

UNIVERSITÉ TÉLUQ

TRAVAIL MTI6012

PRÉSENTÉ À DANIEL LEMIRE

COMME EXIGENCE PARTIELLE DU COURS MTI 6012 - Travail dirigé

Par Achille KAMDEM

16338394

Classification automatique de texte :

Application à la détection de la fraude sentimentale

31-10-2020

Table des matières

Table des matières	ii
Table des tableaux.....	viii
Table des figures	ix
Table des abréviations.....	x
PARTIE 1 : Étude de la fraude sentimentale : Contexte, anatomie et dynamique organisationnelle de la fraude	i
Chapitre 1 : Contexte, enjeux et problématiques sociales	1
Mise en contexte	1
Enjeux sociaux de la fraude sentimentale	4
Problématique de la fraude sentimentale	9
Sensibilisation des masses	10
Difficultés de mutualisation des dispositions juridiques internationales.....	12
L'insuffisance des approches technologiques existantes	13
L'inexistence d'outils spécialisés	15
Objectifs de notre travail	15
Méthodologie	16
Modélisation de la détection de la FS à partir du corpus dynamique	16
Mise en œuvre de la solution	17

Conclusion prévue et impact industriel/social	17
Chapitre 2 : Fraude sentimentale : Anatomie, organisation et modèle de persuasion	19
La rencontre amoureuse.....	19
L'attirance virtuelle et physique	19
L'amour dans l'espace virtuel	21
La fraude sentimentale	22
Genèse et définition	22
La persuasion et les techniques d'influence	24
Le processus de persuasion.....	30
Le profil des victimes	36
Conclusion.....	37
Chapitre 3 : Dynamique organisationnelle des fraudeurs	38
Les réseaux criminels	38
La structure organisationnelle	38
La collusion avec la légitimité	40
Les principaux foyers d'opérations en Afrique	41
Les conditions de l'émergence de la fraude en Afrique.....	42
La dimension sociale	43
La dimension spirituelle ou le cyber-spiritisme	45
Conclusion.....	47

PARTIE 2 : Détection automatique de la fraude sentimentale	48
Chapitre 4 : Classification automatique de texte : État de l'art	49
Détection de la FS : une tâche de classification automatique de texte	50
Définition et méthodologie de classification de textes	52
Définition	52
Méthodologie de classification automatique de textes	53
Préparation des documents du corpus	55
Nettoyage et normalisation	55
Extraction des caractéristiques	60
Pondération des termes	62
La représentation du corpus	64
Sélection des caractéristiques	65
Les méthodes de sélection	66
La détermination du seuil des métriques	69
Réduction des caractéristiques	69
Indexation sémantique latente	70
Analyse en composantes principales	71
Analyse en composantes indépendantes	72
L'analyse discriminante linéaire	73
Conclusion	75

Chapitre 5 : Apprentissage automatique pour la classification automatique de texte	76
Généralités	77
Positionnement de l'apprentissage automatique	77
Définition	78
Typologie et applications	78
Logique de construction d'un modèle de classification	82
Architecture de l'apprentissage automatique	82
Quelques notions de construction du modèle d'apprentissage	85
Les méthodes d'apprentissage	90
Naïve Bayes	90
K-plus proches voisins	91
Machines à vecteur de support	93
Arbre de décision	98
Les méthodes d'ensemble à base d'arbre de décision	100
Évaluation des classifieurs	104
L'erreur totale	104
Le taux de faux positifs et taux de faux négatifs	105
La précision	106
Le rappel	106
La spécificité	106
Le taux de succès	106

F_β score	106
Macro-moyenne et micro-moyenne	107
Coefficient de corrélation de Matthews	107
La courbe ROC	107
Travaux connexes sur la détection de la FS	108
Conclusion	111
 Chapitre 6 : Modélisation de la détection de la fraude sentimentale	 113
Modélisation automatique continue	113
Collecte des données	114
Nature des données	114
Difficulté de collecte de données de qualité	115
Origine des données	115
Constitution et qualification du corpus	117
Analyse exploratoire du corpus	118
Exploration du langage	118
Description statistique	120
Du prétraitement des données à la construction du modèle optimale	126
Distribution des données et validation croisée	127
Expérimentation 1 : Recherche de la combinaison idéale à contribution individuelle des méthodes	 128
Expérimentation 2 : Recherche de la combinaison idéale à contribution optimisée des méthodes	 135

Discussion sur les résultats	143
Conclusion et perspectives	146
Références.....	150
Bibliographie	157
Annexes	158

Table des tableaux

Tableau 1 : Erreurs dans la prise de décision identifiées	26
Tableau 2 : Sept propositions de la théorie de la transition d'espaces du cybercrime ..	44
Tableau 3 : Matrice de confusion	104
Tableau 4 : Description des métriques de la matrice de confusion.....	105
Tableau 5 : Répartition des documents du corpus	119
Tableau 6 : Caractéristiques de tendance centrale des sentiments	124
Tableau 7 : Caractéristiques de dispersion des sentiments	125
Tableau 8 : Cadre d'expérimentation	127
Tableau 9 : Total des expériences en expérimentation 1	129
Tableau 10 : Métriques de l'expérience 1 du classifieur régression logistique.....	131
Tableau 11 : Métriques de l'expérience 1 du classifieur forêt aléatoire	133
Tableau 12 : Métriques de l'expérience 1 du classifieur SVM	134
Tableau 13 : Métriques de l'expérience 1 du classifieur K plus proches voisins	136
Tableau 14 : Métriques de l'expérimentation 2.....	142
Tableau 15 : Meilleures métriques de performance des expérimentations 1 et 2	143

Table des figures

Figure 1 : Processus de changement d'attitude.	29
Figure 2 : Le modèle des techniques de persuasion utilisées par les fraudeurs.	32
Figure 3 : Exemple de fichier modèle utilisé par les fraudeurs.	51
Figure 4 : Un exemple modèle de lettre.	52
Figure 5 : Processus de classification automatique de documents.	53
Figure 6 : Sous-ensemble de l'intelligence artificielle.	77
Figure 7 : Formation d'un modèle d'AA.	83
Figure 8 : Compromis entre la variance et le biais.	88
Figure 9 : Représentation de la classification K plus proches voisins avec $K = 3$ à partir d'un point xi	93
Figure 10 : SVM avec maximisation de marge à séparation linéaire.	94
Figure 11 : gauche : Trois hyperplans possibles pour séparer les points. Droite : impossibilité de séparer tous les points.	95
Figure 12 : SVM à séparation polynomiale.	98
Figure 13 : Entraînement indépendant de trois modèles sur des ensembles identiques.	101
Figure 14 : Vote dans la forêt aléatoire.	102
Figure 15 : Entraînement séquentiel de trois modèles sur des ensembles.	103
Figure 18 : Nuage des entités nommées des messages légitimes (à gauche) et illégitimes (à droite).	120
Figure 19 : Histogramme des parties du discours des messages légitimes et illégitimes.	121
Figure 20 : Nuage de points de la polarité et la subjectivité des messages légitimes (en bleu) et illégitimes (en rouge).	122
Figure 21 : Données structurées du corpus.	123
Figure 22 : Histogramme des métriques essentielles de performance.	138
Figure 23 : Histogramme des temps d'apprentissage et de test.	139
Figure 25 : Histogramme des métriques de l'expérimentation 2.	142

Table des abréviations

AA	Apprentissage Automatique
ACCC	Australian Competition and Consumer Commission
ACP	Analyse en Composantes Principales
BBC	British Broadcasting Corporation
BCC	Bureau de la Concurrence du Canada
CAC	Centre Antifraude du Canada
CT	Classification de textes
DE	Données d'Entraînement
DT	Données de Test
DV	Données de Validation
EFCC	Economic and Financial Crimes Commission
ELM	Elaboration Likelihood Model of persuasion
FBI	Federal Bureau of Investigation
FNR	False Negative Rate
FPR	False Positive Rate
FS	Fraude Sentimentale
FTC	Federal Trade Commission
IA	Intelligence Artificielle
ICA	Independent Component Analysis
LDA	Allocation Dirichlet Latente
LDA	Linear Discriminant Analysis
LSI	Latent Semantic Indexing
NMF, NNMF	Non Negative Matrix Factorization
PCA	Principale Components Analysis
PL	Prolongement Lexical
POS	Part-Of-Speech
ROC	Receiver Operating Characteristics
SEM	Simple Event Model
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TALN	Traitement Automatique du Langage Naturel
TF	Term Frequency
TF-IDF	Frequency-Inverse Document Frequency

PARTIE 1 : Étude de la fraude sentimentale : Contexte, anatomie et dynamique organisationnelle de la fraude

Chapitre 1 : Contexte, enjeux et problématiques sociales	1
Chapitre 2 : Fraude sentimentale : anatomie, organisation et modèle de persuasion	19
Chapitre 3 : Dynamique organisationnelle des fraudeurs	38

Chapitre 1 : Contexte, enjeux et problématiques sociales

Mise en contexte

L'internet est aujourd'hui un outil indispensable pour des organisations, les entreprises et les particuliers. Comme nous avons souvent remarqué, son utilisation n'est pas exempte de menaces pouvant porter atteinte à la sécurité, la prospérité et la qualité de vie des citoyens. À ce propos, Josephine Palumbo, sous commissaire de la direction des pratiques commerciales trompeuses du bureau de la concurrence du Canada affirme que :

L'économie numérique est en train de changer la façon dont les Canadiens travaillent, vivent et interagissent. Le fait d'être connectés signifie aussi que notre exposition à la fraude a augmenté de façon exponentielle. Personne n'est à l'abri de la fraude, donc nous devons tous apprendre à nous protéger, et garder en tête que le signalement peut aider à prévenir d'autres torts. (BCC¹, 2020).

Ces menaces sont consécutives aux fraudes qui peuvent être de types divers² (abonnements piégés, vol d'identité, fraude du faux PDG, fraude médicale ou liée à la santé, fraude relative aux services de rencontre, fraude aux entreprises, hameçonnage, fraude aux contribuables, etc.). Si toutes ces fraudes sont néfastes, car générant des conséquences importantes sur les victimes, la fraude relative aux sentiments l'est encore plus parce qu'elle froisse la dignité humaine des victimes. Au-delà des conséquences financières, la fraude sentimentale (FS) crée chez les victimes des conséquences psychologiques (Life

¹ <https://www.canada.ca/fr/bureau-concurrence/nouvelles/2020/03/luttez-contre-la-fraude-en-mars-le-bureau-de-la-concurrence-lance-sa-campagne-annuelle-du-mois-de-la-prevention-de-la-fraude.html>

² [https://www.bureaudelaconcurrence.gc.ca/eic/site/cb-bc.nsf/vwapj/CB-IBBS2-FR.pdf/\\$file/CB-IBBS2-FR.pdf](https://www.bureaudelaconcurrence.gc.ca/eic/site/cb-bc.nsf/vwapj/CB-IBBS2-FR.pdf/$file/CB-IBBS2-FR.pdf)

After Scams Ltd, 2019; Yates, 2019). La FS est le fait pour une personne mal intentionnée de créer de faux profils dans les sites de rencontre et, par la suite, d'entrer en communication avec de potentielles victimes, les mettent en confiance et fait naître des sentiments en elles pour enfin solliciter de ces dernières, de l'argent sous des prétextes fallacieux.

Dans le registre des statistiques, l'on note qu'en 2019, elle a coûté aux Canadiens plus de 18,3 millions de dollars et figure sur le palmarès des 10 fraudes³ les plus signalées au Canada. Pour la même période, en Australie, la commission de la concurrence et des consommateurs⁴ a recensé environ 28.6 millions de dollars escroqués et un total de 3 948 signalements. Plus impressionnants, sont les chiffres aux États-Unis, où les consommateurs ont déclaré avoir perdu 201 millions de dollars. On note également une progression de 40 % depuis l'année précédente, positionnant la FS au premier rang des fraudes signalées à la Federal Trade Commission⁵. En Europe, notamment aux Pays-Bas, le bureau d'aide⁶ contre la fraude a obtenu 639 plaintes avec des pertes évaluées à près de 3,74 millions d'euros pour 259 victimes. Enfin, en 2018 et selon la BBC, Action Fraud au Royaume-Uni a enregistré 4 555 signalements de FS, avec⁷ des pertes totales qui ont augmenté de 27 % par rapport à l'année précédente.

³ <http://www.antifraudcentre-centreantifraude.ca/features-vedette/10-frauds-fraudes-fra.htm>

⁴ <https://www.scamwatch.gov.au/types-of-scams/dating-romance>

⁵ <https://www.ftc.gov/news-events/press-releases/2020/02/new-ftc-data-show-consumers-reported-losing-more-200-million>

⁶ <http://www.rfi.fr/fr/technologies/20200125-internet-fraude-sentimentale-ph%C3%A9nom%C3%A8-global>

⁷ <https://www.bbc.com/news/business-47176539>

Compte tenu de l'impact psychologique subi par des victimes, les signalements ne sont pas tous effectifs, cela contribue à minorer l'ampleur du phénomène. Beaucoup d'entre elles ne souhaitent pas divulguer leur cas de fraude, car elles éprouvent de la honte et craignent le jugement de leurs proches.

Voici quelques cas de fraudes sentimentales qui ont défrayé la chronique :

- Sarah⁸, 46 ans, admet avoir envoyé plus d'un million de dollars à son petit ami, « Chris », qu'elle a rencontré en ligne il y a un peu plus d'un an, mais ne l'a pas encore vu en personne.
- Elle croyait qu'« Éric »⁹ était son prince charmant, malheureusement, elle s'est fait dépouiller une somme de 1 000 000 \$.
- L'histoire de Susie, résidant en Caroline du Sud qui s'est fait spolier la bagatelle somme de 99 000 \$ sur le réseau social Facebook par son prétendant virtuel, Howard (Yates, 2019).
- Cecilie Fjellhoy¹⁰ domiciliée à Londres pensait avoir rencontré son prince charmant, malheureusement, elle s'est retrouvée embourbée dans un véritable cauchemar psychologique doublé d'une extorsion d'une valeur de 200 000 \$. Elle vit aujourd'hui avec des dettes et dans l'insécurité.
- Revivons l'histoire de Joy¹¹ qui a consenti 510 000 au nom de

⁸ <https://www.youtube.com/watch?v=Ud8vm01nzMw>

⁹ <https://www.youtube.com/watch?v=g0Uar7glYi8>

¹⁰ <https://abcnews.go.com/US/woman-man-met-tinder-swindled-200k-didnt-dump/story?id=62806053>

¹¹ <https://www.youtube.com/watch?v=EWOmPKZ4tnw&t=67s>

l'amour ou de Juile Ericksen¹² qui a perdu 500 000 \$.

Enjeux sociaux de la fraude sentimentale

L'émergence de la FS est à la base de plusieurs enjeux sociaux qui s'articulent autour de cinq impacts de nature financière, émotionnelle, sociale, physique et politique.

Sur le plan financier, les victimes peuvent perdre une partie ou la totalité de leurs épargnes et actifs, les conduisant à faire face aux dettes, se voir refuser des prêts ou du crédit, avoir besoin de retourner à la population active après l'âge de la retraite, perdre sa propre maison et être sans-abri (Life After Scams Ltd, 2019). De l'aveu de Susie, une victime américaine : « Les arnaqueurs font en sorte que tu tombes amoureuse d'eux. Et puis, finalement, il ne reste plus rien. Tu n'as pas juste perdu ton argent, ton cœur est également brisé. » (Yates, 2019).

Sur le plan émotionnel, les victimes souffrent de détresses émotionnelles qui s'expriment à travers trois émotions dominantes : le chagrin, la honte et l'anxiété.

- **Le chagrin**

Le dépouillement psychologique constitue une perte qui déclenche une tristesse intense. Les victimes assimilent la perte au deuil. La fin d'une relation, la perte de biens importants, la perte d'amitiés, la santé et la perte d'un ancien

¹² https://www.youtube.com/watch?v=_6hKErybcBo

mode de vie peuvent être ressenties à la fois physiquement et mentalement. Le rapport (Life After Scams Ltd, 2019) en fait d'ailleurs cas :

Se sentir à la fois agité et épuisé en même temps, un sommeil de moindre qualité et des difficultés de mémoire et de concentration sont des symptômes courants, ainsi que des étourdissements, des palpitations, des tremblements, des maux de tête, des nausées et des changements d'appétit.

En plus du chagrin causé par la perte, les victimes peuvent également souffrir du rejet de la famille et des amis, qui pourraient croire qu'elles en sont la cause, « qu'elles auraient dû faire attention », « qu'elles auraient dû alerter ses proches », etc. Elles peuvent aussi perdre leur estime de soi, leur confiance et leur sécurité en la société. Sans contrôle, le chagrin et la tristesse chroniques peuvent conduire au désespoir ou la dépression, provoquant ainsi de graves problèmes de santé mentale.

- **L'anxiété**

L'anxiété peut entraîner des problèmes de réflexion, de raisonnement et de concentration. L'anxiété crée des niveaux élevés d'hormone du stress, le cortisol, qui est associé à l'hypertension, aux maladies cardiovasculaires et à certains cancers (Life After Scams Ltd, 2019).

L'anxiété peut être l'un des éléments constitutifs de nombreuses affections (Rossant-Lumbrosso, et Rossant, 2020) telles que les maladies psychiatriques (schizophrénie, dépression, etc.), les maladies vasculaires ou dégénératives (parkinson, épilepsie, etc.), les maladies endocriniennes ou métaboliques (hyper

ou hypothyroïdie, hyper ou hypoparathyroïdie, hypercorticisme, hypoglycémie, phéochromocytome, etc.), les maladies organiques (asthme, angine de poitrine, etc.), l'intoxication et les syndromes de sevrage (corticoïdes, alcool, barbituriques, etc.).

Par ailleurs, les victimes de la FS peuvent avoir une crainte réaliste que des images et photos sexuellement explicites qui ont été données au fraudeur puissent être utilisées pour les faire chanter davantage (notamment en les menaçant de les partager sur les réseaux sociaux) .

- **La honte**

Le sentiment de honte peut compromettre le bien-être mental et physique de la victime. La honte est la cause d'une mauvaise santé, comme l'indiquent les chercheurs Lyons et Lyons (2017) cités par (Life After Scams Ltd, 2019), qui soutiennent que

Les personnes qui ont honte ont une santé plus mauvaise et une espérance de vie plus courte. Comme la honte devient chronique, elle peut entraîner divers effets négatifs sur la santé du fait de la pression physiologique sur le corps et ses systèmes en raison de niveaux chroniques élevés de PIC et de cortisol. Une variété de conditions, telles que prises de poids, maladie cardiaque, durcissement des artères et diminution de la fonction immunitaire peuvent en résulter.

Lorsqu'une personne subit une perte majeure, comme la perte de son conjoint en cas de décès ou de divorce, la perte de sa richesse en raison d'une crise financière mondiale ou la perte sa santé à la suite d'une maladie ou d'un accident, le soutien de ses proches est naturel. Dans ces cas de figure, il y a

généralement beaucoup de soutien de la part de la famille, des amis, des professionnels de la santé, etc. Ce n'est la faute de personne. En revanche, pour la victime de la FS, son inattention est mise en cause, ses proches se disent « c'est sa faute ». Dans ce cas, les victimes peuvent éprouver une honte profonde et sont parfaitement conscientes de la stigmatisation négative associée à la fraude. Susie avoue ceci : « Je me sens folle de vous raconter tout cela. » (Jeff Yates).

La réaction naturelle des personnes éprouvant de la honte est de se retirer et d'éviter d'autres personnes (Greenberg et Johnston, et Pascual Leone) cités par (Life After Scams Ltd, 2019), conduisant inéluctablement la personne à se sentir davantage déconnectée et seule, ce qui contribue à aggraver le traumatisme initial.

Sur les plans social et physique, les pertes pécuniaires subies par les victimes les mettent dans une situation d'indigence financière tout en les déconnectant de toute forme de socialisation. Le difficile maintien des loisirs et des engagements sociaux les expose au risque de perdre des amis et des relations avec la communauté, entraînant in fine, l'isolement. Wilson et al (2007) soutiennent que l'isolation augmente le risque de démence ainsi que la régression cognitive. De plus, une victime isolée vit constamment en mode d'hypervigilance permanente, occasionnant des troubles du sommeil, le risque de morbidité et de mortalité (Hawkey et Cacioppo, 2010). Life After Scams Ltd (2019) explique

néanmoins que l'isolement n'est pas nécessairement un problème si la personne a accès à une source de bienveillance et à des activités joyeuses ou significatives.

De fil en aiguille, l'on arrive aux problèmes de santé mentale, car, en effet, l'isolement social et la solitude sont fortement corrélés aux maladies mentales, car l'on démontre une association forte entre les difficultés financières et les problèmes de santé mentale (Sohrab et Rob, 2017).

Sur le plan de la politique publique, comme nous l'avons vu plus haut, la FS contribue aux problèmes et enjeux de santé mentale. Plus spécifiquement, les gouvernements dont le rôle principal est de garantir le bien-être des populations font face aux problématiques et enjeux de l'isolement social, de la prévention du suicide, de la gestion des sans-abris, de l'éducation financière et du numérique. Autant de questions que soulèvent la FS.

Les répercussions économiques sont non moins importantes. Seulement pour la question de la santé mentale, la Commission de la santé mentale du Canada rapporte que « les maladies et les problèmes coûtent à l'économie canadienne plus de 50 milliards de dollars par an, soit près de 1 400 \$ par personne au Canada en 2016 » (CSMC, 2017). En Europe, « les problèmes de santé mentale engendrent un coût total supérieur à 600 milliards EUR – soit plus de 4 % du PIB – dans les 28 pays de l'Union européenne »¹³. Observons

¹³ <https://www.oecd.org/fr/sante/les-troubles-de-la-sante-mentale-representent-un-lourd-fardeau-economique-pour-les-pays-europeens.htm>

notamment que ces « coûts sont liés aux taux d'emplois et à la productivité moins élevés des personnes atteintes de troubles de la santé mentale (1.6 % du PIB ou 260 milliards EUR) ».

Dans le contexte de santé mentale, où les préjudices financiers et psychologiques en sont les déclencheurs (Sohrab et Rob, 2017, p.8), la productivité des victimes est moins élevée, conduisant inéluctablement à une baisse de rendement des organisations, qui par effet d'entraînement fait baisser les recettes fiscales. Par ailleurs, les victimes qui vivent désormais dans la précarité sont prises en charges sur le plan sociale, cela génère une hausse des prestations sociales qui vient ponctionner davantage les finances publiques. À ce propos, Québec a augmenté les prestations d'aide sociale aux personnes incapables au travail, cette décision¹⁴ augmente ainsi l'effort que doivent consentir les citoyens.

Problématique de la fraude sentimentale

Eu égard aux enjeux sociétaux sus-évoqués, les gouvernements et les associations civiles peinent à apporter des solutions idoines à la prévention la FS. Les aspects sensibilisation des masses, déploiements juridiques et orientations technologiques demeurent encore timides à l'effet de contrer le phénomène. Examinons leurs insuffisances.

14 <https://ici.radio-canada.ca/nouvelle/1141579/contraintes-severes-emploi-augmentation-gouvernement>

Sensibilisation des masses

L'un des aspects importants de la lutte contre la FS demeure la sensibilisation des citoyens. Le Centre antifraude du Canada (CAC¹⁵), l'Australian Competition and Consumer Commission (ACCC¹⁶), la Federal Trade Commission (FTC¹⁷) aux États-Unis, FraudeHelpdesk¹⁸ aux Pays-Bas ou encore l'agence ActionFraud¹⁹ du Royaume-Uni, sont autant d'organismes dont l'une des missions principales est de protéger le consommateur/citoyen par la voie de la sensibilisation dans le but de prévenir les pratiques douteuses.

Tenons par exemple, au Canada, le CAC dont les objectifs sont de perturber les activités criminelles, de renforcer le partenariat entre les secteurs privés et publics, et préserver l'économie canadienne permet aux citoyens de :

- Signaler la fraude ;
- Se renseigner sur différents types de fraude ;
- Reconnaître les indices de fraude ;
- Se protéger contre la fraude.

L'une des campagnes phares de sensibilisation est le « Mois de la sensibilisation à la fraude ». À cette période, plusieurs activités de sensibilisation en matière de fraude sont organisées. « La fraude évolue, restons vigilants » est

15 <https://www.antifraudcentre-centreantifraude.ca/about-ausujet/index-fra.htm>

16 scamwatch.gov.au

17 <https://www.consumer.ftc.gov/>

18 <https://www.fraudehelpdesk.nl/>

19 <https://www.actionfraud.police.uk/>

d'ailleurs le thème retenu pour la 16e édition en 2020. À ce propos, la Banque du Canada, en collaboration avec la Sûreté du Québec et plusieurs autres partenaires des forces policières ont contribué à l'édition du livret « La fraude en 3D – détecter, dénoncer, décourager²⁰ ».

Malgré tous ces programmes et activités, force est de constater l'augmentation sans cesse croissante des cas de fraudes en général, mais également les FS en particulier, comme nous l'avons d'ailleurs si bien observé dans la mise en contexte plus haut. Deux hypothèses, dont l'une de nature psychologique et l'autre, liée à l'espace virtuel peuvent être mises en avant pour tenter d'expliquer l'échec de la sensibilisation :

Les victimes potentielles négligent facilement les « signaux d'alarme », car les états psychologiques prévalent au détriment des effets de la sensibilisation.

Whitty (2018) caractérise ces personnes de la façon suivante :

Les victimes sont généralement des femmes d'âges moyens et bien éduquées. De plus, elles ont tendance à être plus impulsives (obtenant un score élevé sur l'urgence et la recherche de sensations), moins gentilles, plus fiables et ont une disposition addictive.

Du fait de la nature virtuelle de la relation, les partenaires ne peuvent se connaître que par le dévoilement de soi (le fait de choisir l'information et de contrôler la manière et le temps auquel ils souhaitent les divulguer ou les dissimuler). C'est d'ailleurs l'avis de Ben-Ze'ev (2004) cité par Marika (2010), qui ajoute trois autres

²⁰ <https://www.banqueducanada.ca/wp-content/uploads/2019/01/fraude-3d.pdf>

caractéristiques de l'espace virtuel : 1) l'anonymat plus grand et la vulnérabilité réduite, 2) l'inexistence de barrières et 3) la facilité à percevoir l'autre comme étant semblable à soi. Ces caractéristiques montrent que l'espace virtuel contribue à réduire la méfiance des partenaires engagés dans une relation amoureuse.

Malgré les aspects sus-évoqués, la sensibilisation est importante, mais doit vraisemblablement s'accompagner de solutions technologiques qui s'occuperont de contrôler la véracité du dévoilement de soi du partenaire autant au début que pendant la relation.

Difficultés de mutualisation des dispositions juridiques internationales

Recouvrer les fonds spoliés est un processus long, fastidieux et infructueux, c'est du moins le calvaire qu'a vécu la Québécoise Ginette Raymond²¹ qui s'est fait spolier plus de 10 000 \$ par un fraudeur de la Côte d'Ivoire et qui expérimente des difficultés dans le processus de recouvrement de ces fonds. Ses difficultés trouvent une explication par le fait que la FS est en général de nature transfrontalière (les fraudeurs sont généralement hors de l'autorité juridique des victimes compte tenu de la nature du cyberspace) et les dispositions juridiques actuelles pour faire face à ce genre de menaces ne sont pas vigoureuses. Le déploiement juridique de dimension internationale se fait aujourd'hui par des mutualisations sporadiques des forces comme on l'a vu pendant l'opération²²

²¹ <https://www.tvanouvelles.ca/2018/03/12/abandonnee-dans-sa-lutte-contre-larnaque>

²² <https://www.tvanouvelles.ca/2019/09/10/le-fbi-et-le-nigeria-arretent-au-moins-281-cybercriminels>

menée conjointement par la police fédérale américaine (FBI) et les autorités nigérianes (EFCC) et qui a conduit à l'arrestation de près de 300 individus.

Au lieu d'investir dans des mécanismes juridiques à l'effet de recouvrer des sommes spoliées ou de mettre hors d'état de nuire les fraudeurs, il serait à notre sens, préférable d'investir suffisamment dans la prévention de ladite spoliation. L'objectif primordial des fraudeurs est de dépouiller les victimes par des demandes périodiques et souvent inopportunes d'argent, sommes habituellement transférées par l'intermédiaire des agences de transfert d'argent (Western Union²³, MoneyGram²⁴, etc.). Une solution technologique pourrait à la demande des potentielles victimes, identifier ces types de demandes et prévenir les autorités compétentes tout en respectant la vie privée de ces victimes. Rappelons tout de même qu'un important nombre de ces victimes sont des personnes âgées qui ont épargné toute leur vie pour enfin prendre leur retraite méritée et qui de plus méconnaissent l'outil numérique.

L'insuffisance des approches technologiques existantes

Plusieurs travaux existent pour la détection des pourriels (Zhang et al, 2004 ; Cormack, 2008; Sanz et al, 2008; Dada et al, 2019; Bhowmick et al, 2016). Ces travaux utilisent l'apprentissage automatique pour reconnaître les courriels illégitimes. Or les données du profil telles que le nom propre, nom utilisateur, âge, genre, localisation, latitude, longitude, pays, ethnie, occupation, état matrimonial,

²³ <https://www.westernunion.com>

²⁴ <https://www.moneygram.com>

courriel, religion, orientation sexuelle, intention de recherche et une description de la personnalité ne sont disponibles que sur des sites de rencontre en ligne. Cette spécificité rend insuffisantes les solutions proposées pour détecter les pourriels vis-à-vis de la FS, car elles se focaliseront davantage sur le contenu du courriel que sur les aspects du profil des personnes impliquées.

Dans le contexte de la FS, plusieurs travaux récents (Suarez-Tangil et al, 2020; Koen de, 2019; Edwards et al, 2018; Huang et al, 2015) ont retenu notre attention, simplement parce qu'ils sollicitent aussi bien i) les techniques d'apprentissage automatique pour détecter les faux profils que ii) les techniques d'ontologie pour organiser les faux profils à l'effet de comprendre les comportements des fraudeurs.

Malgré les résultats probants obtenus de ces travaux, il reste néanmoins la problématique du nombre élevé de faux positifs et faux négatifs. Une approche plus holistique qui prend en compte non seulement les données du profil, mais également les messages échangés entre les fraudeurs et les potentielles victimes augmenterait les performances de la détection des fraudes sentimentales tout en minimisant les faux positif et faux négatifs. Dans ce travail, nous allons explorer les opportunités qu'offre les communications entre les partenaires dans l'optique de construire un classifieur qui, une fois entraîné sur ces dernières permettra de détecter les messages distillés par les fraudeurs.

L'inexistence d'outils spécialisés

L'un des outils de vérification à la disposition des usagers est l'outil de recherche d'image inversée de Google²⁵ ou celui de TynEye²⁶ qui consiste à rechercher une image sur le web. Lorsque les résultats font apparaître plusieurs sites de rencontre pour la même image et avec des noms différents, cela constitue généralement un « signal d'alarme ». Malheureusement, les usagers négligent ou oublient cet exercice de recherche.

Un outil qui permettrait d'analyser le contenu des échanges des partenaires et d'en notifier à la potentielle victime ou même à une tierce personne (membre de famille, autorité, etc.) constituerait une approche de protection beaucoup plus défensive.

Objectifs de notre travail

L'objectif général du projet est de réaliser un système de détection automatique de la fraude sentimentale afin de mieux protéger les citoyens des conséquences psychologiques et financières qu'elle occasionne.

Plus spécifiquement, il est question de mettre en œuvre les missions suivantes :

- L'élaboration d'un modèle prédictif qui permettra la classification automatique des données textuelles issues des échanges entre les personnes impliquées dans une relation sentimentale.

²⁵ <https://www.google.ca/imghp?hl=fr>

²⁶ <https://tineye.com/>

- L'élaboration d'une application internet qui permettra aux usagers de vérifier l'état frauduleux des messages reçus et de recevoir des rapports analytiques de détections y afférents.

Méthodologie

La construction du système de détection automatique de ladite fraude requiert l'usage de 2 modules :

Modélisation de la détection de la FS à partir du corpus dynamique

Nous commencerons par la constitution du corpus documentaire dynamique (en référence aux messages/courriels échangés entre fraudeur et potentielle victime) en collectant des données à l'aide des outils du « web scraping » via internet et plus spécifiquement des sites traitant de la fraude sentimentale²⁷. Nous allons par la suite construire par étapes successives le modèle vectoriel (Salton et al, 1975) : la préparation des données (segmentation/N-grammes (*Kanaris et al, 2006*), la construction des caractéristiques (Varghese et Dhanya, 2017). Ensuite, nous allons sélectionner quelques modèles d'apprentissage automatique les plus couramment utilisés dans ce type de contexte (Koen de, 2019) tels que : forêt de décision, naïve bayésienne, SVM, arbre de décision et régression logistique. Enfin, nous partitionnerons le corpus afin d'expérimenter et évaluer la qualité desdits classifieurs à l'effet de ne retenir que celui qui performera le mieux.

²⁷ stop-scammers.com, romancescam.com

Mise en œuvre de la solution

Nous souhaitons implémenter la solution applicative afin de la mettre à la disposition des usagers. À cet effet, nous envisageons de développer un système, autrement dit d'architecturer, de concevoir et d'implanter une application web qui permettra aux usagers et opérateurs du système d'obtenir quatre fonctionnalités essentielles :

- Un dispositif d'expérimentation des classifieurs ;
- L'orchestration de l'analyse et de la détection des messages frauduleux au moyen du modèle prédictif déterminé ;
- L'édition des rapports analytiques des résultats ;
- L'édition de l'historique des opérations de vérification de l'utilisateur.

Conclusion prévue et impact industriel/social

La solution informatique dont ce travail sera la finalité est destinée à protéger les usagers des conséquences de la fraude sentimentale. Ses bénéfices sociaux s'articulent autour des aspects suivants :

- La protection des aspects psychologiques et financiers des citoyens durant leurs relations amoureuses en ligne.
- La réduction du taux de victimisation, ce qui aura une incidence sur 1) la réduction substantielle des prestations sociales et des coûts de soutien aux victimes (logements sociaux, prévention du suicide, resocialisation,

santé mentale, etc.) et 2) le maintien de la productivité globale des organisations.

- La prévention de la revictimisation des victimes.

Chapitre 2 : Fraude sentimentale : Anatomie, organisation et modèle de persuasion

Avant d'aborder les solutions à la problématique sus-évoquées, nous allons nous attarder à la compréhension du phénomène en question. Il s'agit d'emblée d'une question sociale située à l'intersection des domaines de la cybersécurité et de la psychologie. À cet effet, nous mettrons à contribution les meilleurs experts desdits domaines en épluchant les travaux scientifiques récents en la matière. Nous débuterons en explorant les entrailles de la cyber-relation amoureuse tout en le comparant à son pendant physique. Par la suite, nous nous arrêterons pour définir et expliquer la fraude sentimentale ainsi que l'étude des caractéristiques psychologiques des victimes. Enfin, nous présenterons le modèle de persuasion développé par la cyber-psychologue Monica Whitty qui nous aidera à comprendre les techniques utilisées par des fraudeurs et surtout les erreurs de jugement que font les victimes.

La rencontre amoureuse

L'attirance virtuelle et physique

Selon Levine (2000) cité par Marika (2010), l'attraction comporte généralement 5 composantes : la proximité et la fréquence des contacts, la présentation de soi, la similarité, la réciprocité, les attentes et les idéalizations.

