République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université A. MIRA-BEJAIA
Faculté des Sciences Exactes
Département d'Informatique
Laboratoire d'Informatique Médicale (LIMED)

THÈSE

Présentée par

Khaled BEDJOU

Pour l'obtention du grade de

DOCTEUR EN SCIENCES

Filière : Informatique

Option : Cloud Computing

Thème

Analyse des réseaux sociaux pour détecter et prévenir les menaces terroristes

Soutenue le : 22.02.2024 Devant le Jury composé de :

Nom et Prénom	Grade		
M. Achroufene Achour	**MCA**	Univ. de Béjaia	Président
M. Azouaou Faiçal	**Prof**	ESTIN-Amizour	Rapporteur
M. Amad Mourad	**Prof**	Univ. de Bouira	Examinateur
M. Farah Zoubeyr	**MCA**	Univ. de Béjaia	Examinateur

Année Universitaire : 2023-2024

Remerciements

Je tiens à remercier en premier lieu, mon Directeur de thèse le Professeur AZOUAOU Faiçal et le Dr ALOUI Abdelouhab pour leur suivi, leur encouragement et leur patience avec moi tout au long de cette thèse ;

Je remercie le président de jury Monsieur ACHROUFENE Achour qui m'a honoré d'avoir accepté de présider le jury de ma soutenance ;

Je remercie Messieurs FARAH Zoubeyr & AMAD Mourad d'avoir accepté d'examiner mon travail ;

Je tiens à exprimer ma profonde gratitude envers mes chers parents et l'ensemble de ma famille pour leur soutien constant et leur confiance tout au long de la réalisation de ma thèse ;

Je remercie également mes amis et collègues du département d'informatique et du laboratoire LIMED de l'université de Bejaia, en particulier mon ami AKILAL Karim pour son aide précieuse ;

Je remercie tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

A la mémoire de mon directeur de thèse, collègue et ami le Dr ALOUI Abdelouhab qui a dirigé cette thèse de 2017 à 2020

ALOUI Abdelouhab
1976-2020

Liste des travaux

Revues internationales avec comités de lecture

1. Bedjou, K & Azouaou, F. (2023). Detection of terrorism's apologies on Twitter using a new bi-lingual dataset. *nt. J. Data Mining, Modelling and Management, , Vol. 15, No. 04,*, pp.331–354.

Communications avec acte et comités de lecture

1. Khaled Bedjou, Abdelouhab Aloui et Faical Azouaou : LexD3T : A Lexical Detection Process of Terrorist Threats on Twitter. 7th. International Symposium ISKO-Maghreb'2018 Knowledge Organization in the perspective of Digital Humanities. Novembre 2018, Béjaia. Algérie.

2. Khaled Bedjou, Faical Azouaou et Abdelouhab Aloui : Detection of terrorist threats on Twitter using SVM. The 3rd International Conference on Future Networks and Distributed Systems ICFNDS '19, July 1–2, 2019, Paris, France.

Communications (hors thèse) avec acte et comités de lecture

1. Khaled Bedjou, Lamia Berkani et Faical Azouaou : Prototype d'un système de recherche personnalisée de services web sémantiques. 7th. International Symposium ISKO-Maghreb'2018 Knowledge Organization in the perspective of Digital Humanities. Novembre 2018, Béjaia. Algérie.

Autres évènements scientifiques

a. Membre du comité d'organisation du 1er WorkShop on Deep Learning : Methods and Applications, Novembre 2018. Université de Béjaia.
b. Participation à l'African Machine Learning Summer School, Juin 2018. USTHB – Alger.
c. Présentation d'un sémianaire au Laboratoire LIMED, intitulée « Les réseaux sociaux, les défis » Mai 2017. Université de Béjaia.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

API	Application Programming Interface
AUC	Area Under the ROC Curve
BERT	Bidirectional Encoder Representations from Transformers
Bi-GRU	Bidirectional Gated Recurrent Units
Bi-LSTM	Bidirectional Long Short Term Memory
BOW	Bag Of Words
CBOW	Continuous Bag of Words
CNN	Convolutional Neural Network
DBOW	Distributed Bag of Words
DL	Deep Learning
DM	Distributed Memory
DNN	Deep Neural Network
Doc2Vec	Document to Vector
DT	Decision Tree
GloVe	Global Vector
GPT	Generative Pre-training Transformer
GRU	Gated Recurrent Units
ISIS	Islamic State of Iraq and Sham
KNN	k-nearest neighbor
L-HSAB	Levantine Hate Speech et Abusive
LexD3T	Lexical Detection Process of Terrorist Threats on Twitter
LM	Language Modeling
LR	Logistic Regression
LSTM	Long Short Term Memory
MaxEnt	Maximale Entropie
ML	Machine Learning
MLM	Masked Language Modeling
MLP	Multi-Layer Perceptron
NB	Naïve Bayes
NER	Named-Entity Recognition
NLP	Natural Language Process
OSN	On-line Social Network

PCNN	Pronunciation-based Convolutional Neural Network
POS	Part Of Speech
RecNN	Recursive Neural Network
RF	Random Forest
RNN	Recurrent Neural Network
RS	Réseaux sociaux
SG	Skip Gram
SGD	Stochastic Gradient Descent
SemEval	Semantic textual similarity Evaluation
SMOTE	Synthetic Minority Over- sampling Technique
SMS	Short Message Service
SVC	Support Vector Classifier
SVM	Support Vector Machine
Tf-IDF	Term Frequency-Inverse Document Frequency
URL	Uniform Ressource Locator
Word2Vec	Word to Vector

Introduction générale

L'utilisation des réseaux sociaux est devenue quotidienne, avec plus de 64 % de la population du monde qui est active sur les réseaux sociaux (We-are-social, 2023). La majorité des utilisateurs des réseaux sociaux s'expriment dans un langage non académique et dans des langues différentes, ce qui rend la tâche de contrôle des publications très difficile. L'accès facile et rapide aux réseaux sociaux permet à toutes les catégories de personnes (les personnes à faible personnalité, les populations jeunes et très jeunes) d'être exposés à la violence, au harcèlement et à la radicalisation. Beaucoup de critiques sont faites aux réseaux sociaux, même par leurs propres fondateurs à l'instar de Chris Hughes co-fondateur de Facebook qui a appelé dans une publication dans le New York Times[1] au démantèlement de Facebook pour limiter ses risques et son pouvoir.

Des problèmes liés à la sécurité des personnes sont relevés quotidiennement sur les réseaux sociaux : des problèmes de violation de vie privée, des problèmes de dépendances et risque de l'utilisation de ces réseaux par les mineurs, harcèlements, insultes, messages de haines, contenu obscène, problèmes de sécurité des personnes fragiles exposées aux contenus violents, contenu faisant l'apologie du terrorisme, etc.
La détection des menaces terroristes sur les réseaux sociaux représente l'objet de notre thèse. Nous pouvons trouver sur les réseaux sociaux différents types de ces menaces comme :

- Apologie des opérations terroristes
- Apologie des mouvements terroristes
- Coordination à travers les réseaux sociaux pour préparer des actes terroristes
- Recrutement des nouveaux membres
- Diffusion d'images ou de vidéos violentes

Bien que la plupart des plateformes de médias sociaux ont mis en place des mécanismes permettant aux utilisateurs de signaler des publications dangereuses, cette mesure s'avère insuffisante. Il subsiste néanmoins des problèmes liés à la sécurité des personnes car les contenus ne sont pas supprimés immédiatement. Ils sont plutôt maintenus en ligne jusqu'à ce

[1] https://www.nytimes.com/2019/05/09/business/facebook-response-chris-hughes.html

qu'ils atteignent un seuil de signalements spécifique ou soient examinés par un modérateur. Cette tâche prend beaucoup de temps et permet au contenu d'être vu et partagé, ce qui pose un vrai risque sur les personnes exposées à ces contenus.

A titre d'exemple, en novembre 2016, un journaliste algérien a été arrêté pour apologie du terrorisme[2]. En effet, ce journaliste avait applaudi des deux mains l'exécution par Daesh d'un pilote jordanien en février 2015. Il avait auparavant écrit plusieurs publications en faveur de l'organisation terroriste Daesh sur les réseaux sociaux mais sans être repéré (On-line, 2016). Comme ce journaliste, il y a des dizaines, voire des centaines de personnes qui font l'apologie du terrorisme sur les réseaux sociaux mais qui ne sont pas automatiquement détectées par les services de sécurité, surtout celles écrites en arabe. Notre motivation pour entreprendre cette thèse résidait dans le souhait d'apporter une assistance aux services de sécurité en automatisant la détection rapide de ce type de contenus, ainsi que de leurs auteurs.

Pour remédier à ces problèmes, la détection automatique (par la machine) et immédiate du contenu violent est primordiale. Cependant, vu l'insuffisance des outils informatiques aidant à détecter ce phénomène, il importe de concevoir un système automatique apte à la recherche et l'analyse des données qui circulent sur les réseaux sociaux afin de pouvoir détecter et diminuer les menaces.

Nous visons à avoir la capacité de repérer et détecter les contenus qui représentent de véritables menaces terroristes, de ceux qui ne le sont pas. Par exemple, parmi ces deux publications:

- *Tribute to the Kouachi brothers who have just fallen Martyr. Jazakoum allah khayr we are happy.*
- *#Tunisie was just as victim of Daech attacks! Not coming to blame an entire people for a terrorist crime.*

Comment faire la distinction que la première publication peut présenter une vraie menace terroriste et qu'il faut la détecter et la signaler automatiquement, tant dis que la 2ème ne présente aucun danger. Cette tâche est relativement simple pour un être humain, mais elle est extrêmement complexe pour une machine. C'est pourquoi nous avons proposé trois contributions de recherche pour y remédier.

[2] https://www.lecourrierdelatlas.com/algerie-un-journaliste-d-echourouk-arrete-pour-apologie-au-terrorisme-6733/

La première contribution (Bedjou, et al., 2018) consiste en un processus de détection des menaces terroristes sur Twitter en utilisant une approche lexicale. Il s'agit dans cette approche de détecter tout contenu lié au terrorisme. Nous nous sommes basés dans cette contribution sur les techniques du traitement automatique de la langue (NLP), avec le lexique des deux langues arabe et anglais pour repérer les publications comportant des termes associés au terrorisme. Nous avons opté pour ces deux langues pour les raisons suivantes : (1) l'anglais est la langue la prédominante dans les réseaux sociaux ; (2) il existe peu de recherches menées en langue arabe, et nous aspirons à apporter notre contribution dans ce domaine. Les résultats obtenus dans cette première contribution en termes de f-measure sont 0.035 pour l'arabe et 0.033 pour l'anglais.

La deuxième contribution (Bedjou, et al., 2019) porte sur l'intégration des techniques et algorithmes de l'apprentissage automatique visant à apprendre à la machine à classifier des publications selon deux catégories *Menaçant* ou *Non menaçant,* et cela dans les deux langues, à savoir l'arabe et l'anglais. Il s'agit d'un système conçu pour classer et détecter des publications susceptibles de constituer de réelles menaces terroristes sur le réseau social Twitter. Pour cela, nous avons élaboré un processus en 12 étapes pour analyser, traiter et identifier les tweets menaçants. La classification binaire est réalisée à l'aide du classifieur SVM. Les résultats obtenus dans cette deuxième contribution en termes de f-measure sont comme suit : pour la classes Menaçant, 0.26 en anglais et 0.42 en arabe, et pour la classe non menaçant, 0.70 en anglais et 0.77 en arabe.

Dans notre troisième contribution (Bedjou & Azouaou, 2023), nous avons proposé un processus indépendant de la langue pour détecter et classer les apologies des terroristes sur Twitter en trois catégories (apologie, non apologie et neutre). Pour cela, nous avons construit un ensemble de données comprenant 12 155 tweets annotés manuellement. Nous avons exploré diverses techniques de représentation de texte, notamment : TF-IDF, BERT, et l'incorporation de couches (Layer Embedding), ainsi que divers algorithmes d'apprentissage automatique tels que RF, DT, KNN et NB, et des algorithmes d'apprentissage profond comme GRU, SimpleRNN, LSTM, BiLSTM et BERT pour la classification. En outre, nous avons utilisé l'apprentissage par transfert en ajoutant une couche à un modèle préalablement entraîné sur un ensemble de données massives avec nos propres données. BERT a réalisé les performances les plus élevées en matière de f-mesure dans cette troisième contribution, affichant des taux variants entre 0,57 et 0,90 pour l'arabe, et entre 0,78 et 0,87 pour l'anglais. Il est à noter que, exception faite de la

classe positive en anglais, où Random Forest a surpassé les autres en obtenant la meilleure performance avec un taux de 0,66.

Cette thèse est constituée de quatre chapitres comme suit :

- Dans le premier chapitre, nous exposons une analyse approfondie des réseaux sociaux, mettant l'accent sur les divers types d'analyses, les modalités d'accès aux données, les techniques de représentation de texte les plus couramment employées dans la recherche, ainsi que les défis auxquels nous avons été confrontés au cours de l'élaboration de cette thèse.
- Dans le chapitre 2, nous présentons un état de l'art des travaux de recherche sur la détection des menaces terroristes sur les réseaux sociaux, ainsi que des travaux sur la détection des messages haineux, de l'analyse des sentiments. Nous détaillons pour chaque travail étudié, l'approche et la méthodologie utilisées, les techniques de représentation de texte employées, les ensembles de données utilisés, les expérimentations menées et les résultats de performances obtenues.
- Dans le chapitre 3, nous détaillons notre démarche pour la construction de notre ensemble de données (data set). Nous présentons les techniques employées pour normaliser nos données, puis exposons le processus suivi pour l'annotation manuelle de nos données en trois classes. Enfin, nous discutons des techniques de traitement du langage naturel (NLP) utilisées pour le prétraitement et le nettoyage de nos données en vue de les préparer à un apprentissage machine.
- Dans le chapitre 4, nous exposons de manière approfondie les trois contributions de notre thèse. Pour chacune d'entre elles, nous détaillons la méthodologie adoptée, l'ensemble de données mobilisé, les expérimentations conduites, et nous présentons en détail les résultats obtenus.

Nous terminons cette thèse par une conclusion générale qui résume notre travail, et dans laquelle nous présentons les constats de nos recherches, ainsi que quelques perspectives de recherche.

Partie 1.
État de l’art

Chapitre 1. Analyse des réseaux sociaux

En 2023, plus de 4,9 milliards de la population est active sur les réseaux sociaux, et la majorité des utilisateurs s'expriment dans un langage naturel, non académique et dans plusieurs langues différentes. Ceci rend le traitement et contrôle des publications compliqué et très difficile. Nous présentons dans ce chapitre l'analyse des réseaux sociaux d'une manière générale, puis nous nous focalisons sur les techniques les plus utilisées pour la représentation et traitement du texte extraits de ces réseaux.

1. Introduction

Les réseaux sociaux sont des applications (web ou mobiles) qui permettent aux inscrits de partager des contenus sur leurs comptes ou via des pages et des espaces créés. Parmi les réseaux sociaux les plus connus et les plus utilisés nous citons : Facebook, Youtube, Twitter, Instagram, TikTok… La figure 1 présente les plateformes les plus utilisées avec le nombre d'utilisateurs de chaque plateforme.

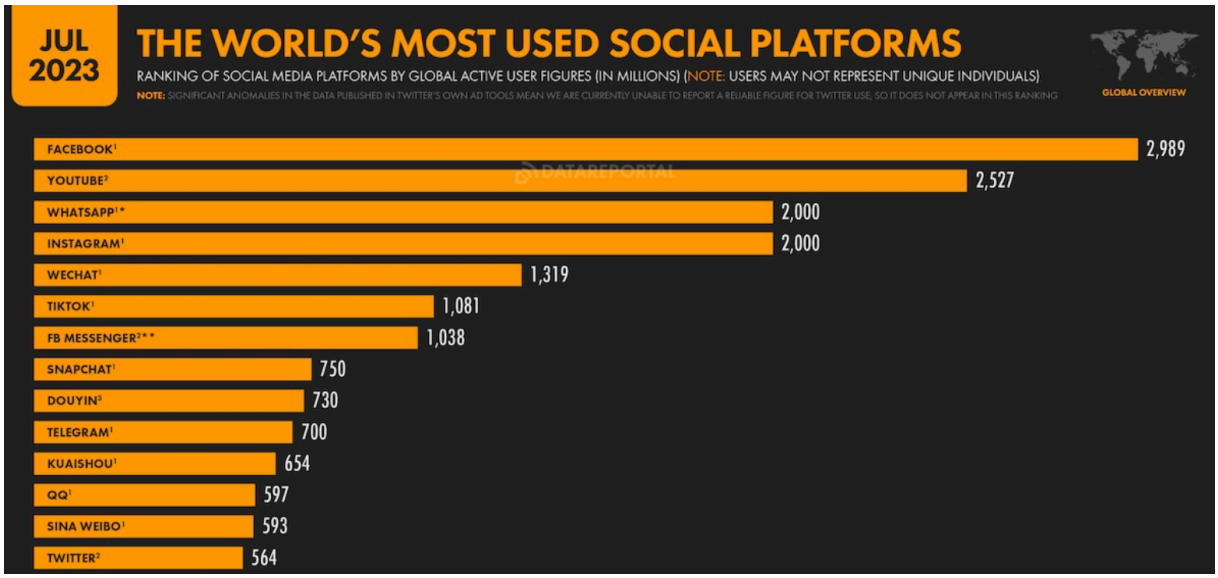


Figure 1. Plateformes de réseaux sociaux (We-are-social, 2023)

La taille de données qui circulent chaque jour sur ces réseaux sociaux est gigantesque avec par exemple plus de 10 milliards de messages échangés sur Facebook ou 500 millions de tweets publiés par jour sur Twitter (We-are-social, 2023).

2. Types d'analyse de réseaux sociaux

Il existe deux types d'analyse de réseaux sociaux qui sont l'analyse structurelle et l'analyse du contenu.

2.1. Analyse structurelle

Consiste à analyser la structure du réseau, c'est-à-dire une cartographie des relations ou liens entre les membres du réseau (nœuds). Il s'agit d'étudier le réseau social comme un graphe où les nœuds sont les personnes ou les groupes tandis que les liens sont les relations ou les flux entre ces nœuds. La figure 2 suivante illustre un exemple de nœuds et de leurs liens dans un réseau social.

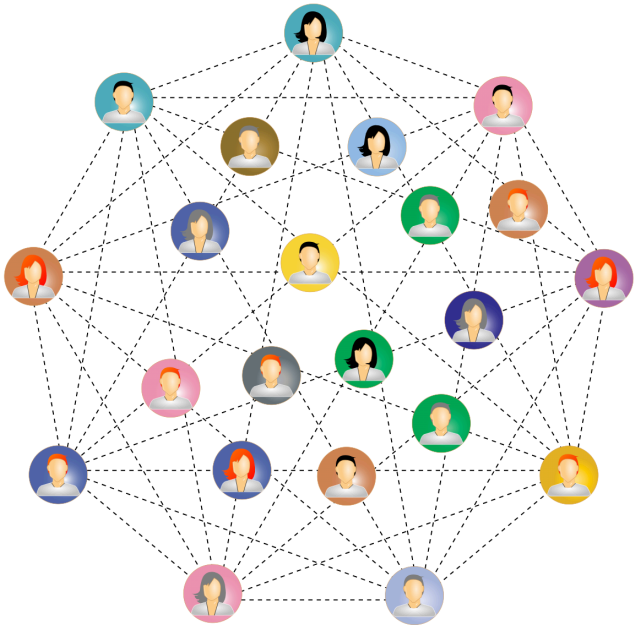


Figure 2. Structure d'un réseau social

Parmi les axes de recherche que nous pouvons classer dans l'analyse structurelle, nous citons :

- Analyse des graphes sociaux : il s'agit de visualiser les connexions et mesurer les propriétés telles que la centralité d'un nœud, la détection et prédiction des liens, identification des groupes et communautés, systèmes de recommandation, etc.
- Analyse d'influence : elle consiste à identifier les individus qui ont un grand impact sur les groupes ou communautés.
- Analyse de réseau de diffusion : c'est étudier comment les informations se propagent dans un réseau avec l'identification des chemins de diffusion et des acteurs qui l'influencent.

2.2. Analyse de contenu

Consiste à analyser les publications, les messages échangées ou toute autre forme de données qui circulent dans le réseau. Ces contenus peuvent être sous plusieurs formats : image, vidéo, texte… Il s'agit donc de récolter des données et à les analyser afin de prendre des décisions stratégiques.

Parmi les axes de recherche que nous pouvons classer dans l'analyse du contenu, nous citons :

1. Analyse de désinformation et de fausses informations : détecter les rumeurs et les fausses informations sur les réseaux sociaux, mécanismes de vérification de la fiabilité d'une information(source).
2. Analyse d'engagement et d'opinion : comprendre la réaction des utilisateurs avec un contenu publié sur les réseaux sociaux, ça inclut les partages, les retweets, les mentions (j'aime, émoticônes…), etc.
3. Analyse de sentiment : il s'agit de détecter le ton émotionnel des messages, des publications ou commentaires. En général trois classes sont utilisées dans l'analyse de sentiment à savoir : positive, négative et neutre.

