figure du poète. De plus, dans cette classe, on trouve des termes qui expriment explicitement des émotions ou des sentiments, comme le nom *amour*, le verbe *sentir*, et l'adjectif *morne*. En revanche, dans la classe Professionnel, les références aux émotions ou sentiments sont moins directes, avec des lemmes comme *cri*, *lourd* ou encore des

termes concrets comme *arbre* et *nuit*. Ces lemmes peuvent être associés à des événements ou objets déclencheurs d'émotions, sans pour autant constituer une mention explicite de ces émotions.

Étant donné que de nombreuses caractéristiques sont de « type » faible ou forte présence du lemme dans le poème, il est possible d'explorer les divers prismes émotionnels des caractéristiques les plus influentes (présentées et analysées ci-dessus) pour chaque modèle. Par exemple, il serait intéressant de déterminer si l'un des modèles utilise davantage de caractéristiques (lemmes) associées à l'émotion de joie. Dans un premier temps, une analyse générale peut être réalisée sans distinction des classes, en prenant en compte tous les meilleurs modèles y compris l'Extra Trees, pour lequel la différenciation des classes n'est pas possible. La Figure 28 illustre ainsi la répartition de chaque catégorie émotionnelle : excitation, dominance, joie, peur, tristesse, colère, surprise et dégoût.

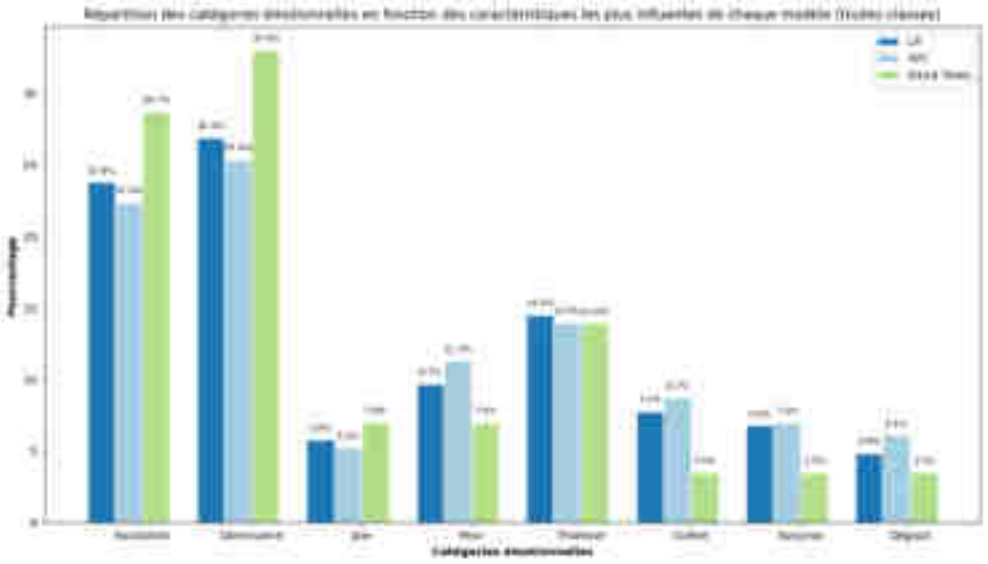

Figure 28 : Répartition des catégories émotionnelles selon les caractéristiques les plus influentes de chaque modèle

De manière générale, on observe que la dominance a un impact plus significatif que l'excitation dans les prédictions des trois modèles. De plus, ces deux catégories émotionnelles influencent davantage le modèle Extra Trees que le SVC et la régression logistique.

En ce qui concerne les émotions de base, les caractéristiques les plus influentes pour les trois modèles sont principalement associées à la tristesse, représentant 14,5 % du total pour le modèle LR, 13,9 % pour les modèles SVC et Extra Trees. La deuxième émotion la plus représentée est la peur, avec 9,7 % pour le modèle LR, 11,3 % pour le SVC, et 7 % pour l'Extra Trees. Cependant, le modèle Extra Trees ne suit pas la même répartition que les deux autres. Alors que les modèles SVC et LR montrent l'émotion de colère comme troisième plus représentée, Extra Trees met en avant la joie (égalité avec la peur). Ensuite, la surprise est la quatrième émotion la plus significative pour les modèles SVC et LR, suivie par le dégoût pour le SVC et la joie pour le LR. Enfin, le SVC classe la joie en dernière position, tandis que le LR place le

dégoût en dernière position. Le modèle Extra Trees utilise quant à lui les émotions de colère, surprise, et dégoût de manière très minimale et égale, chacune représentant seulement 3,5 % de sa répartition totale.