Ces composantes sont toutes aussi valables dans un environnement virtuel à quelques nuances près. Si la proximité se définit dans un lieu géographique, en environnement virtuel, l'on parle davantage de sites de rencontre, des médias

sociaux, des applications de messageries électroniques. La fréquence des contacts (rencontres, clavardages, échanges) dans l'espace virtuel attire tout autant. En revanche, en ce qui concerne la présentation de soi, les personnes sur internet peuvent choisir l'information et contrôler la manière et le temps auquel elles souhaitent la divulguer ou la dissimuler; permettant ainsi de faire intervenir l'anonymat dans le profil et même dans le processus de dévoilement à l'autre. Dans l'espace physique, la dissimulation est plus contraignante. La similarité est le fait de partager les mêmes attitudes et valeurs, or sur internet, la validation du comportement des personnes est complexe; à cet égard, Levine (2000) cité par Marika (2010) conseille de passer de la relation virtuelle à la vie réelle à l'intérieur d'un mois, afin de minimiser les déceptions et fournir une « base de réalité » qui rendra plus probable l'établissement d'une relation à long terme.

Aimer une personne qui nous aime en retour, s'appelle la réciprocité. Elle est tributaire du dévoilement de soi et est intense en contexte virtuel plus que dans la vie réelle. Quant aux attentes et les idéalizations, elles semblent être un indicateur de développement amoureux vis-à-vis d'une personne avec qui l'on communique, car Marika (2020) souligne que le simple fait de penser à une personne qui nous attire a pour effet d'attribuer des qualités à cette personne. Cela devient particulièrement central dans le monde virtuel, où, face à un partenaire, l'attraction est basée exclusivement sur les cognitions, les perceptions et les croyances autogénérées (Levine, 2000) cité par Marika (2020).

Comme nous pouvons le constater, la similarité, la réciprocité, les attentes et les idéalizations sont des facteurs qui se manifestent et s'intensifient au gré du dévoilement de soi, de la proximité et de la fréquence des contacts des partenaires impliqués dans la relation.

L'amour dans l'espace virtuel

La manifestation de la fraude sentimentale sur internet fait intervenir quatre facteurs : un ou plusieurs fraudeurs organisés en réseau, une victime, un espace virtuel complètement anonyme du point de vue des profils des intervenants et le stratagème de fraude. L'espace virtuel et la vie réelle sont autant des espaces propices à l'attraction et à la séduction. Contrairement à la vie réelle, l'environnement virtuel est sujet à l'anonymat du profil tout en permettant la multiplicité des relations.

L'effet de l'anonymat

L'honnêteté et le dévoilement de soi sont deux comportements qui peuvent être mis à mal dans une relation amoureuse sur internet du fait de l'anonymat des échanges.

Concernant l'honnêteté, Whitty (2002) révèle selon son étude, que les hommes étaient plus susceptibles que les femmes de mentir et ce, principalement sur leur statut socio-économique. Wright (1999) cité par Marika (2020), l'explique par le fait que les femmes sont plus attirées par des hommes intelligents, ambitieux et ayant un statut socio-économique élevé. En revanche, les femmes étaient plus susceptibles que les hommes de mentir pour des raisons de sécurité

sans doute pour éviter d'être géographiquement localisées. L'étude de Whitty (2002) affirme également que moins on passe de temps à échanger, plus, nous avons tendance à mentir. Contrairement à ce que nous allons découvrir par la suite, les personnes mal intentionnées investiront au contraire plus de temps pour mentir davantage afin d'arriver à leur fin.

Du fait de la nature virtuelle, les partenaires ne peuvent se connaître que par le dévoilement de soi. C'est d'ailleurs l'avis de Ben-Ze'ev (2004) cité par Marika (2020), qui rajoute trois autres caractéristiques de l'espace virtuel : 1) l'anonymat des caractéristiques physiques, 2) la réduction de la vulnérabilité du fait de l'inexistence de barrières et 3) une plus grande facilité à percevoir l'autre comme étant semblable à nous.

Multiplicité de partenaires

Internet est l'espace par excellence où des personnes peuvent faire cohabiter des relations multiples sans aucune friction. C'est d'ailleurs le média privilégié de ces internautes comme l'indique une fois de plus Levine (2000) cité par Marika (2020), les internautes attirés par plusieurs personnes ou développant plusieurs relations via Internet ont tendance à utiliser régulièrement ce média.

La fraude sentimentale

Genèse et définition

La solitude et le besoin d'aimer et de se sentir aimer poussent aussi bien les femmes que les hommes à aller chercher des solutions sur des sites de

rencontres. Malheureusement, ces sites regorgent des individus dont l'intention première est d'escroquer les âmes sensibles et avides d'affection.

Le phénomène est apparu vers les années 2000 et trouve ses origines dans la fraude par courrier électronique (Whitty and Buchanan, 2012a). La fraude sentimentale fait partir de la fraude par marketing de masse. Cette dernière exploite les techniques de communication de masse (par exemple, courriel, messagerie instantanée, courriels en masse, sites de réseautage social) pour tromper et soutirer de l'argent aux personnes. L'une des plus connues d'entre elles est la 419 (escroquerie par courriel au Nigéria), qui existait sous forme de lettres avant Internet.

Le stratagème est le même, quel que soit le pays. Les fraudeurs créent de faux comptes sur des sites de rencontre et sur les réseaux sociaux en utilisant des photos d'hommes ou de femmes récupérées sur internet. Par la suite ils entrent en communication avec de potentielles victimes, entretiennent ainsi des échanges pendant des semaines (voire des mois) à l'effet de les mettre en confiance et faire naître des sentiments. Puis, ils sollicitent de leur victime de l'argent sous des prétextes fallacieux d'une aide visant à se sortir d'une situation problématique nuisant au bon déroulement de la relation prétendument amoureuse. Ces motifs sont légion, l'on peut citer entre autres : l'achat d'un visa ou d'un billet d'avion, payer des frais médicaux d'urgence ou aider sa famille ou même la location d'un appartement au cas où la victime déciderait d'aller voir son

prétendu amoureux. Ces escrocs maîtrisent bien les outils informatiques et sont spécialistes de la retouche d'images et des flux vidéo.

La fraude sentimentale est donc un vol par la ruse qui s'opère dans l'espace cybernétique via des systèmes de messagerie électronique, médias sociaux, sites de rencontres, et autres sites internet. Elle met en exergue deux entités, l'arnaqueur/fraudeur et sa victime et touche généralement des personnes âgées, à situation familiale séparée ou divorcée.

La terminologie utilisée au Canada notamment par le centre antifraude du Canada est celle de stratagème de rencontre (ou escroquerie relative à un service de rencontre en ligne) ou encore fraude sentimentale. Certains auteurs utilisent les termes escroquerie, fraude ou arnaque pour désigner le délit et les termes sentimentale ou amoureuse pour signifier la typologie. Dans ce travail, nous allons utiliser les termes fraude sentimentale, qui nous paraissent plus appropriés au contexte cybernétique.

La persuasion et les techniques d'influence

Les personnes qui s'adonnent aux pratiques de fraude sur internet utilisent habituellement des techniques de persuasion dont les fondements sont du domaine de la psychologie sociale. Si Whitty (2013) s'accorde à dire qu'il n'existe aucune théorie qui examine explicitement la fraude par marketing de masse et qu'à la place, les chercheurs ont tendance à s'inspirer des théories établies développées par les psychologues sociaux et à les adapter à la fraude par marketing de masse et non spécifiquement à la fraude sentimentale, elle propose

néanmoins quatre aspects qui relèvent des techniques destinées à contraindre les victimes à céder aux pressions des fraudeurs. Ce sont des erreurs dans la prise de décisions, le pari à long terme, le modèle de vraisemblance d'élaboration, et la communication virtuelle.

Les erreurs dans la prise de décisions

Lea et al. (2009a) affirme que les victimes font des erreurs de jugement devant l'offre de fraude et que les fraudeurs multiplient les situations afin d'augmenter la probabilité d'une mauvaise prise de décision. Ils proposent d'ailleurs au terme de leurs études une compréhension plus élaborée des mécanismes psychologiques de la fraude en général. Ils permettent de distinguer comme le montre le tableau 1, deux types de processus : cognitifs (par exemple, lorsqu'on fait excessivement confiance à l'autorité) et motivationnels (lorsqu'on manque de maîtrise de soi).

Le pari à long terme

Miser moins pour gagner plus, tel est le constat que font Lee et al. (2009a) lorsqu'ils observent que :

Certaines personnes considéraient la réponse à une arnaque comme un pari à long terme : elles reconnaissaient qu'il y avait quelque chose de mal avec l'offre, mais la taille du prix possible (par rapport à la dépense initiale) les a incitées à l'essayer. La chance de réussir.

Dans ces conditions, les victimes étaient susceptibles de réguler et de résister aux émotions associées aux offres de fraude. Notons également que ces victimes ne révèlent pas leur implication à leurs proches (familles, amis, etc.), gardant ce

Tableau 1 : Erreurs dans la prise de décision identifiées

Motivationnel	Cognitif
Influences viscérales	Capacités cognitives réduites
Motivation réduite pour le traitement de l'information	Illusions positives
Préférence pour confirmation	Connaissances de base et confiance excessive
Manque de maîtrise de soi	Activation de la norme
Régulation de l'humeur et fixation fantôme	Autorité
La recherche de sensations	La preuve sociale
Aimer et ressemblance	Retransmission
Réciprocité	
Engagement et cohérence	

Source : (Traduit de Lee et al. (2009a; p. 26))

qu'elles croient être une opportunité loin des leurs. Dans le cas de la fraude sentimentale, où l'intimité est de mise, cela peut expliquer le renfermement des victimes et donc la possibilité pour le fraudeur de garder aussi longtemps sa victime dans les mêmes états motivationnels et cognitifs, loin des influences extérieures.

Le modèle de vraisemblance d'élaboration cognitive

Le modèle théorique de la *vraisemblance d'élaboration (ELM en anglais*

pour Elaboration Likelihood Model of persuasion) est un modèle utile pour comprendre pourquoi les personnes sont arnaquées par la fraude par marketing de masse. Il a été formulé dans le but d'offrir une théorie générale du changement d'attitude et décrit les modalités de formation de l'attitude et de la persuasion suivant que la motivation et le degré d'implication du sujet sont importants ou faibles.

Le modèle met en exergue deux voies, la voie centrale et la voie périphérique. Le processus de l'itinéraire central nécessite que l'individu soit suffisamment motivé à traiter l'information et qu'il fasse preuve d'une haute élaboration alors que le processus de la voie périphérique n'implique pas d'élaboration, mais s'appuie davantage sur des signaux périphériques et des raccourcis mentaux pour contourner l'argument logique et le contre-argument et chercher à déclencher l'acceptation sans réfléchir profondément au message. La figure 1 présente le modèle théorique de la vraisemblance d'élaboration cognitive.

Whitty (2013) affirme que l'utilisation de la voie centrale est faite par des personnes qui jugent de l'importance du message et par conséquent sont motivées à y investir des ressources cognitives adéquates pour leur traitement. Et que la faiblesse des canaux d'informations et des raisons personnelles contribue à utiliser les voies périphériques. Barbier et Fointiat (2015, p. 6) soutiennent que :

La persuasion (c.-à-d. le changement d'attitude) obtenue à la faveur d'un traitement central de l'information est plus persistante, plus résistante aux tentatives ultérieures de contre-persuasion, et prédit

mieux les comportements futurs que le changement d'attitude obtenu à la faveur d'un traitement périphérique de l'information.

Dans un contexte de fraude, Rusch (1999) soutient que tout stratagème implique une offre d'une valeur qui dénature leur qualité et leur caractéristique objective. Ce faisant, les fraudeurs ne peuvent pas se permettre d'utiliser la voie centrale. Le chemin périphérique est impérativement privilégié en vue de maximiser plus rapidement les contraintes motivationnelles et cognitives responsables de la mauvaise prise de décision de la victime. Les victimes ne sont donc pas susceptibles de se préoccuper des signaux cognitifs dès lors que les fraudeurs suscitent de fortes émotions, telles que l'excitation et la peur.

La communication virtuelle

La communication virtuelle utilise des dispositifs numériques comme transmetteur et médiateur. Les SMS, la messagerie instantanée, les forums, les jeux en ligne multi-joueurs, le courrier électronique, le web, Wikipédia font partie de cette forme de communication²⁸. L'utilisation intensive de cette forme de communication permet d'expliquer pourquoi des personnes sont susceptibles d'être victime de la fraude sentimentale, c'est la théorie hyperpersonnelle.

La théorie hyperpersonnelle de Walther (Walther 1996, 2007) cité par Whitty (2013) se définit par les termes suivants :

Les récepteurs idéalisent les partenaires en raison des messages qu'ils reçoivent qui, selon eux, démontrent la similitude de leur partenaire en ligne ainsi qu'un caractère hautement souhaitable. En revanche, les expéditeurs exploitent la technologie pour présenter de

²⁸ https://fr.wikipedia.org/wiki/Communication_virtuelle

manière sélective des aspects d'eux-mêmes que l'autre jugerait socialement souhaitables.

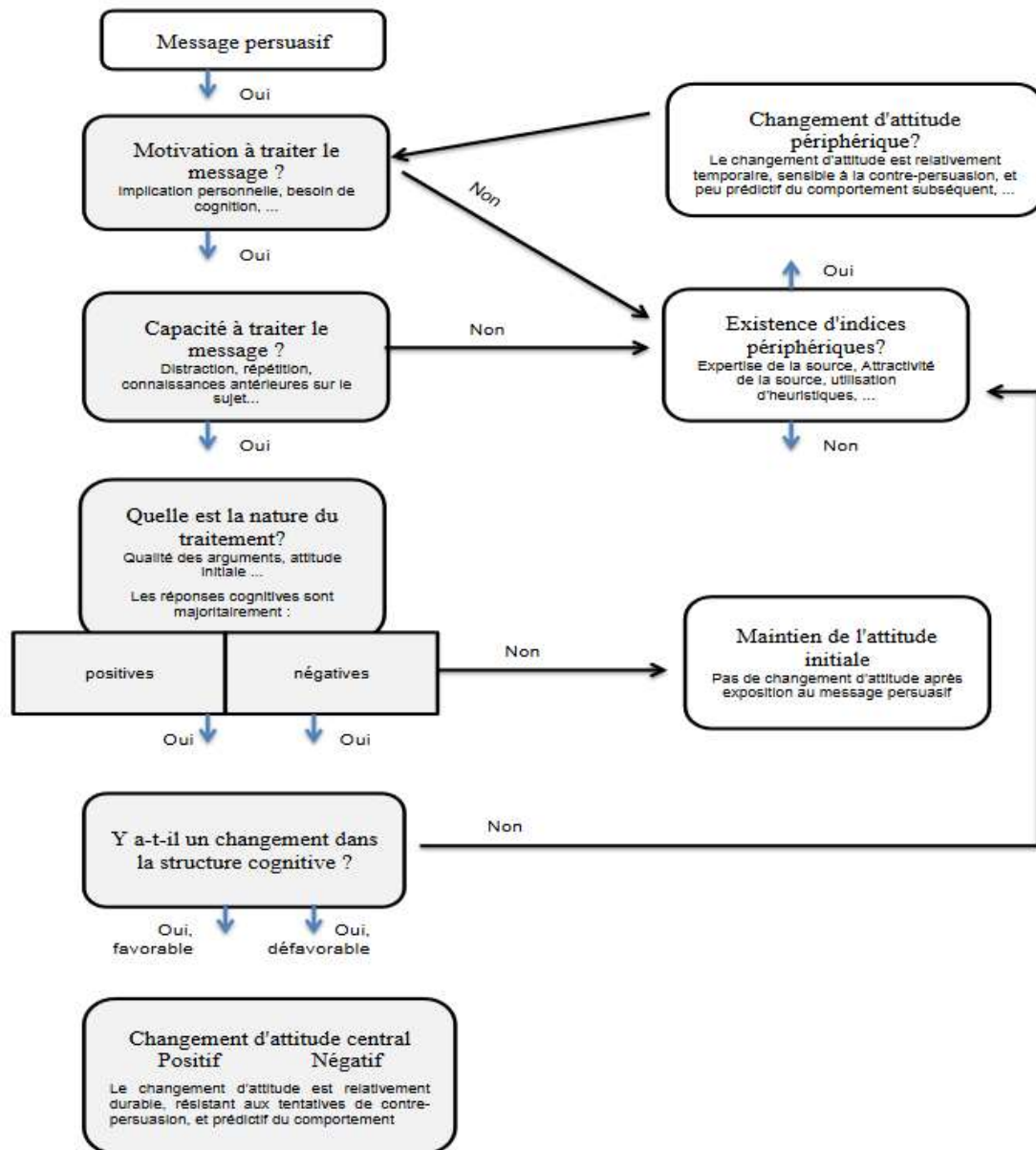


Figure 1 : Processus de changement d'attitude.
Source : (Barbier et Fointiat, 2015, p.7)

Cette théorie permet d'expliquer pourquoi dans certains environnements en ligne, les individus développent des « relations hyperpersonnelles » (c'est-à-dire des

relations plus intimes que celles qui seraient autrement vécues dans un contexte physique). L'opportunité est ainsi donnée aux fraudeurs de construire stratégiquement aussi bien leur présentation que la préparation psychologique de leur potentielle victime.

Le processus de persuasion

Comprendre le processus de la fraude, de sa genèse à son aboutissement, autrement dit le *modus operandi* utilisé par fraudeurs pour opérer en toute confiance, requiert une analyse approfondie des techniques de persuasion utilisées à un moment donné de la communication avec leurs victimes. Tel est le travail effectué sur un ensemble de participants par Whitty (2013). Elle identifie sept étapes au cours desquelles la victimisation se construit (voir la figure 2).

- **Étape 1 : Motivé pour trouver le partenaire idéal**

La première étape exprime le besoin émotionnel des partenaires. L'étude montre que les victimes ont évoqué l'espoir de trouver le partenaire idéal. Certains participants étaient célibataires et cherchaient depuis des années le partenaire idéal, tandis que d'autres n'avaient quitté leur relation que récemment (Whitty, 2013). Chacun d'eux idéalisait un type de partenaire spécifique et semblait motivé à le retrouver.

- **Étape 2 : Présenté le profil idéal**

À ce stade, les fraudeurs mettent alors à la disposition des victimes des profils qui correspondent à la représentation qu'elles se font du partenaire idéal. Dans la présentation des faux profils, la photo joue un rôle prépondérant.

Pour les profils féminins, l'on retrouvait habituellement des photos de femmes légèrement vêtues alors que pour les hommes, des photos d'hommes en uniforme (par exemple, un haut gradé de l'armée). En ce qui a trait à la description des profils, l'on note pour les profils féminins, l'utilisation des emplois peu rémunérés comme infirmière, étudiante, tandis que pour les hommes, la description présente une personne veuve avec enfant, le plus souvent avec un emploi mieux rémunéré (par exemple homme d'affaires, responsable marketing).

Whitty (2013), souligne que les femmes recherchent un partenaire avec un statut socio-économique élevé et que les hommes recherchent une partenaire physiquement attrayante. Les fraudeurs le savent, fabriquent et présentent de faux profils respectueux des codes sociaux et suffisamment attrayants pour attirer les victimes.

- **Étape 3 : La manipulation**

À l'étape précédente, un faux profil est présenté à la victime. À ce stade, le fraudeur va prendre du temps et mettre en œuvre des connaissances nécessaires pour présenter la personne idéale (mentionné dans le faux profil) à la victime. Cette étape est destinée à préparer la victime en vue de lui soutirer de l'argent. Whitty (2013) affirme que cette étape renferme des similitudes avec la façon dont un délinquant sexuel peut soigner un enfant. Elle cite d'ailleurs Gillespie (2002) qui va définir cette préparation comme étant :

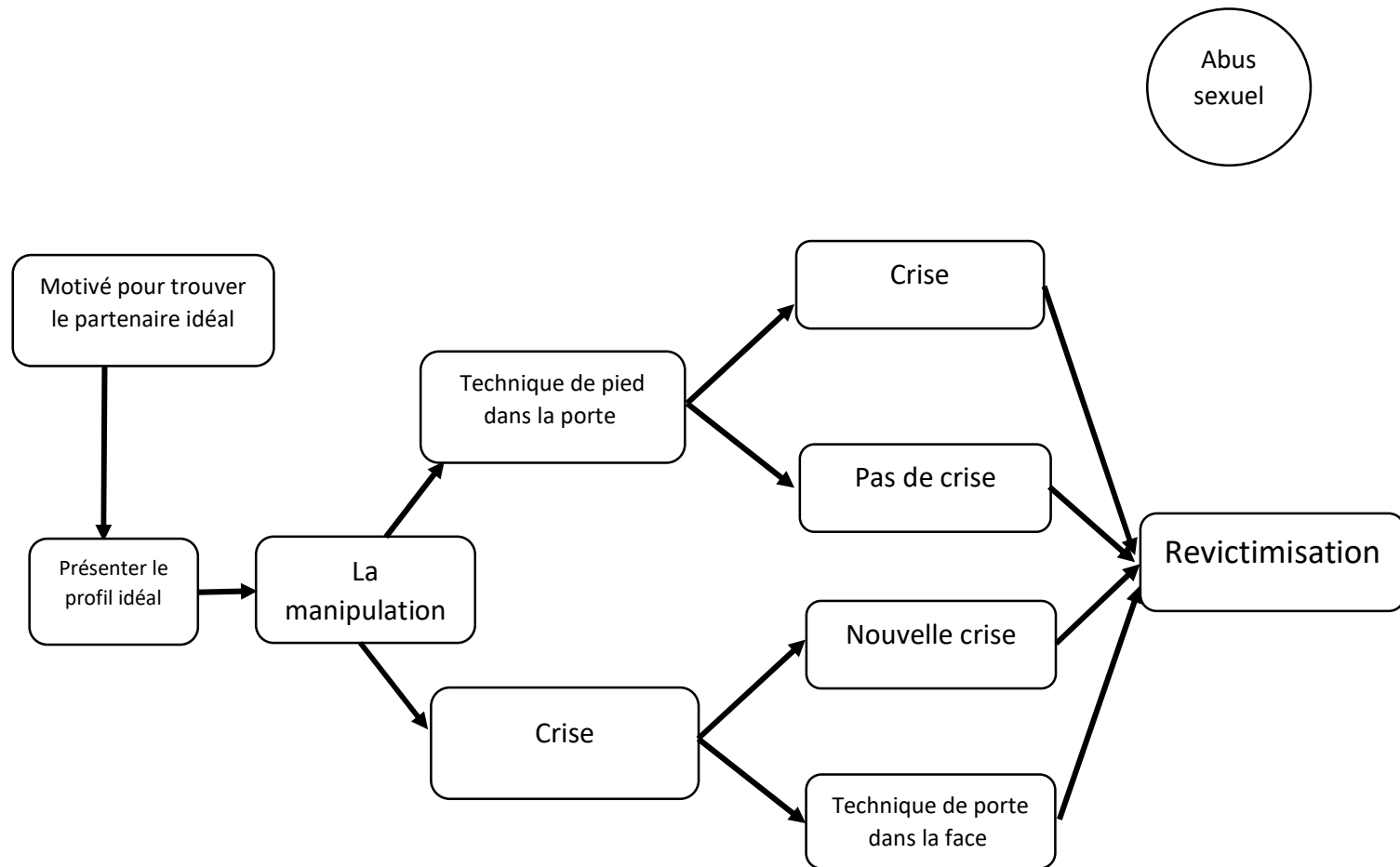


Figure 2 : Le modèle des techniques de persuasion utilisées par les fraudeurs.
Source : traduit de (Whitty, 2013, p 22).

un délinquant sexuel peut soigner un enfant. Elle cite d'ailleurs Gillespie (2002) qui va définir cette préparation comme étant :

Un processus par lequel un enfant se lie d'amitié avec un agresseur potentiel dans le but de gagner la confiance de l'enfant, lui permettant d'obtenir que l'enfant acquiesce à une activité abusive. C'est souvent une condition préalable pour qu'un agresseur ait accès à un enfant. (Whitty, 2013, p. 411)

Les fraudeurs préparent leur victime, gagnent leur confiance en leur fournissant un environnement apparemment sûr pour révéler leurs pensées, secrets et insécurités les plus profonds. Des courriels poétiques ont été utilisés pour courtiser les victimes et la messagerie instantanée a été utilisée pour authentifier la relation et rapprocher la victime d'eux. De plus, en faisant de la relation une routine de leur vie (avec les conversations matinales et nocturnes), un fort attachement se crée, ce qui rend plus difficile la rupture.

Dans cette phase de manipulation, la communication virtuelle engendre une « relation hyperpersonnelle » telle que décrite par Walther (1996, 2007) cité par Whitty (2013) qui se développent, contribuant ainsi à embourber davantage la victime et donc, à diminuer sa capacité à se soustraire de la relation dans les moments de soupçons. Notons que vers la fin de cette étape, le fraudeur pourrait tester la victime en demandant un cadeau. Si la victime se conforme, le fraudeur passera à l'étape suivante (Whitty, 2013). Cette étape peut s'étaler sur plusieurs mois.

- **Étape 4 : La piqure**

À l'issue de l'étape précédente, la victime bien préparée est prête à une

tentative d'extorsion d'argent. Au bon vouloir du fraudeur, il peut emprunter l'une des deux options suivantes : la première est une technique dite de « pied dans la porte » ou « doigt dans l'engrenage » ou encore phénomène du premier pas et la seconde est la provocation d'une crise. Selon Wikipédia²⁹, la première option

est une technique de manipulation décrite par les psychologues sociaux. Elle consiste à faire une demande peu coûteuse qui sera vraisemblablement acceptée, suivie d'une demande plus coûteuse. Cette seconde demande aura plus de chance d'être acceptée si elle a été précédée de l'acceptation de la première, qui crée une sorte de palier et un phénomène d'engagement.

L'étude de Whitty (2013) montre qu'il s'agissait des demandes de cadeaux ou de petites sommes d'argent dans le but de tester la victime en vue des demandes plus importantes, si la victime avait été préparée avec succès, elle se conformait à donner de l'argent, sinon elle tentait de mettre fin à la relation, dans ce cas de figure, le criminel revenait à l'étape 3 et tentait de poursuivre le processus de préparation jusqu'à ce que la victime soit prête à donner de l'argent.

La deuxième option qui est la provocation d'une crise intervient lorsque le fraudeur pense avoir suffisamment bien préparé la victime pour espérer soutirer de gros montants d'argent. L'étude montre qu'après seulement un mois de préparation, une victime a donné au fraudeur un montant de 90 000 £.

- **Étape 5 : La poursuite de la fraude**

Wikipédia³⁰ définit la « **porte dans la face** » ou encore « **porte au nez** »

²⁹ <https://fr.wikipedia.org/wiki/Pied-dans-la-porte>

³⁰ <https://fr.wikipedia.org/wiki/Porte-au-nez>

comme étant une variante inverse de la technique de manipulation du « pied dans la porte » cité plus haut et consiste à faire précéder une demande de comportement plus ou moins coûteuse par une demande beaucoup plus coûteuse, parfois même fantaisiste. Telle est la technique utilisée par les fraudeurs en réponse aux victimes qui ne se sont pas exécutées aux demandes d'argent relatives aux crises à l'étape 4 (la piqure). L'étude révèle qu'à la suite du succès de cette technique, les fraudeurs avaient deux options, soit ils présentaient une nouvelle crise ou qu'ils continuaient à demander des sommes raisonnables.

Les victimes qui avaient succombé aux crises à l'étape précédente se voyaient présenter de nouveau une crise. Quelle que soit l'option choisie, les fraudeurs auront deux voies, la crise ou les demandes de petites sommes d'argent jusqu'à ce que les victimes quittent la fraude ou qu'elles soient abusées sexuellement.

- **Étape 6 : Abus sexuel**

À l'issue de l'étape précédente, une fois que la victime ne donne plus d'argent, le fraudeur amplifie le lien sexuel dans la relation et demande à la victime de se dénuder et de réaliser des actes sexuels devant une webcam. Une fois terminée, une vidéo de la scène est capturée et est utilisée pour faire chanter la victime en menaçant de la partager avec les proches de cette dernière.

- **Étape 7 : La revictimisation**

Whitty (2013) explique que les victimes qui sont arrivées à l'étape 7 ont

sauté l'étape 6. Certaines victimes ont été informées qu'elles étaient impliquées dans un cas de fraude. Malgré cela, pour de multiples raisons, elles sont revenues dans le processus de fraude pour finalement se faire arnaquer de nouveau.

Il pouvait s'agir du fraudeur qui expliquait à sa victime qu'il l'avait bel et bien arnaquée et que dans le processus, il était tombé amoureux d'elle. L'incrédulité de la victime aveuglée par ses sentiments la conduit vers un nouveau cycle de victimisation. Il pouvait s'agir également d'une communication envoyée par le fraudeur lui-même, se faisant passer pour un agent de police, prétextant que le fraudeur avait été appréhendé et qu'il fallait un montant d'argent pour appliquer la loi et récupérer ainsi les sommes soutirées. C'est ainsi que la victime se voyait spolier de nouveau.

Le profil des victimes

Plusieurs des victimes de fraude sentimentale ont des caractéristiques psychologiques qui ont contribué à leur prévalence. Whitty (2018) a effectué des recherches afin de déterminer ces dernières. Précisément au Royaume-Uni où une étude sur 12 060 participants a été menée en comparant les victimes de la fraude sentimentale à celles qui n'avaient jamais été arnaquées. Douze facteurs propices à la fraude sentimentale se sont dégagés, notamment : l'éducation, la connaissance en cybersécurité, le manque de préméditation, l'urgence, la recherche de sensations, le manque de persévérance, le lieu de contrôle, la confiance en autrui, la fiabilité, la gentillesse, la cupidité, la dépendance et la disposition.

De façon descriptive, on avait 60% des femmes qui avaient été victimes d'une arnaque amoureuse, contre 40% des hommes. En ce qui concerne l'âge, 21% des victimes étaient des personnes plus jeunes, 63% d'âge moyen et 16% des personnes âgées.

Finalement, elle a trouvé que les victimes sont généralement des femmes d'âge moyen et bien éduquées. De plus, elles ont tendance à être plus impulsives (obtenant un score élevé sur notion d'urgence et la recherche de sensations), moins gentilles, plus fiables et ont une disposition addictive (Whitty, 2018).

Conclusion

Malgré la distance physique, la relation amoureuse en ligne montre une forme d'attirance sentimentale très forte du fait de la théorie de relation hyperpersonnelle. Nous avons en outre mis en exergue les caractéristiques psychologiques des victimes dont les plus importantes sont la recherche de sensations fortes et le caractère addictif. La détermination des fraudeurs les pousse à mettre en place des stratégies de fraude constituées de plus en plus des techniques relevant du domaine de la psychologie telles que la persuasion et l'influence. Fort des caractéristiques psychologiques de certaines personnes, Nul doute que ce phénomène a encore du chemin à faire.

Chapitre 3 : Dynamique organisationnelle des fraudeurs

L'exercice de la fraude qu'elle soit à caractère sentimental ou non, nécessite l'implication d'un ou plusieurs fraudeurs. Plusieurs d'entre eux travaillent en coordination et en coopération pour créer d'énormes réseaux de fraudes et partager ainsi des données et des techniques éprouvées comme celles présentées au chapitre précédent.

Dans ce chapitre, nous allons aborder deux aspects de la dynamique organisationnelle des fraudeurs. En premier, nous ferons une incursion dans l'organisation des fraudeurs pour expliciter sa structure agile et son caractère hautement criminogène. Dans le deuxième aspect, nous irons visiter les principaux foyers de conception de la fraude dont les plus éminents selon la littérature ambiante se trouvent en Afrique de l'Ouest. Dans ces contextes, nous allons nous apercevoir que l'organisation se résume en une technostructure à très forte densité de spiritisme.

Les réseaux criminels

La structure organisationnelle

Certains chercheurs s'accordent à dire que ce groupe de fraudeurs s'apparente aux réseaux criminels. C'est en effet l'avis de Rege (2009) qui y trouve des similitudes frappantes. Il mentionne par exemple que ces réseaux sont flexibles et n'ont pas de critères d'adhésion ni de codes de groupe. Le recrutement est fait en fonction des besoins et uniquement pour l'opportunité de la fraude dont le seul but est de gagner de l'argent. (Rege, 2009) cite

abondamment le chercheur canadien, Lemieux (2003) spécialiste des réseaux criminels, qui identifie les principaux rôles inhérents à la plupart des réseaux criminels.

Au cœur d'un réseau criminel se trouvent les **organiseurs** qui déterminent la nature et la portée des activités (Lemieux, 2003) cité par Rege (2009).

Les **extensionneurs** sont responsables de l'expansion du réseau criminel. Ils recrutent de nouveaux membres et encouragent la collaboration avec d'autres entreprises illégales, le gouvernement et la justice. Rege (2009) précise en effet, que dans les réseaux de fraudeurs, plusieurs personnes peuvent agir à la fois comme organisateurs et extensionneurs. Les extensionneurs recrutent des membres avec des compétences variées (dactylographie, maîtrise de l'anglais, piratage informatique) selon les besoins pour assurer le bon fonctionnement global de l'entreprise criminelle.

Les **exécuteurs** sont responsables de la réalisation des objectifs de l'organisation. Ils mettent en œuvre des attaques selon les plans établis par les organisateurs et possèdent des compétences spécialisées nécessaires pour mener à bien l'opération (Rege, 2009). Dans les réseaux d'escroqueries sentimentales, les exécuteurs étaient ceux qui pouvaient parler des langues étrangères, rédiger des lettres coquettes, rédiger des courriels et parler au téléphone avec leurs victimes pour les inciter à participer aux escroqueries.

Les **gardiens** sont les protecteurs du réseau criminel et prennent les mesures nécessaires pour assurer la conformité des victimes (Lemieux, 2003)

cité par Rege (2009). Dans le contexte de la fraude sentimentale, l'usage du chantage émotionnel et de l'extorsion est fait afin de contraindre les victimes à se départir de leur argent, tel que nous l'avons observé dans les techniques de persuasion au chapitre précédent.

Les **déménageurs d'argent** collectent de l'argent auprès des victimes et le reversent à leur entreprise criminelle. L'argent transite habituellement par des compagnies de transfert d'argent (Western Union, Moneygram, etc.), et ce, comme nous allons le voir plus bas, avec la complicité des personnes qui exercent dans ces compagnies.

La collusion avec la légitimité

La flexibilité de cette forme d'organisation en réseau contribue à son dynamisme et à son efficacité, lui permettant de mener à bien des activités tout en minimisant les risques d'implosion ou d'influences extérieures. Tade (2011) fait d'ailleurs allusion au réseau informel en soulignant une fois de plus la notion de spécialisation des membres, enjeu fondamental qui permet de maîtriser des techniques de fraude et d'augmenter les chances de succès du processus. Il précise que le réseau est articulé autour des entreprises de transfert d'argent, des agences de sécurité, des co-fraudeurs et parfois, des familles.

Afin de se donner une apparence de légitimité, le réseau n'hésite pas à recruter dans les organisations gouvernementales. Ceux-ci fournissent un accès privilégié et une légitimité inestimable aux faux documents officiels (des passeports, des visas, documents gouvernementaux disposant du papier à

entête, des timbres et des sceaux officiels), contribuant ainsi à l'efficacité de l'organisation criminelle et à la fraude amoureuse (Rege, 2009), car les victimes sont souvent convaincues de leur authenticité.