Important : Notre thèse s'inscrit dans le dernier axe à savoir : l'analyse de sentiments appliquée sur des publications textuelles. Dans toute la suite de cette thèse, nous nous intéresserons qu'aux contenus textuels sur les réseaux sociaux. Cette focalisation s'explique par le fait que ces contenus constituent la majeure partie des données, étant donné que la plupart des utilisateurs expriment leurs opinions et avis sous forme de texte. De plus, la détection automatique des contenus menaçants et faisant l'apologie du terrorisme n'est pas encore entièrement automatisée, représentant ainsi un défi persistant dans le domaine de la recherche.

3. Domaines d'application de l'analyse des réseaux sociaux

A travers les données collectées sur les réseaux sociaux, nous pouvons extraire des informations cruciales que nous pouvons exploiter dans plusieurs domaines (Farzindar, 2013). Parmi ces domaines, nous citons :

- Marketing et veille concurrentielle : c'est avec l'analyse des réseaux sociaux que les entreprises étudient les opinions, avis et réactions des clients à leurs produits. Cela les aide à revoir leurs stratégies de marketing, à identifier les tendances du marché, et être à jour face à la concurrence.

- Politique : l'analyse des réseaux sociaux permet de suivre les opinions politiques, la popularité des personnes politiques et leurs programmes, de prédire les résultats d'élections, etc.
- Gestion de crise : en surveillant les réseaux sociaux, les organisations ou gouvernements peuvent détecter préalablement des crises émergentes, des risques potentiels d'explosion sociales, des manifestations ou émeutes… pour prendre les mesures nécessaires de les éviter ou d'y faire face.
- Sécurité : l'analyse des réseaux sociaux joue un rôle très important dans le domaine de la sécurité en aidant à détecter les comportements malveillants, lutter contre la désinformation et propagation des rumeurs, détecter les menaces et identifier toute forme de radicalisation et d'extrémisme.

Important : Notre thèse s'inscrit dans le domaine de la Sécurité, il s'agit donc d'analyser les contenus textuels des réseaux sociaux pour détecter les menaces terroristes. Notre motivation principale est d'assister les services de sécurité à repérer les contenus (et éventuellement leurs auteurs) pouvant représenter un vrai danger à la population.

4. Techniques d'analyse des réseaux sociaux

Plusieurs techniques sont utilisées dans l'analyse des réseaux sociaux selon l'objectif à atteindre. Batrinca et Treleaven (Batrinca & Treleaven, 2015) ont classé ces techniques en 3 catégories différentes :

a. **Domaine des statistiques et de la Recherche d'information (RI)**: des méthodes statistiques à forte intensité de calcul, y compris les méthodes de ré échantillonnage, les méthodes Monte Carlo de la chaîne de Markov, la régression locale, etc.
b. ***Domaine des physiciens et des mathématiciens*** : des modèles de simulation complexes de systèmes difficiles à prédire, dérivés de la physique statistique, de la théorie de l'information et de la dynamique non linéaire.
c. **Domaine de l'Intelligence Artificielle :** le champ de l'intelligence artificielle englobe une diversité de domaines, tels que l'ingénierie des connaissances, le raisonnement, le traitement automatique du langage, l'apprentissage automatique, les systèmes multi-agents, etc.

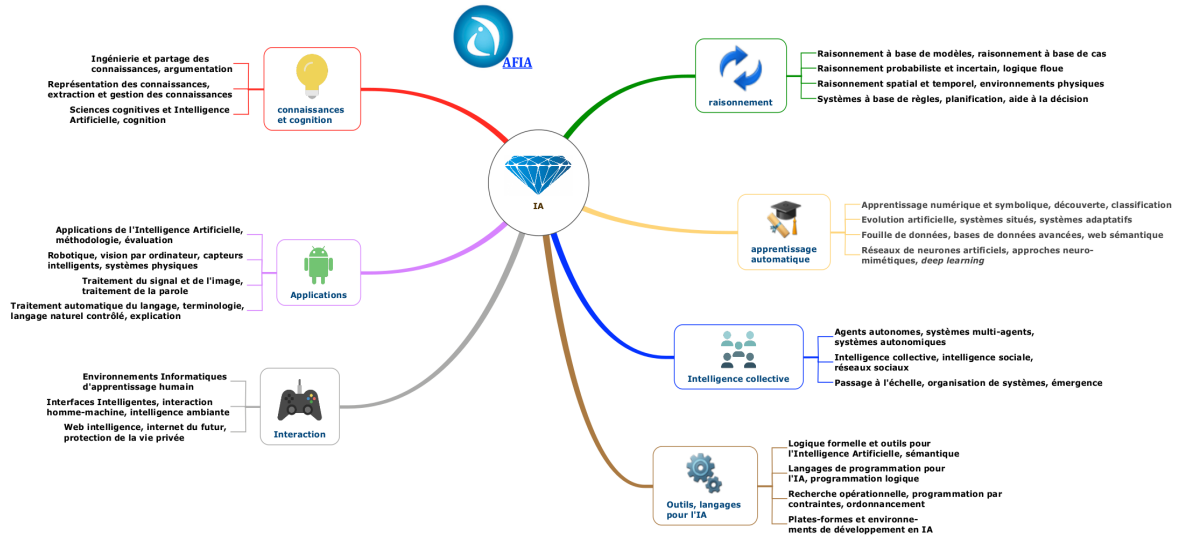


Figure 3. Domaines de l'IA (AFIA, 2023)

La figure 3 ci-dessus présente les différents domaines de l'IA en mettant en évidence leurs axes et leurs portées respectives. Notre thèse s'inscrit dans le domaine de l'apprentissage automatique, avec une attention particulière portée aux axes de classification, d'analyse de sentiment, et de traitement automatique du langage naturel.

L'apprentissage automatique, également connu sous le nom de machine learning en anglais, est une branche de l'intelligence artificielle (IA) qui se concentre sur le développement de techniques permettant aux systèmes informatiques d'apprendre et de s'améliorer de manière autonome à partir de l'expérience.

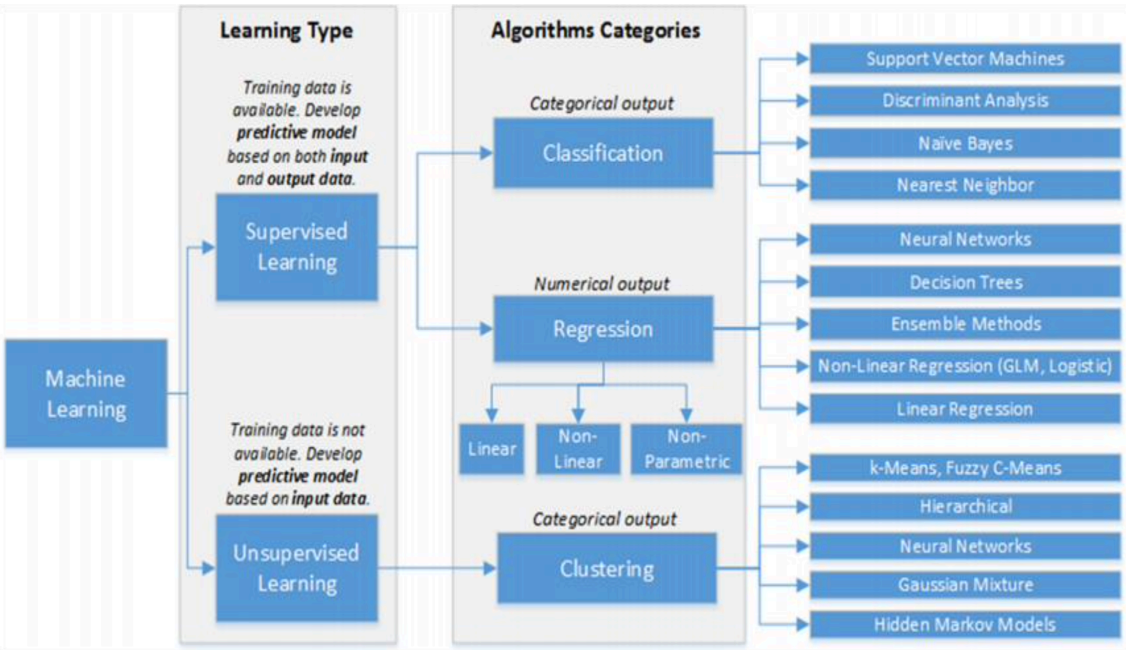


Figure 4. Algorithmes de l'apprentissage automatique (Batrinca & Treleaven, 2015)

Cet apprentissage par la machine peut se faire de 2 manières, un apprentissage supervisé, et un apprentissage non supervisé. La figure 4 illustre les algorithmes utilisés dans les deux types d'apprentissage.

- Apprentissage supervisé : consiste à entrainer un modèle sur des données étiquetées (annotées) en trouvant des liens entre les caractéristiques extraites (entrées) et les étiquettes (sorties). Le but est de créer un modèle qui pourra se généraliser sur des données non vues. Ce type d'apprentissage est beaucoup utilisé dans les problèmes de classification (prédiction de catégories ou classes) et régression (prédiction de valeurs numériques).
- Apprentissage non supervisé : consiste à créer un modèle d'apprentissage sur des données non étiquetées. Le but est de découvrir les structures internes ou d'identifier les caractéristiques des données en entrées sans connaître les sorties au préalable. Ce type d'apprentissage est beaucoup utilisé dans les problèmes de Clustering (regroupement de données similaires).

5. Accès et traitement des données dans les réseaux sociaux

Deux principaux défis se posent lorsqu'on aborde l'analyse des réseaux sociaux. Le premier réside dans la difficulté d'accéder aux données, qui sont de plus en plus de nature commerciale ou privée. Le deuxième défi concerne le manque d'outils gratuits disponibles pour l'analyse et traitement des données, ce qui nécessite souvent le développement personnalisé d'outils et de programmes pour accéder et traiter ces données.

5.1. Accès aux données

Lorsqu'on entreprend l'analyse des réseaux sociaux dans le cadre d'un domaine spécifique ou d'une recherche particulière, il est impératif de collecter une volumineuse quantité de données. En fonction de la thématique que l'on souhaite aborder, il devient essentiel de considérer diverses catégories de données, notamment les données historiques (données anciennes 'Historic Data'), les flux de données en temps réel (données en direct en live 'Stremaing Data'), les données brutes (données non traitées 'Raw Data'), et enfin les données annotées (données étiquetées 'Annotated Data').

Il existe plusieurs catégories d'accès aux données sur les réseaux sociaux à savoir :

- Données libres d'accès : dépôts de données qui peuvent être téléchargés gratuitement (eg, Kaggle Dataset[3], Hugging Face Dataset[4]). En général, ces datasets contiennent des données brutes non traitées et non annotées.
- Données propriétaires : des ensembles de données sont commercialisées et nécessitent un paiement ou un abonnement pour y accéder. (Eg, IEEE Data Port[5]).
- Données via des APIs : des plateformes comme Twitter, Facebook, Google et bien d'autres offrent des API qui permettent d'accéder aux données de leurs services, comme les tweets, les publications, les informations de profil, etc. Ces APIs peuvent être utilisées dans des logiciels (eg, DataSift, RapidMinerStudio) ou dans des lignes de codes des programmes informatiques. Bien que l'utilisation de certaines de ces APIs est gratuite, l'accès aux données, par contre, peut être limité (Exemple de l'API Twitter où la limite des tweets gratuitement accessibles est de 500 000 tweets par mois).

Comme la majorité des données présentes sur les réseaux sociaux émanent des humains et, par conséquent, ne sont pas organisées de manière structurée (c'est-à-dire qu'elles ne suivent pas de structure ou de modèle de données préétablis), il devient essentiel d'entreprendre un processus de traitement de ces données. Ce traitement vise à les transformer en données structurées afin d'améliorer la compréhension globale et faciliter ainsi leur traitement par la machine.

5.2. Prétraitement des données

Après l'accès aux données via une des techniques citées précédemment, on procède à leur prétraitement afin de les purifier, les normaliser et les préparer à un apprentissage automatique.. Pour cela, plusieurs techniques du NLP peuvent être appliquées, à savoir : Tokenisation, Lemmatisation, Suppression des mots vides, Normalisation, Remplacement des abréviations, Suppression des chiffres, etc. La figure 5 suivante illustre le processus de prétraitement de données textuelles.

[3] https://www.kaggle.com/datasets
[4] https://huggingface.co/datasets
[5] https://ieee-dataport.org/datasets

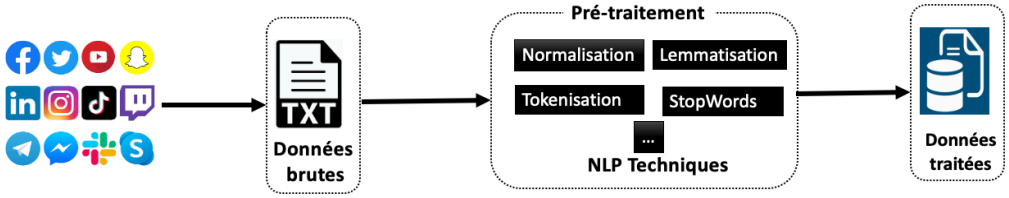


Figure 5. Prétraitement des données

Le prétraitement des données prépare les données à ce qu'elles soient mieux adaptées à leur utilisation dans des modèles d'apprentissage automatique. Pour cela, ces données doivent être transformées en vecteurs numériques compréhensibles par la machine. Il existe plusieurs méthodes et techniques de représentation et de transformation de texte, nous explorerons les plus essentielles et couramment utilisées dans la section suivante.

6. Catégories de représentation des données textuelles

Dans la littérature, Birunda & Devi (Selva Birunda & Kanniga Devi, 2021) distinguent trois catégories de représentation de textes à savoir : 1) Représentation traditionnelle, 2) Représentation statistique, 3) Représentation contextualisée. La figure 6 ci-dessous illustre ces 3 catégories avec un ensemble de techniques pour chaque catégorie.

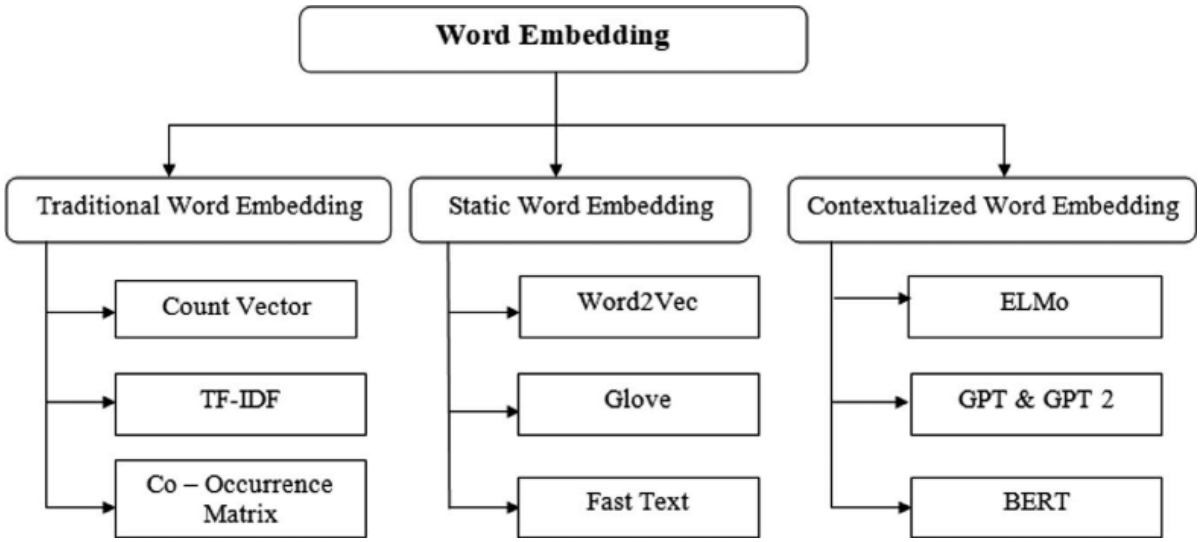


Figure 6. Catégories des techniques de représentation de texte

Nous détaillons dans ce qui suit les trois catégories, en mettant l'accent sur les différentes techniques utilisées dans chacune d'elles.

6.1. Représentation traditionnelle

Cette catégorie regroupe les techniques qui se basent sur les occurrences et les fréquences des mots dans un texte, ainsi que celles qui se basent sur la sémantique et les ontologies. Nous exposons deux techniques basées sur les occurrences et les fréquences des mots, qui sont des standards couramment utilisés dans la classification du texte qui sont : sacs de mots (Bag of Words) et TF-IDF. En outre, nous présentons trois techniques axées sur la sémantique et les ontologies, à savoir : Rada, Resnik et Jiang & Conrath.

6.1.1. Sac de mots (Bag of Words)

Cette technique appelée « sac de mots » proposée par Harris (Harris, 1954), le principe est qu'un document est représenté par l'histogramme des occurrences des mots le composant : pour un document donné, chaque mot se voit affecté le nombre de fois qu'il apparaît dans le document.

Exemple:

*Christian, you're not under attack because you're wea*k => [1, 2, 1, 1, 1, 1, 1, 0, 0]

You're under attack because you're a Threat => [2, 1, 1, 1, 1, 1, 0, 0, 0]

Taille du vecteur = taille du vocabulaire, 9 dans cet exemple (il y' a 9 mots distincts).

6.1.2. Tf-IDF

L'une des techniques les plus utilisées dans le domaine de recherche d'information (RI) pour le calcul de la similarité est la norme TF-IDF. Proposée par G. Salton et C. Buckley (Salton & Buckley, 1988) en 1988 pour accorder une pertinence lexicale à un terme au sein d'un document. Il s'agit d'appliquer une relation entre un document, et un ensemble de documents partageant des similarités en matière de mots clés. C'est une relation de quantité (tf) / qualité lexicale (idf) à travers un ensemble de documents.

Tf-IDF est couramment utilisé pour pondérer chaque mot dans le document texte en fonction de son caractère unique. En d'autres termes, l'approche TF-IDF saisit la pertinence des mots, des documents texte et des catégories particulières (Khan, et al., 2010).

TF-IDF se calcule suivant les formules suivantes :

Le score TF-IDF (appelé par convention ***w***) ***w = TF*IDF*** est donné dans la formule 1.

$$W_{x,y} = tf_{x,y} * \log\left(\frac{N}{df_x}\right) \qquad (1)$$

Où

x est un terme, y est un document.

$tf_{x,y}$ = fréquence de x dans y

df_x = nombre de documents contenant x

N = nombre total de documents

tf se calcule comme suit :

$$tf = \frac{Nombre\ d'occurences du\ terme\ analysé}{Nombre\ des\ termes\ total} \qquad (2)$$

Et ***idf*** se calcule comme suit :

$$idf = \log\left(\frac{Nombre\ total\ de\ documents}{Nombre\ de\ documents\ contenant\ le\ terme\ analysé}\right) \qquad (3)$$

6.1.3. Techniques basées sur la sémantique et les ontologies

Dans la littérature, il existe plusieurs travaux de recherche sur les mesures de similarité sémantiques qui se basent sur les ontologies pour la classification du texte, nous citons :

a. **Mesure de Rada et al** (Rada, et al., 1989) : La distance entre deux concepts est représentée par la distance minimale entre les éléments et le parent commun le plus récent (chemin le plus court dans les liens hiérarchiques).

b. **Mesure de Resnik** (Resnik, 1999) : elle se calcule par la formule suivante :

$$sim_{edge}\left(C_0^x, C_0^y\right) = \frac{2*MAX - len\ (C_0^x, C_0^y)}{2*MAX} \qquad (4)$$

où Max : représente la plus grande distance entre la racine de l'ontologie et les feuilles de l'arborescence.

Len (C_x, C_y) : la plus petite distance entre deux concepts C_x et C_y.

c. **Mesure de Jiang & Conrath** (Jiang & Conrath, 1997) **:** En considérant la hiérarchie de concepts issue de WordNet[6], elle se calcule comme suit :

$$Sim\ (X, Y) = \frac{1}{distance\ (X, Y)} \qquad (5)$$

Avec

[6] WordNet : A lexical database for English https://wordnet.princeton.edu

$$distace\ (X,Y) = E(X) + E(Y) - \Big(2.E\big(CS(X,Y)\big)\Big) \qquad (6)$$

Où

E(X) est la probabilité d'occurrence du concept X dans le corpus de référence

E(CS(X,Y)) est la probabilité d'occurrence du plus petit ancêtre commun de X et Y dans le même corpus.

D'après une étude effectuée par Varelas et al (Varelas, et al., 2005) qui ont comparé les méthodes les plus populaires de mesures de similarité sémantique sur WordNet, la mesure qui permet d'obtenir les résultats les plus proches des jugements humains est la méthode proposée par Jiang & Conrath, avec une corrélation de 83% avec les valeurs affectées par les humains.

D'après (Sarnovský & Paralic, 2008) il existe des problèmes ouverts dans l'extraction de texte en utilisant les ontologies, comme la polysémie et la synonymie. La polysémie fait référence au fait qu'un mot peut avoir plusieurs sens. Il n'est pas facile de distinguer les différentes significations d'un mot (appelé désambiguïsation du sens des mots), ce qui nécessite souvent le contexte dans lequel le mot apparaît. La synonymie signifie que différents mots peuvent avoir le même sens ou une signification similaire.

6.2. Représentation statistique

Cette catégorie regroupe les techniques qui se basent sur la prédication qui fournit des probabilités aux mots et fait correspondre chaque mot à un vecteur. Cette représentation est statique parce qu'elle ne modifie pas le contexte des mots, et que les vecteurs créés pour les mots ne changent pas d'une phrase à l'autre. Parmi les techniques les plus utilisées de cette catégorie, nous présentons : GloVe, Word2Vec et Doc2Vec.