Pour approfondir l'analyse des catégories émotionnelles associées à la prédiction des différentes classes, il est essentiel d'examiner en détail les caractéristiques les plus influentes des modèles SVC et LR. Cette investigation permettra de clarifier la manière dont ces caractéristiques influencent la prédiction pour chaque classe, offrant ainsi une compréhension approfondie de leur rôle dans la classification. Les figures 29 et 30 montrent la répartition des différentes catégories émotionnelles au sein de chaque classe.

Les modèles se basent principalement sur les catégories d'excitation et de dominance pour prédire la classe des amateurs, contrairement à la classe des professionnels. L'émotion de joie émerge comme étant significativement plus présente dans les caractéristiques influentes pour prédire la classe Amateur. En revanche, la peur se révèle plus significative dans les caractéristiques des modèles destinés à prédire la classe Professionnel. De plus, la tristesse apparaît plus fréquemment parmi les caractéristiques influentes pour la classe Amateur, tandis que la colère prédomine pour la classe Professionnel. La surprise, quant à elle, est davantage associée aux caractéristiques influentes des modèles pour la classe Amateur. À l'inverse, le dégoût se retrouve plus souvent dans les caractéristiques influentes des modèles pour la classe Professionnel.

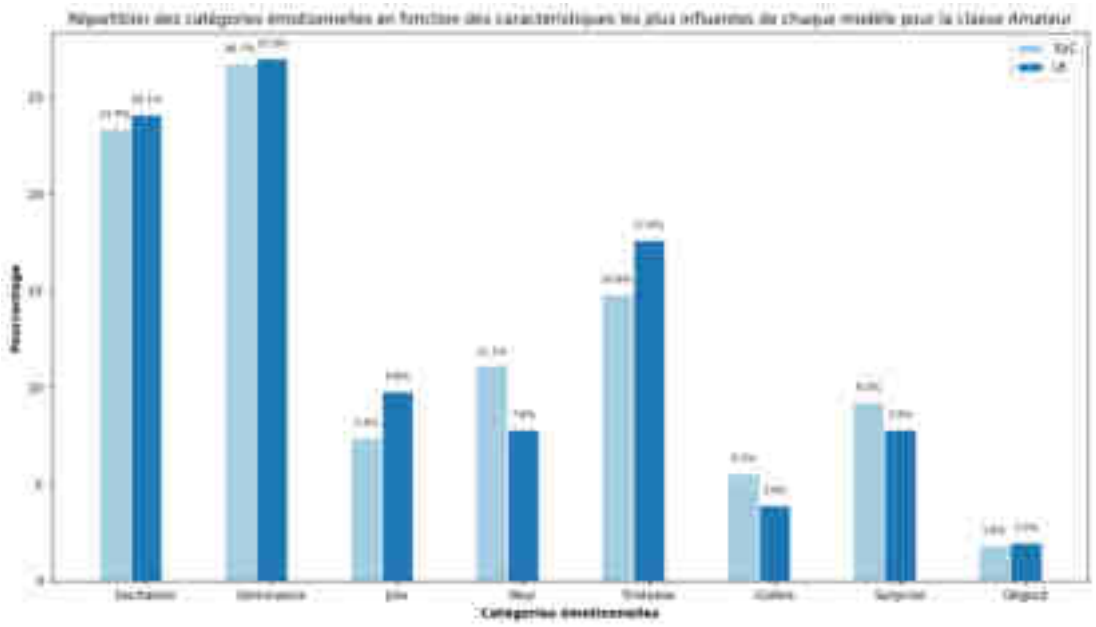

Figure 29 : Répartition des catégories émotionnelles selon les caractéristiques les plus influentes pour prédire la classe Amateur