Tade (2011) souligne le caractère tactique et adaptatif de la fraude en insistant sur le fait qu'une fois que le stratagème est connu du peuple, les changements tactiques sont opérés pour trouver d'autres stratagèmes payants. Ils étudient l'humeur des habitants du pays en termes de vulnérabilité aux nouvelles transactions en ligne afin de décider où planter leurs tentes. Dans son étude sur l'organisation sociologique de la fraude au Nigéria, il note que pour les fraudeurs, l'envoi de messages frauduleux et les rencontres en ligne ont été signalés comme des domaines de spécialisation à faible risque et financièrement rentable.

Les principaux foyers d'opérations en Afrique

La fraude est un phénomène mondial et les fraudeurs sont localisés partout sur le globe. Ibrahim (2016) cité par Whitty, et Ng (2017), souligne que des recherches interculturelles ont révélé que les jeunes d'Afrique de l'Ouest notamment ceux du Nigéria et du Ghana (et dans une certaine mesure la Côte d'Ivoire) sont particulièrement impliqués dans les cyberfraudes. Tandis que les jeunes des pays occidentaux tels que le Canada, les États-Unis, la Grande-Bretagne et la Finlande s'adonnent plus à la cyberintimidation et le cyberharcèlement. Dans notre travail, nous nous intéresserons aux jeunes

fraudeurs d'Afrique de l'Ouest dans l'espoir de comprendre les motivations, les enjeux ainsi que les moyens déployés pour atteindre leurs buts.

Les conditions de l'émergence de la fraude en Afrique

Selon l'un des éminents sociologues africains, Akinsola Akiwowo, la vie sociale en Afrique est inspirée par cinq catégories de valeurs sociales inaliénables qui comprennent: (1) « ire aiku » la valeur de la bonne santé pour la vieillesse (2) « ire owo » la sécurité financière (3) « ire oko-aya » la valeur de la camaraderie intime et de l'amour (4) « ire omo » la valeur de la parentalité et (5) « ire abori ota » la valeur de l'auto-actualisation assurée, Akiwowo (1983, 13-14), cité par Tade (2013). Or, les réalités économiques et culturelles (augmentation du chômage des jeunes, de la corruption et des détournements de fonds, culte de biens matériels et expositions des richesses) en Afrique sont loin de concourir à la satisfaction de l'une des valeurs essentielles qu'est la sécurité financière, d'où la nécessité pour les jeunes de se livrer à des crimes afin d'assouvir leurs fantasmes financiers.

Un tout autre aspect est la valeur de la parentalité, qui pousse ces jeunes vers le cyberspace, comme l'explique Tade (2013) en ces termes « Au Nigéria, un enfant qui va à l'école ou quitte la « sphère de contrôle » de ses parents doit se souvenir de l'enfant dont il « est ». En d'autres termes, parce que l'enfant porte le nom de la famille, il doit jalousement le protégé ». Cela explique donc le fait que les jeunes préfèrent mener des opérations de nature déviante vers le

cyberespace où le nom de famille, le statut des parents et leur identité deviennent anonymes.

Enfin, Tade (2013), explique la migration des fraudeurs au Nigéria de l'espace physique vers le cyberespace par l'utilisation de la théorie de la transition d'espaces du cybercrime (Space Transition Theory en anglais) qui s'articule autour de sept propositions qui sont présentées dans le tableau 2.

La dimension sociale

Qu'il s'agisse des « Yahoo, Yahoo+ » au Nigéria, des « Sakawa » au Ghana, des « Brouteurs » en Côte d'Ivoire et même des « Feymania » au Cameroun, ils ont tous une dénomination. En effet, tels sont les noms attribués aux fraudeurs dans leur pays respectif. Ils sont des jeunes hommes pour la plupart, généralement entre seize et trente ans. Cependant, Rich (2017) cité par Whitty, et Ng (2017), postule que certains garçons « Yahoo » sont généralement des adolescents de douze à seize ans. Elle cite aussi Aghatise (2006) qui rapporte qu'un étudiant sur cinq dans les universités nigérianes est impliqué dans la cybercriminalité. Au Ghana, les « Sakawa » sont généralement sans emploi ou sous-employés et ne fréquentent pas les établissements d'enseignement ou tout autre cursus scolaire.

Sur le plan organisationnel, Rich (2017) cité par Whitty, et Ng (2017) revisite la technostructure de la fraude nigériane, et relève trois rôles essentiels : Au bas de l'échelle, on trouve les garçons « Yahoo », spécialistes de l'accumulation des courriels et de l'envoi des messages frauduleux alors que le niveau suivant gère

Tableau 2 : Sept propositions de la théorie de la transition d'espaces du cybercrime

i.	Les personnes ayant un comportement criminel réprimé (dans l'espace physique) ont une propension à commettre des délits dans le cyberspace, qu'elles ne commettraient pas autrement dans l'espace physique en raison de leur statut et de leur position.
ii.	La flexibilité de l'identité, l'anonymat dissociatif et l'absence de facteurs de dissuasion dans le cyberspace offrent aux délinquants le choix de commettre la cybercriminalité.
iii.	Le comportement criminel des délinquants dans le cyberspace est susceptible d'être importé dans l'espace physique, qui peut également être réexporté vers le cyberspace.
iv.	Les aventures intermittentes des délinquants dans le cyberspace et la nature spatio-temporelle dynamique du cyberspace offrent la possibilité de s'échapper.
v.	(a) Les étrangers sont susceptibles de s'unir dans le cyberspace pour commettre des délits dans l'espace physique. b) Les associés dans l'espace physique sont susceptibles de s'unir pour commettre des délits dans le cyberspace.
vi.	Les personnes issues de sociétés fermées sont plus susceptibles de commettre des délits dans le cyberspace que les personnes issues des sociétés plus ouvertes.
vii.	Un conflit entre les normes et valeurs de l'espace physique et les normes et valeurs du cyberspace peuvent conduire à des cybercrimes.

Source : traduit de (Tade, 2013, p. 694).

les réponses. Plus ils sont élevés dans la hiérarchie, plus ils sont instruits. Rich (2017) mentionne également que « selon Interpol au Nigéria, ces fraudeurs sont

très probablement des spécialistes opérant dans une hiérarchie organisée semblable à d'autres formes de criminalité organisée ». Plus encore, les liens tribaux et linguistiques structurent les relations et favorisent l'organisation du réseau plutôt que l'émergence des fraudeurs indépendants.

Comme le note Tade (2013), il est tout de même intéressant de remarquer qu'il n'y a pas d'hiérarchie permanente entre les membres, les supérieurs hiérarchiques étant constamment changeants en fonction de leurs succès ou de leurs pertes financières. L'argent gagné par les fraudeurs est dépensé pour un niveau de vie extravagant qui comprend la conduite de véhicules coûteux, la fête et la consommation d'alcool, la location d'appartements luxueux, l'achat des bijoux coûteux et des gadgets technologiques. Cette vie ostentatoire exacerbe ainsi les rivalités entre les membres, (Aghatise, 2006; Aransiola et Asindemade, 2011) cités par Whitty, et Ng (2017).

La dimension spirituelle ou le cyber-spiritisme

Au-delà des techniques de persuasion utilisées pour contraindre les victimes à se séparer de leur argent, de plus en plus les fraudeurs sollicitent l'implication des forces invisibles, mystiques ou spirituelles à l'effet d'amadouer leur victime. Si la théorisation occidentale ne parvient pas à expliquer correctement les réalités sociales dans les sociétés africaines, en Afrique, le spirituel constitue la base de l'interprétation des événements sociaux (Akiwowo 1983) cité par Tade (2013).

Bien que la science soit devenue une religion dans le monde, le fait que le microscope de la science ne puisse pas à ce jour saisir les forces ou les éléments spirituels de la sorcellerie, du «juju», des malédictions, des esprits familiaux, des fantômes et autres, ne signifie

pas que ces choses n'existent pas. Nwolise (2012, p. 2) cité par Tade (2013).

Il affirme que la science et la technologie concernent les choses physiques et présentent ainsi une vérité partielle de la réalité sociale. Pour lui, «la science est lâche, pompeuse ou trop arrogante pour insister sur le fait que ce qu'elle ne peut saisir, mesurer ou contrôler n'existe pas ».

Dans les faits, les jeunes fraudeurs « Yahoo+ » au Nigéria et ceux du Ghana les « Sakawa » font abondamment usage des influences mystiques pour accroître leur degré de succès. Whitty, et Ng (2017) qualifient cette pratique de techno-spirituel et précise qu'au Nigéria :

Les fraudeurs « Yahoo+ » emploient « afose », qui est une incantation que la victime ne peut jamais réfuter, « orukaere » qui sont des ornements magiques qui fonctionnent en conjonction avec des incisions sur les corps du fraudeur, et « ijapa » qui font référence aux tortues enchantées sur lesquelles les fraudeurs reposent les pieds lorsqu'ils surfent sur Internet. Ces charmes et matériaux fétichistes peuvent être obtenus auprès d'herboristes, de sorciers « juju » ou d'un prêtre « Yahoo+ » spécifique.

Au Ghana, on constate la même mouvance et le même procédé, avec une différence au niveau des artefacts mystiques, car le pouvoir peut provenir de bibelots magiques, de mouchoirs ou d'ordinateurs portables envoûtants qui permettent à leurs utilisateurs « d'entrer spirituellement sur Internet » ou « d'entrer dans l'esprit des victimes ». Warner (2011) cités par Whitty, et Ng (2017), affirme que « Le Sakawa » est davantage une religion au Ghana qu'une technique, s'étendant au-delà du Ghana jusqu'au Bénin et dans une certaine mesure dans toute l'Afrique de l'Ouest.

Conclusion

Dans ce chapitre, nous avons mis en exergue l'organisation hautement structurée des réseaux de fraudeurs de par le monde, organisation où chacun a un rôle bien défini et qui ne manque pas de tisser des relations de connivence avec les institutions financières et étatiques à l'effet de faciliter les transactions de fraude. Qu'il s'agisse des fraudeurs « Yahoo+ » au Nigéria et « Sakawa » au Ghana, tous agissent selon des logiques construites ou émergées de leur milieu. Aussi dans ces milieux, nous avons illustré le facteur de précarité et l'influence de la spiritualité comme moteurs et amplificateurs du phénomène de la fraude sentimentale.

PARTIE 2 : Détection automatique de la fraude sentimentale

Chapitre 4 : Classification automatique de texte : État de l'art	49
Chapitre 5 : Apprentissage automatique pour la classification automatique de texte	76
Chapitre 6 : Modélisation de la détection de la fraude sentimentale	113

Chapitre 4 : Classification automatique de texte : État de l'art

La dernière décennie a vu l'émergence des mégadonnées, caractéristiques d'une explosion massive de données sur la toile. Les sources de données multiples et diverses (les médias sociaux, la messagerie, l'internet des objets, la blogosphère, Wikipédia, etc.) produisent à la seconde des quantités impressionnantes de données aussi bien structurées et non structurées que semi-structurées. Marr (2015) souligne qu'« Il y a 2,5 quintillions d'octets de données créés chaque jour à notre rythme actuel, mais ce rythme ne fait que s'accélérer avec la croissance de l'Internet des objets (IoT)³¹ » .

Laney (2001) a d'ailleurs conceptualisé ce phénomène sous le triptyque (volume, vitesse et variété). Dans ce contexte d'abondance de données, il est indispensable de se questionner sur leur valorisation, à ce propos Marr (2015) trouve indispensable de compléter les 3V avec deux autres V : la véracité et la valeur. La véracité fait référence à la fiabilité des données alors que la valeur fait allusion à leur utilité.

Les données textuelles représentent une part importante de ce déluge de données, ainsi, pour les valoriser, notamment à travers les objectifs de véracité et valeur. Le traitement de la syntaxe, de la sémantique, du signal, l'extraction d'informations et la bibliométrie ont été explorés par des scientifiques du domaine du traitement automatique du langage naturel (TALN) avec comme corolaire, des

³¹ <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#5b68bd5d60ba>

productions prolifiques, dont les applications aussi nombreuses et variées telles que la traduction automatique, la fouille d'opinion, la recherche d'information, la classification de documents, le résumé automatique, etc.

Dans ce chapitre, nous allons nous intéresser à la classification de documents, domaine qui se rapproche le plus à notre problématique de détection de la fraude sentimentale, où il sera question de classer les messages (informations échangées entre deux personnes impliquées dans une relation sentimentale) selon qu'ils soient frauduleux ou non. Dans un premier temps, nous allons justifier l'apport de la classification automatique de texte (CAT), puis nous abonderons sur l'état des connaissances de la CAT, non sans souligner les différents processus de classification dont les étapes relèvent du domaine du TALN.

Détection de la FS : une tâche de classification automatique de texte

Comme nous avons vu aux chapitres précédents, l'arnaque sentimentale met en exergue deux personnes et une plateforme d'échange d'information internet et/ou mobile. Ce cadre génère une quantité importante de données à mesure que la communication évolue dans le temps. Tel que le montre la figure 3, si ces données sont principalement à prédominance textuelle au début de la relation, les échanges audios et vidéos interviennent au gré de l'intensité de celle-ci.

Nous avons également longuement insisté sur le caractère organisé et criminel de l'organisation des fraudeurs. On assiste aujourd'hui à une automatisation de la fraude. En effet, afin de faire gagner du temps et les coûts

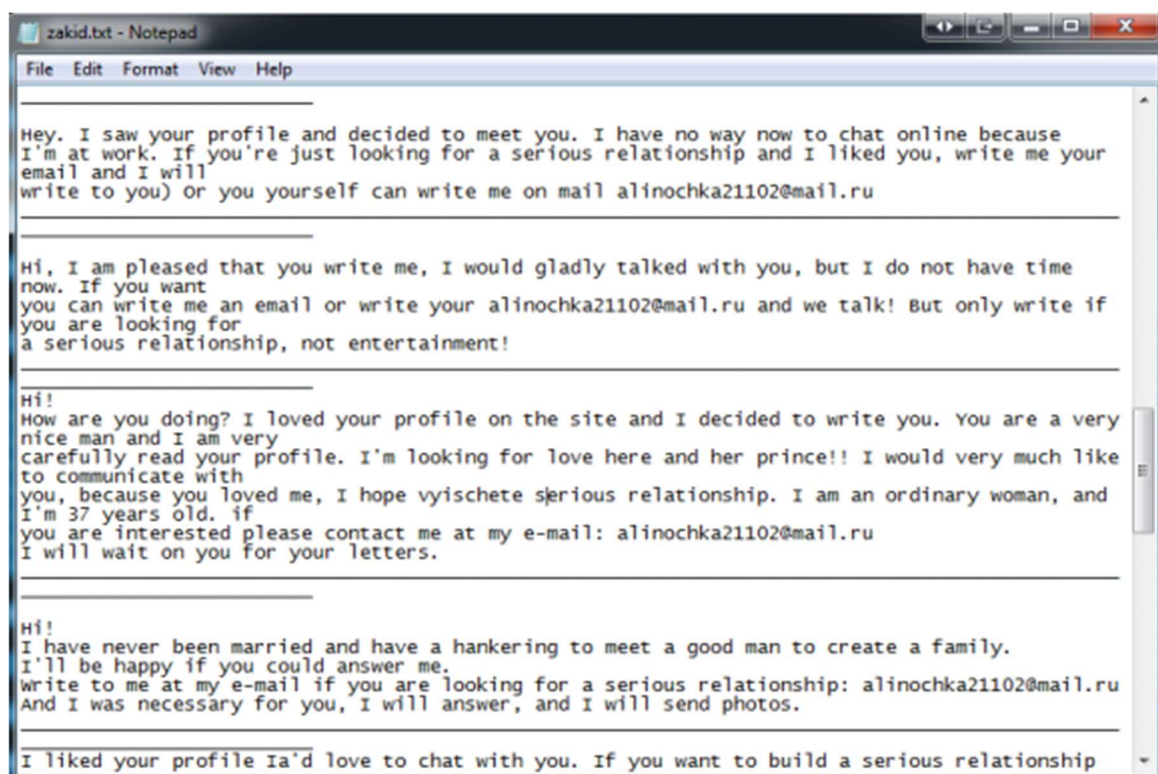


Figure 3 : Exemple de fichier modèle utilisé par les fraudeurs.

Source : tiré du site krebsonsecurity.com³².

d'exploitation, le réseau fabrique des ensembles d'outils à destination des fraudeurs moyennant des coûts. Ce sont des modèles de messages conçus pour appâter tout type de profils d'individu (genre, origine, langue, ethnie et groupe d'âge). A ce propos, la figure 4 présente un exemple de modèle utilisé.

³² <https://krebsonsecurity.com/2016/01/fraudsters-automate-russian-dating-scams/>

Ce sont ainsi des milliers de courriels, sms ou messages disponibles sur des réseaux sociaux qui sont ainsi échangés entre des fraudeurs et leurs victimes. Ces actifs informationnels sont des bons candidats au processus de classification automatique de texte à l'effet de détecter les fraudeurs.

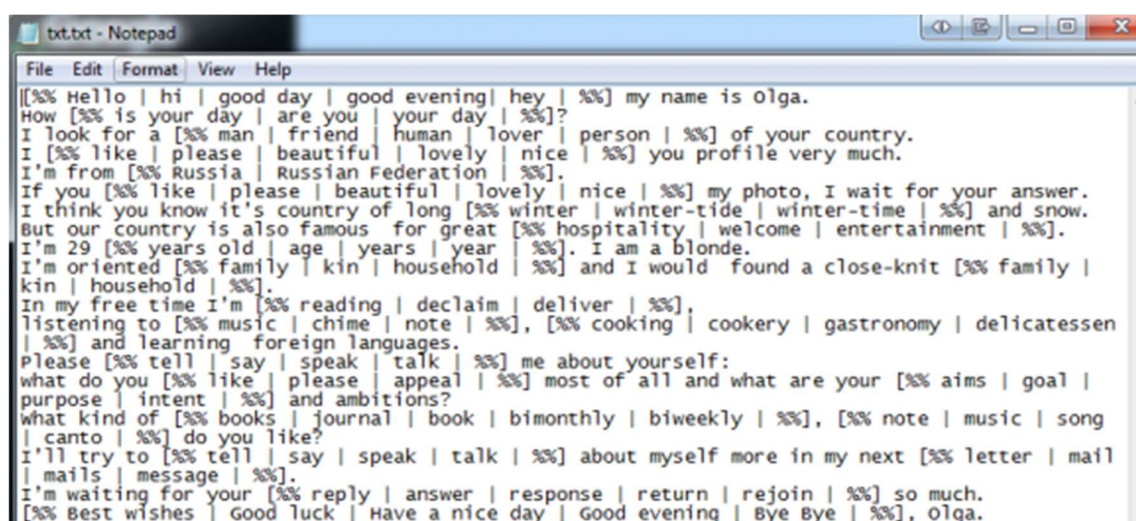


Figure 4 : Un exemple modèle de lettre.

Source : tiré du site krebsonsecurity.com³³.

Définition et méthodologie de classification de textes

Définition

Initialement exploitée au début des années 60, elle a connu un engouement vers le début des années 90 sous l'impulsion de l'augmentation des capacités de calcul et surtout à travers ses multiples applications dont elle permet. Qu'il s'agisse de la recherche d'information, du filtrage des informations, de l'analyse des sentiments, du système de recommandation, de la gestion des

³³ <https://krebsonsecurity.com/2016/01/fraudsters-automate-russian-dating-scams/>

connaissances, du résumé des documents, la CAT offre des mécanismes d'attribution de classes aux documents (Sebastiani, 2002).

Plus formellement le problème de la classification est défini comme suit (Allahyari et al, 2017) : Avec un ensemble de documents $D = \{d1, d2, ..., dn\}$ où d est un document étiqueté suivant les valeurs de l'ensemble $C = \{c1, c2, ..., ck\}$. La tâche de classification consiste à rechercher un modèle ou classifieur (f) tel que l'on pose $f : D \rightarrow C$ et $f(d) = c$ qui attribue la bonne étiquette de classe au nouveau document d .

Méthodologie de classification automatique de textes

Construire un système de classification exige de suivre une démarche méthodologique, comme le montre la figure 5, elle fait état de cinq principales étapes.

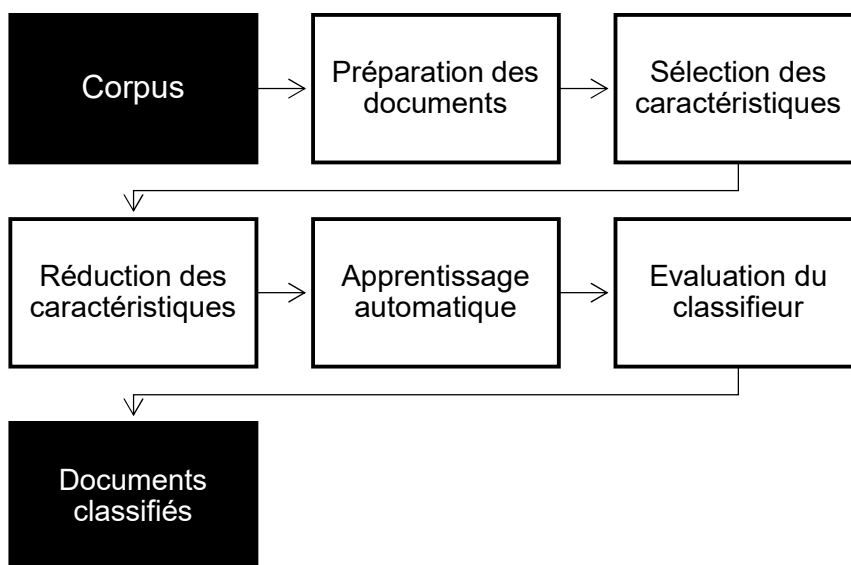


Figure 5 : Processus de classification automatique de documents.

Source : inspiré de (Kowsari et al, 2019).

Afin d'obtenir un lexique caractéristique des documents, leur préparation est nécessaire. Elle consiste d'abord à l'identification des termes (mots, concepts ou suite de caractères à longueur fixe) caractéristiques des documents et par la suite, à appliquer des filtres morphosyntaxiques (suppression des mots vides, lemmatisation, racinisation, etc.). À l'effet de faciliter les opérations aux étapes subséquentes, la représentation des documents est nécessaire. Elle est le procédé de transformation des documents sous la forme matricielle dont les individus sont des documents et les termes issues ou extraites de celles-ci en sont des variables ou des dimensions des caractéristiques

Compte tenu de l'explosion des caractéristiques, la sélection des plus discriminantes et représentatives des documents ainsi que leurs transformations en d'autres caractéristiques nouvelles de taille réduite, mais bénéfiques au regard de la problématique étudiée sont nécessaires.

Puis, arrive finalement, l'entraînement des données issues de la matrice à l'aide d'algorithmes d'AA tels que : forêt de décision, naïve bayes, machines à vecteurs de support (SVM), Arbre de décision, etc. Cette étape permet de construire des classifieurs aptes à classer les documents selon divers degrés de performance. Compte tenu des spécificités des données étudiées, les méthodes d'OA peuvent performer de manière disparate. L'étape d'évaluation vise à fournir des mesures à l'effet de déterminer le classifieur le plus à même de répondre aux besoins de classification avec un haut niveau de satisfaction.

La méthodologie de CAT sera explicitée plus amplement dans les sections suivantes, notamment dans ces étapes de préparation de données, de sélection et de réduction des caractéristiques.

Préparation des documents du corpus

Le corpus documentaire issu de la capture des données de diverses sources n'est généralement pas adapté aux algorithmes de CAT ou tout au moins à l'extraction des caractéristiques optimales. L'objectif de cette étape est le nettoyage et la normalisation du corpus pour le prédisposer aux traitements futurs. À cet effet, nous exposerons quelques techniques telles que : la segmentation, la lemmatisation, la racinisation, le traitement des mots vides, de la capitalisation, de l'argot, de l'abréviation, des bruits et des erreurs d'orthographe.

Nettoyage et normalisation

Segmentation des mots et phrases

Aussi appelée tokenisation, elle consiste à découper de longues chaînes de texte en plus petites parties appelées jetons (Verma et al, 2014). Ces parties peuvent être de tailles variables, aussi bien unitaires (tels que les symboles) qu'au-delà des mots (par exemple : les sections, les paragraphes et les phrases). Par exemple les jetons de la phrase « *les thermes romains sont des établissements abritant les bains publics* » Sont : `{"les", "thermes", "romains", "sont", "des", "établissements", "abritant", "les", "bains", "publics"}`.

En se basant sur le caractère de fin de phrase (le point), l'on peut également découper le corpus en un ensemble de phrases.

Étiquetage des parties du discours

Identifier les parties du discours (ou part-of-speech (POS) en anglais) apparait important comme technique de prétraitement du corpus. Ces parties du discours comprennent le nom commun, le nom propre, l'adjectif qualificatif, le déterminant, le pronom, le verbe, l'adverbe, la préposition et tout autre type appartenant à des classes grammaticales d'un langage. Malheureusement, cette technique tient compte du contexte de la phrase, ainsi une erreur dans la segmentation de phrase entraînerait des erreurs d'étiquetage (Tabassum et Patil, 2008).

Mots vides

Tous les mots du corpus ne sont pas porteurs de sens, ou sont moins significatifs pour les besoins de la classification. Les termes généralement fréquents tels que les articles, les propositions, les verbes auxiliaires, les adverbes et conjonctions feront partie d'un anti-dictionnaire, constituant ainsi des termes à ôter du corpus, comme le précisent d'ailleurs Saif et al (2014). Les mots : le, les, avant, encore, travers, il, de, du, avoir, être sont des exemples de mots vides.

Capitalisation

Parce que le corpus est souvent un ensemble de texte construit suivant les règles grammaticales, il peut donc être constitué de mots aussi bien en

majuscules qu'en minuscules. Dans ce cas de figure, la sémantique est syntaxiquement représentée par plus d'un mot, cela implique que les documents disposants ces mots avec des capitalisations variables ne seront pas représentés dans le même espace de caractéristiques (Kowsari et al, 2019, p. 4). Cette problématique peut être jugulée par la conversion de toutes les lettres en minuscule. Toutefois la problématique d'interprétation des capitonymes³⁴ (mot qui change de sens et parfois de prononciation lorsqu'il est en majuscule) est à considérer. Les exemples de capitonymes en anglais sont les suivants :

March (le mois) et march (action de marcher) –verbe

Titanic (le bateau) et titanic (gigantesque) –adjective

Turkey (le pays) et turkey (l'oiseau)

Argot et abréviation

L'abréviation est la réduction à un ou plusieurs caractères d'un mot ou groupe de mots. Par exemple, ACP : Analyse en Composantes Principales et OMS : Organisation Mondiale de la Santé). Certaines techniques permettent l'extraction des abréviations (Schwartz et Hearst, 2003; Pustejovsky et al, 2001a, 2001b). L'argot fait allusion à un vocabulaire et un style de langage propre à un groupe particulier et qui par ailleurs infiltre le langage commun (Kulkarni et Wang, 2017). L'une de façon de les traiter est de les convertir vers les langages communs (Wu et al, 2016).

³⁴ <https://en.wikipedia.org/wiki/Capitonym>

Les bruits

Certains caractères présents dans le corpus sont de nature bruyante, ce sont particulièrement des ponctuations et caractères spéciaux. Selon le contexte, Ces caractères peuvent être supprimés afin d'éviter les problématiques de performance de la classification (Agarwal et al, 2007).

Erreurs d'orthographe

Des erreurs d'orthographe peuvent se glisser dans le corpus. Ce sont particulièrement des fautes typographiques provenant de diverses sources et plus particulièrement les réseaux sociaux. La correction orthographique consiste à corriger les mots dans les documents. À cet effet, plusieurs solutions ont été utilisées, dont ceux de Schierle et al (2008), qui ont produit un système efficace de correction orthographique sensible au contexte.

Racinisation

Il arrive que l'on retrouve dans un corpus plusieurs mots avec des variations orthographiques différentes pour la même sémantique. Par exemple, les mots « automatique » et « automatiques » ont le même sens, mais orthographiés différemment. Dans ce cas de figure et afin d'éviter d'indexer plusieurs fois le même mot, il faut tenir compte de la variabilité des formes de mots par l'utilisation d'une technique de normalisation qu'est la racinisation ou troncature (stemming, en anglais). Quelques exemples de racinisation :

Consolant → *Consol*, *Courir* → *Cour*, *Coureur* → *Coureur*, *Courant* → *Cour*,
Facilement → *Facil*, *Assez* → *assez*.

Krovetz (2000, p. 279) souligne que, l'un des problèmes de la troncature est qu'il ne gère pas la sémantique du mot, soulignant par exemple que le mot « gravitation » est lié au sens de la force de gravité du mot « gravité » plutôt qu'au sens qui signifie « grave » (par exemple, la gravité du crime), et que si la « gravitation » est tronquée, nous pourrions la confondre avec les sens de « gravité », car ces deux mots ont la même racine, par exemple : *gravité* → *gravit*, *gravitation* → *gravit*.

Lemmatisation

Contrairement à la racinisation, la lemmatisation fait référence à l'analyse morphologique du langage pour trouver la forme de base d'un mot ou « lemme ». Elle s'obtient en remplaçant ou en supprimant le suffixe en tenant compte des règles grammaticales et de sa forme à l'instar de son genre (masculin ou féminin), son nombre (un ou plusieurs), sa personne (moi, toi, eux...), son mode (indicatif, impératif...). Korenius et al (2004, p. 625) soulève deux problèmes que posent cette technique : (i) les formes de mots homographiques³⁵ posent des problèmes d'ambiguïté (et de précision) et (ii) tous les mots ne peuvent pas être lemmatisés, car le dictionnaire du lemmatiseur ne les contient pas. Quelques exemples peuvent être tirés de ces phrases :

« La cour du Québec fait courir les prisonniers, car en courant, ils développent leur forme physique » où les lemmes sont respectivement :

³⁵ <https://www.dcode.fr/generateur-homoglyphes-homographes#q1>

La → le, cour → cour, du → du, Québec → Québec, fait → faire, courir → courir, les → le, prisonniers → prisonnier, car → car, en → en, courant → courir, ils → il, développent → développer, leur → lui, forme → former, physique → physique.

Extraction des caractéristiques

L'extraction des caractéristiques (features, en anglais) est un processus qui permet de déterminer les termes capables de représenter au mieux les structures syntaxiques et sémantiques des documents du corpus. À cet effet, nous disposons de plusieurs techniques telles que : l'extraction des n-grammes, des expressions et des phrases.

N-grammes

Un n-gramme est une tranche de n caractères (ou mots) consécutifs d'une plus longue chaîne (Cavnar et Trenkle, 1994). En fonction du contexte, l'on pourra distinguer les n-grammes de caractères et les n-grammes de mots. Par exemple, pour la phrase suivante : *Nécessité fait loi dans un contexte de crise*. L'on peut extraire les 1-gram (unigramme), 2-grammes (bigrammes) de mots suivants :

Unigramme : *Nécessité, fait, loi, dans, un, contexte, de, crise*

Bigrammes : *Nécessité fait, fait loi, loi dans, dans un, un contexte, contexte de, de crise*

Un exemple pour les n-grammes de caractères, soit l'expression *prix nobel*, déterminons ses unigrammes et bigrammes de caractères.

Unigramme : *p, r, i, x, n, o, b, e, l*

Bigrammes : *pr, ri, ix, xn, no, ob, be, el*

La technique des n-grammes révèle néanmoins quelques insuffisances. Le vocabulaire peut rapidement devenir très volumineux. Étant donné qu'elle est appliquée trop souvent après les activités de nettoyage et de normalisation, une grande partie de l'information du corpus est supprimée, détruisant ainsi la structure morphologique du corpus initial (Scott et Matwin, 1999, p. 3).

Les phrases

Fort des insuffisances des n-grammes abordées précédemment, Scott et Matwin (1999, p. 3-4) soutiennent que l'utilisation des expressions³⁶ comme caractéristiques permet de préserver certaines informations délaissées par les n-grammes. Tout en arguant un potentiel gain de performance dans des systèmes à base de règles. Ces auteurs affirment par exemple que l'expression « apprentissage automatique » a une signification très spécifique qui est séparée et distincte des mots « machine » et « apprentissage ». Il est concevable que l'expression « apprentissage automatique » produise un gain d'information élevé pour une recherche dans un corpus comprenant l'expression « intelligence artificielle » même si les mots individuels donnent un gain faible.

Reconnaissance des entités nommées (REN)

Comme dans le cas de l'étiquetage des parties du discours, la REN extrait les entités nommées telles que noms de personnes, noms d'organisations, noms de lieux, quantités, distances, valeurs, dates, etc. Anđelić et al (2017) démontrent

³⁶ Les auteurs parlent aussi de « Noun Phrases » et de « Key Phrases » comme caractéristique.

une bonne performance grâce à l'utilisation unique de la technique de REN pour la CAT.

Pondération des termes

Après l'extraction des termes du corpus, il convient de les valoriser. À cet effet, quatre techniques se dégagent de la littérature scientifique en CAT, nous pouvons citer : le sac de mots, la TF-IDF, la reconnaissance des entités nommées et le prolongement lexical.

Le sac de mots

Le sac de mots est une technique largement utilisée en extraction de caractéristiques, car elle est simple et flexible. En effet, elle consiste pour chaque mot à spécifier sa présence ou non dans un document. La technique permet de construire un lexique de mots connus ainsi que leur fréquence de présence dans un corpus (Waykole et Thakare, 2018, p. 352). Au-delà de la fréquence de présence, l'on peut également parler de la fréquence des termes du document. Communément appelé TF (ou term frequency, en anglais), il désigne le nombre d'occurrences du terme dans un document.

TF-IDF

L'utilisation du sac de mots ou dans une certaine mesure le TF, révèle une insuffisance importante. En recherche d'information, l'on observe que si un mot est largement présent dans un document (TF important) et que par ailleurs il l'est également dans le reste du corpus présent, alors ce mot n'est plus suffisamment discriminant. Pour corriger ce problème, Jones (1972) a adjoint au TF la

fréquence inverse des documents (IDF) qui correspond à la fréquence du mot dans le corpus. Cette fréquence IDF vise à pénaliser les mots communs et augmenter la pertinence des mots rares. La fréquence combinée TF-IDF (ou term frequency-inverse document frequency, en anglais) est formellement donnée par l'équation $W(d, t) = TF(d, t) * \log(N / df(t))$, où N est le nombre de documents et $df(t)$ le nombre de documents contenant le terme t dans le corpus.

Prolongement lexical (PL)

En plus de supprimer l'ordre des mots dans leurs représentations ainsi que la structure sémantique car les techniques précédentes se préoccupent uniquement à définir un lexique basé sur la syntaxe sans se préoccuper de la sémantique des mots, ou tout au moins de leurs similitudes. Prenons par exemple la synonymie dans un corpus faisant apparaître les mots *vélo* et *bicyclette*, ces deux mots seront représentés indépendamment avec la technique du sac de mots. De plus, du fait de cette redondance, la dimensionnalité des caractéristiques augmente considérablement. D'où la nécessité de trouver des techniques pour adresser ces problèmes. De nos jours, l'on fait de plus en plus référence au prolongement lexical (PL). Le PL est une technique d'apprentissage des caractéristiques dans laquelle chaque mot ou phrase du vocabulaire est mappé à un vecteur de dimension de nombres réels (Kowsari, 2019). Ils sont pour la plupart une extension des fonctionnalités du sac de mots par la prise en charge du contexte et du sens des mots. « Word2Vec », « GloVe » and « FastText » et plus

récemment le « Deep contextualized word representations » en sont les méthodes.