6.2.1. GloVe (Global Vectors)

Proposée par J. Pennington et al (Pennington, et al., 2014), une implémentation du plongement lexical combinant des méthodes exploitant la cooccurrence des mots. GloVe utilise les probabilités que deux mots co-occurrent, c'est à dire la probabilité qu'un mot *k* apparaisse dans le contexte d'un mot *w*. Cette probabilité est calculée en faisant le rapport entre le nombre d'apparitions du mot *k* dans le contexte de *w* et le nombre de fois que chaque mot apparaît dans le contexte de *w*. L'information sémantique est alors obtenue en examinant le rapport entre ces probabilités. L'objectif de l'apprentissage du modèle GloVe va donc être d'apprendre des vecteurs de mots tels que leur produit scalaire est égal au logarithme de leur probabilité de cooccurrence.

6.2.2. Word2Vec

Word2vec est une approche pour représenter le sens des mots, développée par une équipe de recherche de Google (Mikolov, et al., 2013). Elle est basée sur les réseaux de neurones à deux couches et cherche à apprendre les représentations vectorielles des mots composant un texte, de telle sorte que les mots qui partagent des contextes similaires soient représentés par des vecteurs numériques proches.

Word2vec prend comme entrée un grand corpus de texte et produit un espace vectoriel, généralement de plusieurs centaines de dimensions, chaque mot unique du corpus se voyant attribuer un vecteur correspondant dans l'espace. Les vecteurs de mots sont positionnés dans l'espace vectoriel de telle sorte que les mots qui partagent des contextes communs dans le corpus sont situés à proximité les uns des autres dans l'espace.

Il existe 2 architectures ou modèles de Word2vec à savoir : Continuous Bag of Word (CBOW) et Skip-Gram (SG).

6.2.3. Doc2Vec

Développé par la même équipe (que Word2Vec) au sein de Google (Le & Mikolov, 2014). Il s'agit d'une généralisation (extension) du Word2Vec pour le traitement des phrases et paragraphes. Doc2Vec fournit des vecteurs de longueur fixe, et une représentation du sens d'une phrase ou d'un texte. Appelée aussi paragraphVector, elle a 2 modèles d'architectures: Distributed Memory (DM) and Distributed Bag of Words (DBOW). La figure 7 suivante présente l'architecture de Doc2Vec.

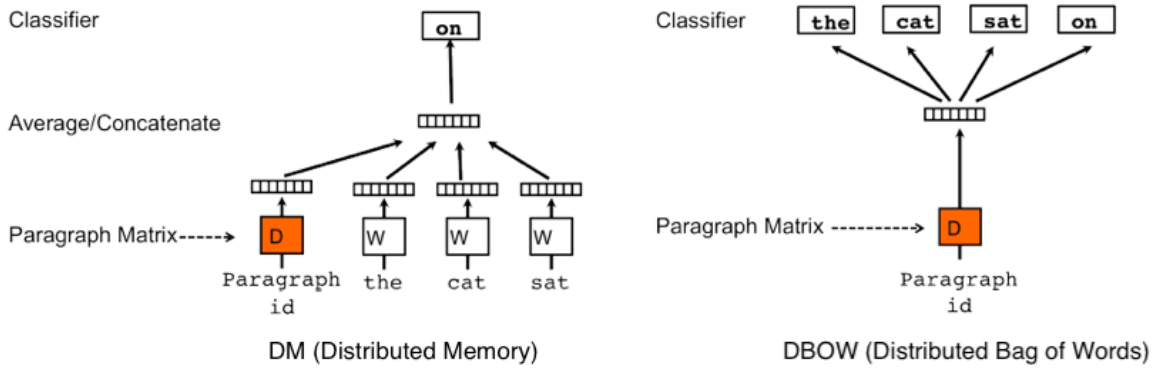


Figure 7. Architecture Doc2Vec (Le & Mikolov, 2014)

Le principe de Doc2Vec est similaire à Word2Vec, mais en plus d'avoir une liste de vecteurs de mot, le modèle va également générer une liste de vecteurs de document. Chaque vecteur de document va représenter le sens général du document auquel il est associé et sera construit

conjointement avec les vecteurs des mots composant ledit document. La tâche donnée au réseau de neurones sera alors de prédire un mot sachant son contexte et le document.

Formellement, le modèle a pour objectif, pour une séquence de mots d'entraînement $w_1, w_2, ... w_T$ de maximiser la moyenne des logs attribués aux fenêtres de mots k :

$$\frac{1}{T}\sum_{t=1}^{T}\left[\sum_{j=-k}^{k}\log P(w_{t+j}|w_t\right]_{(j\neq 0)} \qquad (7)$$

Tel qu'à chaque mot, sont associés deux vecteurs à apprendre : U_w (vecteur d'entrée : input) et V_w (vecteur de sortie : output), puis calculer ainsi la probabilité reliant chaque mot aux mots du vocabulaire par l'équation :

$$P(w_i|w_j) = \frac{\exp(U_{w_i}^T V_{w_j})}{\sum_{l=1}^{V}\exp(U_l^T V_{w_j})} \qquad (8)$$

Où V est la taille du vocabulaire.

6.3. Représentation Contextualisée

Cette catégorie est récente et elle se base sur le contexte des mots, qui changent de sens d'une phrase à une autre. Cette technique permet aux modèles de traiter le texte dans les deux sens, du début à la fin et de la fin au début. C'était un facteur clé des limitations des modèles précédents, qui ne pouvaient interpréter le texte que dans son ensemble. Parmi les techniques les plus utilisées de cette catégorie nous citons : BERT et GPT.

6.3.1. BERT

BERT (Bidirectional Encoder Representations from Transformers) est destiné à préformer des représentations bidirectionnelles profondes à partir d'un texte non étiqueté en conditionnant toutes les couches sur le contexte de gauche comme de droite (Devlin, et al., 2018).

BERT est un modèle qui a considérablement amélioré les performances du traitement automatique des langues (Madabushi, et al., 2020). Les enregistrements d'entrée sont la somme des enregistrements de jetons, des enregistrements de segments et des enregistrements de positions. Un exemple est donné dans la figure 8 suivante.

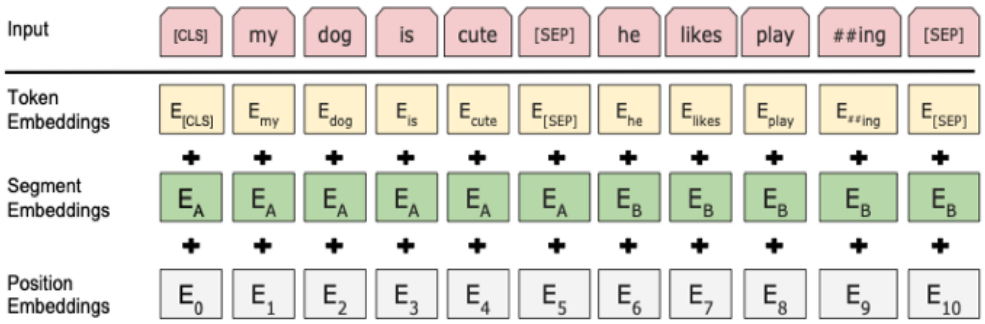


Figure 8. Exemple de représentation de texte avec BERT (Devlin, et al., 2018)

Au lieu de prédire le mot suivant dans une séquence, BERT utilise une nouvelle technique appelée MLM (Masked Language Modeling) : il masque aléatoirement des mots dans la phrase, puis tente de les prédire. Le masquage signifie que le modèle regarde dans les deux sens et utilise le contexte complet de la phrase, à gauche et à droite, pour prédire le mot masqué. Contrairement aux modèles de langage traditionnels (TF-IDF, Word2vec, Glove...), BERT prend en compte les mots précédents et suivants en même temps.

Techniquement, BERT est un modèle des Transformers qui fonctionne en effectuant un nombre restreint et constant d'étapes. À chaque étape, il applique un mécanisme d'attention pour comprendre les relations entre les mots de la phrase, indépendamment de leurs positions respectives.

BERT a un très grand nombre de paramètres (jusqu'à 110 millions de paramètres pour le modèle de base « Bert-base »). La prise en main du modèle est donc difficile : la phase de pré-entraînement, en particulier, nécessite une infrastructure importante.

L'un des principaux avantages de BERT est qu'il peut être adapté à des domaines spécifiques et entraîné sur un certain nombre de tâches différentes. Cela signifie qu'étant donné que BERT a été si bien pré-entraîné sur de gros corpus de données, il peut être appliqué à de petits ensembles de données tout en gardant de bonnes performances. BERT est gratuit et libre (open source) et il est déjà utilisé avec succès sur de petits ensembles de données dans plusieurs études de recherche, comme par exemple (Mansour, et al., 2020) (Chau, et al., 2020).

Selon Devlin & al (Devlin, et al., 2018) les principaux atouts de BERT sont :

- Non séquentiel : les phrases sont traitées dans leur ensemble plutôt que mot par mot.
- Self-Attention : il s'agit de l'unité de nouvellement introduite qui est utilisée pour calculer les scores de similarité entre les mots d'une phrase.

- Incorporations positionnelles : une autre innovation introduite pour remplacer la récurrence. L'idée est d'utiliser des poids fixes ou appris qui encodent des informations liées à une position spécifique d'un mot dans une phrase.

BERT tient compte du contexte d'un mot : en effet, les méthodes précédentes d'intégration de mots renvoient le même vecteur pour un mot, quelle que soit la façon dont il est utilisé, alors que BERT renvoie des vecteurs différents pour le même mot en fonction des mots qui l'entourent.

BERT est un modèle qui a considérablement amélioré les performances du traitement automatique du langage. BERT est une méthode récente qui est largement utilisée dans les travaux récents en analyse des sentiments et en traitement automatique du langage (Madabushi, et al., 2020).

6.3.2. GPT

GPT (Generative Pre-training Transformer) est un modèle linguistique unidirectionnel (de gauche à droite seulement) développé par OpenAI[7] et pré-trainé sur plus de 8 millions de documents web. GPT est basé sur les Transformers qui fournit un mécanisme de codeurs-décodeurs multicouches pour identifier les dépendances d'entrées et de sorties. Il est utile pour diverses tâches notamment la génération de textes, la traduction et les systèmes de questions-réponses. Son approche se compose de deux étapes : la première étape consiste à apprendre un modèle linguistique à haute capacité sur un large corpus de textes ; la deuxième étape est le réglage fin (fine-tunning) qui sert à adapter le modèle à une tâche définie avec des données étiquetées (Radford, et al., 2018). La figure 9 suivante illustre l'architecture de GPT.

[7] https://openai.com/

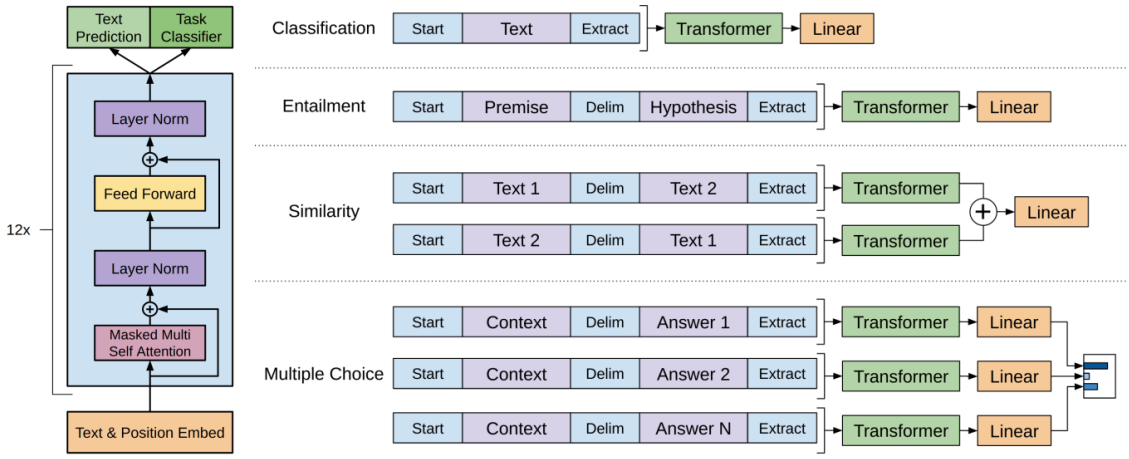


Figure 9. Architecture de GPT

A gauche de la figure, c'est l'architecture du transformer utilisé et à droite ce sont les entrées de transformations pour le réglage fin. Toutes ces entrées sont structurées en jetons (tokens) afin de les traiter par le modèle pré-entrainé.

6.4. Synthèse comparative

Dans le tableau 1 ci-après, une synthèse sous forme de comparaison des techniques étudiées pour la représentation du texte.

Tableau 1. Synthèse des techniques de représentation du texte

Technique		Points forts	Points faibles
Représentation Traditionnelle	TF-IDF (Salton & Buckley, 1988)	- Pertinence lexicale à un terme au sein d'un document - Prise en compte des cooccurrences des mots	- Non prise en compte de l'ordre des termes (mots) - Basée sur le lexique et non la syntaxe - Aucune compréhension des sens des phrases.
	BOW (Harris, 1954)	Prise en compte des occurrences des mots dans le texte	- Le modèle BOW considère uniquement si un mot connu apparaît dans un document ou non. Il ne se soucie pas du **sens**, du **contexte** et de l'**ordre** dans lequel ils apparaissent. - 2 phrases différentes peuvent avoir les mêmes représentations (vecteurs). - Taille du vecteur s'agrandit d'une façon linéaire avec la taille du vocabulaire.
	Rada et al (Rada, et al., 1989)	Simple et facile à calculer	La distance n'est pas normalisée, pouvant aller de 0 à l'infini [0, ∞[
	Resnik (Resnik, 1999)	Distance normalisée entre [0,1]	La similarité entre les termes en haut et en bas de la hiérarchie (ontologie) sont égales peu importe le sens des termes.
	Jiang & Conrath (Jiang & Conrath, 1997)	Combiner l'entropie (contenu informationnel) du concept spécifique à ceux des concepts dont on cherche la similarité	Non prise en compte du contexte

Représentation statique	GloVe (Pennington, et al., 2014)	- Combine les avantages des deux grandes familles de modèles dans la littérature : • factorisation matricielle globale • méthodes de fenêtre de contexte local - Exploite efficacement les informations statistiques en s'entraînant uniquement sur les éléments non nuls dans une matrice de cooccurrence des mots.	- Non prise en compte du contexte des documents. - Limité aux mots et non aux phrases ou paragraphes. - Il peut y avoir des ambiguïtés sur des termes pouvant être présents dans plusieurs domaines
	Word2Vec (Mikolov, et al., 2013)	- Prédiction des mots environnants de chaque mot (contexte). - Apprentissage des vecteurs à partir des données d'entrainement (Training data) - Word similarity = vector similarity	- La qualité dépend des données d'entrée, du nombre d'échantillons et de la taille des vecteurs - Il ne peut pas fournir de vecteurs de longueur fixe pour un texte de longueur variable - Limité aux mots et non aux phrases ou paragraphes.
	Doc2Vec (Le & Mikolov, 2014)	- Traitement des phrases et paragraphes. - Représentation du sens d'une phrase ou d'un texte. - Fournit des vecteurs de longueur fixe, peu importe la taille des phrases et paragraphes. - Prise en compte de l'ordre des mots et de leur contexte.	- Entraînement très long et nécessite un corpus conséquent. - Il n'est pas possible de construire un modèle pour ensuite le réutiliser sur un autre corpus, puisqu'il est nécessaire de construire les vecteurs de documents
Représentation Contextuelle	BERT (Devlin, et al., 2018)	Meilleure prise en compte du contexte d'un mot dans les deux sens de la phrase. Il retient les significations syntaxiques et sémantiques d'un texte.	Ne traite pas les longues phrases (taille limitée)
	GPT (Radford, et al., 2018)	GPT peut prédire un mot d'une phrase avec dix prédictions possibles	Nécessite des calculs trop lourds avec un risque de création de fausses informations

D'après cette étude, nous avons constaté que les techniques les plus couramment utilisées dans les travaux de recherche récents sont celles qui se basent sur le contexte, en raison de leur

contribution à la compréhension du langage naturel. En revanche, les techniques impliquant le calcul de similarités sémantiques avec des ontologies sont de moins en moins employées. Les méthodes basées sur la fréquence et l'occurrence des mots demeurent prédominantes et servent de point de comparaison aux techniques contextuelles.

7. Particularité de la langue Arabe

L'utilisation des réseaux sociaux dans le monde arabe continue de croître. Plus de 55% des utilisateurs utilisent l'arabe dans leurs publications (Salem, 2017). Des millions de publications par jour à vérifier et peu de travaux de recherche ont été menés dans ce domaine (JABER, et al., 2023). L'analyse de texte en arabe est un grand défi (Last, et al., 2006) car l'arabe est basé sur une structure grammaticale unique et très différente des autres langues (latines). En outre, la langue arabe contient des particularités qui la rendent très complexe, telles que la modulation, la forme, les pluriels, la phonétique et les dialectes.

7.1. Modulation (التشكيل)

Mots qui changent de sens en fonction de la modulation ou de la conjugaison. Des exemples sont présentés dans le tableau 2.

Tableau 2. Exemples de modulation en langue Arabe

MOTS EN ARABE	SIGNIFICATION
مَدْرَسَةٌ | مُدَرِّسَةٌ | مُدَرَسَةٌ	Ecole |Enseignante | Etudiée
أَسُودُ | أَسْوَدُ | أُسودُ	Je règne | Noir | Lions

7.2. Forme

Les lettres qui changent de forme en fonction de leur utilisation, comme (ة,ت) ,(ء,ؤ,ئ) ou (ى,ي). Des exemples sont présentés dans le tableau 3.

Tableau 3. Différentes formes de lettres et mots en langue Arabe

MOTS EN ARABE	SIGNIFICATION
راحت | راحة	REPOS | PARTIE
الثرى| الثري	RICHE| ENTEREMENT
أسماء| أسماؤهم, أسمائهم	LEURS NOMS | NOMS

7.3. Pluriels irréguliers

Les pluriels irréguliers qui ne respectent aucune règle. Des exemples sont présentés dans le tableau 4.

Tableau 4. Exemple de particularité du pluriel des mots en langue Arabe

MOTS EN ARABE (SINGULIER)	PLURIEL (EN ARABE)	SIGNIFICATION
نافذة	نوافذ	FENETRES
أسد	أسود	LION
امرأة	نساء	FEMME

7.4. Phonétique

En raison de ses possibilités phonétiques, la langue arabe dispose d'un vaste vocabulaire. Environ 17 positions distinctes dans l'alphabet arabe sont utilisées pour prononcer les lettres. Par conséquent, tout écart dans la prononciation peut entraîner un changement de sens.

Sur la base de leurs similitudes phonétiques, les lettres arabes peuvent être divisées en six groupes.

Tableau 5. Catégories des lettres Arabes selon le son phonétique

No.	Similar Groups	Similarity index
1	{ء- أ - إ – آ – ا}, {ي - ى}	1
2	{ط - ت},{ذ - ض},{ز - ظ - ذ},{س - ص},{ث - س}	0.8
3	{ن - م},{ق - ك}	0.6
4	{ج - د},{ق - ا},{ق - ج}	0.4
5	{ع - ا},{ي - ر}	0.2
6	Any other combination of Arabic letters	0

Le tableau 5 illustre les groupes de lettres arabes pour chaque catégorie sur la base de l'indice de similarité, qui va de 100 % de similarité à 0 % de similarité (Al-Sanabani & Al-Hagree, 2015)

7.5. Dialectes

Bien que l'Arabe standard (العربية الفصحى) soit la langue officielle de la plupart des pays arabes, chacun de ses pays possède une ou plusieurs langues d'usage quotidien connues comme dialectes. Les dialectes sont les différentes variantes linguistiques d'une langue par rapport à la prononciation, le vocabulaire, la grammaire et même la syntaxe.

Le sens des mots utilisés dans les différents dialectes, change d'un pays ou une région à une autre. Par exemple, dans plusieurs pays du Golfe, le mot "طرّش" signifie "envoyer un message ou un texte", au Liban, "peindre quelque chose en blanc", au Maroc, "gifler quelqu'un", et au Yémen, "vomir" (Matrane, et al., 2023).

Malheureusement, beaucoup des dialectes arabes ne sont pas suffisamment étudiés en raison de la rareté des ressources et la complexité des traitements (Fsih, et al., 2022). Parmi les travaux qui sont intéressés à l'analyse de texte dans les dialectes arabes, nous citons : des travaux sur le

dialecte algérien (Benali, et al., 2023) et (Ouchene & Bessou, 2023), des travaux sur la dialecte tunisien (Kchaou, et al., 2023) et (Haddad, et al., 2023).

8. Problème des données déséquilibrées

Sans données de qualité, même les meilleurs algorithmes d'apprentissage automatique ne seront pas très performants (Whang, et al., 2023). Lorsque l'on collecte des données en vue d'une analyse, d'une détection ou d'une classification, il est fréquent que les différentes catégories (ou classes) de données ne soient pas réparties de manière équitable ou équilibrée. On qualifie ces données de "déséquilibrées" lorsque le nombre d'observations dans chaque classe est inégalement réparti. Cela signifie qu'une classe contiendra la plupart des données, tandis que l'autre ne contiendra qu'une minorité des données. Un exemple est donné dans la figure 10 ci-dessous.