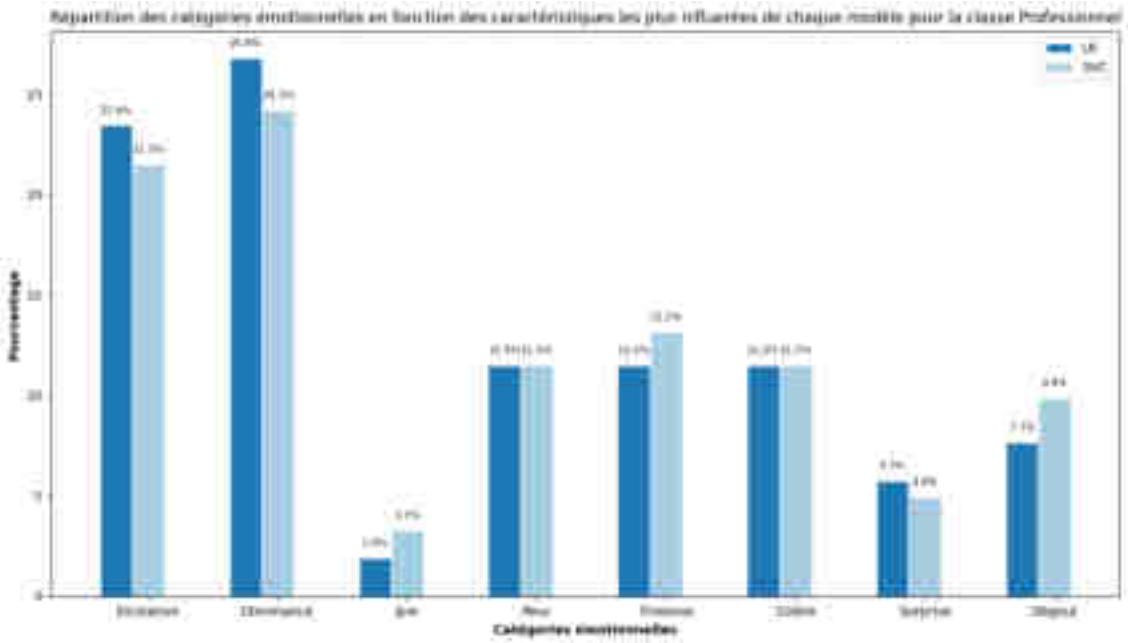

Figure 30 : Répartition des catégories émotionnelles selon les caractéristiques les plus influentes pour prédire la classe Professionnel

Conclusion

Cette étude a exploré la distinction entre poèmes amateurs et professionnels en français en se basant sur les émotions exprimées. Un corpus de 500 poèmes a été constitué, comprenant 250 poèmes amateurs et 250 poèmes professionnels. Des lexiques émotionnels ont été identifiés, fusionnés et enrichis pour une analyse grâce à des méthodes de TAL.

Une analyse statistique approfondie du corpus a été réalisée pour comprendre la distribution des données. Les termes émotionnels ont été particulièrement étudiés pour identifier leur fréquence et leur impact dans les poèmes. Cette étape initiale a permis de dresser un tableau clair des émotions présentes dans les deux groupes de poèmes. Les différents résultats obtenus coïncident avec les travaux antérieurs en langue anglophone des chercheurs américains et canadiens. En effet, Kao et Jurafsky (Kao & Jurafsky, 2012 ; Kao, 2012) ainsi que Dalvean (Dalvean, 2015) avaient observé que les poètes professionnels étaient moins enclins à utiliser des termes émotionnels que les amateurs, mais préféraient davantage dépeindre des situations qui éveilleraient des réponses émotionnelles chez le lecteur. Au sein du corpus étudié dans ce projet, cette conclusion s'applique également. En calculant les moyennes de termes émotionnels par poème pour chaque groupe, les résultats montrent 32,35 mots émotionnels par poème pour les poètes professionnels et 33,43 termes émotionnels par poème pour les amateurs. Cette différence, bien que minimale, concorde avec les précédentes recherches menées sur la langue anglaise.

D'autres résultats significatifs ont également émergé de cette étude. Tout d'abord, il a été constaté qu'il n'y avait presque aucune disparité en ce qui concerne la polarité des mots émotionnels entre les deux groupes de poètes. Cependant, il convient de noter que les poètes, quels qu'ils soient, utilisent davantage de termes émotionnels à polarité positive plutôt que négative. De plus, une observation notable réside dans le fait que les poètes amateurs ont tendance à employer davantage de termes évoquant la joie, la tristesse, la surprise et la colère, tandis que les poètes professionnels manifestent une propension accrue à exprimer le dégoût.