La représentation du corpus

L'idée de représentation du corpus tient du fait que la CAT fait intervenir d'autres composantes dans le processus, au rang desquels, l'utilisation des algorithmes d'AA. Pour mettre ces derniers à contribution, le corpus doit être représenté dans un formalisme propice à son traitement, pour le faire nous allons mobiliser les vertus du modèle de représentation vectorielle. Proposé au milieu des années 70 (Salton et al, 1975), ce modèle permet de représenter sous forme matricielle un corpus, où un document est représenté par un vecteur dont chacune des dimensions est un terme ou caractéristique possédant une valeur (de fréquence, présence, poids, etc.). Plus formellement, en utilisant la pondération TF, le vecteur d'un document d sera donné par l'équation suivante : $V_d = [w_{1,d}, w_{2,d}, w_{3,d}, \dots, w_{N,d}]^T$, avec $w_{t,d} = TF(d, t)$, où $TF(d, t)$ est le nombre d'occurrences du terme t dans le document d .

Par exemple, représentons le corpus de trois documents ($d_1 = La\ vie\ est\ belle$, $d_2 = Il\ faut\ croquer\ la\ vie$, $d_3 = Belle\ est\ la\ vie$) selon le modèle vectoriel.

En ne faisant aucun prétraitement du corpus en dehors du traitement de la capitalisation, on aura une représentation sous la forme suivante :

	<i>la</i>	<i>vie</i>	<i>est</i>	<i>belle</i>	<i>il</i>	<i>faut</i>	<i>croquer</i>
d_1	1	1	1	1	0	0	0
d_2	0	1	0	0	1	1	1
d_3	1	1	1	1	0	0	0

Ce modèle de Salton est donc une représentation mathématique du document. À l'observation, l'ensemble des termes constituant le vocabulaire représente des dimensions tandis que les documents constituent l'espace vectoriel. Dans sa forme la plus simple, il précise pour chaque document l'occurrence des termes. Bien qu'étant facile à appréhender, ce modèle présente néanmoins deux insuffisances majeures : (i) l'explosion du vocabulaire (car la synonymie et d'autres traitements morphologiques des mots ne sont pas exploités) diminue le pouvoir discriminant des documents du fait de l'abondance des vecteurs creux (ii) le modèle privilégie le caractère lexical au détriment de la sémantique des termes en supposant l'indépendance des termes ainsi que leur ordre dans le document. Depuis lors, le modèle originel a suivi plusieurs innovations à l'effet de combler les lacunes observées, notamment à l'aide des techniques de sélection et de réduction des caractéristiques.

Sélection des caractéristiques

L'extraction des caractéristiques produit un nombre assez important de termes caractéristiques du corpus étudié, d'où la nécessité d'opérer des choix en lien avec les objectifs de classification. L'orientation vers la réduction de la

dimensionnalité d'un espace vectoriel est motivée principalement par les besoins d'optimisation et d'efficacité. Dans une représentation vectorielle, les temps et le besoin en mémoire sont tributaires de la taille des caractéristiques.

La sélection des caractéristiques vise à choisir parmi les dernières, les plus pertinentes afin d'accroître l'efficacité et la précision de l'activité subséquente dans le processus de l'AA (Forman, 2003) tout en atténuant la problématique du surapprentissage (Ikonomakis et al, 2005, p. 73). Le principe de sélection consiste à calculer indépendamment pour chaque caractéristique, un score suivant un critère défini pour ne finalement sélectionner qu'un sous-ensemble des caractéristiques possédant les meilleurs scores. À cet effet, plusieurs approches dont celles présentées par Chandrashekar et Sahin (2014) (« filter », « wrapper » et « embedded »), celles issues de la théorie d'information telle que le khi2, le gain d'information (GI), information mutuelle (IM) sont communément exploitées. Nous allons brièvement insister sur quatre d'entre elles afin de mieux comprendre le principe général de la sélection des caractéristiques.

Les méthodes de sélection

Le gain d'information

Le gain d'information (IG, information gain en anglais) désigne la quantité d'information suivant l'absence ou la présence d'un terme. Il est formellement défini par la formule

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) * \log \left(\frac{P(t, c)}{P(t) * p(c)} \right)$$

Information mutuelle

L'information mutuelle capture la dépendance entre une caractéristique et la classe de prédiction, plus le terme appartient à la classe, l'IM sera grand, cependant si un terme n'a aucune information commune avec la classe, cela signifie une indépendance de ces deux variables, et donc l'IM sera zéro. Formellement l'IM sera

$$MI(t_k, c_i) = \frac{\log(P(t_k, c_i))}{P(t_k) * P(c_i)}$$

Khi2

Appelé également test du χ^2 ou « khi carré », ce test vient de la statistique et se prête bien à la sélection des caractéristiques, car il permet de tester l'indépendance entre deux variables. Le Khi2 sera formulé comme suit :

$$\text{Khi2}(t_k, C_i) = N(AD - CB)^2 / ((A + C)(B + D)(A + B)(C+D)) \text{ où}$$

N : Nombre total de documents dans le corpus,

A : Nombre de documents de la classe contenant le terme,

B : Nombre de documents contenant le terme t_k dans d'autres classes,

C : Nombre de documents de la classe qui ne contiennent pas le terme t_k ,

D : Nombre de documents qui ne contiennent pas le terme t_k dans d'autres classes.

Séparation binomiale

Cette métrique, encore appelé BNS (Bi-Normal Separation, en anglais) modifie la représentation vectorielle d'un document en augmentant l'importance

relative des mots utiles. Pendant que la métrique TF-IDF pénalise les mots qui se trouvent dans de nombreux documents au sein d'un corpus, la métrique BNS augmente le score des fonctionnalités qui fournissent plus de vrais positifs que de faux positifs dans une tâche de classification binaire. En d'autres termes, un mot trouvé dans une seule classe obtiendra un score plus élevé qu'un mot trouvé dans toutes les classes, car le premier ne produira que de vrai positif et aucun faux positif (Baillargeon et al, 2019, p. 434).

Étant donné D , un corpus, et $C \subset D$ l'ensemble des documents appartenant à la classe positive et $C' = D \setminus C$ l'ensemble des documents appartenant à la classe négative. La pondération BNS pour chaque mot w_i , $i \in 1, \dots, |V|$ dans le vocabulaire V est défini formellement comme suit :

$$BNS(w_i, D) = \text{abs}(\Phi^{-1}(TPR_C(w_i, D)) - \Phi^{-1}(FPR_C(w_i, D))),$$

où abs correspond à la fonction valeur absolue et Φ^{-1} est la fonction quantile d'une variable aléatoire normale. De plus, $TPRC$ et $FPRC$ sont respectivement le taux de vrais positifs et les taux de faux positifs de classification pour la classe positive C , et

$$TPR_C = \frac{|d_j : w_i \in d_j, d_j \in C|}{|d_j \in C|}; \quad FPR_C = \frac{|d_j : w_i \in d_j, d_j \notin C|}{|d_j \notin C|}.$$

Pour autant, cette métrique n'est utilisable que dans le cadre de la classification binaire et nécessite une extension pour le support des multiclassés (Baillargeon et al, 2019, p. 435).

La détermination du seuil des métriques

La stratégie de détermination du seuil des métriques est essentiellement fonction du contexte d'utilisation, elle peut être définie autant analytiquement qu'expérimentalement (SEBASTIANI, 2002, p. 19). Pour Wang et al (2010, p. 500), le choix du seuil commence par une normalisation des valeurs de chaque caractéristique entre 0 et 1, un appareillage est ensuite effectué entre chacune des caractéristiques et la classe selon les mesures de performance appropriées. De plus ils préconisent d'opter pour un seuil de 0.5 pour une classification binaire et la recherche d'un seuil optimal en cas de déséquilibre de classe.

Réduction des caractéristiques

La réduction des caractéristiques (ou des dimensions) fait référence à la transformation des caractéristiques. La compression significative de données, la découverte de structure, l'élicitation de caractéristiques et la visualisation de données massive en sont quelques exemples d'applications de la réduction des caractéristiques. Elle procède de façon différente à celle de la sélection. Là où la sélection pondère les termes du vocabulaire pour ne retenir que les valeurs les plus élevées, la réduction quant à elle, est la création de nouvelles caractéristiques en utilisant celles existantes. La nécessité de réduction provient du fait que face aux problématiques récurrentes de polysémie, d'homonymie et de synonymie, les caractéristiques initiales ne sont pas souvent adaptées pour la représentation du contenu d'un document. La construction d'un ensemble de

caractéristiques déterminé à partir des celles existantes permet de corriger ces problèmes (SEBASTIANI, 2002, p. 3).

L'indexation sémantique latente, l'analyse en composantes principales, l'analyse en composantes indépendantes, l'analyse discriminante linéaire sont quelques-unes des méthodes populaires dont nous allons brièvement décrire les logiques qui les sous-tendent.

Indexation sémantique latente

Cette méthode d'indexation sémantique latente (LSI en anglais) permet de révéler la structure sémantique « latente » du vocabulaire utilisé dans le corpus. C'est davantage une technique du domaine du TALN brevetée en 1988 (BIRICIK et al, 2012, p. 1142) et qui permet de corriger les problématiques de la synonymie et dans une certaine mesure, la polysémie des termes du vocabulaire. Elle consiste à réduire l'espace de dimension des vecteurs de document en se basant sur la cooccurrence des termes avec l'espace originel. Cela est rendu possible par la mobilisation de la décomposition en valeurs singulières (SVD en anglais).

En effet, la matrice d'occurrence sera décomposée en valeurs singulières pour lesquelles, les plus pertinentes seront considérées comme le nouvel espace vectoriel, avec un nouveau vocabulaire de termes porteurs de sens. Mi (2014) nous donne le principe LSI suivant :

Étant donné une matrice terme-document, le SVD se décompose en un ensemble de 3 composants plus petits, soit $X = U\Sigma V^T$.

Si nous représentons les corrélations entre les termes sur les documents avec XX^T , et les corrélations entre les documents sur les termes avec X^TX , nous pouvons également démontrer ces matrices avec des équations $XX^T = U\Sigma\Sigma^TU^T$ et $X^TX = V\Sigma^T\Sigma V^T$. Lorsque nous sélectionnons des valeurs singulières à partir de Σ et les vecteurs correspondants de U et V matrices, nous obtenons une approximation de rang pour X avec une erreur minimale. Cette approximation peut être vue comme une réduction de dimension.

Toutefois, Labadié (2008, p. 32) dans la thèse, souligne le caractère aveugle des capacités d'inférence utilisée pour identifier les liens sémantiques entre les termes ou documents. De plus, il souligne que la « LSA³⁷ [LSI] peut faire apparaître des relations entre des termes n'ayant aucun point commun en faisant une association par transitivité malheureuse, à cause d'un terme particulièrement polysémique ». SEBASTIANI (2002, p. 18) pour sa part soutient qu'un terme original doté de vertu discriminant par rapport à une classe, pourrait voir cette vertu anéantie dans le nouvel espace vectoriel.

Analyse en composantes principales

L'ACP (ou PCA en anglais) est une méthode descriptive ancienne et encore très utilisée de nos jours. Le principe est de trouver un nouvel ensemble de dimensions corrélé avec l'ensemble d'origine et qui maximisent au mieux la variabilité statistique (Jolliffe, 2016). Les nouvelles dimensions sont appelées composantes principales, obtenues à l'aide de l'exploitation des valeurs et

³⁷ https://en.wikipedia.org/wiki/Latent_semantic_analysis

vecteurs propres. Cette méthode est utilisée dans plusieurs domaines à l'instar de la biologie, l'économie, le traitement d'images avec pour intérêts de décrire, de visualiser et de réduire les bruits dans les données.

Formellement, l'on exprime la PCA de la façon suivante : supposons une matrice X (n, p) où n est le nombre de documents et p , le nombre de caractéristiques. Un vecteur de documents de cette matrice est x_1, \dots, x_p dont la j ème colonne est le vecteur x_j d'observations sur la j ème variable. La combinaison linéaire est donnée par $\sum_{j=1}^m (a_j x_j) = Xa$ où a est un vecteur de constantes a_1, a_2, \dots, a_p . La variance d'une telle combinaison linéaire est donnée par $\text{var}(Xa) = a^T S a$, où S est la matrice de covariance de l'échantillon associée à l'ensemble de données et a^T désigne la transposition de a . En définitive, le but est de trouver la combinaison linéaire avec la variance maximale. Cela se traduit par la maximisation de $a^T S a - \lambda (a^T a - 1)$, où λ est un multiplicateur de Lagrange.

Tout en étant sensible aux points extrêmes, l'ACP est davantage adaptée au phénomène linéaire.

Analyse en composantes indépendantes

L'analyse indépendante des composantes (ICA en anglais) est une technique de séparation aveugle de source³⁸ introduite dans les années 80 par Hérault et al (1985). Largement utilisée pour sa robustesse face au bruit et à la

³⁸ https://en.wikipedia.org/wiki/Signal_separation

carence en information sur les signaux à l'instar du problème du cocktail (Shlens, 2014, p.2). Elle se définit comme suit : supposons que nous observons des mélanges linéaires x_1, \dots, x_n de n composantes indépendantes, soit $x_j = a_{j1}s_1 + a_{j2}s_2 + \dots + a_{jn}s_n$ $\forall j$. Notons A la matrice avec les éléments a_{ij} , en utilisant cette notation à matrice vectorielle, le modèle de mélange ci-dessus s'écrit $X = As$.

Hyvärinen et Oja (2000) notent qu'en désignant par a_i les colonnes de la matrice A , le modèle peut aussi s'écrire $X = \sum_{i=1}^n (a_i s_i)$. Pour deux variables aléatoires différentes s_1 et s_2 , l'on dira que la variable aléatoire s_1 est indépendante de s_2 si les informations sur la valeur de s_1 ne fournissent aucune information sur la valeur de s_2 , et vice versa alors. (Naik et al, 2011, p. 64). Cette méthode tient son intérêt de l'indépendance des variables.

Relativement à l'extraction des caractéristiques, quelques implémentations ICA ont été proposées, à l'instar de FastICA, Infomax, ICA-R (Naik et al, 2011, p. 70).

L'analyse discriminante linéaire

De l'anglais LDA (Linear discriminant analysis), elle est une méthode dans laquelle la réduction de dimension est consécutive à la combinaison linéaire de caractéristiques représentant la séparation de classe (Mi, 2014, p. 1142). Elle est utilisée lorsque les fréquences intra-classe sont inégales et essaye donc de maximiser le rapport de la variance inter-classe à la variance intra-classe dans

tout ensemble de données. Se faisant, elle garantit ainsi une séparabilité maximale (Balakrishnama et Ganapathiraju, 2012).

Mi (2014) nous donne une fois de plus un développement formel de cette méthode. Soit la matrice de dispersion inter-classes Σ_B suivante $S = \frac{W^T \Sigma_B W}{W^T W}$ et celle de dispersion intra-classes $\Sigma_B = \sum_c (\mu_c - u)(\mu_c - u)^T$, la séparation à maximiser des classes peut être calculée suivante $\Sigma_W = \sum_c \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)^T$. La transformation linéaire peut être donnée par une matrice U , où les colonnes sont constituées des vecteurs propres de $\Sigma_B^{-1} \Sigma_W$ dans l'équation

$$\begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_K \end{bmatrix} = \begin{bmatrix} u_1^T \\ u_2^T \\ \dots \\ u_K^T \end{bmatrix} (x - \mu) = U^T (x - \mu)$$

Finalement, les vecteurs propres obtenus en résolvant l'équation $\Sigma_W u_{k_B} = \lambda_k \Sigma_B u_{k_B}$ peuvent être utilisés pour la réduction de dimension.

En résumé, plusieurs autres méthodes telles que : l'autoencodeur, de projection aléatoire (random projection), T-distributed Stochastic Neighbor Embedding (t-SNE) peuvent être exploitées selon le corpus et les besoins de classification.

Conclusion

Nous avons montré le positionnement de notre problématique comme relevant du domaine de CAT. En rapport avec la FS, nous sommes bel et bien dans problématique de classification. À l'issue de ce chapitre, nous nous sommes efforcés de synthétiser les apports de la littérature scientifique du domaine de la CAT. Essentiellement, nous avons présenté l'état de l'art récent des premières étapes du processus de CAT (la préparation des documents du corpus, la sélection et la réduction des caractéristiques).

Chapitre 5 : Apprentissage automatique pour la classification automatique de texte

Les algorithmes d'apprentissage automatique (AA) interviennent à l'issue des activités de prétraitement du corpus, de sélection et de réduction des caractéristiques. À ce stade, les données ont été vectorisées et peuvent donc être utilisées par les algorithmes d'AA. Relativement à la CAT, plusieurs algorithmes ou modèles ont été proposés, tous différents par leurs approches, soient ensemblistes (boosting, bagging, etc.), basés sur les arbres (arbre de décision, forêts aléatoires), ou basés sur des réseaux de neurones (DNN, CNN, RNN, DBN, HAN) et d'autres tels que : naïve bayes, machines à support vectoriel, k-voisins, etc.

Dans un premier temps, nous allons définir le processus de l'AA et son positionnement dans le vaste domaine de l'intelligence artificielle. Ensuite, nous expliquerons la logique de construction d'un modèle de classification, puis nous abonderons sur les différents algorithmes d'AA judicieusement sélectionnés au regard de la littérature du domaine de la CAT. Les algorithmes sont habituellement utilisés en fonction du contexte et des données du problème posé, à cet effet, nous allons aborder ensuite, la question de l'évaluation des algorithmes en explorant les métriques de performance spécialisées. Pour finir, nous allons présenter les travaux scientifiques les plus récents et inhérents à notre problématique.

Généralités

Positionnement de l'apprentissage automatique

L'Intelligence artificielle (IA) est une approche innovante et efficace de résolution de problèmes. Elle regroupe plusieurs sous-ensembles, dont l'apprentissage automatique(AA)/statistique (ou machine learning en anglais), « qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d' « apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune» (Awad et al, 2015, p. 1) . La figure 6 illustre les disciplines de l'intelligence artificielle.

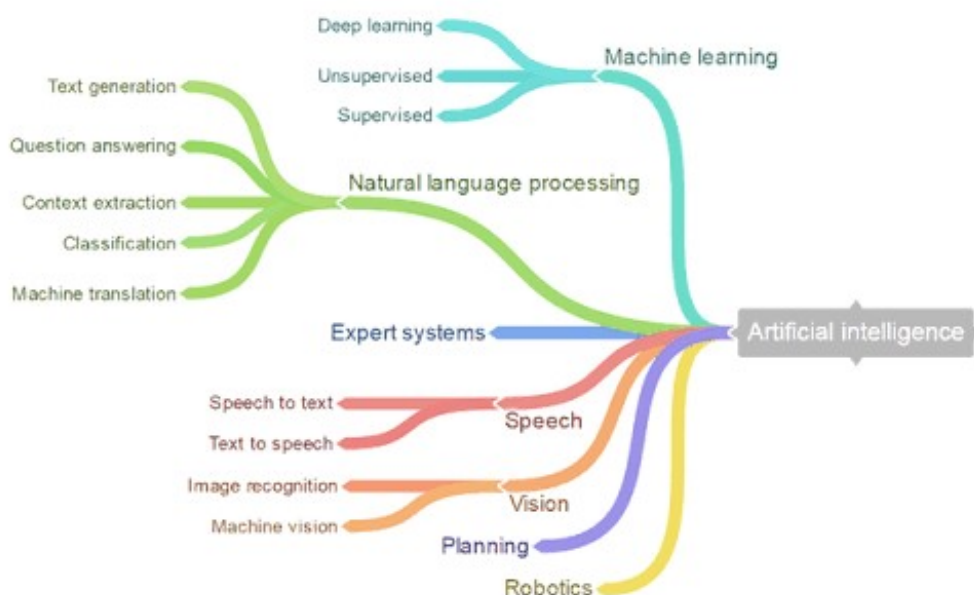


Figure 6 : Sous-ensemble de l'intelligence artificielle³⁹.

³⁹ <https://www.worldforumdisrupt.com/ai-world-forum-san-francisco-2019/artificial-intelligence-understanding-the-hype/>

Définition

Yann Le Cun, scientifique en chef de l'intelligence artificielle chez Facebook déclare qu'« Il n'y a pas d'intelligence sans apprentissage ». L'AA a été inventé en 1959 par Arthur Samuel, elle connaît une émancipation singulière grâce à l'essor des données massives (ou Big data en anglais) avec des applications comme Hadoop, MapReduce, etc. L'AA est une technique d'analyse de données qui diffère des approches informatiques traditionnelles dont le but consiste à faire usage par des ordinateurs, des algorithmes explicitement programmés pour résoudre les problèmes. En AA, les algorithmes permettent aux ordinateurs de s'entraîner sur des ensembles de données à travers des modèles statistiques afin de produire des artéfacts aidant à la décision. La performance de ces algorithmes d'AA s'augmente au gré de la disponibilité des données pour l'apprentissage.

Aujourd'hui, les effets de l'AA font partie de notre quotidien, plusieurs applications en font la manifestation, à l'instar de la prédiction de prix, la détection des pourriels, le diagnostic médical, la recommandation de produits, la détection de fraude, la cybersécurité, la reconnaissance vocale, les chats bots, la conduite autonome. Ces applications servent dans des domaines divers tels que l'économique, la finance, l'industriel, le commerce, la médecine, l'administration publique, etc.

Typologie et applications

En fonction de la problématique étudiée, nous pouvons choisir parmi les quatre catégories de méthodes d'apprentissage suivantes :

L'apprentissage supervisé, non supervisé, avec renforcement et l'apprentissage profond.

Apprentissage supervisé

Dans ce type d'apprentissage, l'algorithme est entraîné par un ensemble de données étiquetées en entrée et une sortie prédéterminée (appelée classe) à produire des prévisions raisonnables pour les réponses aux nouvelles données. Le spécialiste des données joue donc un rôle de guide et apprend à l'algorithme les conclusions qu'il devrait tirer. Elle permet de concevoir des modèles prédictifs à partir des techniques de classification et de régression.

Techniques de classification. La reconnaissance d'écriture manuscrite utilise par exemple la classification pour reconnaître les lettres et les chiffres, l'imagerie médicale, la reconnaissance vocale et l'évaluation de crédit bancaire, la classification d'images, la détection de fraude d'identité en sont des applications classiques. En effet, la classification s'intéresse à prédire la classe des variables discrètes ou qualitatives à partir des données initiales. Les algorithmes couramment utilisés pour la classification incluent notamment les machines à vecteurs de support (SVM), le boosting/bagging d'arbres de décision, la méthode des k plus proches voisins, la classification naïve bayes, l'analyse discriminante, la régression logistique et les réseaux de neurones.

Techniques de régression. Le modèle linéaire, le modèle non linéaire, la régularisation, la régression pas à pas et le boosting/bagging d'arbres de décision sont quelques-uns des algorithmes de régression. Cette technique se

spécialise dans la prédiction des variables continues ou quantitatives. Prédire la croissance de la population, estimer la durée de vie, prévoir les marchés boursiers et la météorologie en sont des exemples d'applications que peuvent produire les algorithmes de régression.

Apprentissage non supervisé

Cette approche se veut libre et sans intervention humaine « qui consiste généralement à identifier des modèles cachés ou structures dans des collections de données non étiquetées » (SUTTON ,1998). Le regroupement et la réduction de dimension sont des techniques utilisées dans cette approche.

Le regroupement. Aussi appelée agrégation (clustering en anglais), cette technique est utilisée pour effectuer une analyse exploratoire des données afin de trouver des modèles cachés ou des agrégats d'objets similaires à partir d'un ensemble de données hétérogènes. Parmi les algorithmes de regroupement, l'on peut citer : la méthode des k-moyennes et k-médoides, le clustering hiérarchique, les modèles de mélanges gaussiens, les modèles de Markov cachés, les cartes auto-organisatrices, le clustering c-moyennes flou et le clustering soustractif. Par ailleurs, l'analyse de séquence génomique, système de recommandation, l'étude de marché et la reconnaissance d'objets en sont des exemples.

La réduction de dimension. Au chapitre précédent, nous avons abondé sur la réduction des caractéristiques comme élément d'optimisation et d'efficacité du processus de classification. De façon générale, les algorithmes

utilisés dans ce contexte sont des réponses à plusieurs types de besoins dont les plus courants sont : la compression significative de données, la découverte de structure, l'élicitation de caractéristiques et la visualisation de données massive en sont quelques exemples d'applications.

Apprentissage profond

Par l'utilisation des réseaux de neurones à plusieurs couches, il est possible de classer les objets d'un ensemble. Cette approche a été développée à la suite des faiblesses des approches supervisées ou non décrites précédemment. Elles ont montré leur limite et plus généralement leur incapacité à bien généraliser certaines tâches d'IA (Goodfellow et al, 2017) telles que la reconnaissance de la parole ou d'objets.

L'apprentissage profond peut, par exemple, aider à : mieux reconnaître des objets hautement déformables, analyser les émotions révélées par un visage photographié ou filmé, analyser les mouvements et les positions des doigts d'une main, ce qui peut être utile pour traduire le langage des signées, améliorer le positionnement automatique d'une caméra, poser un diagnostic médical (ex. : reconnaissance automatique d'un cancer en imagerie médicale), ou encore reproduire une œuvre artistique à partir d'une image.

Apprentissage par renforcement.

L'apprentissage par renforcement a pour idée de base de laisser l'algorithme apprendre de ses propres erreurs, comme l'indique (SUTTON,1998) « les problèmes d'apprentissage par renforcement impliquent d'apprendre quoi

faire - comment adapter des situations à des actions - afin de maximiser un signal de récompense numérique ». Cette approche repose donc sur un système de récompenses et de pénalités afin de permettre à l'ordinateur d'apprendre à résoudre un problème de manière autonome. Monte-Carlo, Q-learning (État – action – récompense – état), SARSA (État – action – récompense – état – action avec traces d'éligibilité), DQN (Deep Q Network) en sont quelques exemples d'algorithme dans cette approche. Elle sert également plusieurs domaines à travers les applications telles que la gestion des ressources dans les clusters informatiques, le contrôle des feux de circulation, l'entraînement des robots en robotique, la configuration du système web, l'optimisation des réactions chimiques, l'optimisation des enchères et publicités dans l'e-commerce et même le développement des jeux dont l'un des plus célèbres est le jeu AlphaGo Zero⁴⁰.

Logique de construction d'un modèle de classification

Nous allons nous intéresser dans cette section au mécanisme de construction d'un modèle d'apprentissage, nous soulignerons également les dilemmes liés à l'optimisation des modèles.

Architecture de l'apprentissage automatique

Les algorithmes d'AA disposent dans leur structure interne des mécanismes leur permettant de construire des modèles à partir des données suivant un processus qui met en évidence deux jeux de données au minimum : les données d'entraînement (DE) et les données de test (DT). Si l'on considère que le jeu de

⁴⁰ <https://deepmind.com/blog/article/alphago-zero-starting-scratch>

données d'entraînement est un ensemble de paires (X, Y) où : X : est un vecteur de caractéristiques pouvant prendre les valeurs quantitatives ou qualitatives et Y : est l'étiquette (label, en anglais) alors l'intérêt de l'AA est de trouver une fonction f telle que $Y = f(X)$ qui associe la valeur de Y aux valeurs de X . Selon le type de valeur que peut prendre Y , l'on peut définir une typologie de principaux modèles associés à f . Lorsque la valeur de Y est quantitative, l'on parle de modèle de régression, si elle est booléenne, il s'agira d'un modèle de classification binaire, l'on parle également de classification multi-classes lorsque les valeurs de Y sont qualitatives.

Les DE et DT sont utilisées pour définir un modèle selon l'architecture d'AA présentée à la figure 7.

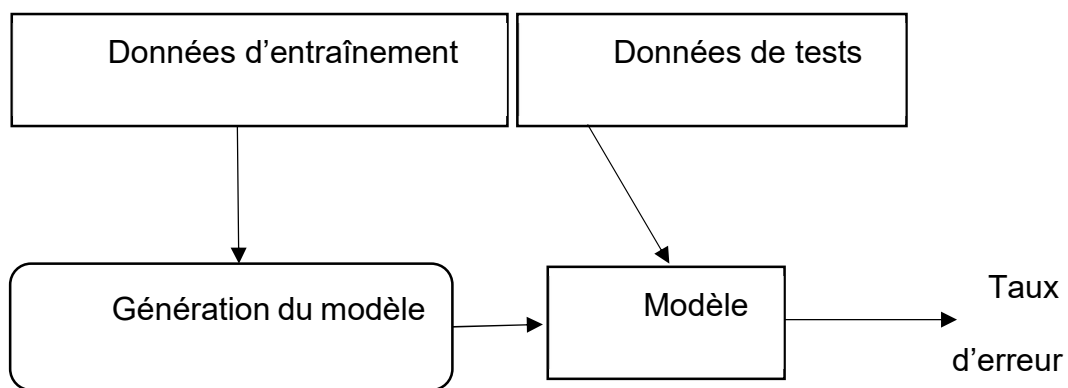


Figure 7 : Formation d'un modèle d'AA.

Source : traduit de (Leskovec et al, 2020, p. 444).

L'ensemble de DE aide à construire le modèle et l'ensemble DT aide à tester le modèle construit, en évaluant la performance ce celui-ci. Nous allons revenir sur

l'activité d'évaluation du modèle dans les sections suivantes, pour l'instant, résumons succinctement le processus de formation et d'évaluation du modèle.

Sur le plan pratique, une fois que le corpus de documents (ou données) est préparé, il est subdivisé selon les portions voulues selon la tâche à accomplir. Habituellement, l'on réserve des proportions de l'ordre de 70% pour les DE et 30% pour les DT, tout en gardant à l'idée que la précision du modèle augmente avec la volumétrie des DE (Baillargeon et al, 2019). Ces ensembles doivent être indépendants et suffisamment distribués. Le processus d'entraînement commence lorsque les DE sont introduites dans l'algorithme, qui va ainsi s'assurer de construire aux moyens des inférences statistiques et des probabilités, des associations entre les entrées (caractéristiques) et les sorties (classes), in fine, nous obtenons un modèle. Le modèle ainsi produit est utilisé avec les DT dans le but d'évaluer sa capacité de prédiction. Les performances d'un modèle sont bonnes lorsqu'il arrive à généraliser efficacement sur les DT.

Toujours dans la pratique, l'on constate l'utilisation d'un ensemble de données de validation (DV). Les DV proviennent de la subdivision des DE, et ont vocation à contribuer à la robustesse du modèle par la prévention ou la minimisation des effets du sur-apprentissage et du sous-apprentissage. On parlera alors de validation croisée.

Quelques notions de construction du modèle d'apprentissage

La performance d'un modèle est tributaire d'un corpus donné. Wolpert (1996, p. 1352) propose le théorème du « No Free Lunch" (ou théorème de l'impossibilité du déjeuner gratuit) qui stipule qu'aucun modèle n'est optimal pour tous les problèmes [ou corpus], de plus, si un modèle performe bien pour un type de problème alors il sera moins performant en moyenne pour tout autre type de problème. Cela implique indéniablement qu'il faille trouver le modèle approprié parmi ceux existants. À cet effet, plusieurs notions nécessitent des élucidations à l'effet de comprendre les tensions paramétriques internes aux modèles ainsi que leurs incidences sur leur performance. Les notions à l'instar de l'erreur du modèle, du biais, de la variance, du paramètre, de l'hyperparamètre, du sur-apprentissage et du sous-apprentissage seront abordées.

Erreur du modèle

Nous avons défini plus haut la fonction (ou modèle) d'apprentissage comme étant $Y = f(X)$, cependant nous devons être plus précis en prenant en compte l'erreur ϵ dite irréductible associée à la prédiction, soit $Y = f(X) + \epsilon$.

Elle est irréductible, car il n'est pas possible de l'éliminer, ce sont en effet des bruits dans les données (erreurs de mesures, valeurs aberrantes ou absentes), ou simplement des problématiques inhérentes aux cas modélisés).

Nous pouvons dès lors estimer le modèle f par $\hat{Y} = \hat{f}(X)$, où \hat{f} représente l'estimation de f et \hat{Y} représente le résultat de la prédiction de f , or \hat{f} ne sera pas

un bon estimateur de f , cette imprécision devra introduire des erreurs qui sont dites réductibles (car il est possible de les minimiser).

Conséquemment, à l'aide de l'équation (5-2), nous pouvons également estimer l'erreur, comme étant le carré de la différence entre la valeur prédite et la valeur actuelle de Y par $E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2$, ce qui équivaut à $E(Y - \hat{Y})^2 = [f(X) - \hat{f}(X)]^2 - Var(\epsilon)$, où $[f(X) - \hat{f}(X)]^2$ est l'erreur réductible et $Var(\epsilon)$, la variance associée à l'erreur ϵ .

En probabilité, cette erreur réductible peut se décomposer par $[f(X) - \hat{f}(X)]^2 = (E[\hat{f}(X)] - f(X))^2 + E[(\hat{f}(X) - E[\hat{f}(X)])^2]$.

Remarquons que le premier terme $(E[\hat{f}(X)] - f(X))$ représente le biais tandis que le deuxième $(E[(\hat{f}(X) - E[\hat{f}(X)])^2])$, désigne la variance. On peut finalement résumer l'erreur du modèle comme étant fonction du biais et de la variance par $Erreur(X) = Biais^2 + Variance + erreur irréductible$.

Le compromis biais et variance

Le biais et la variance sont des formes d'erreurs dans la construction d'un modèle. Ils composent l'erreur réductible dont la minimisation permet de construire un modèle qui se généralise aisément. L'erreur de biais est due à l'incapacité du modèle à apprendre des associations entre les caractéristiques (X) et la classe (Y). Il représente l'écart entre la prédiction attendue ($E[\hat{f}(X)]$) et la valeur correcte de nous essayons de prédire ($f(x)$).

L'erreur de variance est consécutive à la tendance pour un modèle à capturer intensément les relations entre les caractéristiques et leur classe. Poussée à l'extrême, la variance peut absorber toute la variabilité des DE et même les bruits des données plutôt que des classes à prédire, ce qui va conduire au sur-apprentissage et donc nuire au pouvoir de généralisation du modèle sur l'ensemble de tests. L'erreur due à la variance est l'écart quadratique entre la prédiction du modèle sur un seul jeu de données d'apprentissage ($f(X)$) et sa prédiction moyenne sur tous les ensembles d'apprentissages ($E[\hat{f}(X)]$).

Pour obtenir un modèle performant, il est indispensable de minimiser aussi bien l'erreur due au biais que celle due à la variance. Or comme l'illustre la figure 8, la variance et le biais évoluent de façon opposée, nous devons donc trouver un compromis entre la simplicité du modèle (ce qui va entraîner la réduction de la variance, mais introduire de biais) et la complexité de celui-ci (ce qui va entraîner un biais faible, mais va introduire des variances supplémentaires).