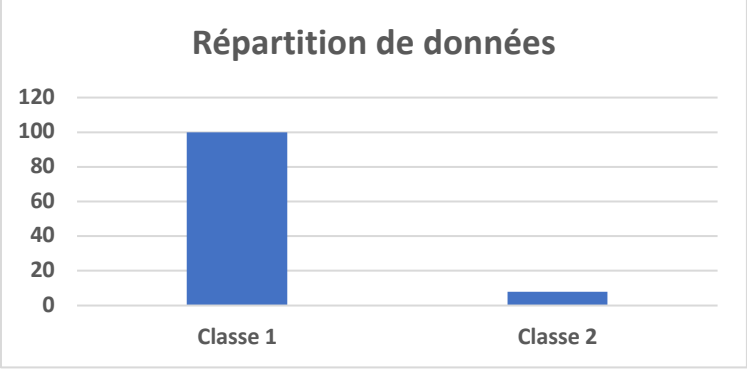


Figure 10. Exemple de données déséquilibrées

Ce déséquilibre est un problème majeur dans le domaine de l'analyse de données. L'apprentissage automatique consiste à s'entrainer sur des données d'entrainement (training data) pour construire un modèle capable de fonctionner sur des données de test (test data). Le déséquilibre de données peut affecter l'apprentissage automatique en forçant le modèle à favoriser les classes majoritaires, ce qui est très préjudiciable au processus d'apprentissage (HALIM, et al., 2023).

Plusieurs approches ont été proposées pour traiter ce problème de déséquilibre de données, ces approches sont divisées en deux catégories : ré-échantillonnage et modification au niveau de l'algorithme.

8.1. Ré échantillonnage

C'est une technique qui consiste à modifier le jeu de données déséquilibré pour le rendre équilibré. On distingue deux types du ré échantillonnage qui sont :

8.1.1. Sous échantillonnage

Elle consiste à supprimer des données de la classe majoritaire pour l'aligner à la classe minoritaire d'une manière aléatoire ou étudiée. Cette technique risque d'affecter la classification vu la perte importante de données.

Une des techniques les plus utilisées de cette catégorie est *NearMiss* qui vise à supprimer les données de la classe majoritaire qui sont les plus proches des classes minoritaires. L'objectif est d'équilibrer les proportions entre les classes tout en gardant la représentativité. Voir plus de détails sur cette technique dans cette étude (Tanimoto, et al., 2022).

8.1.2. Sur échantillonnage

Elle consiste à dupliquer les données de la classe minoritaire pour l'aligner à la classe majoritaire d'une manière aléatoire ou étudiée. Cette technique risque un sur-apprentissage (Santos, et al., 2018).

Une des techniques les plus utilisées de cette catégorie est SMOTE (Synthetic Minority Over-sampling Technique) qui vise à créer des exemples synthétiques de la classe minoritaire en sélectionnant un exemple puis en combinant les caractéristiques de ses k voisins les plus proches. Ce processus est répété pour chaque exemple de la classe minoritaire (Fernández, et al., 2018).

8.2. Modification de l'algorithme

C'est une technique qui consiste à garder le jeu de données déséquilibré sans ré-échantillonnage, et appliquer ensuite des modifications sur l'algorithme de traitement de telle sorte à prendre les données des classes minoritaires au même titre que celles des classes majoritaires.

Parmi les techniques utilisées de cette catégorie, nous citons : la pondération des classes et l'ensemble learning.

8.2.1. Pondération des classes

C'est une technique qui consiste à donner des poids différents aux observations des classes de telle sorte à ce que l'algorithme d'apprentissage accorde plus d'importance aux classes minoritaires. Voir plus de détails de cette technique dans le travail d'Al-Azani El-Alfy & (Al-Azani & El-Alfy, 2017).

8.2.2. Ensemble learning

C'est une technique qui consiste à moyenner les prédictions de plusieurs modèles. Il s'agit en fait de combiner des modèles entrainés sur différents jeux de données en gardant le même ratio entre les classes du jeu de données initial. Un exemple célèbre de cette catégorie est l'algorithme *RandomForest* qui combine plusieurs arbres de décision, et chaque arbre est créé sur un échantillon aléatoire de données avec des sous-ensembles aléatoires de caractéristiques (Gu, et al., 2022).

Il n'y a pas de technique intrinsèquement meilleure qu'une autre (Rout, et al., 2018). Le choix de la technique à utiliser dépendra des caractéristiques des jeux de données sur lesquels elle est appliquée. Il faut noter qu'il est essentiel de vérifier la possibilité d'une perte de données ou d'un sur-apprentissage. En outre, il est courant de combiner ces diverses techniques de manière hybride pour obtenir de meilleurs résultats, comme présenté dans les études de (Hasib, et al., 2022), (Peng & Park, 2022) et (Malek, et al., 2023).

9. Conclusion

Dans ce chapitre, nous avons abordé une variété de sujets liés à l'analyse des réseaux sociaux, notamment les types, les domaines d'application, ainsi que les techniques et outils disponibles. En outre, nous avons examiné les méthodes pour accéder aux données dans les réseaux sociaux, suivi d'une exploration du prétraitement des données et des diverses catégories et méthodes de représentation du texte. L'ensemble de ces approches a été synthétisé en mettant en évidence les avantages et les limitations de chacune. Nous avons également abordé la complexité du traitement de la langue Arabe, en insistant sur sa particularité. De plus, nous avons discuté de problématique liée aux données déséquilibrées, ainsi que les différentes techniques utilisées pour assurer une classification cohérente.

Étant donné que notre objectif est l'analyse des réseaux sociaux en vue de la détection des menaces terroristes, le prochain chapitre se concentrera sur une revue approfondie des travaux de recherche actuels dans ce domaine spécifique.

.

Chapitre 2. Techniques de détection des menaces terroristes sur les réseaux sociaux

Nous présentons dans ce chapitre quelques notions sur les menaces terroristes puis nous mettons l'accent sur les travaux de recherche existants sur la détection de ces menaces sur les réseaux sociaux. Nous étudions les travaux dans deux langues, à savoir l'Anglais et l'Arabe. Nous analysons divers aspects, y compris les techniques d'extraction de caractéristiques, les algorithmes de classification, les ensembles de données utilisés, les résultats obtenus ainsi que les limites et perspectives d'amélioration.

1. Introduction

Au quotidien, des préoccupations relatives à la sécurité des individus se manifestent sur les plateformes de médias sociaux. Cela inclut des questions de confidentialité violée, des problèmes de dépendance et de risque associés à l'utilisation des réseaux sociaux par des mineurs. Bien qu'il y ait des moyens pour signaler des contenus potentiellement dangereux, le retrait ou la suppression de ces contenus n'est pas immédiate et peut parfois être soumise à de longs délais de vérification. Ceci permet au contenu d'être vu et partagé, créant ainsi un vrai risque sur les personnes exposées à de tels contenus. A titre d'exemple la vidéo de l'attentat terroriste de la nouvelle Zélande en 2019 qui est restée diffusée sur Facebook en direct pendant 17 minutes[8].

Le terrorisme est l'emploi délibéré de la violence, de menaces ou d'actes criminels commis par des individus ou groupes pour des fins politiques, idéologiques ou religieux. Il s'agit de créer la peur et la terreur pour perturber l'ordre social ou public, et d'influencer les décisions des gouvernements ou des populations. L'apologie du terrorisme consiste à présenter ou à commenter favorablement des actes ou organisations terroristes. Elle est considérée comme une

[8] https://www.radiofrance.fr/franceinter/podcasts/l-edito-m/massacre-de-christchurch-facebook-s-explique-3120490

forme de soutien direct au terrorisme, car elle contribue à légitimer ou normaliser la terreur, ce qui peut entrainer des conséquences lourdes sur la sécurité publique.

Selon Y. Noguchi & E. Kholmann (Noguchi & Kholmann, 2006) 90% des activités terroristes sur internet se déroulent sur les réseaux sociaux. Parmi ces activités, nous pouvons citer : l'apologie des opérations et mouvements terroristes, la coordination à travers les réseaux sociaux pour préparer des actes terroristes, recrutement des nouveaux membres, appels aux donations et subventions, etc. Ces activités peuvent être présentées sous plusieurs formats : vidéo, image, audio et texte. Dans notre travail, nous nous sommes intéressés aux données sous format textuel. La figure 11 ci-après présente quelques exemples des menaces terroristes sur les réseaux sociaux sous formes de publications textuelles.

Figure 11. Exemples de menaces terroristes sur les réseaux sociaux

Dans la figure 11, des tweets qui sont partagés juste après des attentats terroristes. L'objectif de notre thèse est de détecter toutes ces menaces automatiquement et en temps réel, et détecter également les informations sur leurs auteurs (profils, localisation, …).

Dans les sections suivantes, nous présentons des travaux de recherche sur la détection des menaces terroristes, puis quelques travaux de l'axe de recherche de l'analyse de sentiments dans lequel s'inscrit notre thèse, et enfin des travaux sur la détection des messages haineux qui sont dans la même thématique que la nôtre.

2. Métriques de performances

Les métriques de performances utilisées dans la plupart des travaux de recherche sont : le rappel, la précision, la f-measure et l'exactitude (Accuracy en anglais). Ces métriques sont décrites comme suit :

L'accuracy est une mesure de performance qui calcule le rapport entre les observations correctement prédites et toutes les observations. Elle est calculée comme suit :

$$Accuracy = \frac{True\ Positives+True\ Negatives}{True\ Positives+False\ Positives+True\ Negatives+False\ Negatives} \quad (9)$$

La **précision** est la proportion de résultats trouvés qui sont pertinents. Elle mesure la capacité du système à rejeter les résultats non pertinents. Elle est calculée comme suit :

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Positives} \quad (10)$$

Le **rappel** est la proportion de résultats pertinents qui sont trouvés. Il mesure la capacité du système à fournir tous les résultats pertinents. Il est calculé comme suit :

$$Recall = \frac{True\ Positives}{True\ Positives+False\ Negatives} \quad (11)$$

La **F-measure** (appelée aussi f1 score) est la moyenne harmonique de la précision et du rappel. Elle mesure la capacité du système à donner tous les résultats pertinents et à refuser les autres. Elle est calculée comme suit :

$$F-measure = \frac{2*Recall*Precision}{Recall+\ Precision} \quad (12)$$

3. Travaux de recherche sur la détection des menaces terroristes sur les réseaux sociaux

Il existe de nombreux travaux de recherche sur la détection des menaces terroristes sur les réseaux sociaux. Cependant, il est extrêmement difficile de faire une comparaison directe entre ces travaux à cause de l'absence d'un jeu de données standard (Leenuse & Pankaj, 2023). C'est pour cela, que nous avons opté de présenter ces travaux séparément où nous résumons chaque travail à part. Ensuite, nous donnerons un tableau de synthèse qui compare globalement les travaux présentés par rapport aux ensembles de données utilisés, les techniques d'extraction de caractéristiques choisies, ainsi que les résultats de performances obtenus.

En général, les algorithmes de classification les plus utilisés dans l'apprentissage automatique sont : SVM, Réseaux bayésiens, K-NN, Arbres de décision, RandomForest, etc. Dans l'apprentissage profond, les méthodes les plus utilisées sont : les réseaux neuronaux profonds (DNN), les réseaux neuronaux convolutifs (CNN), les réseaux neuronaux récurrents (RNN), les réseaux neuronaux récursifs (RecNN), l'apprentissage profond hybride qui combine deux ou

plusieurs techniques d'apprentissage profond, telles que les réseaux neuronaux convolutifs (CNN) et les techniques de mémoire à long terme (LSTM), etc.

Mazari et Kheddar (Mazari & Kheddar, 2023) ont étudié la détection des menaces et messages haineux en langue Arabe (dialecte algérien). Ils ont construit et annoté un jeu de données de 14150 commentaires extraits de Facebook, Youtube et Twitter. Trois classes ont été utilisées comme étiquettes : discours haineux, langage offensant et cyberintimidation. Pour tester leur proposition, les auteurs ont appliqué des modèles de l'apprentissage automatique traditionnels (RF, NB, SVC et LR) et des modèles de l'apprentissage profond (CNN, LSTM, Bi-LSTM, GRU et Bi-GRU). Les étapes de l'approche proposée est illustrée dans la figure 12 ci-dessous.

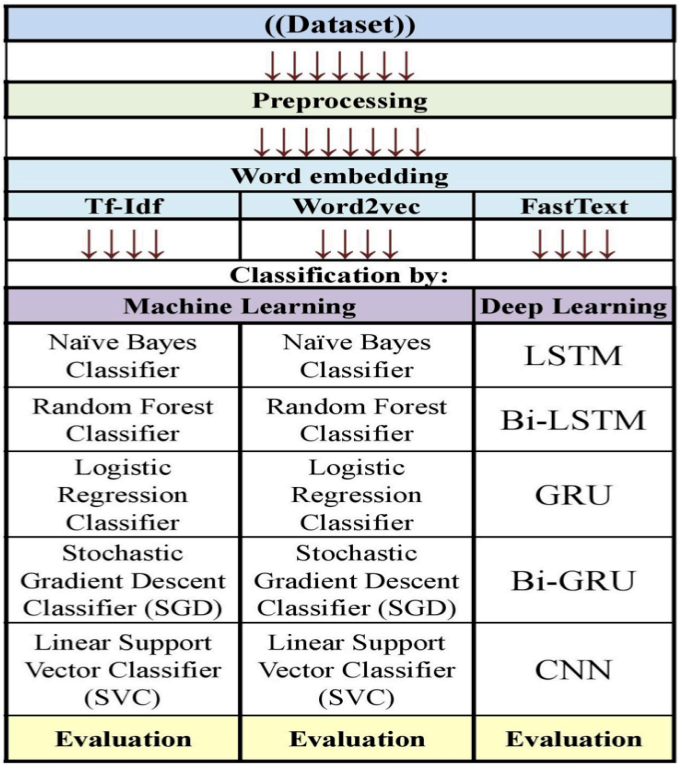


Figure 12. Étapes de l'approche proposée (Mazari & Kheddar, 2023)

Comme nous pouvons le voir sur la figure 12, les auteurs appliquent un prétraitement sur les données, ensuite ils utilisent 3 techniques différentes de représentation de texte : TF-IDF, Word2Vec et FastText, et enfin ils appliquent des algorithmes de l'apprentissage machine et l'apprentissage profond pour la classification. Les résultats des expérimentations ont montré que Bi-GRU obtient la meilleure performance en termes de précision avec un taux de 73,6% et un score f1 de 75,8%.

Aldera et al (a-Aldera, et al., 2021) ont présenté une étude sur la détection de l'extrémisme en ligne dans les contenus textuels. Ils ont étudié 45 travaux de recherche publiés entre 2015 et 2020. Ils ont constaté que la majorité des ensembles de données utilisés sont en anglais (86 %), tandis que 7 % utilisent l'arabe. Les travaux étudiés ont utilisé des approches d'apprentissage automatique avec des techniques d'extraction de caractéristiques (la plus courante étant Tf-IDF), ainsi que des approches d'apprentissage profond (RNN et CNN). L'une des limites relevées par les auteurs est l'absence d'un ensemble de données de référence fiable, la plupart des ensembles de données utilisés étant inaccessibles et privés. Ils ont également constaté qu'il n'y avait pas de comparaison entre les différents travaux pour évaluer chaque approche par rapport aux autres.

O-Theodosiadou et al (Theodosiadou, et al., 2021) ont proposé un framework pour la détection de points de changement significatifs dans les séries temporelles liées au terrorisme qui pourraient indiquer l'occurrence d'événements importants. L'approche proposée consiste à classer les données textuelles en ligne comme étant liées au terrorisme et aux discours de haine, ainsi qu'à analyser les points de changement dans les séries temporelles générées par ces données. Les auteurs ont constaté que leur framework pouvait permettre de détecter des liens entre les incidents terroristes et l'activité en ligne. Ils notent qu'il est difficile de trouver des ensembles de données annotées, en particulier lorsqu'il s'agit de se concentrer sur l'environnement terroriste. L'approche proposée ne peut être comparée à divers scénarios en raison du manque d'ensembles de données adéquats dans le domaine. En outre, la généralisation des résultats est compromise par l'accent mis sur le contenu anglais.

AlGhamedi & Khan (AlGhamdi & Khan, 2020) ont proposé un système de détection des messages arabes suspects sur Twitter. Ils ont utilisé des techniques NLP (tokenize, stemming, lemmatization) pour le prétraitement des données. Ils ont annoté manuellement leurs données en deux classes (Label 0 pour non suspecte et Label 1 pour suspecte). Ils ont testé leur système avec 6 algorithmes de machine learning et ont obtenu une précision de 86% en utilisant SVM. Les résultats présentés dans ce travail sont prometteurs, mais le nombre de tweets sur lesquels les expériences ont été menées est limité (1555 tweets).

G, K. Pitsilis et al (Pitsilis, et al., 2018) ont proposé un système de détection du contenu haineux dans les réseaux sociaux. Cette proposition repose sur l'utilisation des classifieurs de réseaux

de neurones récurrents auxquels s'ajoutent des informations relatives à l'utilisateur (tendance à la violence, au sexisme et à l'historique de publications anciennes). Les auteurs ont expérimenté 16 000 tweets avec une ventilation sur trois classes : racisme, sexisme et neutre. Ils ont réussi à atteindre une précision de 93,2%.

S. Malmasi et M. Zampieri (Malmasi & Zampieri, 2017) ont étudié les méthodes de détection du discours de haine dans les réseaux sociaux. Ils ont utilisé des données Twitter (Tweets) écrites en anglais, annotées avec trois libellés et divisées en trois classes : Haine, Offensant et Ok (non offensant). Les auteurs ont utilisé le classifieur SVM linéaire avec trois caractéristiques : les n-grammes de caractères, les n-grammes de mots et les sauts de mots. Le résultat le plus performant lors des expériences menées sur 14 509 tweets a été atteint en utilisant un modèle de caractères avec des 4-grammes, obtenant ainsi une précision de 78%.

A. Johnston et G. Weiss (Johnston & Weiss, 2017) ont proposé une approche d'identification automatique des pages web et du contenu des médias sociaux liés au contenu extrémiste. Cette approche utilise l'apprentissage machine pour classer le texte en deux classes : extrémiste et bénigne (non extrémiste). Les auteurs ont utilisé la technique doc2vec pour traiter des données (13500 fichiers) dont 65% sont en langue arabe, 20% en anglais, et le reste dans d'autres langues. Ils ont utilisé les réseaux de neurones pour la classification, et les résultats obtenus ont atteint un f1 score de 93%.

B. Iskander (Iskandar, 2017) a proposé un processus de détection des menaces terroristes basé sur l'apprentissage machine. Après la collecte des données et le pré-traitement, il procède au calcul du score (mapping) à l'aide de SentiWordNet[9]. Pour la classification des tweets, l'auteur a utilisé les réseaux bayésiens (naïve Bayes) pour classer l'énoncé comme positif, négatif ou neutre. L'auteur propose l'ajout de l'analyse du comportement des utilisateurs afin de développer une nouvelle classification basée sur l'historique des publications de l'utilisateur afin de déterminer s'il s'agit de menaces réelles ou non.

J. Klausen (Klausen, 2015) a recueillie des informations sur une période de trois mois, à partir des comptes Twitter de 59 combattants d'origine occidentale connus pour être en Syrie. À l'aide

[9] SentiWordNet est une ressource lexicale pour l'exploration d'opinions.

d'une méthode de boule de neige, les 59 comptes de démarrage ont été utilisés pour collecter des données sur les comptes les plus populaires du réseau Twitter. L'auteur met en évidence le rôle joué par le réseau Twitter dans la propagande terroriste, et note qu'il y a même un contrôle joué par des comptes d'alimentation appartenant à des organisations terroristes dans les zones d'insurrection. Il constate également que 85% de tous les flux des 59 comptes utilisent leurs langue maternelle (arabe, et mélange d'arabe et anglais).

P. Burnap et al (Burnap, et al., 2014) ont suivi des publications (Tweets) après l'acte terroriste du Woolwich. Ils se sont intéressés à la taille des publications, nombre de partages (retweet), et à la durée de survie (entre le 1er tweet et le dernier). Ils ont étudié également l'influence des journaux quotidiens qui couvrent un événement, la co-occurrence des contenus (hashtags et liens URLs) et enfin une étude concentrée sur le tweet même (valeur sociale, temporelle et contenu du tweet).

Les auteurs ont identifié trois facteurs important dans la prédiction de diffusion d'un tweet qui sont :

- **Profil de l'auteur** : nombre d'abonnés, nombre de tweets.
- **Contenu** : URLs et Hashtags.
- **Retweets et folowers** : séparer ceux qui sont retweetés plusieurs fois et ceux qui tweetent beaucoup.

M. Cheong et V.C. Lee (Cheong & Lee, 2011) ont proposé un framework permettant de recueillir le sentiment et la réponse des civils sur Twitter lors des attentats terroristes. Leur objectif était de suivre et de visualiser la réaction de la population civile à la suite d'activités terroristes. Le framework proposé consiste en 4 phases :

- Dernières nouvelles (Breaking News)
- Collecte de données et filtre anti-spam
- Détection des sentiments et démographie
- Exploration de données et établissement de rapports

Les limites de ce travail sont la nécessité d'observations par l'humain, l'utilisation de l'Anglais seulement comme langue, et la difficulté de tester les deux premières phases dans le monde réel à cause des défis de prévention des actes terroristes.