Au-delà des analyses fondées sur la présence et la distribution du lexique émotionnel, une chaîne de prétraitement a été développée pour calculer divers scores émotionnels. Ces scores incluent l'excitation, la dominance, la polarité, la joie, la peur, la tristesse, la surprise, la colère et le dégoût. En plus de ces scores, le nombre de mots émotionnels ainsi que la faible ou forte présence de chaque terme émotionnel dans chaque poème ont été calculés. Cette chaîne de prétraitement a permis de normaliser et de préparer les données pour les analyses ultérieures.

Les données prétraitées ont ensuite été utilisées pour entraîner différents classifieurs automatiques. Une recherche d'hyperparamètres a été effectuée pour optimiser les performances de ces classifieurs. Après avoir identifié les paramètres les plus pertinents, les modèles LR, SVC et Extra Trees se sont révélés les plus performants, avec des exactitudes globales respectivement de 0,76, 0,78 et 0,82.

Les caractéristiques les plus influentes pour chacun des trois modèles ont été identifiées et analysées. Cette analyse a montré que les modèles SVC et LR se basaient principalement sur des caractéristiques liées à la distribution des lemmes, tandis que le modèle Extra Trees se fondait davantage sur les scores émotionnels (tels que l'excitation, la dominance, la polarité, la joie, la peur, la tristesse, la colère, la surprise et le dégoût), ainsi que sur le nombre de mots émotionnels de chaque poème. De plus, les 30 caractéristiques les plus significatives pour prédire chaque classe ont été examinées pour deux des modèles capables de cette distinction. Il a été observé que les modèles SVC et LR s'appuient principalement sur des lemmes associés aux émotions de joie, tristesse et surprise pour prédire la classe Amateur, tandis que les lemmes liés aux émotions de colère, peur et dégoût se révèlent plus discriminants pour prédire la classe Professionnel. Il est important de préciser que cela ne signifie pas que les poèmes professionnels sont intrinsèquement plus liés à la colère, la peur et le dégoût, ou que les poèmes amateurs sont davantage associés à la joie, la tristesse et la surprise. Cela indique plutôt que, parmi les caractéristiques disponibles pour la classification, ce sont les indices lexicaux relatifs à ces émotions qui se sont avérés les plus saillants pour les modèles, en raison de l'approche utilisée pour modéliser les poèmes (notamment par l'application des lexiques d'émotions).

Cependant, l'étude présente certaines limites, notamment la taille réduite du corpus, qui ne comprend que 500 poèmes. De plus, l'utilisation d'un lexique émotionnel basé sur l'anglais constitue une autre des limites de cette recherche. Ainsi, pour améliorer ces résultats il serait nécessaire d'augmenter la taille du corpus et de développer un lexique émotionnel spécifique au français ou d'en identifier d'autres plus pertinents. Il serait également pertinent d'explorer l'utilisation de modèles neuronaux pour la classification, ainsi que d'évaluer la tâche en mode zéro-shot en utilisant un large modèle de langage comme Mistral ou GPT-4, afin d'examiner comment leur comportement diffère par rapport aux approches supervisées employées dans cette recherche.

Une idée prometteuse pour de futures recherches serait d'explorer l'influence des rimes sur l'utilisation des mots émotionnels. Par exemple, il serait intéressant d'analyser si les poètes amateurs et professionnels ont des schémas de rimes distincts lorsqu'ils expriment certaines émotions. De plus, cette étude pourrait examiner comment la structure des rimes influence la fréquence des termes émotionnels. Comprendre ces nuances permettrait de mieux saisir comment les rimes contribuent à la distinction entre poèmes amateurs et professionnels.

En conclusion, cette recherche ouvre la voie à de nouvelles études sur la poésie francophone et ses caractéristiques émotionnelles spécifiques, comblant ainsi un vide dans les recherches en TAL sur la poésie francophone. Les méthodes développées et les résultats obtenus peuvent servir de base pour des analyses plus approfondies et pour la création d'outils d'analyse poétique automatisée.

Références bibliographiques

Abdaoui, A., Azé, J., Bringay, S., & Poncelet, P. (2017). FEEL : A French Expanded Emotion Lexicon. *Language Resources and Evaluation*, *51*. https://doi.org/10.1007/s10579-016-9364-5

Abdat, F. (2010). *Reconnaissance automatique des émotions par données multimodales : Expressions faciales et des signaux physiologiques* [Thèse de doctorat, Université de Lorraine]. https://www.theses.fr/2010METZ035S

Augustyn, M., Hamou, S. B., Bloquet, G., Goossens, V., Loiseau, M., & Rinck, F. (2006). *Lexique des affects : Constitution de ressources pédagogiques numériques.* Colloque International des étudiants-chercheurs en didactique des langues et linguistique. https://shs.hal.science/halshs-00418143