Sur-apprentissage et sous-apprentissage

Dans la modélisation prédictive, le signal et le bruit sont deux concepts sous-jacents. Alors que le signal est le modèle d'apprentissage voulu à partir des données, le bruit, en revanche, désigne les informations non représentatives du phénomène étudié ou le caractère aléatoire des données. L'apprentissage va donc consister à séparer le bruit du signal. Dans le cas contraire, le modèle va capturer le bruit, cela a pour corolaire, une haute performance de prédiction sur les DE et de piètres performances sur les DT, on parle alors de sur-apprentissage

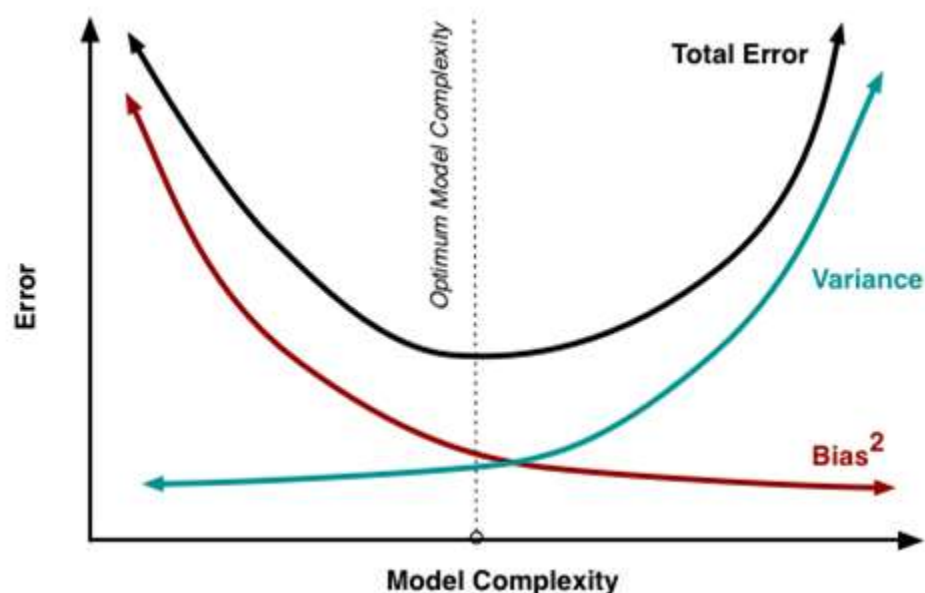


Figure 8 : Compromis entre la variance et le biais.

Source : Scott Fortmann-Roe⁴¹.

sur les DE et de piètres performances sur les DT, on parle alors de sur-apprentissage. En revanche, si le modèle est très simple, ne capturant pas assez la complexité du phénomène étudié, l'on va assister à un sous-apprentissage du modèle.

Le sur-apprentissage et le sous-apprentissage sont étroitement liés aux notions de variance et biais, en effet un biais élevé caractérise un sous-apprentissage alors qu'une variance élevée peut engendrer un sur-apprentissage. Le compromis entre le biais et la variance implique en réalité un équilibrage entre le sur-apprentissage et le sous-apprentissage (Belkin et al, 2019, p. 1).

⁴¹ <http://scott.fortmann-roe.com/docs/docs/BiasVariance/biasvariance.png>

Tandis que le sur-apprentissage peut se corriger en effectuant des activités suivantes : sélection et la réduction des caractéristiques, choisir un modèle plus simple, faire usage de la régularisation, ajuster les hyperparamètres, mettre à contribution des techniques de bagging et de rééchantillonnage et surtout entraîner le modèle sur des DE plus volumineuses. Le sous-apprentissage en revanche peut se résorber par l'ajout de caractéristiques, choisir un modèle plus flexible ou ajuster les hyperparamètres.

Paramètre et hyperparamètre

Dans le but de contrôler le comportement de l'apprentissage du modèle, certaines configurations peuvent être faites au moyen des variables. Lorsqu'une variable fait partie de la structure interne du modèle et est estimée à partir des données, l'on parle de paramètre. Les poids dans un réseau de neurones artificiels, les vecteurs de support dans une machine à vecteur de support et les coefficients dans une régression linéaire sont quelques exemples de paramètres.

À contrario, ils sont appelés hyperparamètres lorsqu'elles sont externes au modèle et non déterminées par ce dernier (Heaton et al, 2017, p. 120). Le taux d'apprentissage pour la formation d'un réseau neuronal, et les k dans l'algorithme des k plus proches voisins en sont des exemples.

Régularisation

Nous avons abordé le théorème du « No Free Lunch » qui souligne la nécessité de trouver le meilleur modèle adapté au phénomène étudié. La régularisation a pour but de contribuer à la détermination dudit modèle. C'est en

effet un ensemble de techniques utilisé pour réduire l'erreur en ajustant un algorithme de manière appropriée sur les DE, afin d'éviter le sur-apprentissage. La régularisation est donc cruciale dans l'apprentissage puisqu'elle permet finalement d'atténuer le dilemme biais-variance.

Les méthodes d'apprentissage.

Nous présenterons brièvement certaines méthodes d'apprentissage sélectionnées à dessein, eu égard à leur pertinence vis-à-vis de notre problématique de travail.

Naïve Bayes

Le classifieur Naïve Bayes met à contribution le domaine de la probabilité pour évaluer les probabilités combinées des caractéristiques et des classes afin d'estimer les probabilités des classes pour un document donné. Il utilise à cet effet, le théorème de Bayes (Berrar, 2018, p. 2) et part du postulat de l'indépendance entre les caractéristiques d'un document (Joachims, 1998). Ce postulat est d'ailleurs ce qui justifie la naïveté aussi bien de son fonctionnement que dans le nom du modèle.

Considérant un corpus constitué de plusieurs documents correspondants à k catégories $\{c_1, c_2, \dots, c_k\}$. L'objectif de l'algorithme Naïve Bayes est de calculer la probabilité conditionnelle d'un individu possédant les caractéristiques vectorielles (x_1, x_2, \dots, x_n) appartient à une classe particulière C_i . Elle est donnée par $P(C_i|x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n|C_i) * P(C_i)}{P(x_1, x_2, \dots, x_n)}$ pour $1 \leq i \leq k$. Or du fait de

l'indépendance des caractéristiques (postulat du classifieur)

$$P(x_1, x_2, \dots, x_n | C_i) = P(x_1, x_2, \dots, x_n, C_i) \quad \text{ou} \quad \text{encore} \quad P(x_1, x_2, \dots, x_n | C_i) = \left(\prod_{j=1}^{j=n} P(x_j | C_i) \right) * \frac{P(C_i)}{P(x_1, x_2, \dots, x_n)} \text{ pour } 1 \leq i \leq k.$$

Puisque $P(x_1, x_2, \dots, x_n)$ est constante quelque soit la classe, simplifions donc l'expression précédente pour aboutir à $P(C_i | x_1, x_2, \dots, x_n) \propto \left(\prod_{j=1}^{j=n} P(x_j | C_i) \right) * P(C_i)$ pour $1 \leq i \leq k$.

Plusieurs variantes de l'algorithme ont été proposées à l'effet de corriger la problématique du déséquilibre de classe (Frank et Bouckaert, 2006).

K-plus proches voisins

Dans cette méthode, le voisinage symbolise la similarité, l'idée est de regrouper les documents en fonction des classes selon leur degré de similarité. Son fonctionnement est le suivant : Étant donné un document de test, le système trouve les K voisins parmi les documents de formation et utilise les catégories des K voisins pour pondérer les catégories candidates. Le score de similarité de chaque document voisin au document de test est utilisé comme poids des catégories du document voisin. Si plusieurs des k voisins les plus proches partagent une catégorie, alors les poids par voisin de cette catégorie sont additionnés, et la somme des poids résultante est utilisée comme score de similitude de cette catégorie par rapport au document de test. Après avoir trié les valeurs de score, l'algorithme attribue à la classe, le candidat détenant le score le plus élevé du document de test (Jiang et al, 2012).

Plus formellement, la règle de décision est définie par $y(x, c_j) = \sum_{d_i \in KNN} \text{sim}(x, d_i) y(d_i, c_j)$ la suivante (Yang et Liu, 1999, p. 44). Où $y(d_i, c_j) \in \{0, 1\}$ est la classification du document d_i par rapport à la catégorie c_j , $\text{Sim}(x, d_i)$ est la similitude entre le document test x et le document d'entraînement d_i .

En guise d'exemple, supposons que l'on veuille classer les points de la figure 9. L'algorithme K plus proches voisins va consister à fixer l'hyperparamètre K, dans notre cas $K = 3$. À partir de cet instant, l'algorithme va identifier les 3 plus proches voisins à partir du point x_i selon le calcul des distances dont le choix du type peut être celui de Minkowski (équation 5-12) ou de ses variantes telles que les distances euclidiennes et de Manhattan. Nous pouvons également utiliser la distance de Mahalanobis (équation 5-13), de Hamming (Leskovec et al, 2020, p. 96), du cosinus (Leskovec et al, 2020, p. 95) et même la distance de Jaccard (Leskovec et al, 2020, p. 94). Une fois les voisins identifiés (le cercle vert de la figure 9), le point x_i est alors attribué à la classe ayant le plus grand nombre d'individus parmi ses voisins. Dans notre exemple, le point est attribué à la classe triangle orange.

Quelques mesures de distances

Étant donné $x_i = (x_i^1, \dots, x_i^K)'$ et $x_j = (x_j^1, \dots, x_j^K)'$ deux variables quantitatives de dimension K, la distance $d(x_i, x_j)$ entre x_i et x_j selon le type de mesure est la suivante :

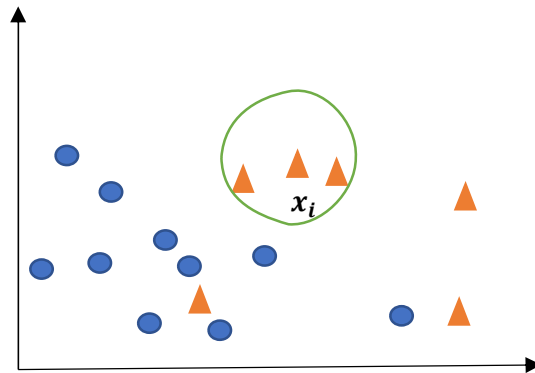


Figure 9 : Représentation de la classification K plus proches voisins avec $K = 3$ à partir d'un point x_i .

La distance de Minkowski est $d^q(x_i, x_j) = \sum_{k=1}^K (x_i^k - x_j^k)^q$.

La distance euclidienne est obtenue en remplaçant q par la valeur 2 et celle de Manhattan est déduite en donnant à q la valeur 1.

La distance de Mahalanobis prend en charge la matrice de covariance Σ et est obtenu par $d^2(x_i, x_j) = (x_i - x_j) \Sigma^{-1} (x_i - x_j)^t$.

La performance de cet algorithme dépendra des choix de la valeur de K ainsi que du type de mesure de distance à utiliser.

Machines à vecteur de support

La machine à vecteur de support (SVM, en anglais) est une méthode d'apprentissage supervisé initié en 1963 par Vapnik et Chervonenki (1964) et consistait à sélectionner un hyperplan séparant à la fois les points de l'ensemble d'apprentissages en deux classes tout en maximisant la distance entre l'hyperplan et les points les plus proches de ce dernier. Dans le contexte de la méthode, cette distance à maximiser est appelée marge (voir la figure 10).

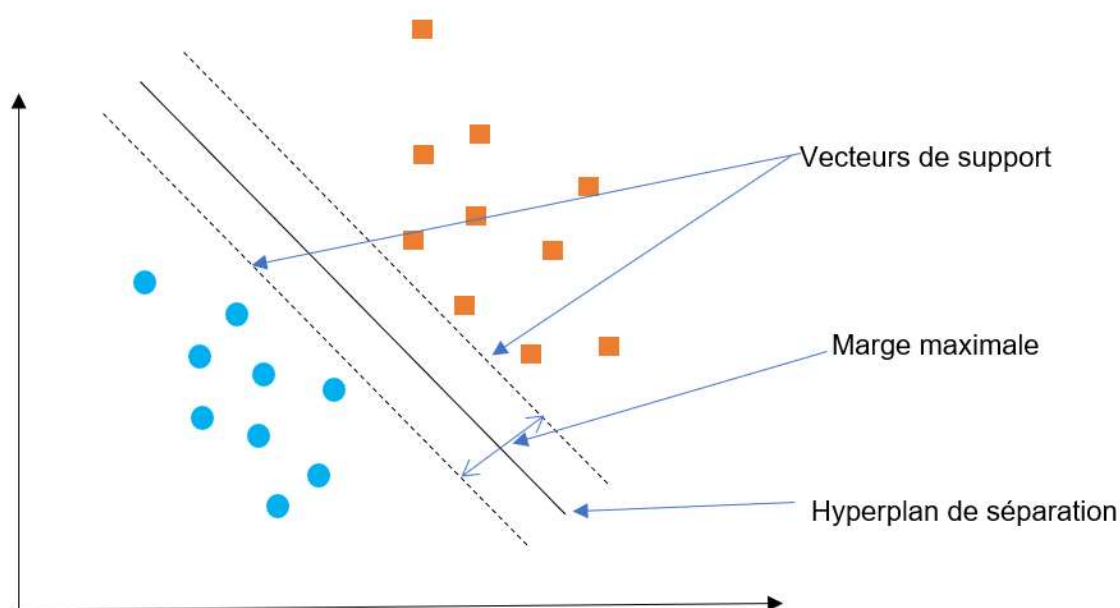


Figure 10 : SVM avec maximisation de marge à séparation linéaire.

Un hyperplan

Considérant un espace à p dimensions, un hyperplan est un sous-espace affine plat de dimension $p - 1$ (Gareth et al, 2013, p. 338). Mathématiquement, un hyperplan est donné par l'équation $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p = 0$, où $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ sont les paramètres et $X = (X_1, X_2, \dots, X_p)^T$ les points sur l'hyperplan. Dire qu'un hyperplan sépare les données en deux parties, revient à dire qu'il existe des points Y tels que $\beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \dots + \beta_p Y_p > 0$ et des points Z tels que $\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \dots + \beta_p Z_p < 0$. On dit alors les points Y sont situés d'un côté de l'hyperplan alors les points Z sont du côté opposé.

Définition de la maximisation de la marge

À ce stade, l'hyperplan permet de séparer les données d'entraînement en deux classes, tout comme le fait le modèle du perceptron (Leskovec et al, 2020, p. 441). Néanmoins, la séparation originelle avec l'hyperplan à ce stade renferme des limitations telles que : l'impossibilité de séparer les données selon une certaine dispersion par un hyperplan (voir image de droite de la figure 11), difficulté de choisir un hyperplan lorsqu'il en existe plusieurs (voir image de gauche de la figure 11) et d'autres limitations algorithmiques (Leskovec et al, 2020, p. 457). La méthode SVM vient corriger ces problèmes, en d'autres termes, elle sélectionne un hyperplan qui sépare les données en deux classes tout en maximisant la distance entre l'hyperplan et les points les plus proches des DE (Demidova et al, 2016).

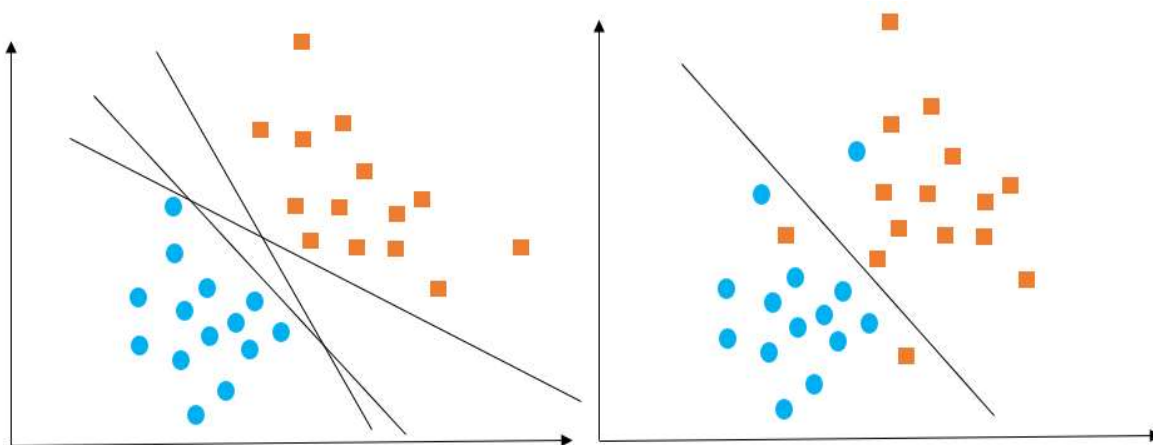


Figure 11 : gauche : Trois hyperplans possibles pour séparer les points. Droite : impossibilité de séparer tous les points.

En supposant que les points doivent être classés dans deux classes, y tel que $y_1, y_1, \dots, y_n \in \{-1, +1\}$ (+1 est la bonne classe et -1 la mauvaise). On définit la capacité de classification de l'hyperplan de la manière suivante : $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p > 0$ Si $y_1 = 1$ et $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p < 0$ Si $y_1 = -1$.

Comme nous l'avons souligné plus haut, à ce stade, l'on peut avoir plusieurs hyperplans. Il convient donc de maximiser la marge de part et d'autre de l'hyperplan en trouvant un hyperplan donné le plus éloigné des DE et déterminer la marge (qui est la distance minimale perpendiculaire entre chaque point des DE et ledit hyperplan).

L'hyperplan à considérer est celui dont la marge est la plus élevée (ou qui a la distance minimale la plus éloignée des DE). La figure X1 montre deux individus des DE à équidistance de l'hyperplan de marge maximale et se trouvant le long des lignes pointillées indiquant la largeur de la marge. Ces deux observations sont appelées les vecteurs de support.

Formulation de la maximisation de marge

Nous allons aborder la formulation mathématique de la maximisation de la marge de l'hyperplan. Soit n DE $x_1, \dots, x_n \in \mathbb{R}^p$ associé aux classes $y_1, y_1, \dots, y_n \in \{-1, +1\}$. L'on définit la maximisation de l'hyperplan comme étant le problème d'optimisation (Gareth et al, 2013, p. 343). Soit *Maximise* M et $\beta_0, \beta_2, \dots, \beta_p, M$ tel que $\sum_{j=1}^p \beta_j^2 = 1$ où $y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \geq M$ (Pour $i = 1, \dots, n$).

Ce problème d'optimisation revient à bien choisir les $\beta_0, \beta_2, \dots, \beta_p$ à l'effet de maximiser la marge M .

Cas linéairement inséparable : vers la séparation non linéaire

Au début des années 90, une approche non linéaire de séparation de marge a été proposée (Bo et Xianwu, 2006). L'astuce consiste à agrandir l'espace des caractéristiques à l'aide des fonctions des prédicteurs, comme les termes quadratiques et cubiques, afin de remédier à cette non-linéarité. Par exemple, plutôt que d'utiliser p caractéristiques, on va doubler le nombre, en effet on passera de X_1, X_2, \dots, X_p à $X_1, x^2_1, X_2, x^2_2, \dots, X_p, X^2_p$.

Les formulations de la section précédente devront être ajustées à la nouvelle réalité d'agrandissement comme suit (Gareth et al, 2013, p. 350) : Soit *Maximise* M et $\beta_0, \beta_{11}, \beta_{12}, \dots, \beta_{p1}, \beta_{p2}, \varepsilon_1, \dots, \varepsilon_n, M$ tel que $y_i(\beta_0 + \sum_{j=1}^p \beta_{j1} x_{ij} + \sum_{j=1}^p \beta_{j2} x^2_{ij}) \geq M(1 - \varepsilon_i)$ où $\sum_{j=1}^p \varepsilon_i \leq C$, $\varepsilon_i \geq 0$, $\sum_{j=1}^p \sum_{k=1}^2 \beta^2_{jk} = 1$

La frontière de décision dans l'espace des caractéristiques d'origine est un polynôme quadratique, et ses solutions sont généralement non linéaires. Un exemple de cette transformation permet d'apercevoir des séparations polynomiales à la figure 12.

En résumé, dans un contexte d'explosion des caractéristiques comme dans les cas de la CAT, la SVM est particulièrement intéressante. Le modèle gère assez bien les caractéristiques non pertinentes tout en étant autant robuste vis à vis du sur-apprentissage car son entraînement ne dépend guère de l'espace de caractéristiques, relève Joachims (1998).

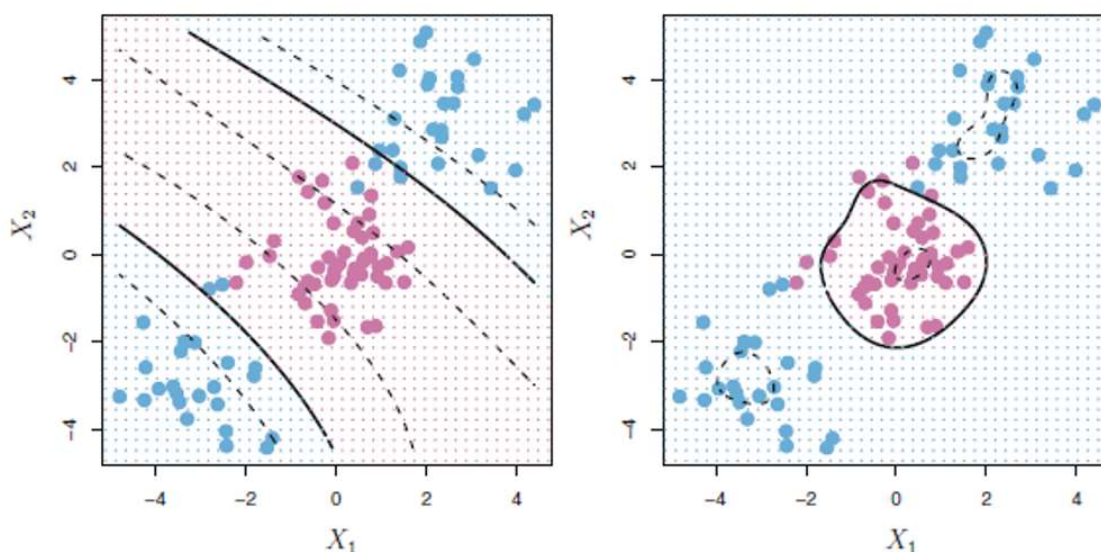


Figure 12 : SVM à séparation polynomiale.

Source : tiré de (Gareth et al, 2013, p. 353)

Arbre de décision

Les arbres de décisions permettent de construire une structure hiérarchique des données à partir de leurs caractéristiques. Ils comportent des nœuds et de feuilles reliées par des branches. La construction de l'arbre (les choix des nœuds et des feuilles) est basée sur les concepts provenant de la théorie de l'information, soit l'entropie et le gain d'information.

L'entropie

L'entropie d'une distribution de probabilité est la quantité d'information qu'elle apporte. Soit un ensemble de données D de n classes (C_1, C_2, \dots, C_n), la quantité d'information nécessaire pour identifier la classe d'un individu de l'ensemble T est obtenue par l'entropie $E(P)$, avec P la distribution de la

probabilité de la partition (C_1, C_2, \dots, C_n) , $P = (\frac{C_1}{|T|}, \frac{C_2}{|T|}, \dots, \frac{C_n}{|T|})$, où C_n correspond au nombre d'éléments de la classe i . Or, pour une distribution $P = (p_1, p_2, \dots, p_n)$, l'entropie est obtenue par $E(P) = -\sum_{i=1}^n p_i \log p_i$. L'entropie de T est alors formulée par $E(T) = -\sum_{i=1}^n \frac{|C_i|}{|T|} \log \frac{|C_i|}{|T|}$. Si en plus, l'ensemble T est partitionné en (T_1, T_2, \dots, T_m) où m est le nombre de valeurs de l'attribut X , l'identification de la classe d'un individu de T_i est quantifié par l'entropie $E(X, T) = -\sum_{j=1}^m \frac{|T_j|}{|T|} E(T_j)$.

Le gain d'information

Le gain d'information de T par rapport à T_j représente la variation de l'entropie causée par la partition de T selon T_j . Le choix du nœud est attribué à la caractéristique ayant le gain le plus élevé. Le gain d'information est formulé par $Gain(X, T) = E(T) - E(X, T)$, soit $Gain(X, T) = E(T) - \sum_{j=1}^m \frac{|T_j|}{|T|} E(T_j)$.

L'indice de Gini

Nous pouvons possiblement utiliser l'indice d'impureté, qui mesure la qualité d'un nœud et son pouvoir discriminant, à cet effet, cet indice aussi appelé le gain de Gini est obtenue par l'équation $Gain\ de\ Gini(X, T) = \sum_{j=1}^m \frac{|T_j|}{|T|} Gini(T_j)$, où Gini est l'indice de Gini dont la formulation est $Gini(T) = 1 - \sum_{j=1}^m (\frac{|C_j|}{|T|})^2$.

En résumé, outre la simplicité des arbres de décision quant à leur interprétation du fait de leur représentation graphique, ils peuvent également prendre en charge des caractéristiques qualitatives sans avoir à les transformer.

Toutefois, ils sont peu robustes et très sensibles au nombre de classes, voyant ainsi la dégradation de leur performance avec un nombre important de classes. De plus, le changement des DE nécessite de reconstruire l'arbre de décision.

Les méthodes d'ensemble à base d'arbre de décision

Les inconvénients liés à la performance et à la sensibilité aux DE des arbres de décisions peuvent être résorbés en agrégeant de nombreux arbres de décision et en utilisant des méthodes ensemblistes qui permettent de combiner plusieurs modèles d'apprentissage (Zhou, 2012, p. 15). Quelques-unes de ces techniques sont le bagging, la forêt aléatoire et le boosting.

Bagging

Le bagging (ou bootstrap aggregation) est une technique destinée à réduire la variance dans les modèles d'AA afin d'accroître la précision des prédictions. Pour ce faire, cette technique postule qu'il est indispensable de déterminer plusieurs ensembles d'entraînement du corpus initial et construire des modèles de prédiction sur chacun de ces ensembles, pour enfin agréger les prédictions (Breiman, 1996). La figure 13 en illustre d'ailleurs le principe. De plus, l'on observe que pour un ensemble de n observations indépendantes E_1, \dots, E_n , chacune des variances σ^2 , la variance de la moyenne \bar{E} des observations est donnée par σ^2 / n , ce qui montre bien que la moyenne de l'ensemble des observations réduit la variance.

L'on définit formellement le bagging de la façon par $\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x)$, où B est le nombre d'ensembles de DE obtenu par la technique de bootstrap en

répétant les exemples de DE. $\hat{f}^{*b}(x)$ est l'estimateur obtenu en entraînant la méthode sur ces ensembles identiques.

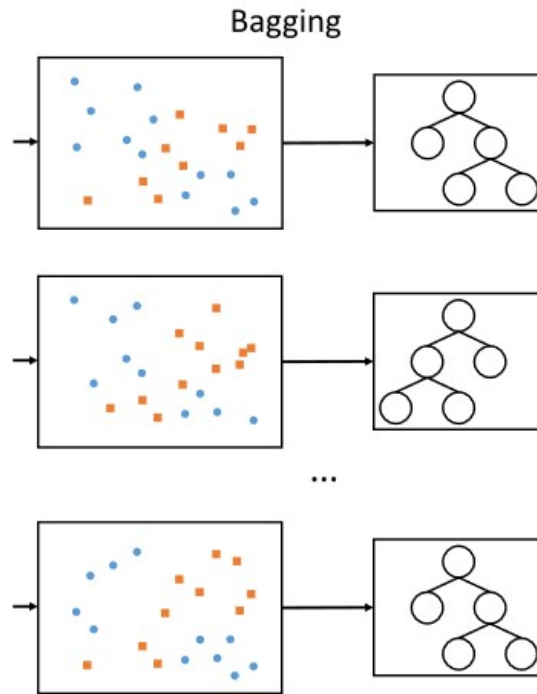


Figure 13 : Entraînement indépendant de trois modèles sur des ensembles identiques.

Source : image tirée de (González et al, 2020, p. 6).

Forêt aléatoire

Sous l'inspiration de Amit et Geman, Leo Breiman propose la forêt aléatoire (FA), qui est en effet une extension de la méthode de bagging (Zhang et al, 2012, p. 157). La FA, autant utilisé pour la classification que pour la régression, est une méthode d'ensemble d'apprentissage supervisé qui entraîne un certain nombre d'arbres de décision afin de pallier aux problématiques du sur-apprentissage observées dans le cadre du modèle d'arbre de décision. La construction des

arbres se fait à l'aide de chacun des sous-espaces sélectionnés aléatoirement dans l'espace des caractéristiques (Ho, 1995). Les arbres ainsi générés sont ensuite combinés pour décider suivant un processus de vote majoritaire de quelle classe sera l'objet de test (voir la figure 14).

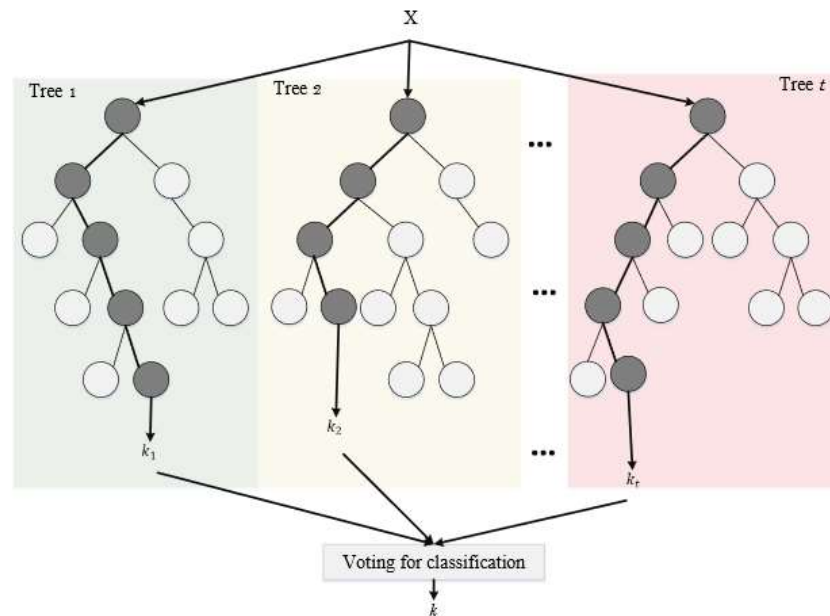


Figure 14 : Vote dans la forêt aléatoire.

Source : image tirée de (Kowsari, 2019, p. 33).

La FA diffère du bagging, tandis que la méthode bagging considère pour ses ensembles d'entraînement l'ensemble des caractéristiques (c), la FA en choisit un nombre plus réduit (m) (Gareth et al, 2013, p. 320). Autrement dit, la FA est une extension de la méthode de bagging qui réduit son espace des caractéristiques pour ses ensembles d'entraînement (habituellement $m = \sqrt{c}$).

Boosting

Tout comme le bagging qui combine plusieurs classifieurs sur des

ensembles de données, le boosting opère ce type de combinaison d'une autre manière. Premièrement, le boosting n'utilise pas l'échantillonnage bootstrap, au lieu de cela, une version modifiée de l'ensemble de données d'origine est utilisée pour générer les arbres de manière séquentielle. L'illustration du principe est présentée à la figure 15. Plusieurs déclinaisons algorithmiques ont été proposées dont les plus éminents sont : adaboost (González et al, 2020, p. 7) et gradient boosting et sa version extrême (González et al, 2020, p. 11, 20).

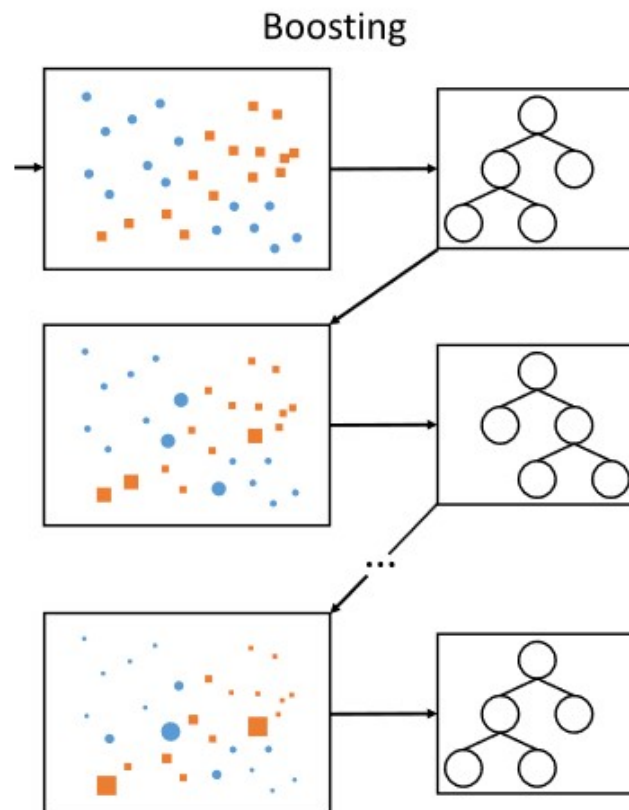


Figure 15 : Entraînement séquentiel de trois modèles sur des ensembles.

Source : image tirée de (González et al, 2020, p. 6).

Évaluation des classifieurs

Nous allons à présent évaluer les performances du modèle de classification de documents construit au terme des étapes précédentes du processus de classification. À cet effet, Nous présenterons succinctement les différentes mesures de performance telles que la précision, le rappel, la spécificité, le taux de succès, la F-mesure, la micro-moyenne et la macro-moyenne qui sont par ailleurs basées sur la matrice de confusion (Ruuska, 2018, p. 57) dont l'illustration est faite au tableau 3, suivie de sa description au tableau 4.

Tableau 3 : Matrice de confusion

		Valeurs réelles	
		-	+
Valeurs prédites	-	TN	FN
	+	FP	TP

Nous allons présenter quelques mesures les plus utilisées dans la classification de documents. Chacune d'elle dispose des avantages par rapport aux autres (Lever et al, 2016) et donc leur combinaison révèle davantage d'information sur la performance des classifieurs. L'indice i représente une classe.

L'erreur totale

L'erreur totale commise par un classifieur est le nombre de documents

Tableau 4 : Description des métriques de la matrice de confusion

Mesure	Définition
TP (vrai positif)	Documents qui doivent être marqués comme appartenant à une classe particulière et le sont effectivement.
FP (faux positif)	Souvent appelée « fausses alarmes », elle détermine le nombre de documents qui serait incorrectement lié à la classe.
FN (faux négatif)	Elle caractérise les erreurs de non-détection et représente ainsi le nombre de documents qui n'est pas marqué comme lié à une classe, mais qui devrait l'être.
TN (vrai négatif)	Documents qui ne doivent pas être marqués comme appartenant à une classe particulière et ne le sont effectivement pas.

incorrectement classés. Elle est donnée par l'équation $E_i = FP_i + FN_i$.

Le taux de faux positifs et taux de faux négatifs

Ils désignent l'incapacité du classifieur à catégoriser un document. Comme mentionné dans le tableau 4, l'on distingue, le taux des « fausses alarmes » ou faux positifs ($FPR_i = \frac{FP_i}{FP_i + TN_i}$) et le taux de « non-détection » ou faux négatifs

$$(FNR_i = \frac{FN_i}{FN_i + TP_i}).$$

La précision

Elle représente le taux de prédiction positive, soit $\pi_i = \frac{TP_i}{TP_i + FP_i}$.