O. Oh et al (Oh, et al., 2010) définissent le réseau social Twitter comme un outil très dangereux surtout quand il est utilisé avec GPS, des logiciels de changement de voix, des RS mobiles, etc. Ils avancent que Twitter est un système d'aide à la prise de décision des terroristes en temps

réel. Les auteurs ont étudié l'impact des publications sur Twitter pendant l'attaque terroriste de Mumbai, ils ont analysé le contenu des tweets et ont montré l'utilisation par les terroristes des informations publiées en temps réel par des utilisateurs qui étaient aux alentours de l'attaque (prise d'otages dans 2 hôtels de luxe). Il s'avère que les terroristes utilisaient les informations publiées sur les réseaux sociaux, notamment Twitter, pour identifier leurs otages (Origine, Appartenance politique, …) ce qui a permis aux terroristes de connaître l'identité de tous les otages et ainsi de changé de stratégies selon les informations collectées.

M. Last & al (Last, et al., 2006) ont proposé un processus de classification multilingue pour détecter le contenu terroriste des pages web. Ils se sont concentrés sur les titres des pages, le contenu textuel et les liens hypertextes. Ils ont utilisé la méthode de représentation théorique des graphes de documents web. Ils ont expérimenté leur processus sur 648 pages web en arabe (dont 200 pages liées au terrorisme et 448 pages non liées au terrorisme) en utilisant le classifieur d'arbre de décision C4.5.

4. Travaux de recherche sur l'analyse de sentiment sur les réseaux sociaux

L'analyse de sentiment est l'une des applications importantes de l'apprentissage automatique dans le domaine du traitement du langage naturel (NLP). Elle a pour objectif de déterminer si un texte exprime un sentiment positif, négatif ou neutre. Pour l'appliquer d'une manière efficace, on utilise souvent des techniques de l'apprentissage machine (extraction de caractéristiques, entrainement du modèle, classification, évaluation et amélioration). Différents algorithmes d'apprentissage machines sont utilisés dans l'analyse des sentiments, on peut les diviser en deux catégories : Apprentissage automatique (Machine learning) et Apprentissage Profond (Deep learning). L'apprentissage machine nécessite une extraction de caractéristiques des données en amont pour que le modèle puisse fonctionner efficacement. Tandis que l'apprentissage profond peut apprendre à extraire les caractéristiques des données d'une manière automatique, grâce aux couches de neurones intermédiaires. Un exemple de la différence entre l'utilisation de l'apprentissage machine et de l'apprentissage profond dans l'analyse de sentiments, est donné dans la figure 13 ci-après.

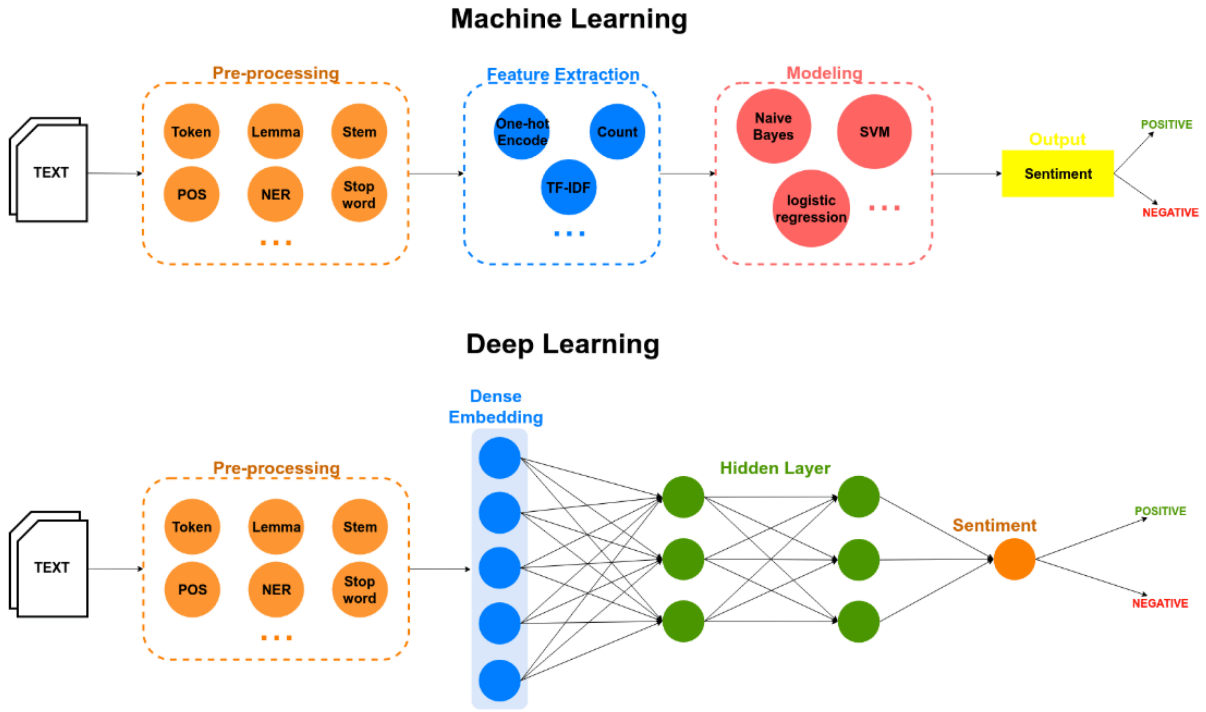


Figure 13. Apprentissage automatique et apprentissage profond dans l'analyse de sentiments (Dang, et al., 2020)

Plusieurs études et recherches sur la classification et l'analyse des sentiments dans les réseaux sociaux ont été effectuées à l'aide des différentes techniques de représentation de texte présentées dans le chapitre 1. L'étude de (Rodríguez-Ibánez, et al., 2023) souligne que les technologies traditionnelles (lexiques, jetons, sacs de mots) sont encore largement utilisées dans les travaux de recherche sur l'analyse de sentiment. Tant dis que, les technologies plus récentes (basées sur les Transformers) comme BERT ou GPT ne sont pas encore très répandues mais elles émergent de plus en plus.

Les auteurs (Dang, et al., 2020) ont réalisé une étude comparative des performances des modèles d'apprentissage profond les plus populaires (CNN, RNN, DNN) sur 8 ensembles de données. Deux techniques de traitement de texte (Word embedding et TF-IDF) ont été utilisées pour la représentation des données textuelles d'entrées (d'initialisation) au réseau de neurones. Ils ont utilisé la validation croisée k-fold avec k=10, et comme métriques : la précision, le rappel, le score F et l'aire sous la courbe (AUC). Les résultats obtenus montrent que le word embedding est meilleur que le TF-IDF pour les trois modèles d'apprentissage profond. En particulier pour RNN qui est la méthode fournissant les meilleurs résultats.

A, M. Founta & al (Founta, et al., 2019) ont proposé une architecture d'apprentissage profond, qui utilise une grande variété de métadonnées et les combine avec des modèles cachés extraits automatiquement du texte des tweets, pour détecter plusieurs comportements abusifs comme les discours de haines, le sexisme, le racisme, l'intimidation et le sarcasme. Ils ont testé leur approche sur Twitter et ont obtenu des résultats prometteurs avec une valeur AUC comprise entre 92 % et 98 %.

Zhang & Luo (Zhang & Luo, 2019) ont proposé de nouvelles structures de réseaux neuronaux profonds en tant qu'extracteurs de caractéristiques efficaces pour la détection des discours haineux dans les ensembles de données Twitter à longue traîne. Ils ont utilisé un modèle de classification basé sur un réseau neuronal profond (DNN) avec des enchâssements de mots en NLP. Les résultats obtenus dépassent de 4 à 16% les performances rapportées dans la littérature en ce qui concerne le score F.

Soliman et al (Soliman, et al., 2017) ont proposé six modèles d'intégration de mots pour la représentation de textes arabes. Ils ont collecté des données à partir de Twitter, Word Wilde Web et Wikipedia. Les auteurs ont utilisé des millions de jetons (tokens) pour construire les modèles avec Word2Vec comme technique de représentation de texte. Pour chaque collection, deux modèles sont créés, l'un avec les techniques CBOW et l'autre avec les techniques Skip-Gram. Les modèles générés ont été évalués selon deux modes : quantitatif et qualitatif. Pour le mode qualitatif, ils ont mesuré la similarité entre les mots, ainsi que sur les Entités Nommées (Named Entity). Pour le mode quantitatif, ils ont utilisé SemEval[10] pour calculer le degré d'équivalence entre deux textes. La performance optimale est atteinte avec l'utilisation de SkipGram, affichant une précision de 0,58.

Les auteurs P. Bourgonje & al. (Bourgonje, et al., 2017) ont évalué un ensemble d'algorithmes de classification pour deux types de contenus en ligne créés par les utilisateurs (Wikipedia Talk, Tweets et Comments) dans deux langues (anglais et allemand). Ils ont regroupé les données en 4 classes : racisme, sexisme, haine, harcèlement et attaques contre des personnes. Ils ont obtenu un score f allant jusqu'à 81,58 % pour les données en anglais en utilisant les réseaux bayésiens.

[10] https://alt.qcri.org/semeval2017/task1/

Mohamed Al-Smadi & al (Al-Smadi, et al., 2017) ont travaillé sur l'identification des paraphrases (si deux textes ont le même sens) et l'analyse de la similarité sémantique entre deux textes (avec un degré de similarité entre 0 et 5). Ils ont collecté les données sur Twitter en suivant les Breaknig News de deux chaînes d'information, Al-Arabiya et Aljazeera. Ils ont collecté puis sélectionné un total de 2 493 tweets, puis ont procédé à l'extraction de caractéristiques lexicales, syntaxiques et sémantiques. À l'aide de ces caractéristiques, des classifieurs à entropie maximale (MaxEnt) et à régression vectorielle de support (SVR) ont été formés et évalués. L'approche proposée a obtenu score f de 0,872 pour l'identification des paraphrases et de 0,912 pour l'analyse de la similarité sémantique.

5. Travaux de recherche sur la détection des messages haineux (hate speech) sur les réseaux sociaux

Le discours de haine consiste à utiliser des expressions ou des phrases violentes, offensantes ou insultantes à l'égard d'une ou plusieurs personnes. De nombreuses recherches se sont penchées sur la détection et la classification du discours haineux (hate speech) sur les médias sociaux. Nous exposons ici un échantillon de ces travaux.

H. Saleh et al (Saleh, et al., 2023) ont étudié la détection des messages haineux en utilisant BERT et BiLSTM. Deux approches sont utilisées dans cette étude pour trouver la meilleure performance de classification comme présentée dans la figure 14 ci-dessous.

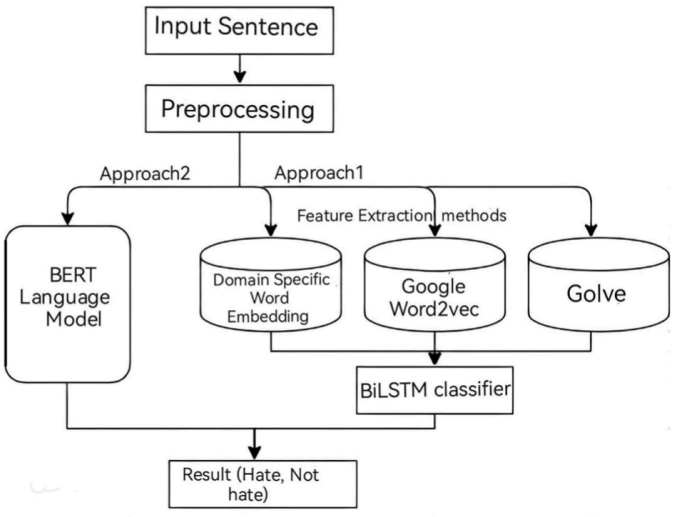


Figure 14. Méthodologie de détection des messages haineux (Saleh, et al., 2023)

Dans cette étude, la détection est faite à base d'une classification binaire (haineux ou non haineux). Les résultats obtenus sont très satisfaisants avec une F1-score de 93% avec BiLSTM

et de 96% avec BERT sur un ensemble de données équilibré combiné à partir de jeux de données existants sur les messages haineux. Les auteurs ont conclu qu'il est très utile de construire de grands modèles pré-entraînés à partir de données spécifiques à des domaines riches dans les plateformes de médias sociaux actuelles.

Huang et al (Huang, et al., 2023) ont étudié et comparé l'annotation faite par ChatGPT avec celle faite par des êtres humains sur des messages haineux implicites. Sur 795 instances de l'ensemble de données, ils ont constaté que ChatGPT a bien annoté 636 comme messages haineux, c'est-à-dire près de 80% d'exactitude. Les résultats démontrent le grand potentiel de ChatGPT comme outil d'annotation de données.

Imane Guellil & al (Guellil, et al., 2020) ont proposé une approche pour détecter les discours de haine contre les politiciens de la communauté arabe sur les médias sociaux. Les auteurs utilisent Word2vec et FastText pour extraire des caractéristiques. Ils ont utilisé des algorithmes classiques (Gaussian Naive Bayes, Logistic Regression, Random Forest, SGD, Linear SVC) et des algorithmes d'apprentissage profond (CNN, Multi-Layer Perceptron[11] (MLP), LSTM, BiLSTM). Les résultats montrent que les algorithmes Linear SVC, BiLSTM et MLP sont les plus performants, avec une précision allant jusqu'à 91 % avec le modèle Skip Gram.

Z. Waseem et al (Waseem, et al., 2017) ont proposé une typologie du langage abusif (Explicite-Implicite, Direct-Généralisé). Cette typologie capture les similitudes et les différences centrales entre les sous-tâches(subtask) et leurs implications pour l'annotation des données et la l'extraction des caractéristiques (Features).

J.H. Park & P. Fung (Park & Fung, 2017) ont exploré une approche en deux étapes pour effectuer une classification sur un langage abusif, puis une classification multi-classes pour détecter les messages sexistes et racistes. En utilisant un corpus public de 20 000 tweets en anglais portant sur le sexisme et le racisme, ils ont obtenu une performance de 0,827 en termes de score F avec HybridCNN, et 0,824 en termes de score F avec la régression logistique.

N. Vishwamitra et al (Vishwamitra, et al., 2017) ont conçu un système (framework) de défense contre le harcèlement mobile innovant appelé MCDefender qui peut détecter et prévenir

[11]MLP : un type de réseau de neurones artificiels composé de plusieurs couches de neurones, avec des connexions entre chaque neurone d'une couche et tous les neurones de la couche suivante.

efficacement la cyberintimidation dans les réseaux sociaux en ligne (OSN) mobiles. Leur système se déclenche avant qu'un message de cyberintimidation ne soit envoyé via un appareil mobile et des attaques de cyberintimidation cachées peuvent également être détectées via une approche plus fine et sensible au contexte. Ils ont utilisé CNN basée sur la prononciation (PCNN) qui est une technique d'apprentissage approfondi (deep learning) pour développer un classifieur de cyberintimidation afin d'améliorer la précision de la détection de cyberintimidation textuelle dans les données bruyantes. Ils ont atteint une exactitude de 98,9%.

6. Synthèse des travaux étudiés

Nous présentons dans le tableau 6 suivant une synthèse de certains travaux étudiés, offrant une vue d'ensemble des techniques employées pour l'extraction de caractéristiques (Features selection), les algorithmes utilisés pour la classification (classifieurs), des ensembles de jeux de données utilisés (DataSet) et enfin les résultats de performances obtenues.

Tableau 6. Synthèse des travaux de recherche

Travaux	Features Selection	Classifieur	DataSet	Résultats (Performances)
(Al-Smadi, et al., 2017)	Lexical, syntactic et semantic features	Maximum Entropy (MaxEnt) et SVR	2493 tweets	F-measure de 0.872 pour paraphrase identification, et 0.912 pour l'analyse de similarité sémantique
(Bourgonje, et al., 2017)	Bag of Words (BOW) feature set (word unigrams)	Bayes, C4.5 DT, Multivariate LR, Maximum Entropy et Winnow2	15,979 tweets anglais, 469 tweets allemands, 11,304 commentaires (Wikipedia Talk)	F-measure de 81.58%
(Pitsilis, (2018, Jan 13).)	Word-based frequency	LSTM	16,000 tweets over three classes: racism, sexism et neutralism	F-measure de 0.93
(Mulki, et al., 2019)	Unigrams, uni- grams+ bigrams et unigrams+ bigrams+ trigrams	NB, SVM	L-HSAB dataset (Twitter) annoté comme Abusif, haine ou normal	NB, accuracy 90.3 pour classification binaire et 88.4 pour 3 classes
(Founta, et al., 2019)	Word Embedding layer	Deep learning (RNN)	Twitter avec 4 classes: haineux (24783), offensif (16059), harcèlement (6091), sarcasme (61075)	AUC allant de 92% à 98%
(Zhang & Luo, 2019)	Word embedding, Word2Vec, Glove	Deep Neural Network (DNN), CNN, GRU	7 jeux de données utilisés dans la détection des messages haineux	F-measure entre 0.83 et 0.94 selon le jeu de données utilisé
(Alshalan & Al-Khalifa, 2020)	Char, n-grams, Bert	SVM, LR, CNN, GRU, BERT	Jeu de données Arabe qui contient 9316 tweets annotés	CNN, F-measure de 0,79 et AUC de 0,89
(b-Aldera, et al., 2021)	Tf-IDF, BERT	SVM, LR, NB, RF, BERT	89,816 Tweets arabes annotés comme extremiste ou non-extremiste	SVM avec TF-IDF, accuracy (0.9729), et avec BERT accuracy de (0.9749)

Comme illustré dans ce tableau, les ensembles de données utilisés diffèrent d'une étude à l'autre en termes de taille, de nombre de classes, de réseau social et de langue des publications. De même, les approches de représentation du texte varient considérablement d'une étude à l'autre : certaines privilégient les techniques traditionnelles tandis que d'autres mettent en avant les méthodes contextuelles. Les algorithmes de classification présentent également une grande diversité, certains recourant à des méthodes d'apprentissage automatique, d'autres à des techniques d'apprentissage profond, voire les deux simultanément.

7. Conclusion

Dans le présent chapitre, nous avons exposé les travaux de recherche existants dans la littérature, portant sur la détection des menaces terroristes sur les réseaux sociaux, l'analyse de sentiments, ainsi que la détection des messages haineux. Nous avons mis l'accent sur les méthodes employées, les résultats obtenus ainsi que les limites et défis qui ont orienté nos propositions de solutions qui seront détaillées dans la partie 2 de cette thèse. Le chapitre suivant se penchera en détail sur le processus de collection de données entrepris pour la création de notre ensemble de données (dataset).

Partie 2. Contributions

Chapitre 3. Construction, normalisation, annotation et prétraitement de données

Nous traitons dans cette thèse la problématique de présence des publications sur des réseaux sociaux pouvant être des menaces terroristes comme les appels à meurtre, les recrutements de nouveaux membres, la propagande, les apologies des actes ou mouvements terroristes... Notre objectif est de proposer une solution capable de détecter ces menaces automatiquement par la machine. Cette approche vise à restreindre la diffusion des contenus dangereux sur les réseaux sociaux et à aider les services de sécurité dans l'identification des auteurs.

Après avoir exploré les concepts liés à l'analyse des réseaux sociaux, et étudié les travaux de recherche portant sur les menaces terroristes, l'analyse de sentiment, ainsi que la détection de l'extrémisme et des discours haineux, nous présentons dans les chapitres suivants nos 3 contributions de de recherche. Dans le présent chapitre, nous commençons par la description des étapes communes à nos 3 contributions, à savoir la collecte, la normalisation, l'annotation et le prétraitement de données. Par la suite, chaque contribution sera détaillée dans le chapitre suivant.

Nous

1. Introduction

Avant de procéder à l'apprentissage machine en vue d'une analyse ou d'une classification, il est nécessaire de suivre plusieurs étapes préliminaires. La première étape consiste à recueillir des données à partir des réseaux sociaux. Comme nous l'avons présenté dans le chapitre précédent, il n'y a pas de jeu de données standard disponible pour notre étude. Bien qu'il existe déjà des ensembles de données collectés comme celui de Fifth Tribe qui contient plus de 17.000 tweets provenant des partisans de l'État islamique (ISIS) après l'attaque terroriste de novembre 2015 à Paris (Fifth Tribe, 2018), ou celui de S. Aldera qui contient plus 89.000 tweets annotés comme extrémiste ou non (b-Aldera, et al., 2021). Cependant, ces ensembles de données présentent des limitations telles que des annotations incomplètes ou une diversité insuffisante dans le premier cas, ou une indisponibilité en raison de leur caractère privé dans le second. C'est pourquoi nous avons opté à la construction d'un nouveau jeu de données adéquat à notre étude. Une fois que

les données sont collectées, il convient de les normaliser, puis de les annoter manuellement. Les données annotées sont ensuite soumises à un prétraitement afin de les préparer à la dernière phase, qui consiste à les entraîner à l'aide de l'apprentissage machine. La figure 15 suivante illustre l'enchainement de ces étapes.