Bafna, P., Pramod, D., & Vaidya, A. (2016). *Document clustering : TF-IDF approach*. 61-66. https://doi.org/10.1109/ICEEOT.2016.7754750

Baudelaire, C. (2018). *Baudelaire : L'intégrale des oeuvres : Tout Baudelaire en 1 volume*. Books on Demand. https://books.google.fr/books?id=EztiDwAAQBAJ

Beausoleil, C. (1991). Que le silence des rues. In *Une certaine fin de siècle: Vol. Tome 2*. La Castor Astral.

Belin, O. (2020). Vers une poésie commune ? Les poètes amateurs de Twitter, Instagram et Wattpad. *Nouvelle revue d'esthétique*, *25*(1), 57-66. https://doi.org/10.3917/nre.025.0057

Bouillot, F., Poncelet, P., & Roche, M. (2014). *De nouvelles pondérations adaptées à la classification de petits volumes de données textuelles*. 131-142. https://hal-lirmm.ccsd.cnrs.fr/lirmm-01054903

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cassina, A. (2021). *La poésie du web : Ethnographie en ligne des poètes ordinaires* [Thèse de doctorat, Université Bourgogne Franche-Comté]. https://theses.hal.science/tel-03681348

Christophe, V. (2019). 3. Les théories néo-darwiniennes. In *Les Émotions : Tour d'horizon des principales théories* (p. 61-73). Presses universitaires du Septentrion. https://doi.org/10.4000/books.septentrion.51006

Communauté Tropes. (s. d.). *Tropes* (Version v7) [Microsoft Windows]. Acetic. http://www.tropes.fr/

Courgeon, M. (2011). *Marc : Modèles informatiques des émotions et de leurs expressions faciales pour l'interaction Homme-machine affective temps réel* [Thèse de doctorat, Université Paris Sud - Paris XI]. https://theses.hal.science/tel-00651467

Dalvean, M. (2015). Ranking contemporary American poems. *Digital Scholarship in the Humanities*, *30*(1), 6-19. https://doi.org/10.1093/llc/fqt036

Dalvean, M. (2016). Ranking Canonical English Poems. *Empirical Studies of the Arts*, *34*(1), 103-125. https://doi.org/10.1177/0276237415621188

Donnat, O. (2009). Les pratiques culturelles des Français à l'ère numérique. Éléments de synthèse 1997-2008. *Culture études*, *5*(5), 1-12. https://doi.org/10.3917/cule.095.0001

Dubois, S. (2006). The French Poetry Economy. *International Journal of Arts Management*, *9*(1), Article 1.

Dujin, A. (2016). Où est passée la poésie française ? Portrait d'un univers paradoxal. *Revue du Crieur*, *5*(3), Article 3. https://doi.org/10.3917/crieu.005.0062

Ekman, P., Freisen, W. V., & Ancoli, S. (1980). Facial signs of emotional experience. *Journal of Personality and Social Psychology*, *39*(6), 1125-1134. https://doi.org/10.1037/h0077722

Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion. In *Handbook of affective sciences* (p. 572-595). Oxford University Press.

Evgeniou, T., & Pontil, M. (2001). *Support Vector Machines : Theory and Applications*. *2049*, 257. https://doi.org/10.1007/3-540-44673-7_12

Fauconnier, J.-P. (2015). *French Word Embeddings models*. http://fauconnier.github.io

Ferdi, D. (2021). *Sélection des caractéristiques basée sur le plongement lexical pour la classification des textes* [Mémoire master 2]. Université de Guelma.

Forsyth, R. S. (2000). Pops and Flops : Some Properties of Famous English Poems. *Empirical Studies of the Arts*, *18*(1), Article 1. https://doi.org/10.2190/E7Q8-6062-K6H4-XFRW

Forum poésie et écriture Poèmes et Poètes—JePoemes.com. (2023, mars 31). JePoèmes. https://www.jepoemes.com/

Gala, N., & Brun, C. (2012, juin 1). *Propagation de polarités dans des familles de mots : Impact de la morphologie dans la construction d'un lexique pour l'analyse d'opinions*.