Le rappel

Encore appelé sensibilité, il représente le taux de vrais positifs, soit la capacité à détecter correctement les points positifs, $\rho_i = \frac{TP_i}{TP_i + FN_i}$.

La spécificité

Elle représente le taux de vrais négatifs, soit le pouvoir de détection des points négatifs, elle s'obtient par $SP_i = \frac{TN_i}{TN_i + FP_i}$.

Le taux de succès

Ce taux indique la capacité de classification du classifieur. Il est à 100% lorsque tous les documents ont été classés dans la bonne classe, il est défini par

$$A_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}.$$

F_β score

Dans un contexte de déséquilibre des classes, la précision s'avère insuffisante, d'où la nécessité de le combiner avec le rappel afin d'obtenir une meilleure image des performances du classifieur. Cette mesure représente la moyenne harmonique du rappel et de la précision, $F_\beta = \frac{(\beta^2 + 1)\pi\rho}{\beta^2\pi + \rho}$, où β est un paramètre positif, π et ρ désignent respectivement la précision et le rappel.

L'on peut choisir de privilégier la précision ou le rappel en fonction des besoins de classification, β peut être fixé à 1 pour donner à la précision et au

rappel la même importance. En supposant que $\beta = 1$, l'on obtient $F_\beta = \frac{2\pi\rho}{\pi+\rho}$. En remplaçant la précision et le rappel pour leurs valeurs respectives, l'on déduit finalement $F_1 = \frac{2TP_i}{2TP_i+FP_i+FN_i}$.

Macro-moyenne et micro-moyenne

Les scores de moyenne macro et micro permettent de voir dans quel rapport les classes impactent individuellement les performances. Soit B , la mesure d'évaluation binaire, C , le nombre de classes et i est une classe quelconque, la macro moyenne se calcule par l'expression $B_{macro} = \frac{1}{|C|} \sum_{i=1}^{|C|} B(TP_i + FP_i + TN_i + FN_i)$. La micro moyenne se calcule comme étant $B_{micro} = B(\sum_{i=1}^{|C|} TP_i, \sum_{i=1}^{|C|} FP_i, \sum_{i=1}^{|C|} TN_i, \sum_{i=1}^{|C|} FN_i)$.

Dans un contexte multi-classe, la micro-moyenne est importante en cas de déséquilibre de classe.

Coefficient de corrélation de Matthews

Utilisé dans la classification binaire, le coefficient de corrélation de Matthews (MCC, en anglais) mesure la qualité du classifieur. Il est particulièrement efficace dans les situations de déséquilibre de classes. Ce coefficient $F_1 =$

$$\frac{TP_i + TN_i - FP_i * FN_i}{\sqrt{(TP_i + FP_i) * (TP_i + FN_i) * (TN_i + FP_i) * (TN_i + FN_i)}}.$$

La courbe ROC

Sensible au déséquilibre de classes, la courbe ROC (Receiver Operating

Characteristics) (Fawcett, 2006) représente sous forme de courbe le taux de vrais positifs ($TPR_i = \frac{TP_i}{TP_i + FN_i}$) par rapport aux taux de faux positifs ($FPR_i = \frac{FP_i}{FP_i + TN_i}$).

Une mesure en relation avec la courbe ROC est l'aire sous ladite courbe appelée AUC (Area Under ROC Curve, en anglais). En effet, elle mesure l'aire sous la courbe ROC. Elle explique la probabilité que le modèle classe un exemple positif aléatoire plus haut qu'un exemple négatif aléatoire. Pour une classification binaire, l'AUC est défini (Kowsari, 2019, p. 4) l'expression suivante $AUC = \int_{-\infty}^{\infty} TPR(T)FPR'(T)dT$, qui est équivalente à $AUC = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(T > T')f_0(T)dTdT'$ ou encore à $AUC = P(X_1 > X_0)$.

Une classification multi-classes où C est le nombre de classes, on définit l'AUC par l'égalité $AUC = \frac{2}{|C|(|C|-1)} \sum_{i=1}^{|C|} AUC_i$.

Travaux connexes sur la détection de la FS

L'on peut aisément imaginer que la détection de la FS s'apparente à la détection des pourriels dans le cadre de la messagerie électronique. Plusieurs travaux s'y sont consacrés (Zhang et al, 2004 ; Cormack, 2008; Sanz et al, 2008; Dada et al, 2019; Bhowmick et al, 2016). Ces travaux utilisent l'apprentissage automatique pour extraire les courriels illégitimes de l'ensemble des courriels.

Toutefois à l'observation, la nature des données du courriel n'est pas similaire à celle des données de la FS. L'on dispose sur un site de rencontre, les données du fraudeur communément appelé profil (nom propre, nom utilisateur, âge, genre, localisation, latitude, longitude, pays, ethnie, occupation, état

matrimonial, courriel, religion, orientation sexuelle, intention de recherche et une description de sa personnalité) et les données échangées avec les victimes. Les travaux cités ci-dessus ne prennent pas en compte le profil du fraudeur (du moins dans son exhaustivité) et se focalisent principalement sur le contenu des courriels échangés. En revanche, Li et Shen (2011) détectent les pourriels en rajoutant à leur approche, l'enrichissement du profil avec le contexte social de l'utilisateur.

Au-delà des solutions liées à la messagerie électronique, nous avons trouvé quatre travaux récents dédiés à la FS qui proposent des solutions de nature quantitative et qualitative :

- Quantitativement, les deux travaux ci-après fournissent des résultats quantifiables quant aux résultats des expériences de modélisation en apprentissage automatique menées.
- Suarez-Tangil et al (2020) ont pour ambition de fournir un système de détection précoce pour arrêter les fraudeurs lorsqu'ils créent des profils frauduleux ou avant de s'engager avec des victimes potentielles. De ce fait, ils ont mis en place un modèle destiné au fournisseur de site de rencontres qui combinent plusieurs éléments du profil tels que : les caractéristiques démographiques, les images et la description de la personnalité. Ces trois types de données sont prétraités suivant leur nature pour ensuite être utilisés pour entraîner un classifieur. Finalement, il en ressort un modèle qui performe autour d'une précision de 97% avec une nette capacité de robustesse face au profil incomplet.

- Koen de (2019) utilise l'apprentissage automatique pour détecter de faux profils uniquement avec les photos du profil. Il entraîne quatre classifieurs (Naïve Bayes, SVM, arbre de décision et la forêt aléatoire) sur un ensemble d'images étiquetées en deux classes de profil (vrai et faux). Sa meilleure performance est réalisée avec le classifieur forêt aléatoire où il obtient une précision de 92,4% et un taux de faux négatifs de 19,7%.
- Qualitativement, Edwards et al (2018) et Huang et al (2015) traitent des aspects ontologiques de la FS. Alors que le premier travail, se focalise à déterminer et à analyser les caractéristiques géographiques (localisation et adresse IP du fraudeur) afin de permettre une meilleure compréhension des origines de la FS, le deuxième travail analyse les comptes frauduleux des sites de rencontre à l'effet de fournir une taxonomie des différents types de fraudeurs. Le bénéfice de ces travaux réside dans le fait qu'ils montrent que différents types d'arnaqueurs ciblent une démographie différente sur le site de rencontre, et créent donc des profils avec des caractéristiques appropriées.

In fine, les apports de ces travaux se focalisent soit, à organiser les profils ou à déterminer la véracité de ces derniers. En effet, plusieurs indices peuvent augmenter la suspicion de faux profil, l'on retrouve habituellement une contradiction entre le lieu de présence du fraudeur et l'origine de son adresse IP, une adresse IP derrière un proxy, un langage suspect d'écriture, une photo de

profil déjà utilisé par d'autres fraudeurs, etc. Nous pensons que cette approche est insuffisante, car en se basant par exemple sur les images de profil, les fraudeurs peuvent s'adapter en vérifiant l'occurrence de l'image sur les médias sociaux ou dans les résultats de recherche d'image inversée de Google avant de l'utiliser. De même, la description du profil peut être soignée de façon à enlever tout doute.

Une approche plus holistique qui prend en compte non seulement le profil des utilisateurs, mais également les communications échangées entre les partenaires augmenterait les performances de la détection de la FS tout en minimisant aussi bien le taux de faux positifs que le taux de faux négatifs. Dans le cadre de notre travail, nous allons explorer les communications échangées sur lesquelles nous allons construire un classifieur.

Conclusion

Dans ce chapitre, nous avons exploré la construction d'un modèle de classification en insistant sur les notions de bruit et de signal dans les données à étudier. Ces notions sont à la base même de l'apprentissage automatique et sont donc inhérentes aux questions de sous/sur-apprentissage et du compromis biais-variance. Nous avons étudié les algorithmes habituellement utilisés dans le champ de la CAT (Naïve Bayes, k-plus proches voisins, machines à vecteur de support et arbre de décision) et les principales métriques d'évaluation. Nous constatons pour finir, que l'ossature des travaux connexes est essentiellement

basée sur l'approche de détection par le profil des utilisateurs plutôt que par les communications échangées entre utilisateurs.

Chapitre 6 : Modélisation de la détection de la fraude sentimentale

Plusieurs avenues sont possibles dans l'élaboration d'un modèle d'apprentissage automatique performant. Le modèle, ainsi que son processus de construction sont inhérents aussi bien à la quantité qu'à la qualité des données recueillies. En nous basant sur la méthodologie présentée au chapitre 4 (section 4.2.2), nous allons bâtir un modèle de détection automatique de la fraude sentimentale. Dans un premier temps, il sera question de la collecte de données de diverses sources, ensuite, nous allons mener des statistiques descriptives portant sur le langage et les sentiments inhérents à notre corpus. Pour finir, nous nous attellerons à construire notre modèle par étape successive et par expérimentation des modèles et méthodes existantes.

Modélisation automatique continue

Comme le montre la figure 16, l'implémentation du modèle suit en effet plusieurs étapes pratiques qui se déclinent suivant cinq activités : la collecte, l'analyse exploratoire et le prétraitement de données, l'application et l'évaluation des modèles d'apprentissage automatique.

L'une des contraintes que nous nous sommes imposés était la construction d'un modèle qui s'adaptera aux stratégies changeantes des fraudeurs. Cela a été rendu possible à travers l'automatisation du processus de CAT. En effet, nous avons mené le développement d'une application internet qui prend en charge l'ensemble des composantes nécessaires à la modélisation continue.

Nous qualifions d'expérimentation, le processus de classification. Les

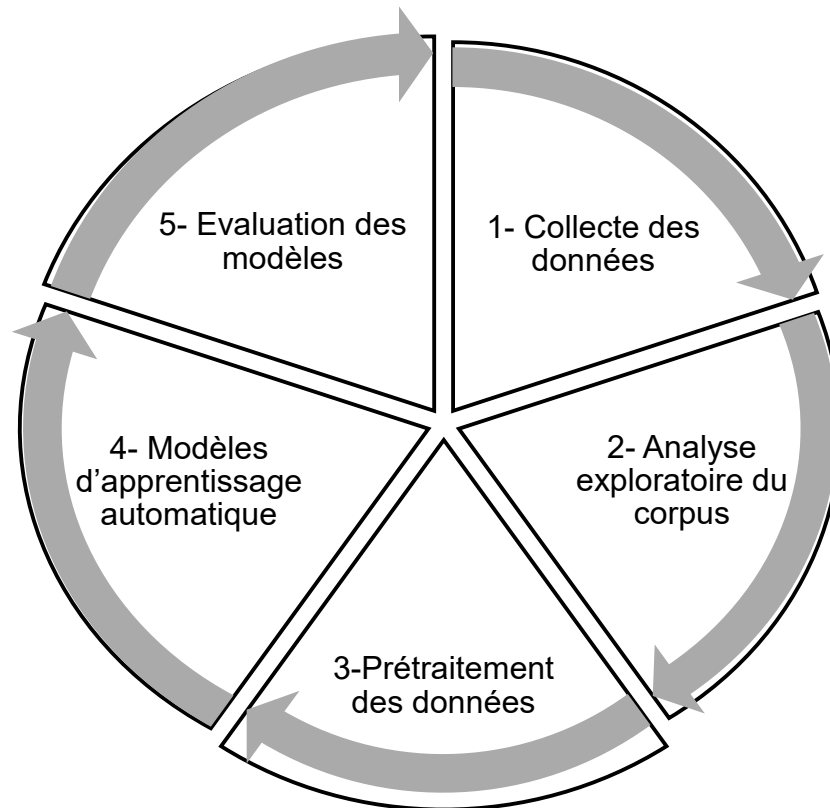


Figure 16 : Processus de modélisation en apprentissage automatique.

expérimentations sont alors menées plusieurs fois dans le temps au gré de l'actualisation des données. Le but est essentiellement de déterminer les hyperparamètres des méthodes qui assurent des performances supérieures.

Collecte des données

Nous allons aborder l'activité de collecte de données, non sans indiquer la nature, l'origine ainsi que leur qualité. Nous présenterons également les embûches liées à ladite collecte.

Nature des données

L'information de la fraude sentimentale a essentiellement pour support le

courriel. Bien que la rencontre entre la victime et son fraudeur débute sur les sites de rencontres (et dans une moindre mesure, sur des messageries et des réseaux sociaux), elle se développe au travers des échanges de courriels. L'information est donc principalement de nature non structurée et semi structurée. Un courriel contient le corps du message qui fait office d'information non structurée et un entête de courriel représentant l'information structurée (date, auteur, destinataire, etc.).

Difficulté de collecte de données de qualité

Comme nous l'avons indiqué au chapitre 2 de ce travail, la fraude sentimentale survient avec des conséquences psychologiques chez les victimes. Ces victimes sont alors moins enclins à signaler ces cas de fraudes, au risque d'être jugé, aggravant ainsi leur état émotionnel. L'information recherchée se trouve dans des boîtes de courriels des victimes et des fraudeurs et donc difficiles d'accès. Quand bien même les victimes décident de signaler l'information aux autorités compétentes, ceci dans le but de mettre en branle le pouvoir judiciaire à l'effet de recouvrir les sommes spoliées et mettre aux arrêts les fraudeurs, l'information est alors disponible chez l'autorité en question, mais pas nécessairement disponible aux sollicitations extérieures. Malgré nos demandes et nos relances, nous n'avons pas pu avoir accès à ces données privilégiées du centre antifraude du Canada.

Origine des données

Ayant éprouvé des difficultés à la collecte des données reflétant des cas

réels entre les fraudeurs et leurs victimes, nous nous sommes résolus à glaner sur internet des données de nature diverse et variée dont les spécifications et origines sont présentées ci-après.

Les modèles de courriel des réseaux criminels

Ces documents sont des modèles de courriels utilisés par des fraudeurs. Ils sont généralement achetés sur des sites des organisations spécialisées pour ce type de fraude. Le premier document a été cité comme preuve dans le procès d'Olayinka Sunmola, condamné à 27 ans de prison pour fraude postale, fraude électronique, complot et extorsion entre États.

- <https://www.bbb.org/stlouis/get-consumer-help/romance-scam-scripts/>

Le deuxième document est de source inconnue dont le lien est le suivant :

- <https://assets.documentcloud.org/documents/6544402/Nigerian-Scammers-Playbook.pdf>

Les documents relatifs au thème de l'amour

Ces documents sont extraits pour la plupart des sites proposant des textes de chanson d'amour. Ce sont des déclarations d'amour en chanson, elles sont intéressantes, car elles permettent de simuler les communications amoureuses.

- <https://matchlessdaily.com/love-songs-lyrics-for-your-boyfriend/>
- <https://www.yourtango.com/2018318763/best-love-quotes-song-lyrics-have-romantic-meanings>
- <https://www.scriptsbug.com/script/working-girl-1988>

Les courriels de la base de données Enron

Nous avons également intégré les courriels en provenance d'une base de données publiques <https://www.cs.cmu.edu/~.enron/> .

Les documents de domaine divers.

Ce sont des documents de nature variée qui ont une distance sémantique à priori lointaine par rapport à problématique étudiée.

Ce sont des articles économiques et politiques concernant l'Afrique :

- <https://www.latimes.com/california/story/2020-08-21/lori-loughlin-mossimo-giannulli-college-admissions-scandal-sentencing>
- <https://www.washingtonpost.com/>
- <https://academic.oup.com>

Des articles des organismes de lutte contre la fraude sentimentale

- <https://www.fbi.gov/video-repository/fbi-statement-on-the-arrest-of-former-uber-cso-for-covering-up-2016-hack/view>
- <https://www.scamwatch.gov.au/types-of-scams/dating-romance>
- <https://www.ic3.gov/media/2011/110429.aspx>

Constitution et qualification du corpus

Le corpus ainsi constitué fait référence à 655 documents, dont les deux moitiés sont respectivement des documents légitimes et illégitimes.

Exemple de document illégitime : « *Simply said... I love you... Being with you is like having every single one of my wishes come true. Loving you has been the*

best thing to ever happen to me! Just had to let you know... you're the best! I love you! There is no long distance about love; it always finds a way to bring hearts together, no matter how many miles are between them. »

Exemple de document légitime: *« Every night in my dreams I see you, I feel you, That is how I know you go on Far across the distance And spaces between us You have come to show you go on. »*

Les proportions de données suivant la nature et le volume sont présentées dans le tableau 5.

Analyse exploratoire du corpus

Nous allons découvrir les principales caractéristiques du corpus ainsi constitué. Trois aspects seront considérés dans cette analyse, premièrement nous réaliserons une exploration des caractéristiques langagières, ensuite nous identifierons les déterminants statistiques et enfin nous pratiquerons une analyse des sentiments de chacun des documents dudit corpus tout en utilisant la visualisation des données comme support à nos analyses.

Exploration du langage

Nuage de mots

Le nuage de mot permet d'observer la fréquence d'utilisation des mots dans un texte. Nous avons obtenu deux nuages des 100 mots les plus fréquemment utilisés, l'un représentant l'ensemble des messages légitimes (voir figure 17, image de gauche) et l'autre, les messages illégitimes (image de droite). On peut aisément constater l'utilisation abondante du mot « *love* » par les fraudeurs,

Tableau 5 : Répartition des documents du corpus

Nature des données	Volume des documents
Modèle de courriel de la fraude sentimentale	50%
Courriels de la base de données d'Enrol	5%
Articles politiques et économiques en Afrique	15%
Article fraude sentimentale	10%
Lyriques de chansons d'amour	20%

cependant, ce mot est également utilisé dans le corpus des messages légitimes sans doute avec l'introduction des lyriques de chansons d'amour. Cette explication peut également valoir pour les mots : « like », « want », « know », « time » qui apparaissent dans les deux classes, mais avec des fréquences



Figure 17 : Nuage de mots des messages légitimes (à gauche) et illégitimes (à droite).

sensiblement différentes. Si les mots « *make* », « *time* », « *need* » sont davantage utilisés dans les messages légitimes, « *life* » et « *thank* » le sont dans messages illégitimes.

PUNCT (ponctuation), SCONJ (conjonction de subordination), SYM(symbole), VERB(verbe), X(autre), SPACE(espace).

La figure 19 présente la distribution des parties du discours du corpus. À l'observation, les messages illégitimes utilisent plus de pronoms et d'adverbes que ceux des messages légitimes, qui sont particulièrement composés de noms propres et plus d'espace et de symboles.

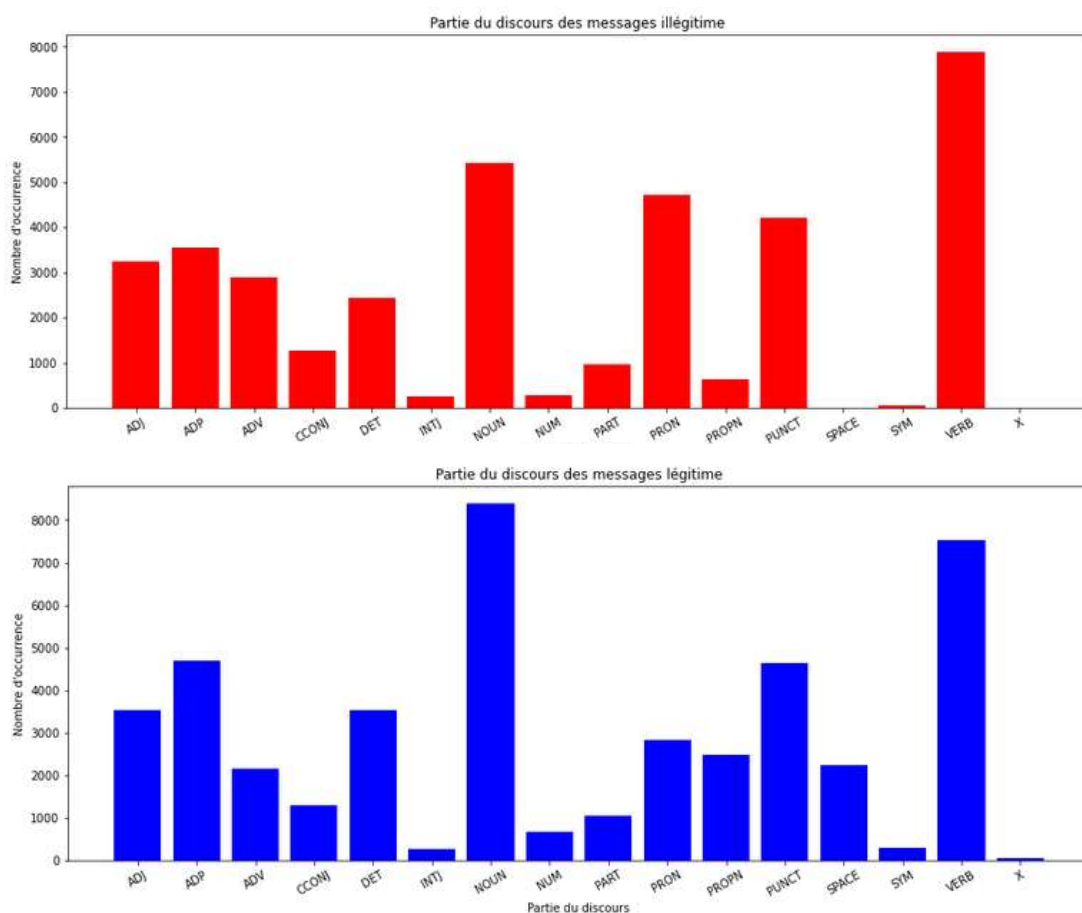


Figure 19 : Histogramme des parties du discours des messages légitimes et illégitimes.

Exploration des valeurs sentimentales

La notion de sentiment a trait à notre sujet, nous allons explorer les

constituants de la valeur sentimentale notamment en nous intéressant à la polarité (Devitt et Ahmad, 2007) et à la subjectivité (Liu, 2010). Nous représentons à la figure 20 le nuage de points la polarité et la subjectivité des messages du corpus. La polarité désigne à quel point un mot est positif ou négatif, sa valeur tend vers -1 si le message est très négatif et vers +1 si elle est très positive. La subjectivité signifie quant à elle, le degré d'opinion d'un mot, elle marque la différence entre les faits et les opinions et tend vers la valeur 0 quand c'est un fait et vers +1 lorsqu'il s'agit d'une opinion.

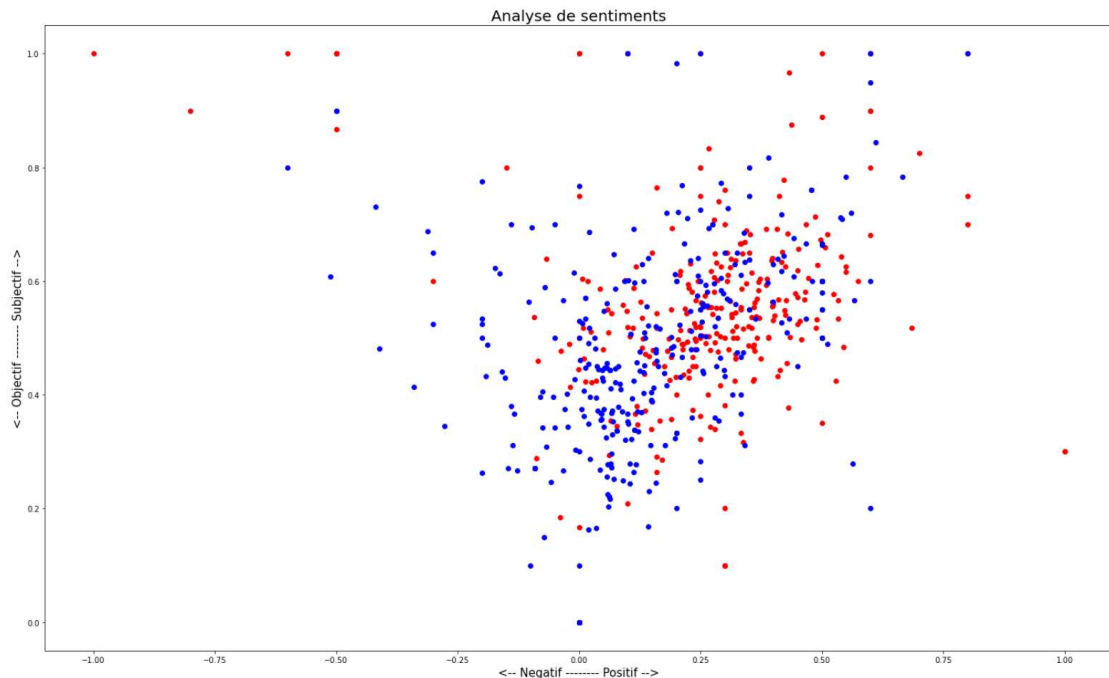


Figure 20 : Nuage de points de la polarité et la subjectivité des messages légitimes (en bleu) et illégitimes (en rouge).

Les messages ne sont généralement pas dans les extrêmes en termes de polarité (entre -0.25 et 65%) et subjectivité (0.2 à 0.8). À l'évidence, les messages

légitimes sont pourvus des faits par rapport aux messages illégitimes. Si un très grand ensemble des messages légitimes sont positifs, les messages illégitimes sont principalement positifs.

Nous nous proposons à présent de réaliser une statistique descriptive concernant l'analyse de sentiments. Pour ce faire, nous avons extrait la structure du corpus (voir figure 21) et disposons de 324 messages légitimes et 331 messages illégitimes dans le corpus, où la colonne (alabel) représente la classe à prédire, elle est composée de deux modalités (illégitime et légitime). La colonne message comporte tous les messages. Nous pouvons également distinguer deux variables

	alabel	message	polarite	subjectivite
0	illegitime	– beautiful -sexy – sensual – and all the good...	0.304167	0.650000
1	illegitime	Thank-you for opening your Heart to Love agai...	0.338542	0.516146
2	illegitime	[Kian:] So here we stand In our secret place W...	0.159259	0.764550
3	illegitime	"I smiled to many people for more than a billi...	0.355655	0.536310
4	illegitime	46 REASONS WHY I LOVE YOU. 1. I love the way w...	0.449474	0.619363

Figure 21 : Données structurées du corpus.

quantitatives continues, la polarité et la subjectivité. Ces variables ont été choisies pour conduire nos analyses descriptives effectuées à la section suivante.

Les caractéristiques de tendances centrales. Le tableau 6 fournit les éléments statistiques qui caractérisent la tendance centrale des données du corpus du point de vue de l'analyse des sentiments. Si les moyennes de la

subjectivité sont identiques, quelle que soit la nature des messages, la polarité des messages illégitimes semble plus marquée par rapport aux messages légitimes. En d'autres termes en moyenne, les messages illégitimes sont positifs. La médiane qui correspond également au quartile 2 (Q2) montre que 50% des messages légitimes et illégitimes sont polarisés respectivement à moins de 0.13 et 0.26, contre environ moins de 0.48 et 0.49 pour la subjectivité. Au regard du mode, l'on peut souligner la neutralité des messages illégitimes justifiée par la présence de la valeur à zéro. À la différence des messages légitimes qui sont davantage subjectifs et positifs.

Tableau 6 : Caractéristiques de tendance centrale des sentiments

		Polarité		Subjectivité	
		Légitime	Illégitime	Légitime	Illégitime
Tendance centrale	Moyenne	0.15	0.24	0.48	0.49
	Médiane	0.13	0.26	0.48	0.52
	Q1	0.02	0.1	0.36	0.42
	Q2	0.13	0.26	0.48	0.52
	Q3	0.28	0.37	0.6	0.60
	Mode	0.5	0	0.6	0

Les caractéristiques de dispersions. L'étendue est l'écart entre la plus grande et la plus petite des valeurs. Ils sont particulièrement élevés dans la

polarité. Cela montre une dispersion de la polarité au sein du corpus (voir le tableau 7). Néanmoins au vu de l'écart-type représentant la variabilité des observations par rapport à la moyenne, les messages ne sont pas assez éloignés de la moyenne des valeurs sentimentales. D'ailleurs la figure (Figure 20) montre bien une concentration du nuage de points autour de la moyenne. De plus, comparativement à la subjectivité, la polarité est fortement dispersée avec les coefficients de variation à 144.57% et 96.67% respectivement pour les messages légitimes et illégitimes contre 30.3% et 48.10% pour la subjectivité des messages légitimes et illégitimes.

Tableau 7 : Caractéristiques de dispersion des sentiments

		Polarité		Subjectivité	
		Légitime	Illégitime	Légitime	Illégitime
Dispersion	Variance	0.046	0.052	0.034	0.056
	Écart-type	0.216	0.229	0.184	0.237
	Étendu	1.4	2	1	1
	Coefficient de variation (écart-type / moyenne) *100	144.57%	96.67%	30.30%	48.10%

Du prétraitement des données à la construction du modèle optimale

Dans cette section, nous allons construire notre modèle de classification. Pour y parvenir, nous expérimenterons successivement plusieurs méthodes et approches standards, et ce, à chaque étape du processus de classification afin d'obtenir in fine le modèle qui offre les meilleures performances en rapport avec notre corpus jadis constitué.

Cette approche de construction s'articulera autour de deux types d'expériences suivant un cadre d'expérimentation défini tel que le montre le tableau 8. Sur la base des données d'entraînement, la première expérimentation consiste à effectuer les apprentissages en combinant les méthodes sélectionnées relativement à chaque étape d'expérimentation.

En d'autres termes, il s'agira de déterminer la combinaison de méthodes la plus performante. La deuxième expérimentation quant à elle, ne se fera non plus sur les combinaisons de méthodes, mais davantage sur l'ensemble des méthodes pris collectivement à chaque étape du processus. Autrement dit, les méthodes sont mises en concurrence à chaque étape, pour ne retenir que la plus performance. C'est la combinaison des méthodes ayant obtenu le meilleur score à chaque étape du processus qui constituera le modèle optimisé. À l'issue de ces deux types d'expérimentation constitués de plusieurs expériences, l'on retiendra comme modèle de notre problème de classification de la FS, le modèle optimisé de l'expérience ayant le meilleur taux de succès.

Tableau 8 : Cadre d'expérimentation

Étapes d'expérimentation	Activités	Méthodes et approches
Étape I	Extraction et pondération des caractéristiques	TF, TF-IDF et Tri-gramme de caractères
	Sélection des caractéristiques	Information mutuelle et Khi2,
Étape II	Réduction des caractéristiques	Analyse en composantes principales, indexation sémantique latente et la factorisation par matrices non négatives
Étape III	Apprentissage automatique	K-plus proches voisins, machines à vecteur de support, forêt aléatoire et la logistique régression

Distribution des données et validation croisée

Pour les deux types d'expérimentation, nous disposons d'un corpus de messages de données composées de 655 messages, réparties pour 70% destinés à entraîner les modèles et 30% pour le test du modèle. Nous avons par ailleurs choisi de mettre en place des validations croisées (Xiong et al, 2020) sur

ces données d'entraînement et ce, sur 3 partitions. L'intérêt pour cette approche nous permet d'adresser la problématique de la variance par rapport au biais aborder dans la section (5.2.2.2) à l'effet de mieux contrôler les enjeux de sur-apprentissage et sous-apprentissage

Bien que le corpus actuel soit moins propice à la validation croisée du fait du faible nombre de messages de son contenu, nous l'avons tout de même mise en place, car le processus d'acquisition de données devra acquérir plus de données à l'avenir, ce qui devra augmenter considérablement la taille du corpus, et aussi, nous permettre de passer à 5, 10 voire davantage de partitions.

Expérimentation 1 : Recherche de la combinaison idéale à contribution individuelle des méthodes

Hypothèse de l'expérimentation

L'on part de l'intuition qu'il est possible de combiner les méthodes spécialisées à chaque étape du processus de classification de manière à identifier la combinaison la plus performante. Cela suppose aussi que chaque étape du cycle d'apprentissage est indispensable à l'optimisation de la classification. Par exemple, s'il est possible d'omettre la sélection des caractéristiques pour se focaliser uniquement sur la réduction, l'on pense qu'il est préférable de commencer par la sélection afin d'optimiser progressivement la performance du processus.

De plus, nous imaginons que l'on peut déterminer la meilleure performance globale du processus grâce à des micro-optimisations de méthodes prises

individuellement. Par exemple, l'on pourra agir sur les hyperparamètres tels que le nombre de caractéristiques à sélectionner pour déterminer la méthode de sélection qui performe le plus ou ajuster le nombre de composante principale en réduction pour identifier la méthode de réduction qui optimise plus la performance globale du processus.

Déroulement de l'expérimentation

À chaque étape du processus, nous avons identifié les méthodes communément utilisées dans les problématiques de CAT, que nous combinerons de manière croisée. L'idée est de trouver la meilleure combinaison possible, autrement dit, la combinaison ayant le meilleur taux de succès à l'issue des expériences. Compte tenu des méthodes choisies, nous aurons de manière exhaustive soixante-douze expériences à évaluer, tel que le montre le tableau 9.

Tableau 9 : Total des expériences en expérimentation 1

Méthodes et approches	
TF, TF-IDF et Tri-gramme de caractères	3
Information mutuelle et Khi2,	2
Analyse en composantes principales, Indexation sémantique latente et la factorisation par matrices non négatives	3
K-plus proches voisins, Machines à vecteur de support, forêt aléatoire et la logistique régression	4
Nombres d'expériences ($3 * 2 * 3 * 4$)	72

L'évaluation des expériences se fera en utilisant les métriques ayant une incidence notable sur l'utilisation des modèles telles que les taux de succès, taux de faux positifs et faux négatifs.

Résultat de l'expérimentation

À l'issue de ces expériences, les métriques d'évaluation ont été capturées, ce sont : la matrice de confusion, la précision, le rappel, la f1-score, le support et le taux de succès.

Les soixante-douze expériences micro-optimisées ont été menées sur la base des hyperparamètres qui varient entre deux valeurs (2000, 2500 caractéristiques) pour la sélection et (200, 250 composantes principales) pour la réduction. Concernant les classifieurs, nous avons choisi les valeurs 500, 750 et 1000 pour le nombre d'arbres de la forêt aléatoire, les valeurs (linéaire, polynomial, rbf et sigmoïdale) pour le classifieur SVM et finalement les valeurs (3 et 5) pour le nombre des plus proches voisins pour le classifieur des K plus proches voisins. Ces valeurs initiales ont été choisies au regard des expériences passées rapportées dans la littérature scientifique (Koen de, 2019). Au final, nous obtenons suivant chaque modèle sélectionné, les résultats ci-après.