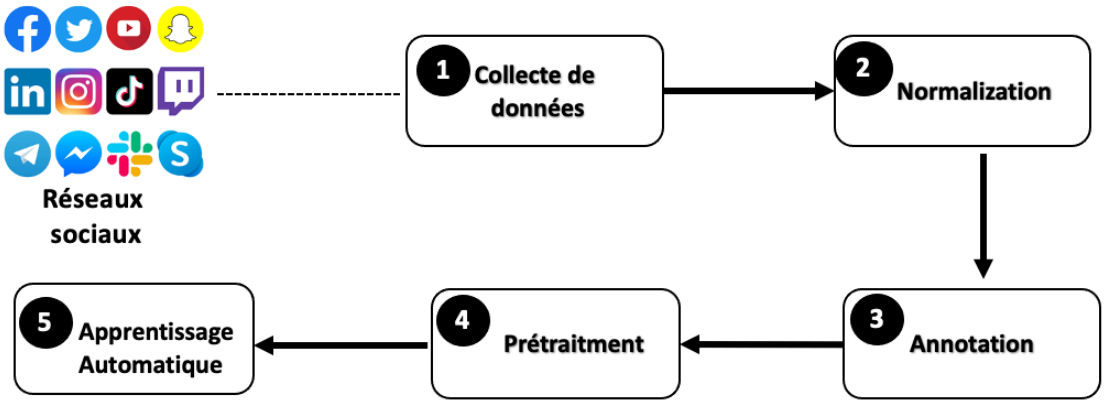


*Figure 15. Étapes de collecte et préparation de données*Dans ce qui suit, nous détaillons chaque étape en fournissant des explications détaillées sur les outils et techniques que nous avons employées dans notre travail.

2. Construction de l'ensemble de données

Comme nous l'avons expliqué dans le chapitre 1 section 5.1, l'accès aux données des réseaux sociaux peut se faire de 3 façons différentes : publiques, privées ou construction de nouveaux jeux de données. Dans notre travail, nous avons opté à la dernière catégorie, c'est à dire la collecte de données via des APIs pour créer notre propre data set. Cette décision est justifiée par le fait que l'utilisation des données privées entrainent des coûts, et que nous n'avons pas trouvé de données publiques adaptées à notre travail. Ainsi, pour la collecte de données et la création de notre data set, nous avons choisi d'utiliser le réseau social *Twitter* pour sa disponibilité, sa facilité d'utilisation et son accès gratuit aux tweets (jusqu'à 500k tweets librement accessibles par mois). Nous avons collecté les tweets à l'aide de l'outil RapidMiner Studio développé par Mierswa et Klinkenberg (Mierswa & Klinkenberg, 2018) que nous avons relié à un compte Twitter que nous avons déjà créé (Voir la figure 16).

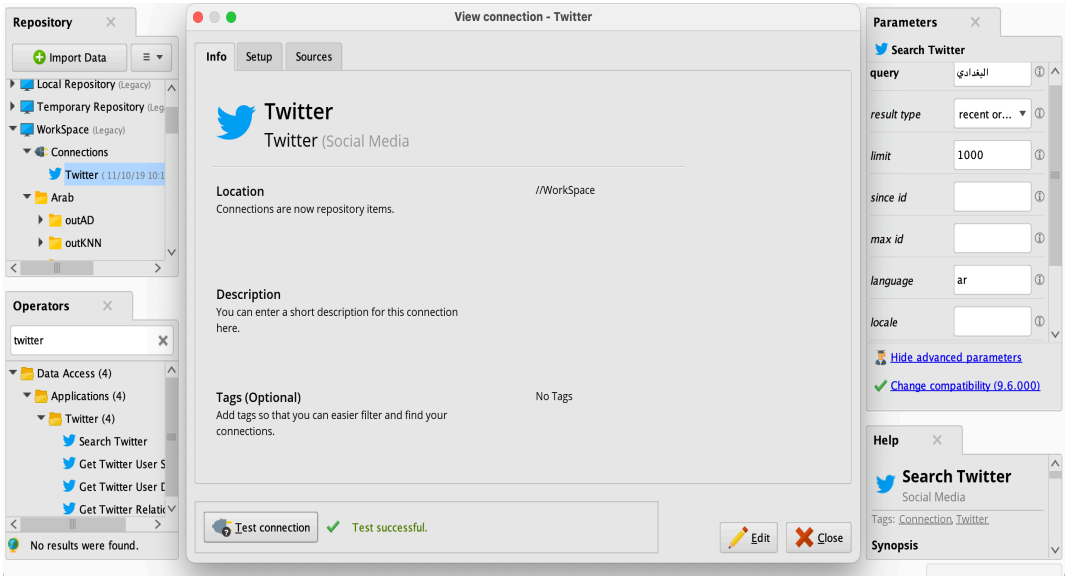


Figure 16. API Twitter sous Rapid Miner Studio

Étant donné que notre étude se concentre sur les menaces terroristes, nous avons décidé de construire un ensemble de données comprenant des tweets abordant ce sujet spécifique. Ainsi, nous avons utilisé des requêtes par mots-clés et des hashtags pour récupérer les tweets liés au terrorisme. Les paramètres que nous avons utilisés dans nos recherches sont :

- *Requête* : mettre un mot-clé à la fois, exemple : البغدادي

- *Limite* : 1000, pour collecter un maximum de 1000 tweets pour chaque mot clé. Nous avons mis ce plafond à 1000 afin de diversifier le jeu de données en incluant de nombreux tweets à l'aide de divers mots-clés.

- *Langue* : 'ar' pour l'arabe, 'en' pour l'anglais.

La figure 17 illustre un exemple d'extraction des tweets sur RapidMiner Studio avec les paramètres renseignés sur la droite de la figure.

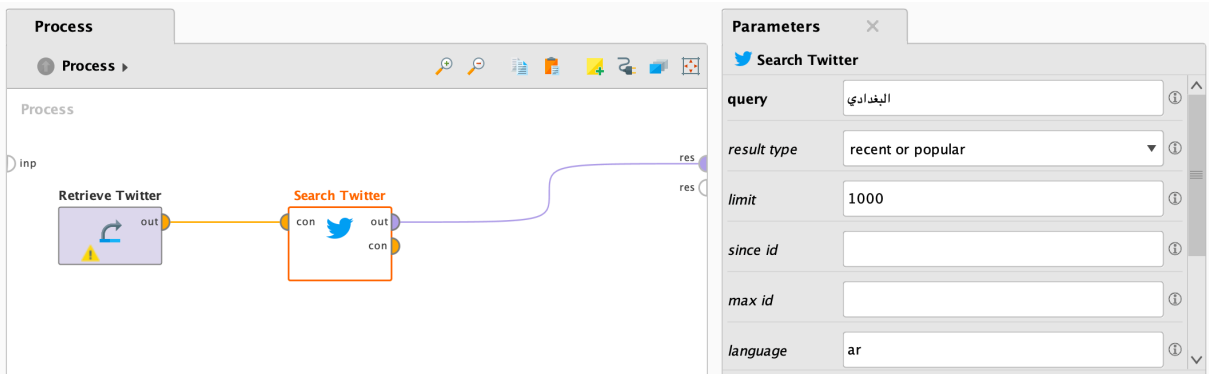
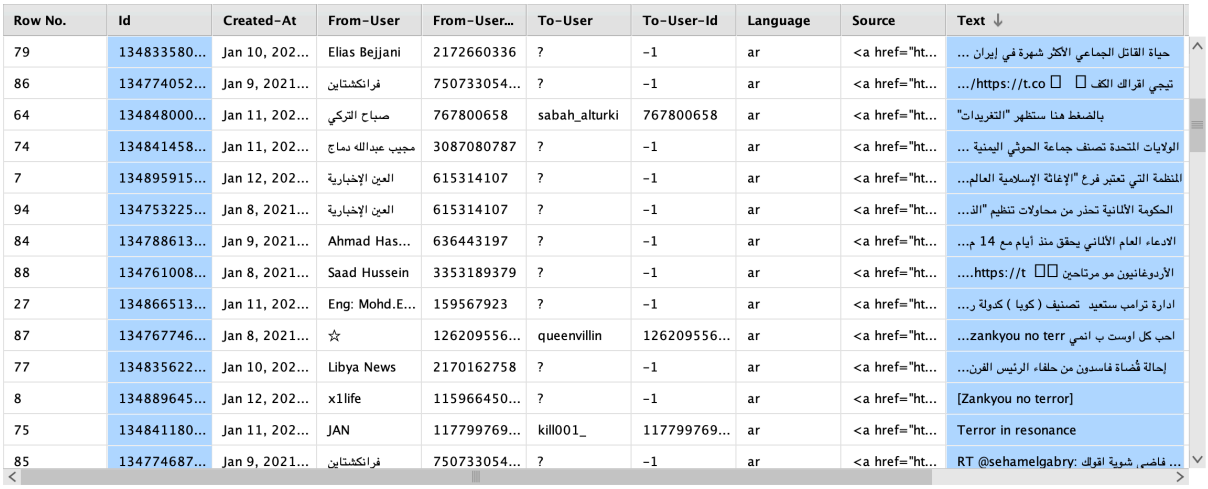


Figure 17. Un exemple d'extraction des tweets dans RapidMiner Studio

Une fois tous les paramètres nécessaires sont renseignés, on exécute le processus d'extraction qui est connecté au compte Twitter déjà créé. Ainsi, les tweets contenant le mot clé saisi dans la requête seront aspirés. Un aperçu de cette extraction est donné dans la figure 18 ci-dessous. Comme on peut noter, outre le contenu des tweets (colonne Text), chaque extraction de tweets nous donne un ensemble important d'informations exploitables comme l'identifiant de l'utilisateur, l'auteur du tweet, la date de publication, le lien, la géolocalisation...



Row No.	Id	Created-At	From-User	From-User...	To-User	To-User-Id	Language	Source	Text
79	134833580...	Jan 10, 202...	Elias Bejjani	2172660336	?	-1	ar	<a href="ht...	حياة القاتل الجماعي الأكثر شهرة في إيران ...
86	134774052...	Jan 9, 2021...	فرانكشتاين	750733054...	?	-1	ar	<a href="ht...	.../https://t.co □ □ تيجي اقراك الكف
64	134848000...	Jan 11, 202...	صباح التركي	767800658	sabah_alturki	767800658	ar	<a href="ht...	بالضغط هنا ستظهر "التغريدات"
74	134841458...	Jan 11, 202...	مجيب عبدالله دماج	3087080787	?	-1	ar	<a href="ht...	الولايات المتحدة تصنف جماعة الحوثي اليمنية ...
7	134895915...	Jan 12, 202...	العين الإخبارية	615314107	?	-1	ar	<a href="ht...	المنظمة التي تعتبر فرع "الإغاثة الإسلامية العالم...
94	134753225...	Jan 8, 2021...	العين الإخبارية	615314107	?	-1	ar	<a href="ht...	الحكومة الألمانية تحذر من محاولات تنظيم "الذ...
84	134788613...	Jan 9, 2021...	Ahmad Has...	636443197	?	-1	ar	<a href="ht...	الادعاء العام الألماني يحقق منذ أيام مع 14 م...
88	134761008...	Jan 8, 2021...	Saad Hussein	3353189379	?	-1	ar	<a href="ht...	....https://t □□ الأردوغانيون مو مرتاحين
27	134866513...	Jan 11, 202...	Eng: Mohd.E...	159567923	?	-1	ar	<a href="ht...	ادارة ترامب ستعيد تصنيف (كوبا) كدولة ر...
87	134767746...	Jan 8, 2021...	☆	126209556...	queenvillin	126209556...	ar	<a href="ht...	...zankyou no terr احب كل اوست ب انمي
77	134835622...	Jan 10, 202...	Libya News	2170162758	?	-1	ar	<a href="ht...	إحالة قُضاة فاسدون من حلفاء الرئيس الفرن...
8	134889645...	Jan 12, 202...	x1life	115966450...	?	-1	ar	<a href="ht...	[Zankyou no terror]
75	134841180...	Jan 11, 202...	JAN	117799769...	kill001_	117799769...	ar	<a href="ht...	Terror in resonance
85	134774687...	Jan 9, 2021...	فرانكشتاين	750733054...	?	-1	ar	<a href="ht...	RT @sehamelgabry: شوية اقوك فاضي ...

Figure 18. Informations extraites sur les tweets

Toutes les informations extraites des tweets sont d'abord enregistrées dans un fichier CSV (hors ligne). A partir de ces fichiers, nous avons créé une base de données SQL dans laquelle nous stockons toutes ces informations de manière structurée. Ensuite, nous avons supprimé les doublons afin qu'ils n'influencent pas l'apprentissage machine. Le tableau 7 montre le nombre de tweets extraits pour chaque collection (mot-clé) en arabe et en anglais.

Tableau 7. Nombre de tweets extraits pour chaque collection

Arabic			**English**	
Collection	***Meaning in English***	***Number of tweets***	***Collection***	***Number of tweets***
البغدادي	Albaghdadi	509	Albaghdadi	484
القاعدة	Alqaida	417	Attack	954
شهيد	Martyr	610	Bomb	1038
داعش	Daech	1059	Daesh	439
حماس	Hamas	665	Hamas	460
حزب الله	Hizbollah	559	Hizbollah	314
ارهاب	Terrorism	369	#IS	415
مفخخ	Booby-trapped	239	Isis	545
نصرالله	Nasrallah	453	IslamicState	350
قصف	Bombing	641	Jihad	490
TOTAL		**5521**	Nasrallah	371
			Terror	774
			TOTAL	**6634**

A la fin de cette étape, nous disposons d'un jeu de données composé de 12155 tweets bruts et uniques. Pour garantir la cohérence et l'uniformité de nos données, nous devrons les normaliser et standardiser. Nous détaillons les techniques que nous avons utilisées pour cela dans la section qui suit.

3. Normalisation

Pour l'étape de normalisation, nous avons appliqué un ensemble de règles pour chaque langue. Par exemple, une des caractéristiques de la langue arabe est que ses lettres sont liées les unes aux autres dans les textes imprimés. Il y a 22 lettres en arabe sur 28 qui ont quatre formes : Lettre seule, comme première lettre d'un mot, à l'intérieur du mot entre deux autres lettres, ou comme dernière lettre d'un mot (Al-Sanabani & Al-Hagree, 2015). Les six lettres arabes restantes (ز, ر, ا و, ذ, د) n'ont que deux formes (Comme première lettre dans un mot, comme dernière lettre dans un mot).

Afin de mieux normaliser nos données et de faciliter leur manipulation, nous leur avons appliqué deux types de normalisation : La normalisation des lettres et la normalisation des mots.

3.1. Normalisation des lettres

Certaines lettres de la langue arabe peuvent se présenter sous plusieurs formes d'écriture différentes. Nous normalisons les formes d'écriture de ces lettres dans cette phase, en ne retenant qu'une seule forme (standard). Le tableau 8 montre quelques exemples de normalisation de lettres où nous avons choisi une forme de lettre parmi ses variétés possibles.

Tableau 8. Différentes formes de lettres arabes

Letter	Name	Different shapes	Chosen shape
أ	Alif	ا,أ,إ,آ	ا
ي	Ya	ي,ى	ي
ء	Hamza	ء، ؤ، ئ	ء

En outre, pour faciliter le traitement des données par la machine, ainsi que la vérification des mots dans un dictionnaire, toute modulation sur les lettres a été supprimée.

3.2. Normalisation des mots

Même avec la normalisation des lettres, nous obtenons occasionnellement des tweets qui contiennent des termes qui ne font pas partie du lexique arabe ou anglais (mots d'un dialecte, erreurs typographiques, abréviation de certaines lettres...). Cela peut rendre difficile l'interprétation du sens d'un tweet. Pour résoudre ce problème, nous avons choisi d'utiliser la distance de Levenshtein (Konstantinidis, 2005) qui est une technique syntaxique, décrite comme suit :

Premièrement, et comme indiqué dans (El-Shishtawy, 2013), la distance de Levenshtein est définie à l'origine pour faire correspondre deux chaînes de longueur arbitraire. Elle garde la trace des plus petites variations entre les chaînes de caractères en termes de nombre d'insertions, de suppressions ou de substitutions nécessaires pour transformer une chaîne en l'autre. Une correspondance parfaite est représentée par un score de zéro.

Nous appliquons cette distance de la manière suivante : lorsque l'on cherche le mot w dans un lexique L,

- si w est dans L : d (w, w) = 0
- si w n'est pas dans L, nous pouvons utiliser la distance de Levenshtein pour trouver les mots w' dans L qui sont les plus proches de w, tels que d (w, w') < k (k est un nombre entier). L'un de ces mots peut être l'orthographe correcte de w. Plus k se rapproche de 0, plus les deux mots sont similaires.

Exemple : pour le mot 'ازهابى ' qui n'est pas dans le lexique arabe, on peut trouver le mot 'ارهابى ' avec d (ازهابى, ارهابى) =1.

4. Annotation

L'apologie du terrorisme consiste à présenter ou à commenter de manière positive les faits et les actes terroristes commis (l'approbation d'une attaque ou la défense des auteurs). Nous avons traité, dans ce travail, l'apologie du terrorisme qui est faite de manière claire, non équivoque et publique. Dans cette phase, nous annotons manuellement les tweets normalisés en 3 classes (positif, négatif et neutre), *positif* si le tweet est considéré comme une véritable apologie du terrorisme (soutien, propagande...), *négatif* si le tweet contient un contenu lié au terrorisme mais ne représente pas une menace réelle, *neutre* si le tweet n'a aucune relation avec les menaces terroristes. Des exemples de tweets, qui sont classés respectivement comme positifs, négatifs ou neutres, sont donnés dans le tableau 9.

Tableau 9. Exemples de tweet annotés pour chaque classe

Language	*Tweet*	*Sens en Anglais*	*Classe*
Arabe	*البغدادي ليس جبانا كي يهرب من الموت وانما هناك رءوس كبار اكبر منه في التنضيم فا الهيدرا لها رءوس عديدة ولكن لديها عقل مدبر يتحكم بهذه الرءوس وقطع راس واحد من رءوس الهيدرا لايعني موتها وسوف ينمو لديها راس اخر واخر ...الخ*	*Al-Baghdadi is not a coward in order to escape from death, but there are big heads that are bigger than him in the organization. The hydra has many heads, but it has a mastermind that controls these heads and cutting off one of the heads of the hydra does not mean its death, and it will grow another head and another...etc.*	***Positif***
	عنصر سابق في داعش تحدثت عن نفاق عناصر التنظيم والفرق الكبير بين اقوال وأفعال قادة داعش الدولة_الاسلامية #العراق #بغداد #البصرة#	*A former ISIS member spoke about the hypocrisy of the organization's elements and the big difference between the words and actions of ISIS leaders #Islamic_State #Iraq #Baghdad #Basra*	***Négatif***
	لا تحكم على الناس قبل ان تسمع منهم . مباشرة القاعده تقول اسمع مني ولا تسمع عني	*Don't judge people before you hear directly from them. Rule says listen to me and don't hear from me*	***Neutre***
Anglais	When the world calls Muslims extremists, why don't we teach them a lesson, we will have to adopt the principles of dissent again, many countries have been destroyed by the label of infidel terrorism and we even cut off the name of jihad, a sword is necessary to protect.		***Positif***
	Mumbai. 26/11. Never Forget. My tributes to our brave men from Mumbai Police, NSG who laid down their lives to keep us safe.		***Négatif***
	A #ransomware #attack hitting Las Cruces Public Schools forced the district to shut down the entire computer system to contain the infection		***Neutre***

L'annotation des 12155 tweets est effectuée manuellement (Voir Figure 19). En d'autres termes, l'annotateur (une personne) lit les tweets un par un et les annote en fonction de sa compréhension du sens du tweet. Outre la fluidité, un discernement objectif est bien sûr requis. Par exemple, l'annotateur doit être capable de faire la différence entre un reportage et une opinion personnelle sur un acte terroriste. De plus, par souci de précision, chaque tweet est annoté par deux personnes. En cas de désaccord (par exemple, un tweet identique est annoté dans deux classes différentes), une troisième personne intervient pour confirmer l'une des annotations.

Figure 19. Annotation manuelle des tweets

Une fois qu'un tweet est annoté, il sera déplacé vers un dossier en fonction de sa classe d'annotation. À la fin de cette étape, nous aurons 3 classes (dossiers) qui contiennent les tweets annotés. Le tableau 10 montre le nombre de tweets Anglais par mot-clé dans chaque classe. Le tableau 11 montre le nombre de tweets Arabes par mot-clé dans chaque classe.

Tableau 10. Nombre de tweets par classe pour l'anglais

English				
Collection	***Number of Tweets***	***Positive class***	***Negative Class***	***Neutral Class***
Albaghdadi	484	118	215	151
Attack	954	28	272	654
Bomb	1038	21	164	853
Daesh	439	30	280	129
Hamas	460	14	120	326
Hizbollah	314	14	93	207
#IS	415	29	118	268
Isis	545	38	396	111
IslamicState	350	34	282	34
Jihad	490	52	303	136
Nasrallah	371	6	44	321
Terror	774	19	392	363
Total	**6635**	**403**	**2679**	**3553**

Tableau 11. Nombre de tweets par classe pour l'arabe

Arabic					
Collection	***Meaning in English***	***Number of Tweets***	***Positive class***	***Negative Class***	***Neutral Class***
البغدادي	Albaghdadi	509	43	242	224
القاعدة	Alqaida	417	73	106	238
شهيد	Martyr	610	21	215	374
داعش	Daech	1059	87	774	198
حماس	Hamas	665	11	131	523
حزب الله	Hizbollah	559	62	462	35
ارهاب	Terrorism	369	50	262	57
مفخخ	Booby-trapped	239	2	85	152
نصرالله	Nasrallah	453	140	244	69
قصف	Bombing	641	13	208	420
Total		**5521**	**502**	**2729**	**2290**

Comme nous pouvons le voir dans les tableaux 10 et 11, la distribution des tweets dans les trois classes est très déséquilibrée pour les deux langues Arabe et Anglais. La figure 20 suivante illustre la répartition des tweets annotés sur les trois classes pour chacune des langues.