Gala, N., & Rey, V. (2008). *POLYMOTS : Une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques*. 91-100. https://aclanthology.org/2008.jeptalnrecital-court.10

Goriachun, D. (2020). *Construction d'une base de données lexicale pour les mots français abstraits et concrets* [Mémoire master 2, Aix-Marseille Université]. https://dumas.ccsd.cnrs.fr/dumas-03024192

Grutschus, A., Sascha, D., Novakova, I., Goossens, V., Kern, B., Kraif, O., & Melnikova, E. (2014). Traitement des lexies d'émotion dans les corpus et les applications d'EmoBase. *Corpus*, *13*, 269-293. https://doi.org/10.4000/corpus.2537

Hamon, T., Fraisse, A., Paroubek, P., Zweigenbaum, P., & Grouin, C. (2015, juin 22). *Analyse des émotions, sentiments et opinions exprimés dans les tweets : Présentation et résultats de l'édition 2015 du défi fouille de texte (DEFT)*. Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015). https://hal.science/hal-01617180

Heejoung Shin. (2022). Analyzing the Positive Sentiment Towards the Term "Queer" in Virginia Woolf through a Computational Approach and Close Reading. *Journal of Computational Literary Studies*, *1*.

Hernandez, N., Jadi, G., Lark, J., & Monceaux, L. (2015). *Exploitation de lexiques pour la catégorisation fine d'émotions, de sentiments et d'opinions*.

Huilgol, P. (2020, février 27). Quick Introduction to Bag-of-Words (BoW) and TF-IDF for Creating Features from Text. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/

Jacobs, A. M. (2019). Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics. *Frontiers in Robotics and AI*, *6*(53). https://www.frontiersin.org/articles/10.3389/frobt.2019.00053

James, W. (1884). What is an Emotion? *Mind*, *9*(34), 188-205.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing* (2e éd.). Prentice-Hall, Inc.

Kao, J., & Jurafsky, D. (2012). A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. *Workshop on Computational Linguistics for Literature*, 8-17. https://aclanthology.org/W12-2502

Leadbeater, C., & Miller, P. (2004). *The Pro-Am Revolution : How Enthusiasts are Changing Our Society and Economy*. Demos.

L'éditeur inventeur du Distributeur d'Histoires Courtes ! (s. d.). Short Edition. Consulté 14 mars 2023, à l'adresse https://short-edition.com/fr/

Les prix de poésie. (s. d.). Le Printemps des Poètes. Consulté 15 octobre 2023, à l'adresse https://www.printempsdespoetes.com/Les-prix-de-poesie

Lewis, M. (2022). The Self-Conscious Emotions. In *Encyclopedia on Early Childhood Development [online]*. Tremblay RE, Boivin M, Peters RDeV, dir. Lewis M. https://www.child-encyclopedia.com/emotions/according-experts/self-conscious-emotions

LIF-TALEP & CNRS. (s. d.). *Polarimots*. Consulté 16 mars 2023, à l'adresse https://polarimots.lis-lab.fr/

Lin, Y. (2002). Support Vector Machines and the Bayes Rule in Classification. *Data Mining and Knowledge Discovery*, *6*(3), 259-275. https://doi.org/10.1023/A:1015469627679

Microsoft, A. (s. d.). *Vision Studio*. Extract Text from Images. Consulté 15 octobre 2023, à l'adresse https://portal.vision.cognitive.azure.com/demo/extract-text-from-images

Mikolov, T., Chen, K., Corrado, G. s, & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*, *2013*.

Mohammad, S. (2018). *Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words*. 174-184. https://doi.org/10.18653/v1/P18-1017

Mohammad, S. (2022, octobre 15). *Word Affect Intensities*. Onzième conférence internationale sur les ressources linguistiques et l'évaluation (LREC 2018), Miyazaki, Japon. http://arxiv.org/abs/1704.08798

Mohammad, S., & Turney, P. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, *29*. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Nouvelles, poésie, romans d'auteurs amateurs—Oniris. (s. d.). Oniris. Consulté 31 mars 2023, à l'adresse http://www.oniris.be/

Ortony, A., & Turner, T. (1990). What's Basic About Basic Emotions? *Psychological Review*, *97*, 315-329. https://doi.org/10.1037/0033-295X.97.3.315

Outils de recherche, Emotaix_Free. (2013, octobre 15). Centre PsyClé. https://centrepsycle-amu.fr/outils-recherche/

Pasquier, A. (2021). Approches des émotions et des affects. In *Psychologie et psychopathologie des émotions: Vol. 2e éd.* (p. 21-59). Dunod. https://www.cairn.info/psychologie-et-psychopathologie-des-emotions--9782100799138-p-21.htm