Classifieur régression logistique. Le tableau 10 présente les résultats pour ce classifieur, la meilleure performance, soit 87.80 % est obtenue en combinant la vectorisation par termes de référence (TF), la sélection des caractéristiques par information mutuelle (IM) ainsi que l'analyse en composante principale (IPCA) pour la réduction des caractéristiques.

Tableau 10 : Métriques de l'expérience 1 du classifieur régression logistique

code_experience	score	tn	fp	fn	tp	fpr	fnr	precision	rappel	Spécificité	Erreur model	k	n_components	kernel	n_estimators	k_neighbors	t_training	t_test
tf_mic_ipca_lg	0.878	97	14	10	76	0.126	0.116	0.844	0.884	0.874	24	2000	250		0	0	70.889	0.03
tf_ki2_lsa_lg	0.873	98	13	12	74	0.117	0.14	0.851	0.86	0.883	25	2500	200		0	0	4.941	0.058
tf_mic_lsa_lg	0.868	99	12	14	72	0.108	0.163	0.857	0.837	0.892	26	2000	200		0	0	72.954	0.031
tfidf_ki2_lsa_lg	0.853	90	21	8	78	0.189	0.093	0.788	0.907	0.811	29	2000	250		0	0	2.58	0.047
tf_ki2_ipca_lg	0.848	99	12	18	68	0.108	0.209	0.85	0.791	0.892	30	2000	200		0	0	4.056	0.065
tfidf_ki2_ipca_lg	0.838	87	24	8	78	0.216	0.093	0.765	0.907	0.784	32	2000	200		0	0	1.734	0.035
tfidf_mic_ipca_lg	0.838	88	23	9	77	0.207	0.105	0.77	0.895	0.793	32	2000	250		0	0	71.174	0.033
tfidf_mic_lsa_lg	0.832	87	24	9	77	0.216	0.105	0.762	0.895	0.784	33	2000	200		0	0	72.17	0.03
ngram3_mic_lsa_lg	0.792	93	18	23	63	0.162	0.267	0.778	0.733	0.838	41	2500	250		0	0	48.926	0.099
ngram3_ki2_ipca_lg	0.787	88	23	19	67	0.207	0.221	0.744	0.779	0.793	42	2500	200		0	0	4.131	0.078
ngram3_mic_nmf_lg	0.787	89	22	20	66	0.198	0.233	0.75	0.767	0.802	42	2500	200		0	0	651.084	4.497
ngram3_ki2_nmf_lg	0.782	84	27	16	70	0.243	0.186	0.722	0.814	0.757	43	2000	200		0	0	525.89	2.157
ngram3_ki2_lsa_lg	0.782	86	25	18	68	0.225	0.209	0.731	0.791	0.775	43	2500	200		0	0	8.756	0.075
ngram3_mic_ipca_lg	0.756	90	21	27	59	0.189	0.314	0.738	0.686	0.811	48	2000	200		0	0	42.74	0.082
tf_ki2_nmf_lg	0.711	71	40	17	69	0.36	0.198	0.633	0.802	0.64	57	2000	200		0	0	351.618	0.26
tfidf_mic_nmf_lg	0.655	45	66	2	84	0.595	0.023	0.56	0.977	0.405	68	2000	250		0	0	353.352	0.266
tf_mic_nmf_lg	0.645	66	45	25	61	0.405	0.291	0.575	0.709	0.595	70	2000	200		0	0	530.253	0.275
tfidf_ki2_nmf_lg	0.467	17	94	11	75	0.847	0.128	0.444	0.872	0.153	105	2000	200		0	0	312.692	0.161

L'on note néanmoins un taux élevé de faux positifs (12.60%) et de faux négatifs (11.60%). La recherche des hyperparamètres optimaux a produit une sélection de 2000 caractéristiques, qui a été réduite à 200 au terme de la réduction des caractéristiques.

Classifieur forêt aléatoire. Le tableau 11 présente les résultats pour ce classifieur, la meilleure performance, soit 87.80 % est obtenue en combinant la vectorisation par les termes de référence- fréquences inverse des documents (TF-IDF), la sélection des caractéristiques Ki2 (qui a produit 2000 caractéristiques) ainsi que l'analyse en composante principale (IPCA) pour la réduction des caractéristiques à 250. L'on note aussi un taux élevé de faux positifs (10.60%) et de faux négatifs (14 %).

Classifieur SVM. Le tableau 12 présente les résultats pour ce classifieur, la vectorisation par les termes de référence-fréquences inverse des documents (TF-IDF), la sélection des caractéristiques Ki2 (2500 sont les caractéristiques résultantes) et la réduction à 250 composantes principales. L'on note aussi un taux assez élevé de faux positifs (17.10%) et de faux négatifs (10.50%).

Classifieur K-plus proches voisins. Le tableau 13 présente les résultats pour ce classifieur, la meilleure performance, soit 81.70 % est obtenue avec l'hyperparamètre $K = 3$ voisins, en combinant la vectorisation à travers le tri-gramme de caractères, la sélection des caractéristiques à travers l'information

Tableau 11 : Métriques de l'expérience 1 du classifieur forêt aléatoire

code_experience	score	tn	fp	fn	tp	fpr	fnr	precision	rappel	Spécificité	Erreur modèle	k	n_components	kernel	n_estimators	k_neighbors	t_training	t_test
tfidf_ki2_ipca_rf	0.878	99	12	12	74	0.108	0.14	0.86	0.86	0.892	24	2500	200		500	0	91.984	0.108
tfidf_mic_nmf_rf	0.873	96	15	10	76	0.135	0.116	0.835	0.884	0.865	25	2000	200		750	0	950.1	0.235
ngram3_mic_ipca_rf	0.868	91	20	6	80	0.18	0.07	0.8	0.93	0.82	26	2000	200		750	0	160.857	0.117
ngram3_mic_nmf_rf	0.863	103	8	19	67	0.072	0.221	0.893	0.779	0.928	27	2500	200		500	0	1900.591	3.983
tfidf_ki2_nmf_rf	0.858	101	10	18	68	0.09	0.209	0.872	0.791	0.91	28	2000	200		750	0	661.68	0.233
ngram3_mic_lsa_rf	0.853	86	25	4	82	0.225	0.047	0.766	0.953	0.775	29	2000	200		750	0	231.688	0.125
ngram3_ki2_lsa_rf	0.853	88	23	6	80	0.207	0.07	0.777	0.93	0.793	29	2500	200		1000	0	64.492	0.123
tfidf_ki2_lsa_rf	0.853	90	21	8	78	0.189	0.093	0.788	0.907	0.811	29	2000	200		750	0	94.471	0.123
tf_ki2_ipca_rf	0.853	95	16	13	73	0.144	0.151	0.82	0.849	0.856	29	2000	200		750	0	123.984	0.175
tfidf_mic_lsa_rf	0.848	89	22	8	78	0.198	0.093	0.78	0.907	0.802	30	2000	200		1000	0	290.091	0.155
ngram3_ki2_ipca_rf	0.843	88	23	8	78	0.207	0.093	0.772	0.907	0.793	31	2000	250		500	0	63.716	0.117
tf_ki2_lsa_rf	0.838	89	22	10	76	0.198	0.116	0.776	0.884	0.802	32	2000	200		750	0	117.723	0.156
ngram3_ki2_nmf_rf	0.838	100	11	21	65	0.099	0.244	0.855	0.756	0.901	32	2000	250		1000	0	1588.502	2.94
tf_mic_lsa_rf	0.827	85	26	8	78	0.234	0.093	0.75	0.907	0.766	34	2500	200		750	0	294.28	0.124
tf_mic_ipca_rf	0.827	89	22	12	74	0.198	0.14	0.771	0.86	0.802	34	2000	200		750	0	288.867	0.165
tfidf_mic_ipca_rf	0.802	79	32	7	79	0.288	0.081	0.712	0.919	0.712	39	2000	250		500	0	290.298	0.113
tf_ki2_nmf_rf	0.797	101	10	30	56	0.09	0.349	0.848	0.651	0.91	40	2500	250		1000	0	1129.226	0.365
tf_mic_nmf_rf	0.766	75	36	10	76	0.324	0.116	0.679	0.884	0.676	46	2500	250		750	0	1190.329	0.331

Tableau 12 : Métriques de l'expérience 1 du classifieur SVM

code_experience	score	tn	fp	fn	tp	fpr	fnr	Precision	rappel	Spécificité	Erreur model	k	n_components	kernel	n_estimators	k_neighbors	t_training	t_test
tfidf_ki2_lsa_svm	0.858	92	19	9	77	0.171	0.105	0.802	0.895	0.829	28	2500	250	linear	0	0	11.948	0.049
tfidf_mic_ipca_svm	0.853	90	21	8	78	0.189	0.093	0.788	0.907	0.811	29	2500	200	linear	0	0	302.625	0.058
tfidf_mic_lsa_svm	0.853	90	21	8	78	0.189	0.093	0.788	0.907	0.811	29	2500	200	linear	0	0	269.635	0.045
tfidf_ki2_ipca_svm	0.843	90	21	10	76	0.189	0.116	0.784	0.884	0.811	31	2500	250	linear	0	0	8.777	0.05
tf_mic_ipca_svm	0.838	94	17	15	71	0.153	0.174	0.807	0.826	0.847	32	2000	250	linear	0	0	265.265	0.045
tf_ki2_ipca_svm	0.832	98	13	20	66	0.117	0.233	0.835	0.767	0.883	33	2000	200	linear	0	0	18.245	0.08
ngram3_ki2_lsa_svm	0.817	81	30	6	80	0.27	0.07	0.727	0.93	0.73	36	2000	250	rbf	0	0	21.62	0.097
tf_ki2_lsa_svm	0.817	99	12	24	62	0.108	0.279	0.838	0.721	0.892	36	2000	200	linear	0	0	21.025	0.084
ngram3_ki2_ipca_svm	0.812	81	30	7	79	0.27	0.081	0.725	0.919	0.73	37	2000	250	rbf	0	0	25.601	0.186
tf_mic_lsa_svm	0.802	95	16	23	63	0.144	0.267	0.797	0.733	0.856	39	2000	250	linear	0	0	266.663	0.04
ngram3_mic_lsa_svm	0.772	75	36	9	77	0.324	0.105	0.681	0.895	0.676	45	2000	250	rbf	0	0	221.637	0.11
tf_mic_nmf_svm	0.772	88	23	22	64	0.207	0.256	0.736	0.744	0.793	45	2000	250	linear	0	0	1421.241	0.269
ngram3_mic_ipca_svm	0.766	75	36	10	76	0.324	0.116	0.679	0.884	0.676	46	2000	250	rbf	0	0	150.313	0.1
ngram3_ki2_nmf_svm	0.756	81	30	18	68	0.27	0.209	0.694	0.791	0.73	48	2000	200	linear	0	0	2050.858	3.884
tf_ki2_nmf_svm	0.711	70	41	16	70	0.369	0.186	0.631	0.814	0.631	57	2500	200	linear	0	0	1787.494	0.355
ngram3_mic_nmf_svm	0.68	67	44	19	67	0.396	0.221	0.604	0.779	0.604	63	2500	250	linear	0	0	2473.181	6.401
tfidf_mic_nmf_svm	0.563	26	85	1	85	0.766	0.012	0.5	0.988	0.234	86	2000	200	linear	0	0	6195.304	0.149
tfidf_ki2_nmf_svm	0.437	0	111	0	86	1	0	0.437	1	0	111	2000	250	rbf	0	0	867.997	0.176

mutuelle (IM) ainsi que l'analyse en composante principale (IPCA) pour la réduction des caractéristiques. L'on note finalement un taux assez élevé de faux positif (11.70%) et de faux négatif (26.70%). La sélection de 2000 caractéristiques a été projetée sur 200 composantes principales afin d'obtenir ce classifieur.

Nous remarquons que le modèle ayant le meilleur taux de succès est la forêt aléatoire entraînée avec le TF-IDF, le khi2 et l'analyse en composante principale, avec un taux de 87.80%, ce taux est identique à celui du classifieur régression logistique, qui lui est entraîné avec le TF, l'information mutuelle et l'analyse en composante principale (voir la figure 22). Cependant la performance la moins bonne est le classifieur K plus proches voisins, entraîné avec le tri-gramme de caractères, l'information mutuelle et l'analyse en composante principale, avec un taux de succès 81.70%.

Nous pouvons apprendre davantage sur le mécanisme de construction du classifieur, notamment le temps d'entraînement et le temps de test du classifieur. La figure 23 montre que la performance d'un classifieur n'est pas corrélé au temps d'apprentissage, ni aux temps de test.

Expérimentation 2 : Recherche de la combinaison idéale à contribution optimisée des méthodes

Hypothèse de l'expérimentation.

Cette expérimentation se base sur le fait qu'en mettant en concurrence les méthodes sur le même ensemble de données d'entraînement, l'on pourra

Tableau 13 : Métriques de l'expérience 1 du classifieur K plus proches voisins

code_experience	score	tn	Fp	fn	tp	fpr	fnr	Precision	rappel	Spécificité	Erreur model	k	n_composants	kernel	n_estimateurs	k_neighbors	t_training	t_test
ngram3_mic_ipca_kn	0.817	98	13	23	63	0.117	0.267	0.829	0.733	0.883	36	2000	200		0	3	151.185	0.159
ngram3_ki2_lsa_kn	0.802	96	15	24	62	0.135	0.279	0.805	0.721	0.865	39	2000	200		0	3	15.091	0.146
ngram3_ki2_ipca_kn	0.797	96	15	25	61	0.135	0.291	0.803	0.709	0.865	40	2000	200		0	3	14.798	0.183
ngram3_mic_lsa_kn	0.797	99	12	28	58	0.108	0.326	0.829	0.674	0.892	40	2500	200		0	5	151.765	0.135
ngram3_mic_nmf_kn	0.68	88	23	40	46	0.207	0.465	0.667	0.535	0.793	63	2500	200		0	3	1097.015	2.186
tfidf_ki2_nmf_kn	0.675	81	30	34	52	0.27	0.395	0.634	0.605	0.73	64	2000	200		0	5	461.366	0.127
tfidf_ki2_ipca_kn	0.675	108	3	61	25	0.027	0.709	0.893	0.291	0.973	64	2000	200		0	5	4.959	0.066
tfidf_mic_nmf_kn	0.655	73	38	30	56	0.342	0.349	0.596	0.651	0.658	68	2500	200		0	3	608.985	0.22
tfidf_ki2_lsa_kn	0.645	106	5	65	21	0.045	0.756	0.808	0.244	0.955	70	2000	200		0	5	6.336	0.062
tfidf_mic_lsa_kn	0.64	104	7	64	22	0.063	0.744	0.759	0.256	0.937	71	2000	200		0	5	137.797	0.056
tfidf_mic_ipca_kn	0.64	106	5	66	20	0.045	0.767	0.8	0.233	0.955	71	2000	200		0	5	138.188	0.07
tf_ki2_nmf_kn	0.604	80	31	47	39	0.279	0.547	0.557	0.453	0.721	78	2000	200		0	3	800.169	0.314
ngram3_ki2_nmf_kn	0.594	71	40	40	46	0.36	0.465	0.535	0.535	0.64	80	2500	250		0	3	944.181	2.874
tf_mic_nmf_kn	0.589	82	29	52	34	0.261	0.605	0.54	0.395	0.739	81	2000	200		0	3	759.357	0.179
tf_mic_lsa_kn	0.533	71	40	52	34	0.36	0.605	0.459	0.395	0.64	92	2000	200		0	3	136.103	0.058
tf_ki2_lsa_kn	0.523	65	46	48	38	0.414	0.558	0.452	0.442	0.586	94	2000	200		0	3	9.438	0.057
tf_ki2_ipca_kn	0.523	65	46	48	38	0.414	0.558	0.452	0.442	0.586	94	2000	200		0	3	4.94	0.064
tf_mic_ipca_kn	0.523	71	40	54	32	0.36	0.628	0.444	0.372	0.64	94	2000	200		0	3	137.003	0.058

mettre en évidence la méthode la plus performante. Elle devra se focaliser en outre à la recherche les hyperparamètres à très forte valeur ajoutée eu égard à la performance. À l'issue de cette expérimentation, nous aurons la combinaison de méthodes la plus optimale ainsi que les hyperparamètres qui le sont tout autant.

Déroulement de l'expérimentation.

Nous conservons les méthodes précédemment sélectionnées, et effectuons à l'étape de sélection des caractéristiques, des exécutions en parallèle des méthodes de sélection sur le même ensemble de données d'entraînement à l'effet de ne retenir que la plus performante. La méthode la plus performante ainsi obtenue, est utilisée pour sélectionner les caractéristiques qui feront l'objet de réduction à l'étape subséquente. À l'étape de réduction, les méthodes sont encore une fois mises en concurrence afin d'identifier la plus optimisée. La démarche expérimentale est représentée à la figure 24.

À ce stage, nous avons une méthode de sélection et de réduction optimisées. Ces méthodes sont par la suite utilisées pour entraîner plusieurs classifieurs toujours dans le but ultime de repérer le classifieur qui performe le mieux.

À cet effet, nous avons considéré deux hyperparamètres généraux et plusieurs autres spécifiques aux classifieurs. Généralement nous avons opté pour le nombre de caractéristiques à sélectionner et le nombre de composantes principales pour la réduction, les choisissant parmi les valeurs suivantes :

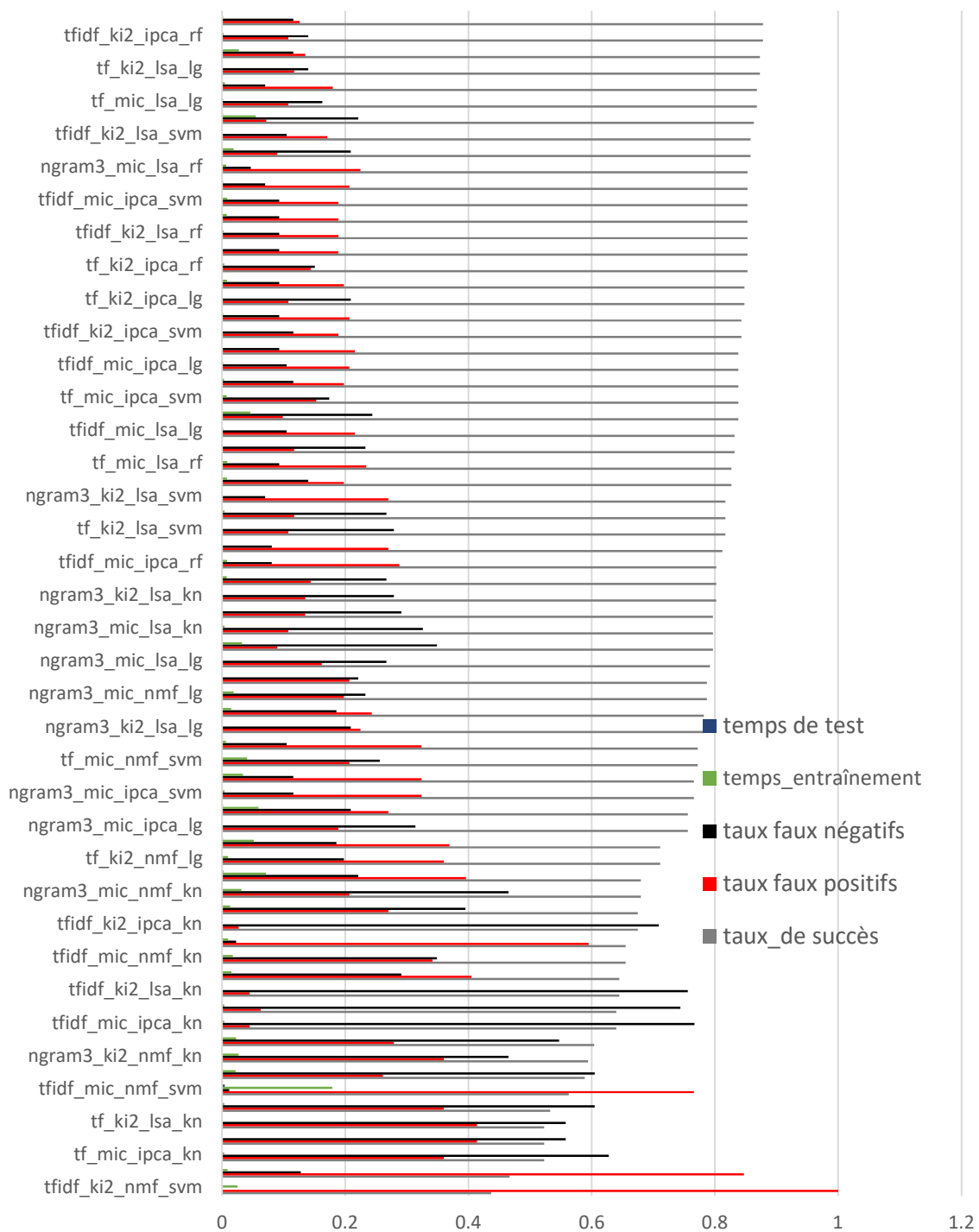


Figure 22 : Histogramme des métriques essentielles de performance.

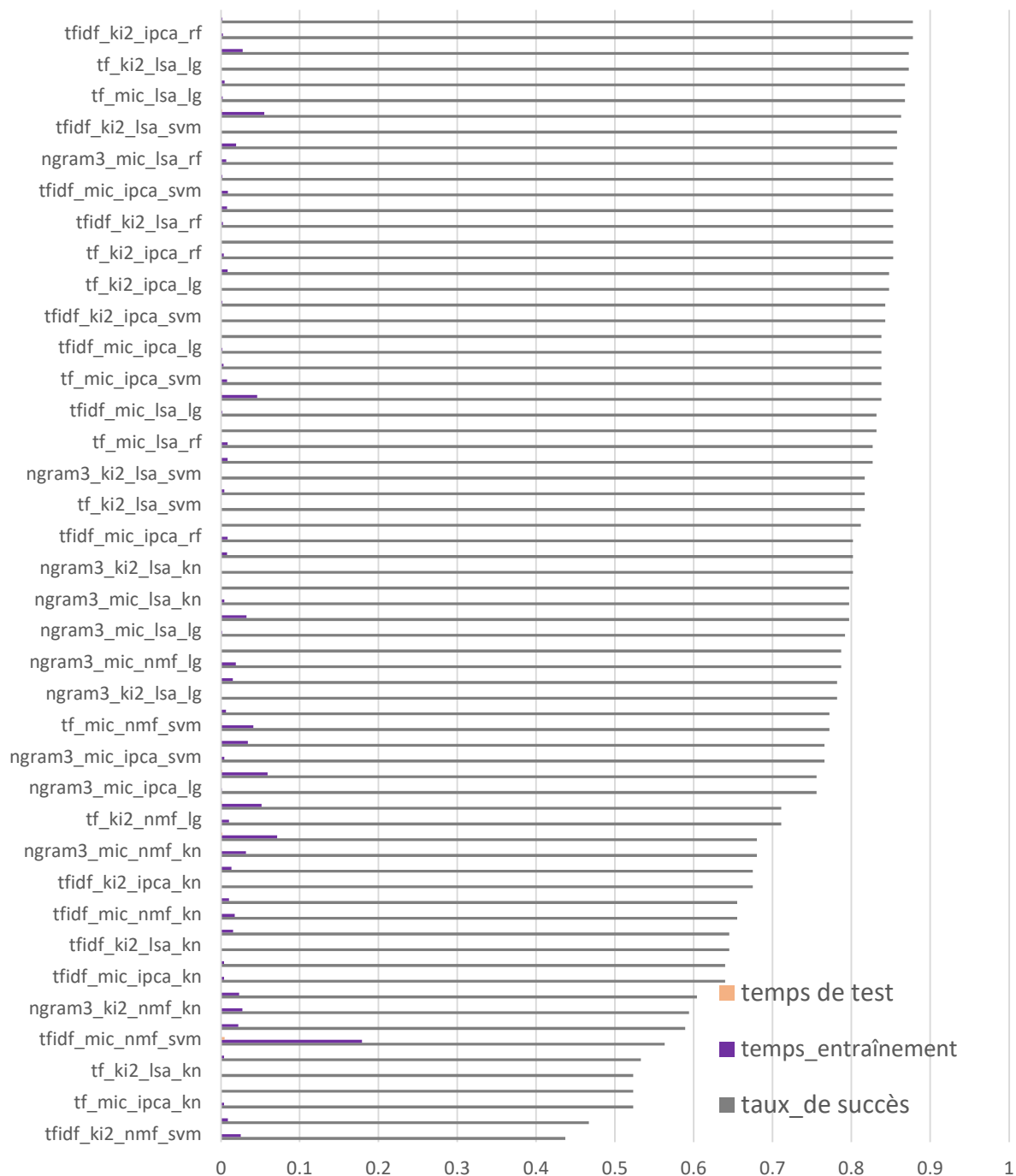


Figure 23 : Histogramme des temps d'apprentissage et de test.

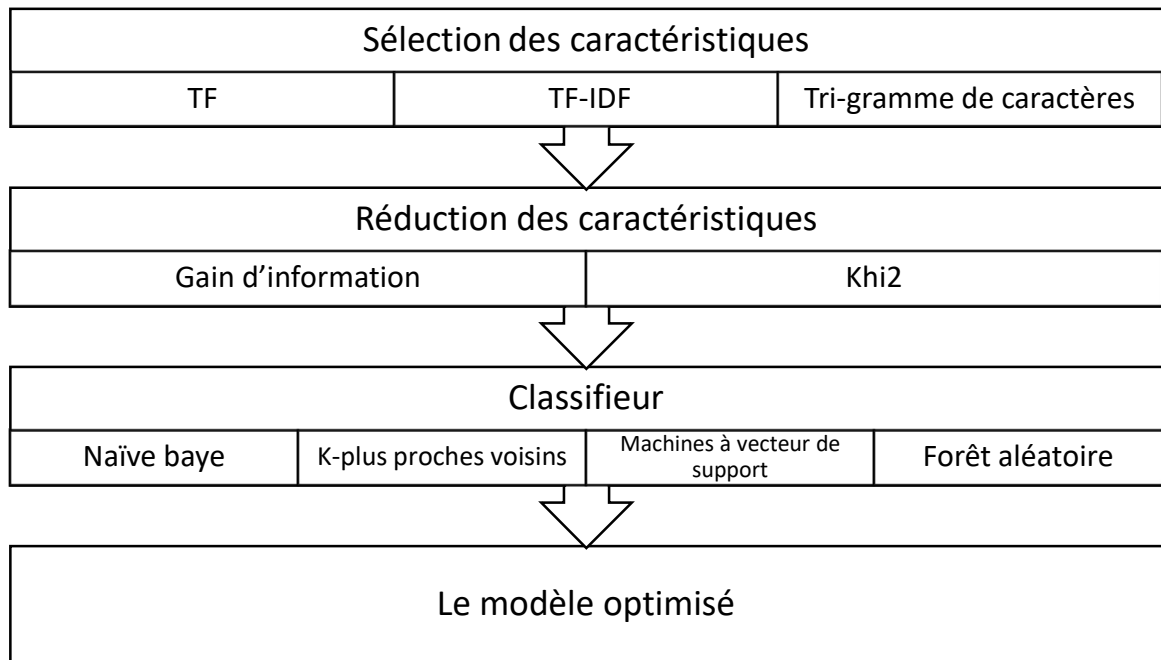


Figure 24 : Processus de l'expérimentation 2.

2000, 2500 caractéristiques pour la sélection et 200, 250 composantes principales. Spécifiquement pour les modèles, nous avons retenu pour la forêt aléatoire, le nombre d'arbres, le situant parmi les valeurs 500, 750 et 1000. Le kernel linéaire, polynomial, sigmoïdale et rbf font partie des valeurs utilisées pour optimiser la machine à vecteur de support. Pour le K plus proches voisins, nous avons utilisé 3 et 5 voisins comme hyperparamètres. Enfin pour apprécier les performances des expériences, nous avons sollicité les mesures suivantes : taux de succès, taux de faux positifs/négatifs et le temps de prédiction sur les données de test.

Résultat de l'expérimentation

Au terme de cette expérimentation, le meilleur score a été obtenu avec le classifieur SVM (kernel polynomial), qui performe le plus avec un taux de succès à 89.80 % pour un taux de faux positifs de 8.10 % et faux négatifs de 12.80% (voir tableau 14). La forêt aléatoire a un taux de succès approximativement identique, et dispose d'un temps de test plus bas à près de 3.713 secondes. Comme nous allons le voir à la section suivante, le temps de réponse est une variable déterminante dans un contexte de production, car il influe sur la qualité de service usager. Nous serons donc portés à choisir un modèle qui performe de façon acceptable mais qui offre des temps de réponse relativement bas. La figure 25 présente le rapport des mesures de performances suivant les classifieurs.

Tableau 14 : Métriques de l'expérimentation 2

code_experience	Score	tn	fp	fn	tp	fpr	fnr	Precision	rappel	Spécificité	Erreur model	k	n_composants	kernel	n_estimateurs	n_neighbors	t_test
optimal_svm	0.898	102	9	11	75	0.081	0.128	0.893	0.872	0.919	20	2000	200	poly	0	0	6.297
optimal_forêt aléatoire	0.893	98	13	8	78	0.117	0.093	0.857	0.907	0.883	21	2000	200		750	0	3.713
optimal_K-plus proches voisins	0.812	100	11	26	60	0.099	0.302	0.845	0.698	0.901	37	2000	200		0	3	5.158
optimal Regression logistique	0.792	91	20	21	65	0.18	0.244	0.765	0.756	0.82	41	2000	200		0	0	3.763

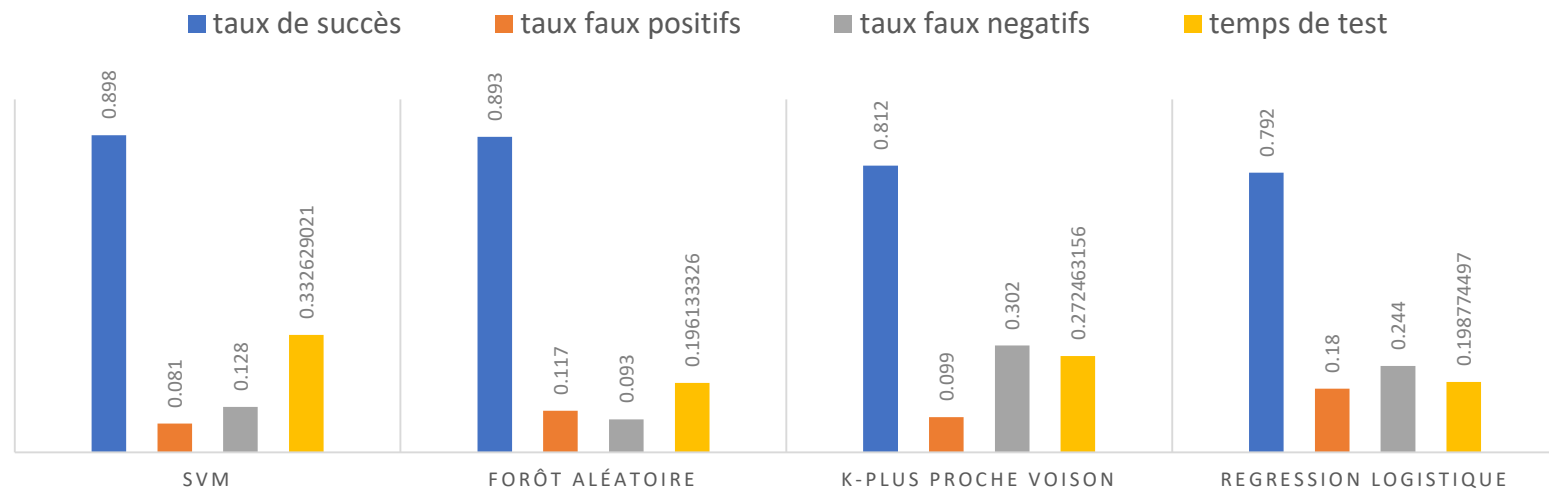


Figure 25 : Histogramme des métriques de l'expérimentation 2.

Discussion sur les résultats

Le meilleur modèle

Les deux expérimentations ont produit plusieurs expériences. La première consistait à observer la contribution de chaque méthode d'étape par rapport à la performance de la combinaison des méthodes. Tandis que la deuxième expérimentation avait pour ambition de produire le modèle le plus optimal possible grâce à la mise en concurrence des méthodes et une obsession marquée à la recherche des hyperparamètres les plus optimisés. Les deux expériences offrent les quatre meilleures performances suivantes :

Tableau 15 : Meilleures métriques de performance des expérimentations 1 et 2

Code_expérience	Taux de succès	Faux positifs	Faux négatifs	Temps de test (s)
optimal_svm	0.898	0.081	0.128	6.297
optimal_forêt aléatoire	0.893	0.117	0.093	3.713
tf_mic_ipca_lg	0.878	0.126	0.116	0.03
tfidf_ki2_ipca_rf	0.878	0.108	0.14	0.108

Selon le taux de succès, le classifieur SVM sera à privilégier avec un taux de 89.80%. Ce choix est néanmoins à discuter compte tenu des impératifs de déploiement. À ce propos, la section suivante s'y intéresse.

Limitations et considérations de déploiement du modèle

Modélisation continue. L'une des considérations du modèle est qu'il est logé dans une application spécialement conçue pour lui permettre de s'entraîner sur de nouvelles données. Autrement dit, le processus de classification est une activité continue. Ce qui nous amène à une recherche

continue de la performance la plus optimale au fil du temps. Le système est conçu pour toujours sélectionner la meilleure performance de sa base de données des expériences qu'elles soient passées ou présentes.

Le taux de faux positifs/négatifs. En contexte de production, les taux de faux positifs et faux négatifs ont une signification importante. Alors que les faux positifs montrent une erreur de détection, les faux négatifs montrent plutôt une erreur de non-détection. Par exemple dans notre contexte de la fraude et en usant notre meilleur modèle (`optimal_svm`), nous pouvons dire qu'environ 8 personnes sur 100 seront considérées comme des fraudeurs alors qu'ils ne le sont pas, pendant que 12 personnes sur 100 seront de fraudeurs non détectés. Compte tenu de l'impact de ces taux pour un système en production, il y a un compromis à faire pour privilégier l'un ou l'autre de ces erreurs, cela aura notamment une incidence autant sur le choix du modèle à considérer parmi plusieurs que sur les usagers dudit modèle.

Le temps de test. Le système conserve également les temps de test (ou temps de détection/prédiction) sur le jeu des données de test. Cela nous donne un aperçu des temps de réponse du modèle une fois en production. Cet historique des temps permet de sélectionner un modèle non seulement sur la base de son taux de succès, mais aussi en tenant compte de l'expérience de l'utilisateur en situation d'utilisation. Dans cette optique, l'on pourra opter pour

l'expérience « `tf_mic_ipca_lg` », car elle offre un meilleur temps de réponse (0.03 seconde) et un taux de succès acceptable.

Conclusion et perspectives

Encore aujourd'hui, la fraude sentimentale (FS) gangrène la société. Les fraudeurs continuent à sévir, usant des techniques de plus en plus sophistiquées relevant des domaines de la psychologie et de la cybersécurité. Comme nous l'avons décrit dans le deuxième chapitre, la FS cause des dommages psychologiques et financiers importants, ayant des conséquences sociales considérables. Notre travail visait à élaborer un modèle d'AA à l'effet d'aider les usagers/citoyens à mieux se prémunir face à ce fléau.