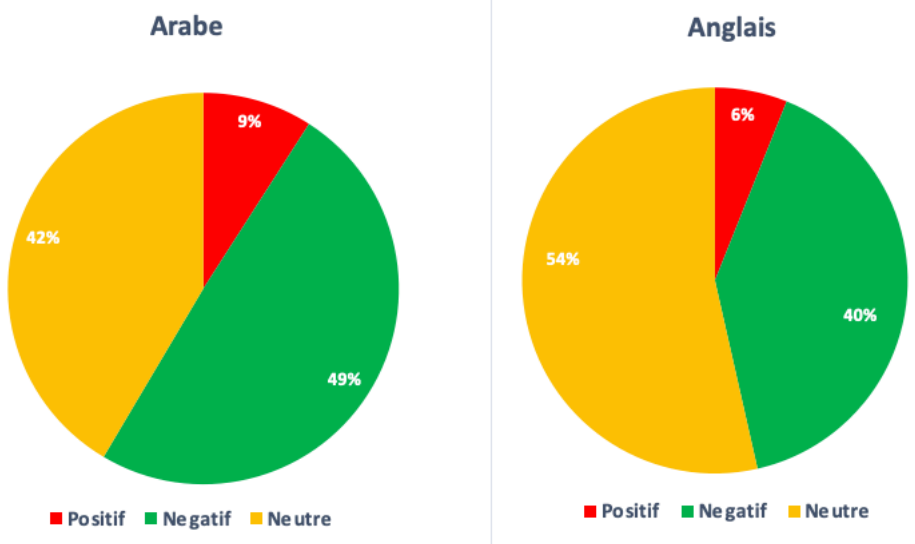


Figure 20. Distribution des tweets annotés dans les trois classes

Comme on peut le constater d'après la figure 20, la classe positive est minoritaire par rapport aux deux autres classes Négatives et Neutre.

5. Prétraitement

Le prétraitement de nos données est effectué pour préparer les tweets à l'apprentissage machine. Les trois techniques (tokenization, lemmatisation et stop words) sont les algorithmes les plus efficaces et les plus utilisés pour le prétraitement des tweets dans les travaux de recherche sur l'analyse des sentiments (Al-Khafaji & Habeeb, 2017). La figure 21 illustre le processus d'application de ces techniques sur nos données en utilisant Rapid Miner Studio.

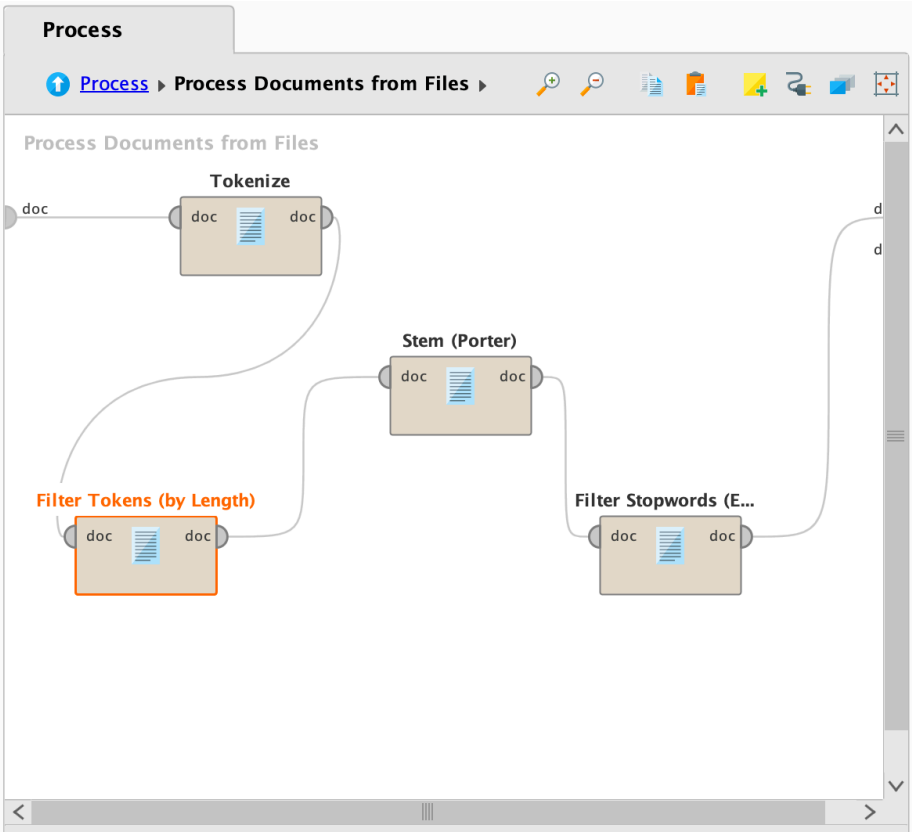


Figure 21. Techniques de prétraitement sur RapidMiner Studio

Ces étapes jouent, en plus de leur rôle de préparation, le rôle de réduction des données en excluant les mots inutiles (mots vides qui n'apportent pas de valeur sémantique pour la compréhension du contenu), et en préservant les mots significatifs (mots importants qui portent le sens principal du texte). Ainsi, en utilisant ces techniques, les tweets deviennent plus courts en n'incluant que les termes les plus importants, et c'est exactement ce que nous voulons avoir avant de faire la classification.

Nous prenons comme exemple un tweet publié en réaction aux attaques terroristes des frères Kouachi contre Charlie Hebdo (attaque terroriste à Paris en janvier 2015) et nous appliquons

les trois étapes (tokenization, lemmatisation et stop words) de manière séquentielle afin de mieux les comprendre.

1- La première étape de prétraitement est la **Tokenisation** qui sert à segmenter notre corpus en unités "atomiques" : les tokens. Il s'agit dans cette opération de découper nos tweets en unités (mots) pour faciliter par la suite la recherche des tweets qui sont en relation avec le contenu terroriste.

Un exemple d'application de la tokenisation sur un tweet est donné comme suit :

Tribute	the	Kouachi	brothers	who	have	just	fallen	martyrs

Comme on peut le voir dans cet exemple, le tweet est découpé en tokens, et chaque token est un mot.

2- La deuxième étape de prétraitement est la **Lemmatisation (stem)** qui sert à regrouper les mots d'une même famille dans un seul lemme, ce dernier va remplacer les différentes formes que peut revêtir un mot, soit : le nom, le pluriel, le verbe à l'infinitif, etc. Comme exemple de cette opération, le lemme terror remplacera les mots : terrorisme, terroriste, terroriser, terroristes, etc.

Pour les tweets en Anglais, nous avons utilisé l'algorithme de stemming de Porter (Porter, 2001) pour dériver les mots en appliquant un remplacement itératif, basé sur des règles, des suffixes des mots pour réduire leur longueur jusqu'à ce qu'une longueur minimale soit atteinte. Nous avons utilisé l'algorithme Light stemming Arabic (Saad, 2016) pour la langue arabe, qui prend une liste de mots et effectue un stemming léger pour chaque mot arabe.

En appliquant la lemmatisation sur l'exemple précédent, nous aurons :

Tribute	the	Kouachi	brother	who	have	just	fall	martyr

Comme on peut le voir sur notre exemple, chaque mot est remplacé par son lemme correspondant (fallen => fall).

3- La troisième et dernière étape de prétraitement est **Stop words** ou suppression de mots vides qui sert à enlever du corpus les mots courts qui sont non significatifs dans un texte et qui ont plus un rôle syntaxique qu'un sens en eux même. Comme exemple de mots vides, nous trouvons les prénoms personnels, les articles, les prépositions, etc.

En appliquant cette technique sur notre exemple, nous aurons :

Tribute	Kouachi	brother	fall	martyr

Comme on peut le voir sur notre exemple, les mots (the, who, have, just) ont été supprimé. Il n'y a que les mots importants qui restent.

Une fois ces trois opérations appliquées à notre corpus enrichi, nous obtenons un corpus traité prêt à être utilisé dans l'indexation, la recherche et la classification.

Il convient de noter qu'il existe d'autres techniques de prétraitement telles que le remplacement des abréviations et la suppression des chiffres. Pour plus de détails, voir (Gupta & Kumari, 2019).

6. Conclusion

Dans ce chapitre, nous avons décrit notre processus de collecte et de création de notre jeu de données, ainsi que les méthodes que nous avons employées pour normaliser les lettres et les mots, en particulier pour la langue arabe. De plus, nous avons exposé le processus d'annotation que nous avons réalisé pour nos tweets, couvrant les langues arabe et anglaise. Nous avons également présenté les diverses méthodes de prétraitement de texte (NLP) que nous avons utilisées sur nos données afin de les préparer en vue de la classification.

La collecte, la normalisation, l'annotation et le prétraitement des données sont des phases partagées par les trois contributions de notre thèse. C'est la raison pour laquelle nous avons réservé un chapitre entier spécifiquement pour les détailler et les présenter de manière exhaustive.

Dans le prochain chapitre, nous détaillerons les trois contributions majeures de notre thèse.

Chapitre 4. Contributions et évaluations

Dans ce chapitre, nous mettons en avant les trois contributions fondamentales de notre thèse en décrivant la méthodologie suivie, les ensembles de données utilisés ainsi que les algorithmes et techniques employés pour chacune des contributions. De plus, nous examinons en détail les expérimentations réalisées pour les trois contributions. Nous analysons les résultats obtenus pour chaque contribution, effectuons des comparaisons entre les divers algorithmes de classification, et évaluons les différentes techniques de représentation de texte que nous avons employées.

1. Introduction

Les trois contributions phares de notre thèse sont décrites comme suit :

- La première contribution est un travail de recherche sur la détection des menaces terroristes sur Twitter en utilisant une méthode lexicale. Ce travail est publié et présenté dans une conférence internationale (Bedjou, et al., 2018).
- La deuxième contribution est un travail de recherche sur la détection des menaces terroristes sur Twitter en utilisant une classification binaire avec l'algorithme du machine learning SVM. Ce travail est publié et présenté dans une conférence internationale (Bedjou, et al., 2019).
- La troisième contribution est un travail de recherche sur la détection des apologies du terrorisme sur Twitter en utilisant des techniques du machine et deep learning. Ce travail est publié dans une revue internationale (Bedjou & Azouaou, 2023).

Nous détaillons dans ce qui suit chacune de ces contributions.

2. Contribution 1 (Recherche lexicale)

Dans cette première contribution, nous nous sommes intéressés à la détection lexicale des contenus relatifs au terrorisme sur le réseaux social Twitter. Nous nous intéressons au contenu des tweets (messages, hashtags et url) que nous avons analysés et traités pour détecter tout tweet susceptible de représenter une menace terroriste, qu'il s'agisse d'une menace explicite, d'une apologie du terrorisme ou de messages de propagande. Pour cela, nous proposons un processus

nommé LexD3T (Lexical Detection process of Terrorist Threats on Twitter). Ce processus est basé sur une recherche lexicale des tweets qui peuvent représenter des menaces terroristes. Pour cela, nous avons utilisé des techniques du traitement du langage naturel (NLP) comme la tokenisation et la lemmatisation pour le nettoyage des tweets, et l'API Lucene Apache pour leur indexation. Pour détecter les tweets menaçants, nous calculons la similarité entre des requêtes générées (à l'aide d'un ensemble de mots-clés) et l'index des tweets créé avec l'API Lucene.

2.1. Méthodologie

La majorité des plateformes de médias sociaux ont mis en place des mécanismes de détection de certains mots dans les publications, ce qui entraîne automatiquement le blocage des publications les contenant. Les utilisateurs, quant à eux, ont recours à diverses méthodes pour contourner ces restrictions, telles que l'espacement entre les lettres des mots ou l'ajout d'une lettre au milieu, par exemple pour le mot *Daesh* ils l'écrivent de plusieurs façons : Da3sh, Da-esh, Da_esh, etc.

La particularité de cette contribution réside dans sa capacité à repérer des tweets contenant des termes liés au terrorisme, qu'ils soient délibérément mal orthographiés ou non, et ce, dans les deux langues : Arabe et Anglais.

Dans le processus LexD3T proposé, nous avons utilisé des tweets extraits du corpus fourni par Fifth Tribe, qui contient des tweets bruts publiés par des partisans de l'État islamique (ISIS) après l'attaque terroriste de novembre 2015 à Paris (Fifth Tribe, 2018).

Les étapes de notre processus LexD3T sont illustrées dans la figure 22 ci-après.

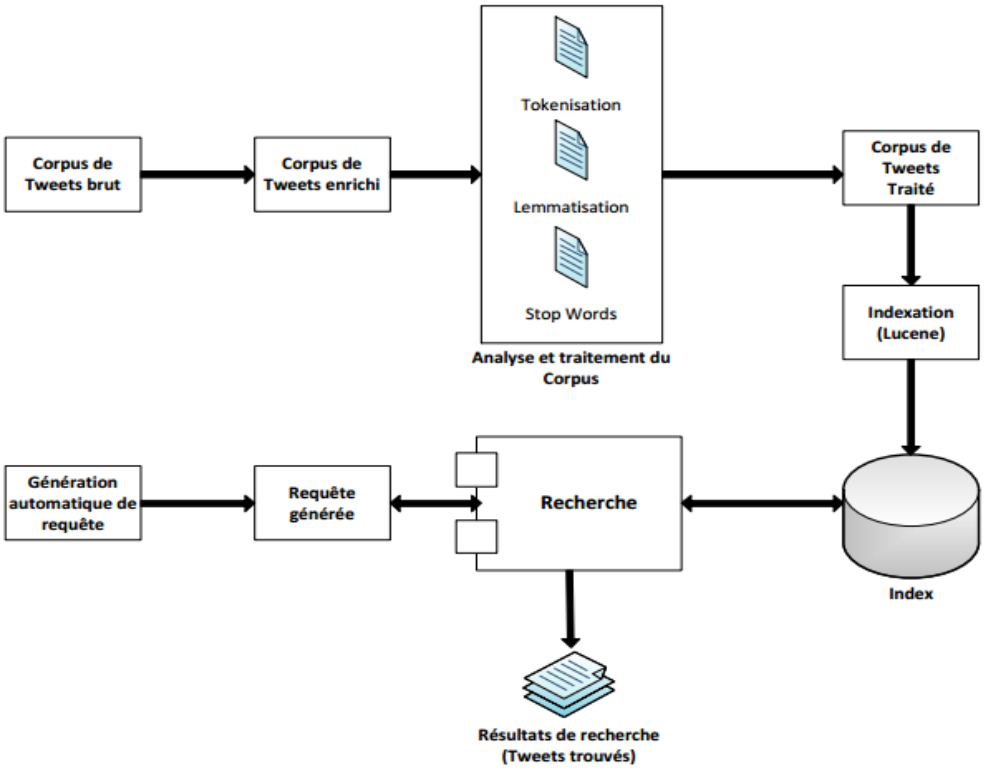


Figure 22. LexD3T processus de détection lexicales des menaces terroristes sur Twitter

Dans cette proposition, nous démarrons avec un corpus d'environ 4000 Tweets, auquel nous injectons 10 Tweets qui contiennent des mots et des expressions liées aux contenus terroristes. Nous avons réparti ces 10 Tweets injectées en deux parties, 5 tweets numéroté de 1 à 5, qui représentent de vraies menaces terroristes, et 5 autres numéroté de 6 à 10, qui ne représentent pas de vraies menaces. Des exemples de tweets injectés sont donnés dans le tableau 12 suivant.

Tableau 12. Exemples de tweets injectés au jeu de données

Tweet injecté	Traduction	Classe
Tribute to the Kouachi brothers who have just fallen Martyrs. We are happy	Hommage aux frères Kouachi qui viennent de tomber Martyrs. Nous sommes heureux.	Vraie menace terroriste
Tunisie was just as victim of Daech attacks! Not to blame an entire people for a terrorist crime.	*La Tunisie a été tout autant victime des attaques de Daech ! Il ne faut pas accuser tout un peuple d'être responsable d'un crime terroriste.*	Fausse menace terroriste

Avec ces injections, nous obtenons un corpus enrichi. Sur ce dernier, nous effectuons des opérations d'analyse et de traitement afin d'éliminer le maximum de mots inutiles à notre recherche.

- La première étape de traitement est la **Tokenization**.
- La deuxième étape de traitement est la **Lemmatisation**.

- La troisième et dernière étape de traitement est **Stop words** ou suppression de mots vides.

Ces trois opérations sont décrites avec des exemples dans le chapitre 3, section 5 (Voir les pages 54 à 56).

Une fois ces trois opérations appliquées à notre corpus enrichi, nous obtenons un corpus traité prêt à être utilisé dans l'indexation et la recherche. Pour interroger ce corpus, nous générons des requêtes qui contiennent un ensemble de mots liés au contenu terroriste comme : terror, daech, bomb, isis, etc.

Pour le choix de la mesure de similarité entre une requête et l'index, nous avons opté pour l'utilisation des techniques et mesures intégrées dans l'API Apache Lucene (Gospodnetic, et al., 2010). Cette mesure utilise la technique de représentation de texte TF-IDF (voir chapitre 1).

La formule de recherche est donnée comme suit :

$score\ (q,d) = coord(q,d) * queryNorm(q) \sum(tf(t\ in\ d) * idf(t) * t.getBoost(\ ) * norm(t,d))$

Où :

tf est la racine carré du *tf* usuel soit tf (d,q)= $\sqrt{tf}$ (Voir la formule 2 du chapitre 1)

idf (t) $=1 + \log(\frac{N}{df} + 1)$ (Voir la formule 3 du chapitre 1)

coord (d, q) est le score calculé en fonction du nombre d'apparitions de d (document) dans q (query). Ce type de score est issu du modèle booléen et est spécifique à Lucene.

queryNorm (q) est égal à la somme des carrés du poids de la requête et se calcule de cette façon : $\frac{1}{\sqrt{W^2+W^2+W^2}}$

t.getBoost () = boost attribué au cours de l'indexation, par défaut le boost est de 1.

norm (t, d) qui englobe un coefficient d'importance des mots, documents et longueur des documents.

Le choix d'utiliser Apache Lucene se justifie par le fait qu'il offre des fonctionnalités très avancées en matière de recherche lexicale, telles que : la recherche de termes ('jihad'), la recherche floue ('jihad', 'jehad', '' gihad '), la recherche de termes similaires (' jihad ',' jiahd '), la recherche booléenne (' jihad AND martyr '), et la recherche avec des jockers sur un ou plusieurs caractères ('jiha?', jihad *).

2.2. Évaluation

Pour évaluer notre processus de détection proposé, nous avons créé un prototype de moteur de recherche qui détecte les tweets susceptibles d'être des menaces terroristes. Nous avons utilisé le corpus de Fifth Tribe (Fifth Tribe, 2018) pour cette évaluation. Ce corpus est un tableau de tweets ISIS, chaque ligne correspondant à un tweet. Il comporte les colonnes suivantes : Nom, Nom d'utilisateur, Description, Emplacement, Followers (Nombre de followers au moment où le tweet a été téléchargé), NumberStatuses (Nombre de statuts de l'utilisateur au moment où le tweet a été téléchargé), Time (Date et horodatage du tweet), Tweets (Le contenu du tweet). Pour notre ensemble de données, nous avons utilisé la dernière colonne "Tweets" qui contient tous les contenus textuels des tweets dont nous avons extrait 4000 de manière aléatoire afin de garantir la représentativité des résultats. Nous nous sommes limités à 4000 tweets en raison des limites techniques des machines (ordinateurs) utilisées dans nos expériences.
Nous avons testé notre proposition sur trois catégories de tweets :
- des tweets rédigés en anglais
- des tweets rédigés en arabe
- des tweets écrits dans les deux langues (anglais et arabe)

Dans ce qui suit, nous présentons les expérimentations que nous avons menées ainsi que les résultats obtenus.

Pour l'Anglais, nous générons des requêtes qui contiennent les mots : bomb, jihad, terror, qaida, die, dead, death, explos, daech, isis, attac, ei, khilafa, baqia, tatamadad, allah, martyr. La figure 23 présente un aperçu du prototype développé et expérimenté sur les tweets en Anglais.

Figure 23. Prototype de détection lexicale des menaces terroristes en Anglais

Nous trions les tweets trouvés en fonction de leur score de similarité, puis nous affichons les tweets trouvés dans leur intégralité (version brute), comme le montre la figure 23. Pour les 4000 tweets utilisés dans notre recherche, nous trouvons 124 tweets qui correspondent aux mots-clés de la requête générée. Nous avons constaté que sur les 124 tweets trouvés, 68 tweets représentent de vraies menaces terroristes; les autres tweets ne représentent aucune menace.
Nous remarquons que les tweets numérotés (injectés auparavant) apparaissent dans les résultats de recherche. Un aperçu des taux de similarité est donné dans la figure 24 suivante.

Figure 24. Aperçu des résultats de recherche lexicale

Nous avons mené la même expérience avec la langue Arabe, nous générons des requêtes qui contiennent les mots : إرهاب , داعش, خلافة, باقية, تتمدد, تفجير, مفخخ, ناسف, جهاد, القاعدة, قنبلة, فتوى
La figure 25 présente un aperçu du prototype développé et expérimenté sur les tweets en Arabe.

Figure 25. Prototype de détection lexicale des menaces terroristes en Arabe

Sur les 4000 tweets utilisés pour la recherche, nous trouvons 110 tweets qui correspondent aux mots-clés de la requête générée. Pour chaque tweet trouvé, un degré de similarité est calculé à l'aide de la formule précédente. Nous notons que sur les 110 tweets trouvés, 75 tweets représentent de vraies menaces terroristes, le reste ne représentant aucune menace.