Petit, E. (2009). Le rôle des affects en économie. *Revue d'économie politique*, *119*(6), 859-897. https://doi.org/10.3917/redp.196.0859

Piolat, A., & Bannour, R. (2009a). EMOTAIX : Un scénario de Tropes pour l'identification automatisée du lexique émotionnel et affectif. *L'Année psychologique*, *109*(4), 655-698. https://doi.org/10.3917/anpsy.094.0655

Piolat, A., & Bannour, R. (2009b). Emotions et affects : Contribution de la psychologie cognitive. In *Le sujet des émotions au Moyen Age* (p. 53-83).

Poèmes-AZ.com—De A à Z. (s. d.). Consulté 15 octobre 2023, à l'adresse https://www.poemes-az.com/de-a-a-z.html

Saif, M. M. (s. d.). *Sentiment and Emotion Lexicons*. Saif M. Mohammad. Consulté 10 mars 2023, à l'adresse http://saifmohammad.com/WebPages/lexicons.html

Soëlie, L., Patrice, B., & Emmanuel, B. (2020). *Influence des lexiques d'émotions et de sentiments sur l'analyse des sentiments*.

Vishnubhotla, K., & Mohammad, S. M. (2022). Tweet Emotion Dynamics : Emotion Word Usage in Tweets from US and Canada. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 4162-4176. https://doi.org/10.48550/arXiv.2204.04862

Vogels, C. (2003). Aude Mouaci, Les poètes amateurs, approche sociologique d'une conduite culturelle. Paris, L'Harmattan, Logiques Sociales, 2001, 364 p. *Sociologie de l'Art*, *PS12*, 195-198. https://doi.org/10.3917/soart.001.0195

Wolff, M. (2023). *Poème—Le vent faisait rage*. Forum poésie et écriture Poèmes et Poètes - JePoemes.com. https://www.jepoemes.com/poeme/le-vent-faisait-rage.23589/

Annexes

Annexe 1. Auteurs professionnels et prix reçus

Auteurs	Prix
Charles Le Quintrec Alain Bosquet Claude Michel ClunyJean FOLLAIN André Pieyre de MandiarguesJean Grosjean Francis Ponge Franck Venaille Guy Goffette Henri Thomas Jacques Darras	Grand prix de la poésie de l'Académie Française
Charles Dobzynski	Grand Prix de la Société des Gens de Lettres
Aimé Césaire Jacques Roubaud Anne Perrier	Grand Prix national de Poésie
Emeric de Monteynard Gilles Sicard	Prix Amélie-Murat
Alain Jouffroy Alain Lance Denise Desautels François de Cornière Frédéric Jacques Temple Gaston Miron Jacques Ancet	Prix Apollinaire
Jacques Chessex	Prix Goncourt de la poésie
Abdellatif Laâbi Claude Roy Georges-Emmanuel Clancier	Prix Goncourt de la Poésie Robert Sabatier
Hermine Venot-Focké Claude Beausoleil	Prix Heredia
Antoine Emaz	Prix International de la Poésie francophone Yvan-Goll
Gabrielle Althen Hughes Labrusse	Prix Louis-Guillaume
Jean Berthet	Prix Maïse Ploquin-Caunan
André Schmitz André Velter Annie Salager Bernard Hreglich Henri Meschonnic Hubert Haddad	Prix Mallarmé
Claude de Burine Alexandre Voisard Armen Lubin Bernard Mazo Daniel Boulanger Étienne Faure Gérard Noiret	Prix Max Jacob
Emmanuel Moses Georges Lauris Henri Droguet	Prix Théophile Gautier de l'Académie Française