Nous avons pris la peine d'étudier la question de la FS en recourant aux travaux des scientifiques spécialisés en la matière. De sa définition, ses caractéristiques, les techniques utilisées ainsi que les structures organisationnelles qui l'élaborent et la mettent en œuvre. Ensuite, nous avons abordé toutes les étapes préconisées à la construction de notre modèle de classification de texte. La collecte des données a été menée sur les sites appropriés avec une bonne dose de diversification d'information, s'en est suivi, l'analyse exploratoire et statistique de ces données avec une primauté sur l'analyse des sentiments (sujet en lien avec notre problématique).

Du prétraitement des données à la construction d'un modèle optimale, deux expérimentations ont été menées. Il s'agissait essentiellement des expériences d'optimisation dont le but était double : premièrement de rechercher les méthodes (autant du point de vue de la sélection et de la réduction des caractéristiques que de l'algorithme d'AA) qui se combinent adéquatement pour influencer agréablement

sur la performance globale du processus de classification. Deuxièmement de rechercher les hyperparamètres optimaux des algorithmes utilisés.

Le modèle obtenu offre une performance se situant à environ 89.80% de taux de succès. Ce modèle a été introduit dans une application informatique développée à cet effet. L'idée est de continuellement expérimenter le modèle dans le but de bonifier ses capacités de détection sur de nouveaux jeux de données obtenus en continu au fil du temps. Malgré tout, ce modèle fait des erreurs de près 11.20%, qui mettent en lumière la nécessité d'amélioration à l'effet de les minimiser. Dans les sections suivantes, nous allons explorer quelques avenues possibles d'amélioration de la performance.

- La qualité et la volumétrie des données

Dans ce travail, compte des difficultés d'accès à l'information, nous avons collecté une quantité relativement faible de données (324 messages légitimes et 331 messages illégitimes) au regard de la volumétrie observé dans les travaux similaires. Nous pensons que l'accès à ces données et de leur utilisation dans notre processus continu d'optimisation, devrait avoir un apport indéniable quant à l'amélioration de notre modèle.

- Le rapport entre l'analyse des sentiments et la détection de la FS

Nous avons abordé la statistique descriptive de quelques-uns des éléments de l'analyse des sentiments (polarité et subjectivité) de notre corpus. Nous avons observé une polarité généralement positive pour l'ensemble du corpus. Parlant

de la subjectivité, les messages légitimes sont plus objectifs tandis que les messages illégitimes sont davantage subjectifs. Toutefois, nous n'avons pas évalué le rapport entre la polarité, la subjectivité et la classe de prédiction. Il serait souhaitable d'explorer cet aspect car il est possible de découvrir les opportunités d'amélioration de notre modèle.

- Apport d'une ontologie de faux profils

L'une des avenues de performance du modèle peut être la combinaison des approches existantes (la détection des fraudeurs à travers leur profil obtenu des sites de rencontre) et la nôtre. Nous pouvons construire une ontologie de profil avec les profils de fraudeurs et élaborer des capacités d'inférence qui nous permettraient de détecter les faux profils ou tout au moins d'indiquer les soupesions. Ces résultats seront combinés à notre processus de classification pour aider à améliorer notre modèle.

- Automatiser le modèle de persuasion pour une détection précoce

Au chapitre 2, nous avons fait allusion au modèle de persuasion de Whitty (2013, p 22), qui présentait les techniques de persuasion utilisées par les fraudeurs pour structurer leur démarche. Nous pouvons imaginer une autre approche de détection qui se baserait sur la temporalité des messages échangés en temps réel de leur l'arrivée dans le système, pour construire un modèle de reconnaissance des phases du modèle de persuasion. La proactivité de ce modèle sera l'avantage principal, car il sera capable de détecter dès les premiers

messages échangés si la personne est en adéquation avec le modèle de persuasion. L'autre avantage est le suivi de la progression de la fraude suivant les phases du modèle de persuasion avec comme corolaires, des descriptions explicatives y afférentes qui seront données à la potentielle victime.

Techniquement, cette modélisation peut être abordée par une description ontologique à l'aide des vocabulaires SEM⁴² (Simple Event Model) qui traitent des problématiques événementielles. Il peut également s'agir de mettre à contribution des modèles heuristiques à l'instar des chaînes de Markov afin d'adresser le caractère dynamique du modèle de persuasion.

⁴² <http://semanticweb.cs.vu.nl/2009/11/sem/>

Références

- Agarwal, S., Godbole, S., Punjani, D., & Roy, S. (2007). How much noise is too much: A study in automatic text classification. In Seventh IEEE International Conference on Data Mining (ICDM 2007) (pp. 3-12). IEEE. doi:10.1109/icdm.2007.21
- Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.
- Anđelić, S., Kondić, M., Perić, I., Jocić, M., & Kovačević, A. (2017). Text classification based on named entities. In ICIST (pp. 23-28).
- Awad, M., & Khanna, R. (2015). *Efficient learning machines: theories, concepts, and applications for engineers and system designers* (p. 268). Springer Nature.
- Baillargeon, J. T., Lamontagne, L., & Marceau, É. (2019, May). Weighting Words Using Bi-Normal Separation for Text Classification Tasks with Multiple Classes. In Canadian Conference on Artificial Intelligence (pp. 433-439). Springer, Cham. doi :10.1007/978-3-030-18305-9_41.
- Balakrishnama, S., & Ganapathiraju, A. (1998, March). Linear discriminant analysis-a brief tutorial. In Institute for Signal and information Processing (Vol. 18, No. 1998, pp. 1-8).
- Barbier, L., & Fointiat, V. (2015). Persuasion et Influence : changer les attitudes, changer les com-portements. Regards de la psychologie sociale. Journal d'Interaction Personne-Système, AssociationFrancophone d'Interaction Homme-Machine (AFIHM), 2015, 4 (1), pp.1-18. hal-01207402
- Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences, 201903070. doi:10.1073/pnas.1903070116
- Berrar, D. (2018). Bayes' theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 403. doi:10.1016/b978-0-12-809633-8.20473-1
- Bhowmick, A., & Hazarika, S.M. (2016). Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends. ArXiv, abs/1606.01042.
- BİRİCİK, G., Diri, B., & SÖNMEZ, A. C. (2012). Abstract feature extraction for text classification. Turkish Journal of Electrical Engineering & Computer Sciences, 20(Sup. 1), 1137-1159.
- Bo, G., & Xianwu, H. (2006). SVM multi-class classification. Journal of Data Acquisition & Processing, 21(3), 334-339.
- Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.
- Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval (Vol. 161175).

- Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16-28.
<https://doi.org/10.1016/j.compeleceng.2013.11.024>.
- Cormack, G. V. (2008). Email Spam Filtering: A Systematic Review. *Foundations and Trends® in Information Retrieval*, 1(4), 335–455.
 doi:10.1561/15000000006. Repéré à
<https://www.nowpublishers.com/article/Details/INR-006>
- CSMC. (2017). Faire valoir les arguments en faveur des investissements dans le système de santé mentale du Canada à l'aide de considérations économiques, Commission de la santé mentale du Canada. Repéré à
https://www.mentalhealthcommission.ca/sites/default/files/2017-03/case_for_investment_fr.pdf
- Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6), e01802.
 doi:10.1016/j.heliyon.2019.e01802
- Dean, A., Hendler, J. (2011). *Semantic web for the working ontologist : effective modeling in RDFS and OWL* (2nd edition). Morgan Kaufmann
- Demidova, L., Nikulchev, E., & Sokolova, Y. (2016). The svm classifier based on the modified particle swarm optimization. *arXiv preprint arXiv:1603.08296*. <http://dx.doi.org/10.14569/IJACSA.2016.070203>
- Devitt, A., & Ahmad, K. (2007, June). Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 984-991).
- Edwards, M., Suarez-Tangil, G., Peersman, C., Stringhini, G., Rashid, A., & Whitty, M. (2018). The geography of online dating fraud. In *Workshop on Technology and Consumer Protection (ConPro)*. IEEE, San Francisco (Vol. 7).
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861-874. doi:10.1016/j.patrec.2005.10.010
- Forman, G. (2003). An experimental study of feature selection metrics for text categorization. *Journal of Machine Learning Research*, 3(1), 1289-1305.
- Frank, E., & Bouckaert, R. R. (2006, September). Naive bayes for text classification with unbalanced classes. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 503-510). Springer, Berlin, Heidelberg.
- Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Springer.
- González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, 64, 205-237.
 doi:10.1016/j.inffus.2020.07.007

- Hawkey, L. C., & Cacioppo, J. T. (2010). Loneliness Matters: A Theoretical and Empirical Review of Consequences and Mechanisms. *Annals of Behavioral Medicine*, 40(2), 218–227. doi:10.1007/s12160-010-9210-8
- Heaton, J. (2017). Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning. *Genetic Programming and Evolvable Machines*, 19(1-2), 305–307. doi:10.1007/s10710-017-9314-z
- Hérault, J., Jutten, C., & Ans, B. (1985). Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In 10 Colloque sur le traitement du signal et des images, FRA, 1985. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images.
- Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE. doi:10.1109/icdar.1995.598994
<https://pdfs.semanticscholar.org/d024/33ad6b77f740fb4f43673eed9b80b0ccb199.pdf>
- Huang, J., Stringhini, G., & Yong, P. (2015). Quit playing games with my heart: Understanding online dating scams. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment* (pp. 216-236). Springer, Cham.
- Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5), 411-430.
- Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS transactions on computers*, 4(8), 966-974.
- Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, 39(1), 1503–1509. doi:10.1016/j.eswa.2011.08.040
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* (pp. 137-142). Springer, Berlin, Heidelberg. 137–142. doi:10.1007/bfb0026683
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.
- Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.
- Kanaris, K., Houvardas, I., & Stamatatos, E. (2006). WORDS VS. CHARACTER N-GRAMS FOR ANTI-SPAM FILTERING. Repéré à <http://www.icsd.aegean.gr/lecturers/stamatatos/papers/ijait-spam.pdf>
- Koen de, J. (2019). Detecting the online romance scam: Recognising images used in fraudulent dating profiles. Repéré à http://essay.utwente.nl/80084/1/Jong_de_MA_EEMCS.pdf

- Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Stemming and lemmatization in the clustering of finnish text documents. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (pp. 625-633).
- Kowsari, K., Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text Classification Algorithms: A Survey. *Inf.*, 10, 150. doi:10.3390/info10040150
- Krovetz, R. (2000). Viewing morphology as an inference process. *Artificial intelligence*, 118(1-2), 277-294. [https://doi.org/10.1016/S0004-3702\(99\)00101-0](https://doi.org/10.1016/S0004-3702(99)00101-0).
- Kulkarni, V., & Wang, W. Y. (2017). Tfw, damngina, juvie, and hotsie-totsie: On the linguistic and social aspects of internet slang. *arXiv preprint arXiv:1712.08291*.
- Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, 6(70), 1
- Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). *Mining of massive data sets*. Cambridge university press. Repéré à <http://infolab.stanford.edu/~ullman/mmds.html#latest>.
- Lever, J., Krzywinski, M., & Altman, N. (2016). Classification evaluation.
- Li, Z., & Shen, H. (2011). SOAP: A Social network Aided Personalized and effective spam filter to clean your e-mail box. *2011 Proceedings IEEE INFOCOM*. doi:10.1109/infcom.2011.5934984
- Life After Scams Ltd. (2019). Life After Scams Ltd Submission to Productivity Commission in response to The Social and Economic Benefits to Improving Mental Health Issues Paper. Australia government productivity commission. Repéré à https://www.pc.gov.au/__data/assets/pdf_file/0011/240869/sub319-mental-health.pdf
- Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing*.
- Marika, J. (2010). Étude exploratoire des rencontres amoureuses via internet. (Thèse de doctorat inédite). Université du Québec à Montréal. Repéré à <https://archipel.uqam.ca/3914/1/D2062.pdf>
- Marr, B. (2015). *Big Data: Using SMART big data, analytics and metrics to make better decisions and improve performance*. John Wiley & Sons.
- Mi, J. X. (2014). A novel algorithm for independent component analysis with reference and methods for its applications. *PloS one*, 9(5), e93984. Doi:10.1371/journal.pone.0093984
- Naik, G. R., & Kumar, D. K. (2011). An overview of independent component analysis and its applications. *Informatica*, 35(1).
- Pustejovsky, J., Castano, J., Cochran, B., Kotecki, M., & Morrell, M. (2001). Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Studies in health technology and informatics*, (1), 371-375..

- Pustejovsky, J., Castano, J., Cochran, B., Kotecki, M., Morrell, M., & Rumshisky, A. (2001). Extraction and disambiguation of acronym-meaning pairs in medline. *Medinfo*, 10(2001), 371-375.
- Rege, A. (2009). What's Love Got to Do with It? Exploring Online Dating Scams and Identity Fraud. *International Journal of Cyber Criminology*, 3(2):494–512. Repéré à https://www.cybercrimejournal.com/aunshulregedec2009.htm#_ftn1
- Rossant-Lumbrosso, J., Rossant, L. (2020). L'anxiété : symptômes, causes et traitements. Repéré à https://www.doctissimo.fr/html/sante/encyclopedie/sa_781_anxiete.htm
- Rusch, J. J. (1999). The “social engineering” of internet fraud. In *Internet Society Annual Conference*, http://www.isoc.org/isoc/conferences/inet/99/proceedings/3g/3g_2.Htm
- Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P., & Mononen, J. (2018). Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. *Behavioural processes*, 148, 56-62.. DOI: 10.1016/j.beproc.2018.01.004.
- Saif, H., Fernández, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In N. Calzolari, K. Choukri, T. Declerck, & et al (Eds.), *LREC 2014, Ninth International Conference on Language Resources and Evaluation. Proceedings* (pp. 810-817) http://www.lrec-conf.org/proceedings/lrec2014/pdf/292_Paper.pdf
- Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620. doi:10.1145/361219.361220
- Sanz, E. P., Gómez Hidalgo, J. M., & Cortizo Pérez, J. C. (2008). Chapter 3 Email Spam Filtering. *Software Development*, 45–114. doi:10.1016/s0065-2458(08)00603-7
- Schwartz, A.S., & Hearst, M. (2003). A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 451-62 .
- Scott, S. and Matwin, S. (1999) Feature engineering for text classification, *Proceedings of ICML-99, 16th International Conference on Machine Learning*. Repéré à http://www.site.uottawa.ca/~stan/papers/1999/icml99.2_word6.pdf
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1), 1-47.
- Shlens, J. (2014). A tutorial on independent component analysis. *arXiv preprint arXiv:1404.2986*.
- Sohrab, K., Rob, L. (2017). Employee Financial Health: How Companies Can Invest in Workplace Wellness. Repéré à <https://s3.amazonaws.com/cfsi-innovation-files/wp-content/uploads/2017/05/26183930/2017-Employee-FinHealth.pdf>.

- Suarez-Tangil, G., Edwards, M., Peersman, C., Stringhini, G., Rashid, A., & Whitty, M. (2020). Automatically Dismantling Online Dating Fraud. *IEEE Transactions on Information Forensics and Security*, 15, 1128-1137. doi:10.1109/tifs.2019.2930479
- Tabassum, A., & Patil, R. R. (2008). A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing.
- Tade, O. (2013). A spiritual dimension to cybercrime in Nigeria: The 'yahoo plus' phenomenon. *Human Affairs*, 23(4), 689-705. Repéré à https://www.academia.edu/25701080/a_spiritual_dimension_to_cybercrime_in_nigeria_the_yahoo_plus_phenomenon
- Tade, O., & Aliyu, I. (2011). Social organization of Internet fraud among university undergraduates in Nigeria. *International Journal of Cyber Criminology*, 5(2). Repéré à https://www.academia.edu/30925682/Social_Organization_of_Internet_Fraud_among_University_Undergraduates_in_Nigeria 860–875
- Vapnik, V., & Chervonenkis, A. Y. (1964). A class of algorithms for pattern recognition learning. *Avtomat. i Telemekh*, 25(6), 937-945.
- Varghese, R., & Dhanya, K. A. (2017). Efficient Feature Set for Spam Email Filtering. 2017 IEEE 7th International Advance Computing Conference (IACC). doi:10.1109/iacc.2017.0152
- Verma, T., Renu, R., & Gaur, D. (2014). Tokenization and Filtering Process in RapidMiner. *International Journal of Applied Information Systems*, 7, 16-18. Repéré à
- Wang, H., Khoshgoftaar, T. M., & Van Hulse, J. (2010). A comparative study of threshold-based feature selection techniques. In 2010 IEEE International Conference on Granular Computing (pp. 499-504). IEEE.
- Waykole, R. N., & Thakare, A. D. (2018). A Review of feature extraction methods for text classification. *IJAERD*, 4(4), 351-354.
- Whitty, M. T. (2013). The Scammers Persuasive Techniques Model: Development of a Stage Model to Explain the Online Dating Romance Scam. *British Journal of Criminology*, 53(4), 665–684. doi:10.1093/bjc/azt009
- Whitty, M. T. (2018). Do You Love Me? Psychological Characteristics of Romance Scam Victims. *Cyberpsychology, Behavior, and Social Networking*, 21(2), 105–109. doi:10.1089/cyber.2016.0729
- Whitty, M. T., & Ng, M. (2017). Literature Review for UNDERWARE: UNDERstanding West African culture to pRevent cybercrimEs. Report for the National Cyber Security Centre as part of a group of studies funded in the Research Institute in Science of Cyber Security.
- Wilson, R. S., Krueger, K. R., Arnold, S. E., Schneider, J. A., Kelly, J. F., Barnes, L. L., ... Bennett, D. A. (2007). Loneliness and Risk of Alzheimer Disease. *Archives of General Psychiatry*, 64(2), 234. doi:10.1001/archpsyc.64.2.234

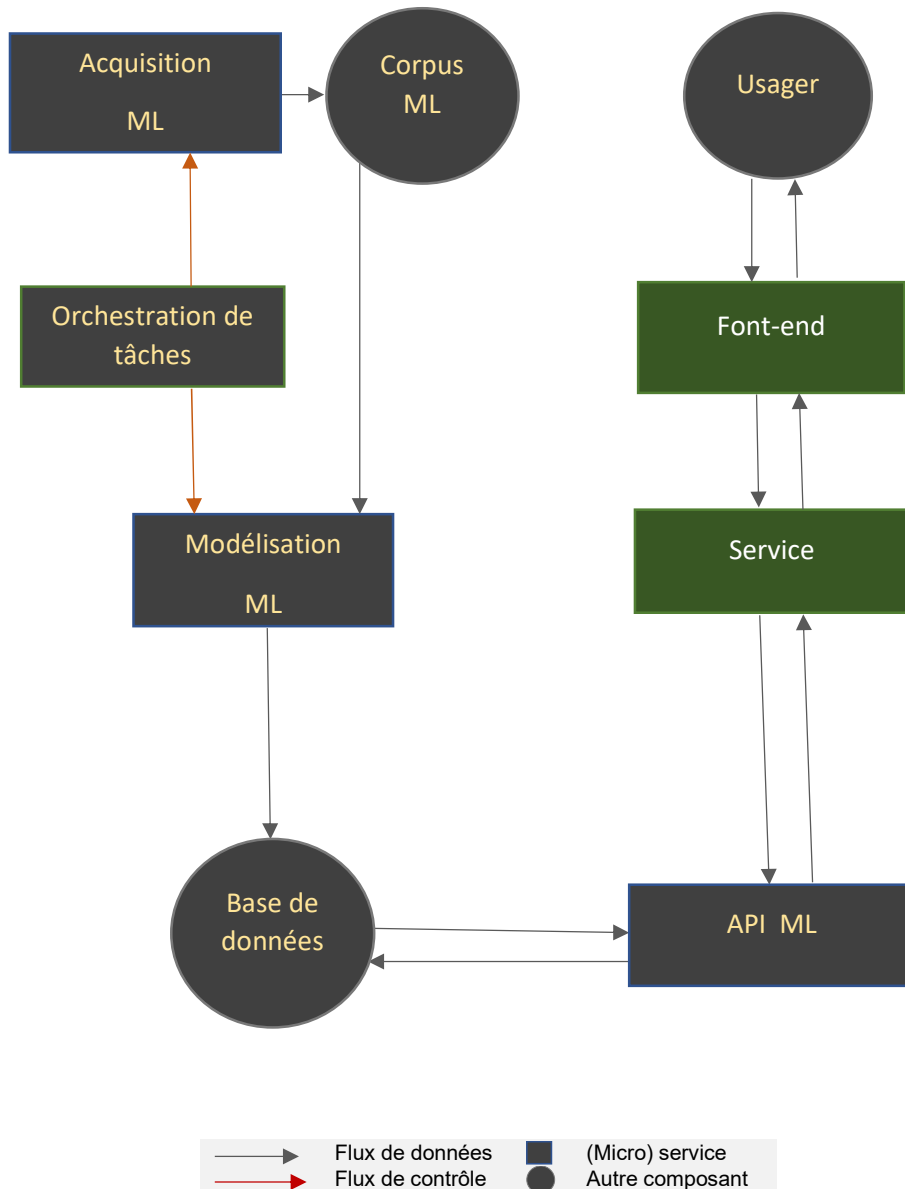
- Wilson, R. S., Krueger, K. R., Arnold, S. E., Schneider, J. A., Kelly, J. F., Barnes, L. L., ... Bennett, D. A. (2007). Loneliness and Risk of Alzheimer Disease. *Archives of General Psychiatry*, 64(2), 234. doi:10.1001/archpsyc.64.2.234 .
- Wu, L., Morstatter, F., & Liu, H. (2016). Slangs: Building and using a sentiment dictionary of slang words for short-text sentiment classification. *arXiv preprint arXiv:1608.05129*.
- Xiong, Z., Cui, Y., Liu, Z., Zhao, Y., Hu, M., & Hu, J. (2020). Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Computational Materials Science*, 171, 109203. Doi:10.1016/j.commatsci.2019.109203
- Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '99*. doi:10.1145/312624.312647
- Yates, J. (2019). Cœurs sensibles s'abstenir, Les ravages des arnaques amoureuses sur le web, Extrait le 2020-03-03, <https://ici.radio-canada.ca/special/arnaques-amoureuses-femmes-victimes-vol-web-facebook/>
- Zhang, C., & Ma, Y. (Eds.). (2012). *Ensemble machine learning: methods and applications*. Springer Science & Business Media.
- Zhang, L., Zhu, J., & Yao, T. (2004). An evaluation of statistical spam filtering techniques. *ACM Transactions on Asian Language Information Processing*, 3(4), 243–269. doi:10.1145/1039621.1039625
- Zhou, Z. H. (2012). *Ensemble methods: foundations and algorithms*. CRC press.

Bibliographie

Ihadjadene, M. (2004). Les systèmes de recherche d'informations: modèles conceptuels. Hermès Science.

Annexes

Annexe 1 : Architecture de la solution applicative



Acquisition ML :
Module de capture de données.

Modélisation ML :
Construction du modèle par réalisation des expériences d'optimisation des hyperparamètres.

API ML : Service REST d'interrogation des expériences.

Orchestration de tâches : Organise, planifie et exécute les tâches ML.

Corpus ML : données d'entraînement et de test.

Base de données : Point de persistance des expériences ML et usagers.

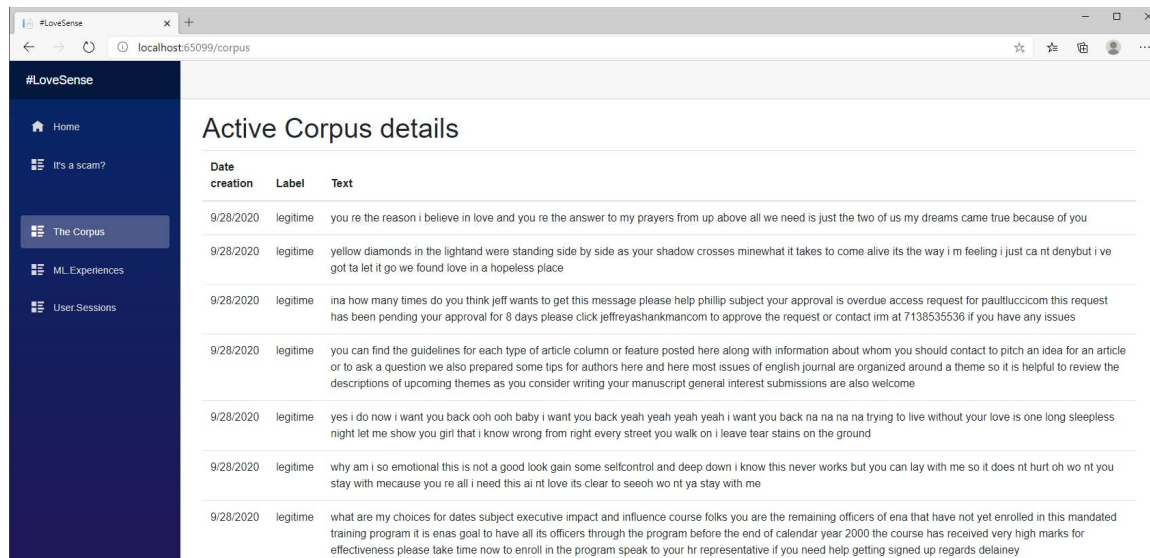
Service : Interface de contrôle des besoins de service.

Front-end : Module d'interaction avec l'utilisateur.

Annexe 2 : Les fonctionnalités du système

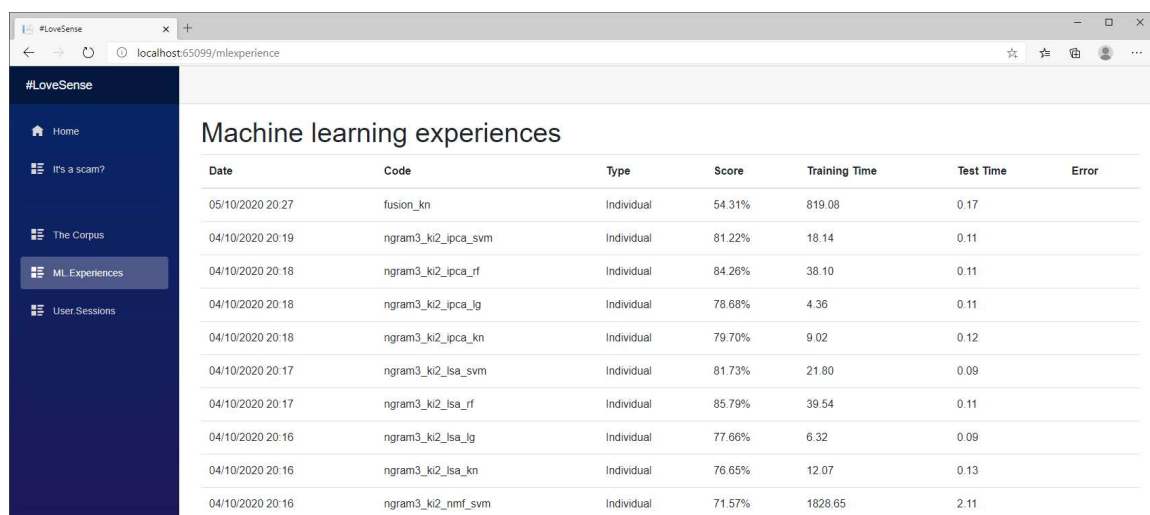
Le système dispose de quatre fonctions :

La grille des messages du corpus.



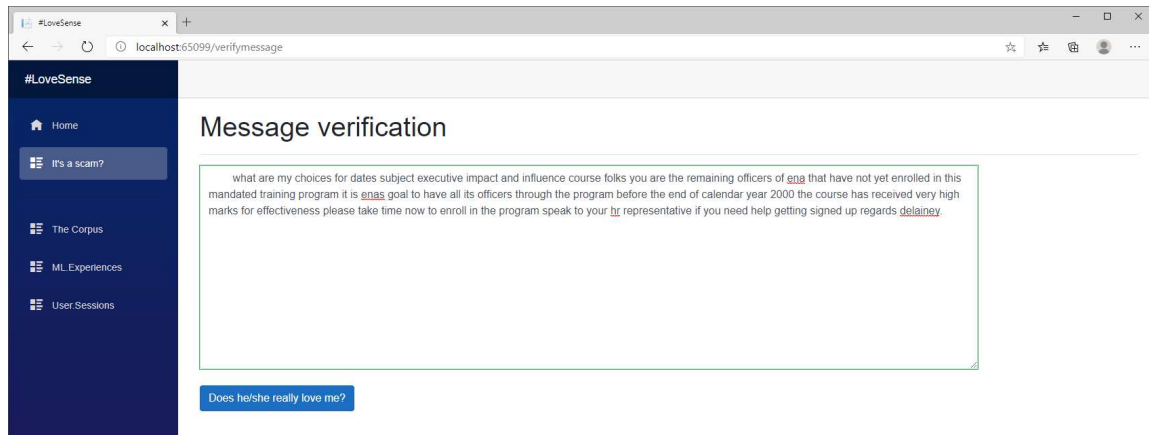
Date	Label	Text
9/28/2020	legitime	you re the reason i believe in love and you re the answer to my prayers from up above all we need is just the two of us my dreams came true because of you
9/28/2020	legitime	yellow diamonds in the lightand were standing side by side as your shadow crosses minewhat it takes to come alive its the way i m feeling i just ca nt denybut i ve got ta let it go we found love in a hopeless place
9/28/2020	legitime	ina how many times do you think jeff wants to get this message please help philip subject your approval is overdue access request for paulluccicom this request has been pending your approval for 8 days please click jeffreyashankmancom to approve the request or contact irm at 7138535536 if you have any issues
9/28/2020	legitime	you can find the guidelines for each type of article column or feature posted here along with information about whom you should contact to pitch an idea for an article or to ask a question we also prepared some tips for authors here and here most issues of english journal are organized around a theme so it is helpful to review the descriptions of upcoming themes as you consider writing your manuscript general interest submissions are also welcome
9/28/2020	legitime	yes i do now i want you back ooh ooh baby i want you back yeah yeah yeah i want you back na na na na na trying to live without your love is one long sleepless night let me show you girl that i know wrong from right every street you walk on i leave tear stains on the ground
9/28/2020	legitime	why am i so emotional this is not a good look gain some selfcontrol and deep down i know this never works but you can lay with me so it does nt hurt oh wo nt you stay with mecause you re all i need this ai nt love its clear to seeoh wo nt ya stay with me
9/28/2020	legitime	what are my choices for dates subject executive impact and influence course folks you are the remaining officers of ena that have not yet enrolled in this mandated training program it is enas goal to have all its officers through the program before the end of calendar year 2000 the course has received very high marks for effectiveness please take time now to enroll in the program speak to your hr representative if you need help getting signed up regards delainey

La grille des expériences destinées à rechercher le meilleur modèle.

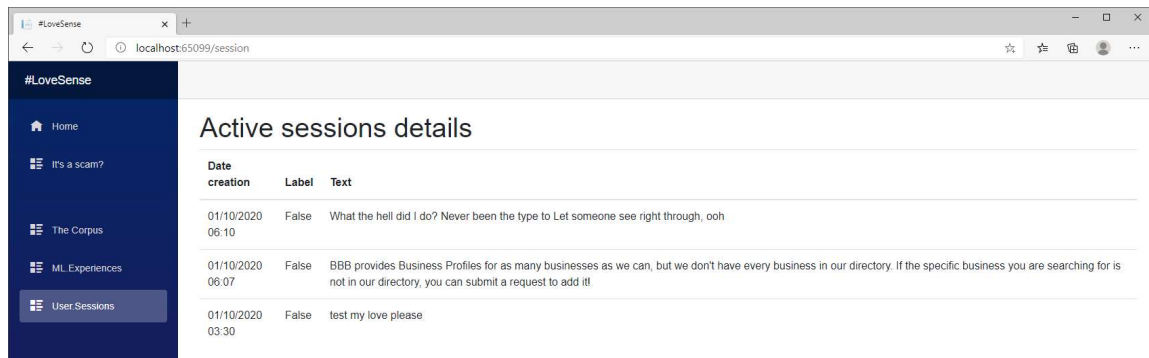


Date	Code	Type	Score	Training Time	Test Time	Error
05/10/2020 20:27	fusion_kn	Individual	54.31%	819.08	0.17	
04/10/2020 20:19	ngram3_ki2_ipca_svm	Individual	81.22%	18.14	0.11	
04/10/2020 20:18	ngram3_ki2_ipca_rf	Individual	84.26%	38.10	0.11	
04/10/2020 20:18	ngram3_ki2_ipca_lg	Individual	78.68%	4.36	0.11	
04/10/2020 20:18	ngram3_ki2_ipca_kn	Individual	79.70%	9.02	0.12	
04/10/2020 20:17	ngram3_ki2_lsa_svm	Individual	81.73%	21.80	0.09	
04/10/2020 20:17	ngram3_ki2_lsa_rf	Individual	85.79%	39.54	0.11	
04/10/2020 20:16	ngram3_ki2_lsa_lg	Individual	77.66%	6.32	0.09	
04/10/2020 20:16	ngram3_ki2_lsa_kn	Individual	76.65%	12.07	0.13	
04/10/2020 20:16	ngram3_ki2_nmf_svm	Individual	71.57%	1828.65	2.11	

La fonction de vérification des messages permet aux usagers de vérifier la véracité de la relation amoureuse.



L'historique des sessions usagers.



Annexe 3 : Implémentation du modèle de classification

Dans la recherche du modèle optimal, nous avons effectué deux types d'expérimentation dont l'implémentation ainsi que le code source complet de la solution a été hébergée sur GitHub (<https://github.com/kamdemize/LoveSense>).

Pour ces deux types d'expériences, le corpus est importé, et prétraité par des opérations du traitement automatique du langage naturel, puis les données issues sont partitionnées en deux catégories (70% apprentissage et 30% test). Nous définissons ensuite des expériences qui sont au fait un ensemble de méthodes à combiner suivant le processus de classification. Cela est fait par la méthode *definir_expériences()*. Ensuite, ces expériences et les données du corpus prétraitées sont utilisées dans la fonction *conduire_expériences()*, qui va se charger de mener les expériences combinatoires de classification. Le processus de classification de chaque expérience est évalué et les résultats sont persistés dans une base de données MongoDB.

La fonction *conduire_expériences()* est implémentée suivant les spécificités du type d'expérience. À l'expérimentation 1, l'on recherche la combinaison idéale à contribution individuelle des méthodes (<https://github.com/kamdemize/LoveSense/blob/main/LoveSense.ML.Modelisation/ExperimentModel.py>) alors qu'à l'expérience 2, il est question de mettre en concurrence les méthodes de même nature afin d'obtenir la meilleure optimisation possible (<https://github.com/kamdemize/LoveSense/blob/main/LoveSense.ML.Modelisation/OptimalModel.py>).

Annexe 4 : Outillage d'implémentation de la solution

Pour mettre en place l'application informatique, nous avons utilisé les outils suivants :

Application Web	Blazor(shorturl.at/biqtL) Asp.net Core 3.1.8 (shorturl.at/twJY3) .NetCore 3.1 (shorturl.at/clSZ0)
Prétraitement (TALN)	Spacy (spacy.io)
Construction du classifieur	Scikit-learn (scikit-learn.org)
API du modèle	Flask (flask.palletsprojects.com)
Test de l'API	Postman(postman.com)
Analyse de données	Anaconda Python (python.org)
Persistance	MongoDB (mongodb.com)
Éditeur de développement	Visualstudio 2019 Jupyter notebook (Jupyter.org)