Annexe 2. Extrait du tableau des métadonnées

id	nom	titre_poeme	fichier_txt	nature_poem	prix_recu	titre recueil	date publication	edition	site	date telechargement	correction
247	Jacques Darras	Prière à saint Antoine	corpus_pro\txt\247.txt	Professionnel	Grand prix de la poésie de l'Académie Française	La Maye réfléchit	2009	Le Cri Edition			
248	Jacques Darras	Passe-crassane	corpus_pro\txt\248.txt	Professionnel	Grand prix de la poésie de l'Académie Française	La Maye réfléchit	2009	Le Cri Edition			
249	Jacques Darras	Le bruit des pages	corpus_pro\txt\249.txt	Professionnel	Grand prix de la poésie de l'Académie Française	La Maye réfléchit	2009	Le Cri Edition			
250	Jacques Darras	Pain d'épices	corpus_pro\txt\250.txt	Professionnel	Grand prix de la poésie de l'Académie Française	La Maye réfléchit	2009	Le Cri Edition			
251	Oyem	Brillances océanes	corpus_amateur\txt\251.txt	Amateur	Aucun		2023		JePoèmes.com	31/07/2023	
252	Oyem	L'imposible dérobade	corpus_amateur\txt\252.txt	Amateur	Aucun		2023		JePoèmes.com	31/07/2023	
253	Oyem	Couleurs d'Armorique	corpus_amateur\txt\253.txt	Amateur	Aucun		2023		JePoèmes.com	31/07/2023	
254	Oyem	Volubile	corpus_amateur\txt\254.txt	Amateur	Aucun		2023		JePoèmes.com	15/09/2023	
255	Oyem	Tu le sais	corpus_amateur\txt\255.txt	Amateur	Aucun		2023		JePoèmes.com	15/09/2023	
256	Carnicella	Il est venu	corpus_amateur\txt\256.txt	Amateur	Aucun		2023		JePoèmes.com	31/07/2023	oui

Annexe 3. Rapports de classification

	Rapport de classification				
		Précision	Rappel	Score F1	Support
Baseline	Amateur	0,45	1	0,62	45
	Professionnel	0	0	0	55
	Exactitude			0,45	100
	Macro-moyenne	0,23	0,5	0,31	100
	Micro-moyenne	0,2	0,45	0,28	100
SVC	Amateur	0,44	0,51	0,47	45
	Professionnel	0,54	0,47	0,5	55
	Exactitude			0,49	100
	Macro-moyenne	0,49	0,49	0,49	100
	Micro-moyenne	0,5	0,49	0,49	100
Multinomial NB	Amateur	0,67	0,87	0,76	45
	Professionnel	0,86	0,65	0,74	55
	Exactitude			0,75	100
	Macro-moyenne	0,76	0,76	0,75	100
	Micro-moyenne	0,77	0,75	0,75	100
LR	Amateur	0,71	0,78	0,74	45
	Professionnel	0,8	0,75	0,77	55
	Exactitude			0,76	100
	Macro-moyenne	0,76	0,76	0,76	100
	Micro-moyenne	0,76	0,76	0,76	100
KNN	Amateur	0,43	0,51	0,47	45
	Professionnel	0,53	0,45	0,49	55
	Exactitude			0,48	100
	Macro-moyenne	0,48	0,48	0,48	100
	Micro-moyenne	0,49	0,48	0,48	100
Radom Forest	Amateur	0,74	0,76	0,75	45
	Professionnel	0,8	0,78	0,79	55
	Exactitude			0,77	100
	Macro-moyenne	0,77	0,77	0,77	100
	Micro-moyenne	0,77	0,77	0,77	100
Decision Tree	Amateur	0,47	0,53	0,5	45
	Professionnel	0,57	0,51	0,54	55
	Exactitude			0,52	100
	Macro-moyenne	0,52	0,52	0,52	100
	Micro-moyenne	0,53	0,52	0,52	100
Extra Trees	Amateur	0,71	0,71	0,71	45
	Professionnel	0,76	0,76	0,76	55
	Exactitude			0,74	100
	Macro-moyenne	0,74	0,74	0,74	100
	Micro-moyenne	0,74	0,74	0,74	100

Annexe 3bis. Rapports de classification (Suite)

	Rapport de classification				
		Précision	Rappel	Score F1	Support
AdaBoost	Amateur	0,59	0,71	0,65	45
	Professionnel	0,72	0,6	0,65	55
	Exactitude			0,65	100
	Macro-moyenne	0,65	0,66	0,65	100
	Micro-moyenne	0,66	0,65	0,65	100
Gradient Boosting	Amateur	0,67	0,64	0,66	45
	Professionnel	0,72	0,75	0,73	55
	Exactitude			0,7	100
	Macro-moyenne	0,7	0,69	0,7	100
	Micro-moyenne	0,7	0,7	0,7	100
Naive Bayes	Amateur	0,6	0,6	0,6	45
	Professionnel	0,67	0,67	0,67	55
	Exactitude			0,64	100
	Macro-moyenne	0,64	0,64	0,64	100
	Micro-moyenne	0,64	0,64	0,64	100