Nous avons mené la même expérience avec la langue Arabe, nous générons des requêtes qui contiennent les mots : إرهاب , داعش, خلافة, باقية, تتمدد, تفجير, مفخخ, ناسف, جهاد, القاعدة, قنبلة, فتوى, bomb, jihad, terror, qaida, die, dead, death, explos, daech, isis, attac, ei, khilafa, baqia, tatamadad, allah, martyr.

La figure 26 présente un aperçu du prototype développé et expérimenté sur les tweets en bilingue (arabe & anglais).

Figure 26. Prototype de détection lexicale des tweets en bilingue (arabe & anglais)

Sur les 4000 tweets utilisés pour la recherche, nous trouvons 72 tweets qui correspondent aux mots-clés de la requête générée. Pour chaque tweet trouvé, un degré de similarité est calculé à l'aide de la formule précédente. Nous notons que sur les 72 tweets trouvés, 41 tweets représentent de vraies menaces terroristes, le reste ne représentant aucune menace.

Nous constatons que la recherche lexicale fonctionne de la même façon sur les deux langues traitées (Arabe et Anglais).

2.3. Résultats obtenus

Pour le calcul des performances, nous avons calculé les taux de Rappel, Précision et F-measure (appelée aussi F score ou F1 score).

Dans notre cas, un document est représenté par un tweet. Donc en appliquant les formules du rappel, précision et F-measure, nous obtenons les résultats présentés dans le tableau 13 ci-après.

Tableau 13. Résultats obtenus pour la recherche lexicale

	Rappel	Précision	F-measure
Anglais	0.017	0.548	0.033
Arabe	0.018	0.681	0.035
Bilingue	0.024	0.664	0.046

Comme nous pouvons le constater, les taux de rappel sont très faibles. Ceci se justifie par le fait que les tweets de menaces terroristes représentent une partie très minime des tweets en général. De plus, la spécification du type de menace recherchée a considérablement réduit le nombre de tweets détectés. Pour la précision, nous constatons que la recherche lexicale n'est pas adéquate à la problématique de recherche des contenus liés aux menaces terroristes. Bien que nous trouvions ces menaces, le bruit (tweets trouvés mais qui ne sont pas de vraies menaces) est trop élevé, ce qui peut poser problème dans le cas de traitement de corpus volumineux. C'est pourquoi nous proposons, afin d'améliorer ces résultats notamment la précision de la recherche, d'intégrer l'aspect sémantique à la détection des tweets de menaces terroristes.

3. Contribution 2 (Classification binaire - SVM)

Nous proposons dans cette deuxième contribution, un système de détection des publications liées au terrorisme dans le réseau social Twitter basé sur SVM. Nous avons établi un processus de 12 étapes pour l'analyse, le traitement et puis la détections des tweets menaçants. Nous avons élaboré 2 scénarios d'utilisation de ce processus, dans le 1er scénario, nous effectuons un apprentissage machine sur 4000 tweets écrits en Anglais, puis 4000 tweets écrits en Arabe et enfin 4000 tweets écrits en bilingue (dans les deux langues). Dans le 2ème scénario, nous effectuons un apprentissage machine sur l'ensemble des tweets à la fois (12000 tweets).

3.1. Méthodologie

La particularité de cette contribution est l'intégration des techniques et algorithmes classiques du Machine Learning pour apprendre à la machine à classifier des tweets selon deux catégories : *Menaçant* ou *Non menaçants* (classification binaire).

Le processus proposé est composé de 12 étapes pour la détection des menaces terroristes sur Twitter. Les 5 premières étapes du processus sont liées au domaine du NLP qui permettent une analyse et un traitement des tweets en appliquant trois principales méthodes : la Tokenization, la Lemmatisation et Stopwords. (Voir chapitre 3 pour plus de détails).

Les 7 étapes qui suivent sont liées au domaine du machine learning, en commençant par une création de deux catégories de données (Données d'entrainement et Données de tests) puis en appliquant un algorithme de l'apprentissage supervisé à savoir SVM afin de générer un modèle de classification des menaces en deux classes : Menaçants et Non menaçants.

Nous avons opté pour SVM vu que notre classification est binaire et que SVM est l'un des meilleurs classifieurs binaires dans l'analyse de texte (Kumari & Srivastava, 2017).
Le processus que nous proposons est illustré dans le schéma de la figure 27 ci-dessous.

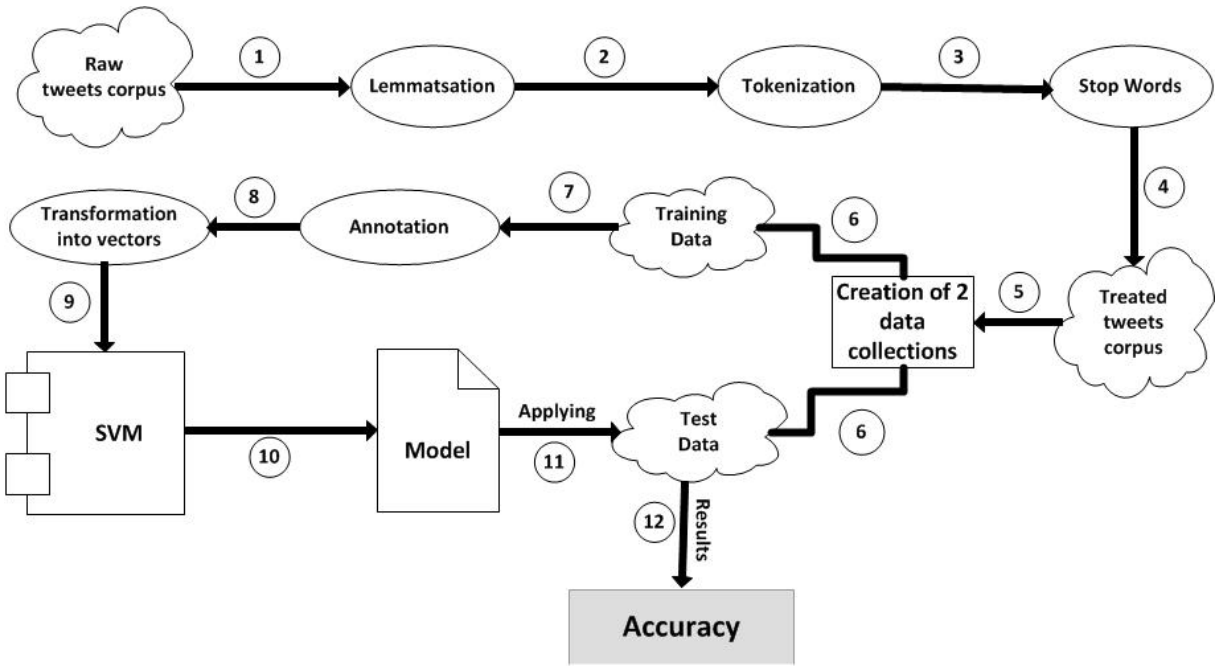


Figure 27. Processus de détection de menaces terroristes sur Twitter avec SVM

Dans ce qui suit, nous détaillons les étapes de ce processus.

- **ETAPE 1** : dans notre proposition, nous démarrons avec un corpus de Tweets brut (tweets entiers), auquel nous appliquons des opérations d'analyse et de traitement de texte afin d'éliminer le maximum de mots inutiles à notre recherche.
- **ETAPES 2,3, et 4** : ces étapes ont été déjà bien détaillées dans le chapitre 3. En bref, la Tokenization sert à segmenter notre corpus en unités "atomiques", la Lemmatisation sert à regrouper les mots d'une même famille dans un seul lemme, quant à Stop words ou suppression de mots vides, elle sert à enlever du corpus les mots courts qui sont non significatifs dans un texte et qui ont un plus rôle syntaxique qu'un sens en eux même.
  Une fois ces trois opérations, de traitement de texte, appliquées à notre corpus, nous obtenons un corpus traité prêt à être utilisé dans notre recherche.
- **ETPAES 5 et 6** : dans cette étape, nous procédons à la séparation de nos données en deux catégories de données : Training Data et Test Data.
  2 ensembles de données sont créés :

  - *Training Data* : ce sont les données (tweets) que nous utilisons pour l'entrainement et apprentissage par la machine, elles représentent deux tiers des données 2/3 de l'ensemble des tweets.
  - *Test Data* : ce sont les données (tweets) sur lesquelles porteront nos tests pour voir la précision du modèle généré après apprentissage par la machine.

- **ETAPE 7** : Annotation : il s'agit à travers cette étape d'annoter manuellement les tweets qui seront utilisés dans la phase d'apprentissage en les classant dans deux Classes : *Menaçants* et *Non Menaçants*. La Classe *Menaçants* regroupe l'ensemble des tweets qui représentent une vraie menace terroriste, tandis que la classe *Non Menaçant* regroupera les tweets qui ne représentent pas de vraies menaces terroristes.
- **ETAPE 8 :** Transformation en vecteurs **:** il s'agit dans cette étape de procéder à une transformation des tweets annotés dans l'étape précédente (format textuel), en vecteurs numériques pour pouvoir les utiliser dans la phase d'apprentissage. Nous avons utilisé la technique TF-IDF (voir chapitre 1) pour cette transformation.
- **ETAPE 9 :** SVM**,** nous appliquons SVM sur les vecteurs générés par l'étape précédente, afin d'apprendre à la machine à classer les tweets dans deux classes (*Menaçant* / *Non Menaçants)*.
- **ETAPE 10 :** une fois que l'apprentissage est fini, un modèle de classification sera créé. Ce modèle représente la connaissance acquise par la machine, et peut être utilisé dans des futurs tests.
- **ETAPE 11 :** Phase de Test**,** il s'agit dans cette étape de procéder aux tests du modèle créé précédemment sur des données non classées dans le but de les classer en deux classes *Menaçants* et *Non Menaçants*.
- **ETPAE 12 :** Résultat**,** à la fin de l'étape précédente, nous obtenons une classification des données de test dans les deux classes. Pour évaluer la précision du modèle créé, il faut calculer le nombre de tweets qui sont bien classés par rapport à l'ensemble des tweets utilisés dans la phase de test.

Pour récapituler, dans notre démarche, nous cherchons d'apprendre à la machine à classifier des tweets (nouvelles données) en deux catégories (*Menaçant* et *Non menaçant*). Le but est de pouvoir détecter avec une précision élevée les tweets qui constituent de véritables menaces terroristes, en s'appuyant sur des modèles de classification obtenus avec l'apprentissage automatique.

3.2. Évaluation

Pour évaluer notre proposition, nous avons développé 2 scénarios d'utilisation de notre processus. Dans le premier scénario, nous appliquons l'apprentissage machine sur 4000 tweets écrits en anglais, puis sur 4000 tweets écrits en arabe, et enfin sur 4000 tweets écrits en bilingue (dans les deux langues). Dans le deuxième scénario, nous appliquons l'apprentissage machine sur l'ensemble des tweets à la fois (12000 tweets). Ces deux scénarios nous permettent de comparer les résultats obtenus, en utilisant TF-IDF comme technique de représentation de texte et SVM comme classifieur, dans les deux langues Arabe et Anglais.

Nous détaillons dans ce qui suit les deux scénarios d'évaluation.

1. **Scénario 1 :** Utilisation de 4000 tweets écrits en arabe, 4000 en anglais et 4000 en bilingue. Dans ce scénario, nous procédons à l'extraction des tweets en utilisant l'API Twitter avec l'outil RapidMiner Strudio.

   La figure 28 ci-apèrs illustre le scénario 1 de notre évaluation.

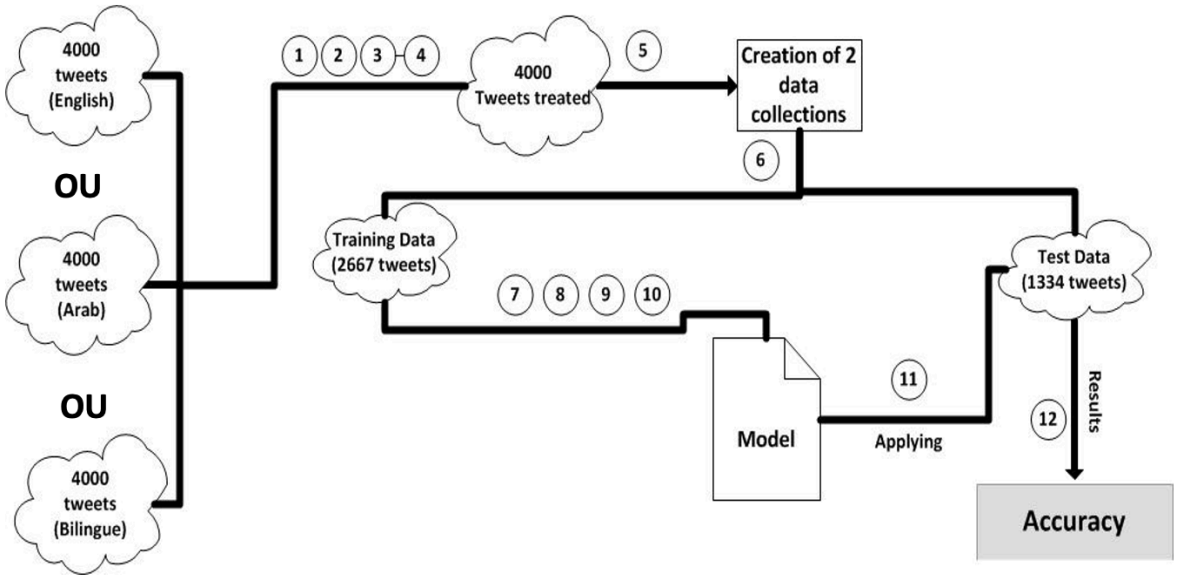


Figure 28. Scénario 1 d'utilisation du processus (classification binaire)

Nous appliquons les étapes 1 à 5 de notre processus sur les tweets extraits pour nettoyer et préparer nos données (ces étapes sont déjà détaillées dans le chapitre 3). Pour l'étape 6, nous créons deux catégories de données comme suit :

a. Données d'apprentissage: nous mettons 2/3 des 4000 tweets à savoir 2667 tweets dans la partie *Données d'apprentissage*. Nous procédons par la suite à l'annotation manuelle de ces 2667 tweets en deux classes *Menaçants* et *Non Menaçants*. Nous

appliquons SVM sur ces 2667 tweets classés et nous obtenons un modèle pour chaque corpus utilisé (un modèle pour l'Arabe, un autre pour l'Anglais, et un autre pour le Bi-lingue).

b. Données de test: nous appliquons le modèle obtenu précédemment sur les 1/3 des 4000 tweets à savoir 1333 tweets non encore classés pour distinguer les tweets menaçants de ceux qui ne le sont pas. Une fois que les 1333 tweets sont classés, nous procédons à la vérification des tweets qui sont bien classés et nous obtenons une précision de classement que nous comparons avec celle obtenue lors de l'expérimentation de la contribution 1 en utilisant l'approche lexicale (Bedjou, et al., 2018).

2. **Scénario 2 :** Utilisation de 12000 tweets à la fois, à travers ce scénario nous souhaitons faire une expérimentation sur l'ensemble des tweets des trois ensembles précédents. Nous regroupons les tweets écrits en Anglais, ceux écrits en Arabe et ceux écrits en bilingue dans un seul ensemble de données, ce qui nous donne un corpus de 12000 tweets réparties en 8000 pour l'apprentissage et 4000 pour le test. La figure 29 ci-dessous illustre le scénario 2 de notre évaluation.

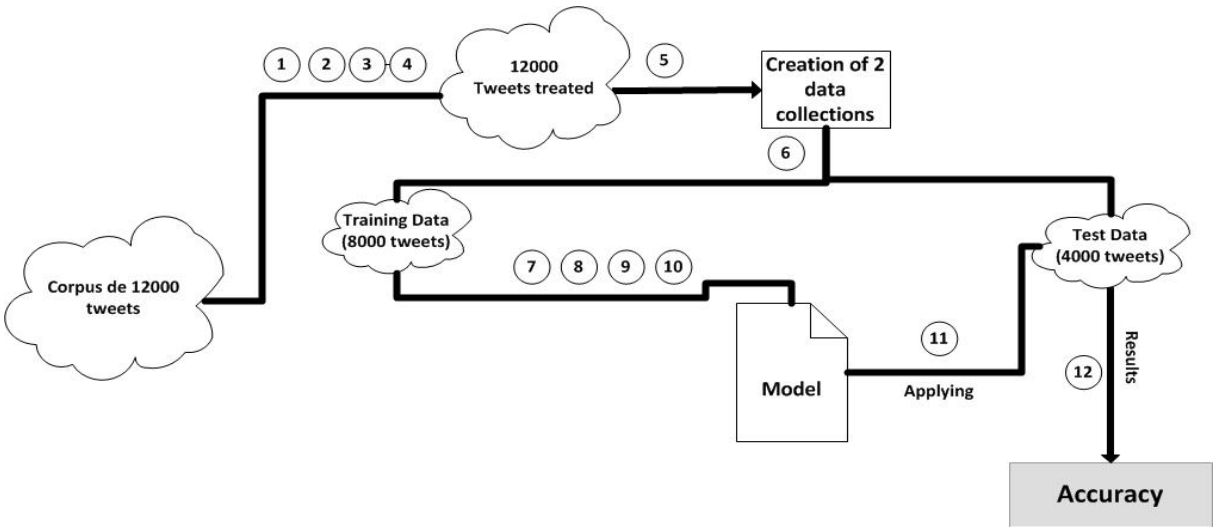


Figure 29. Scénario 2 d'utilisation du processus (Classification binaire)

Dans ce 2ème scénario, nous faisons passer un corpus de 12000 tweets à la fois, et nous obtenons un modèle différent des trois modèles obtenus dans le scénario 1, avec un seul taux de précision. Le but de l'utilisation de ce 2ème scénario est de voir s'il y a une amélioration de la précision de la classification par rapport au nombre de tweets utilisé (4000 puis 12000) et surtout de vérifier la perte ou non de la précision en utilisant SVM sur un corpus multi-langues (tweets écrits dans des langues différentes).

3.3. Implémentation des scénarios d'évaluation

Après avoir collecté, standardisé, annoté et prétraité nos données, elles sont prêtes à être utilisées pour la création d'un modèle de classification avec un entraînement à l'aide de l'algorithme d'apprentissage machine SVM. Nous présentons dans cette partie l'implémentation des deux scénarios d'évaluation.

Pour effectuer un apprentissage automatique avec SVM, nous avons utilisé l'outil Rapid Miner Studio (Mierswa & Klinkenberg, 2018). Nous expliquons dans ce qui suit l'implémentation du 1er scénario en utilisant 4000 tweets en Anglais, et cette implémentation est presque identique pour les autres scénarios, en adaptant juste les données et les techniques NLP de chacune des langues (Arabe et Anglais). Ainsi, nous avons classé les 4000 tweets dans deux ensembles différents (*Menaçant*, *Non menaçant*), puis nous les avons importés dans l'outil comme la montre la figure 30 suivante.

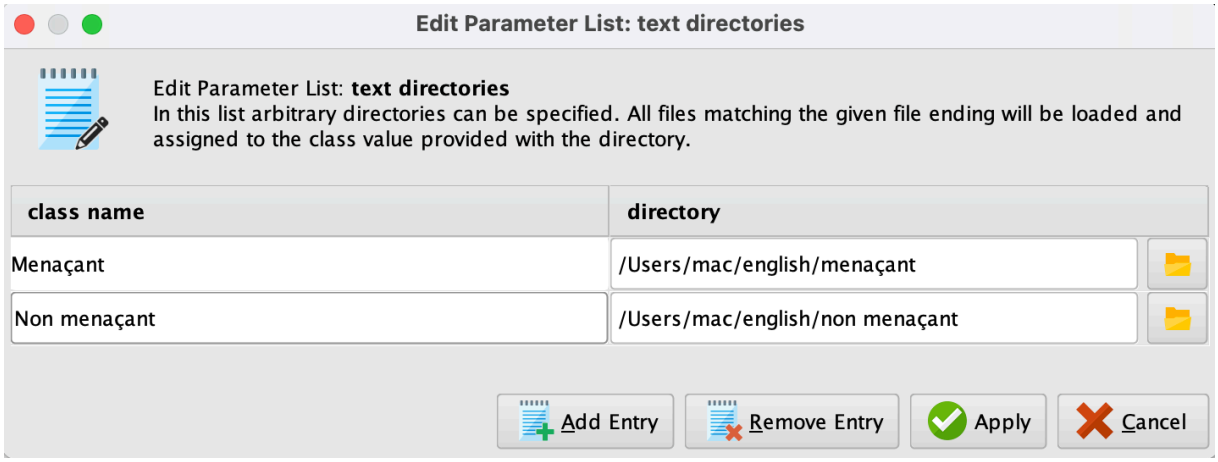


Figure 30. Répartition en 2 classes (Menaçant, Non menaçant)

Une fois que les tweets annotés sont répartis en deux classes, nous procédons à la classification en faisant une cross-validation pour répartir les données en plis (folds) de données (des plis de données d'apprentissage et des plis de données de tests). Le modèle de classification est formé sur des combinaisons aléatoires des plis des données d'apprentissage, puis testé sur des plis des données de tests. On répète ce processus en boucle (epochs) pour avoir un modèle robuste et fiable. La figure 31 suivante présente l'utilisation de la cross-validation sur nos données.

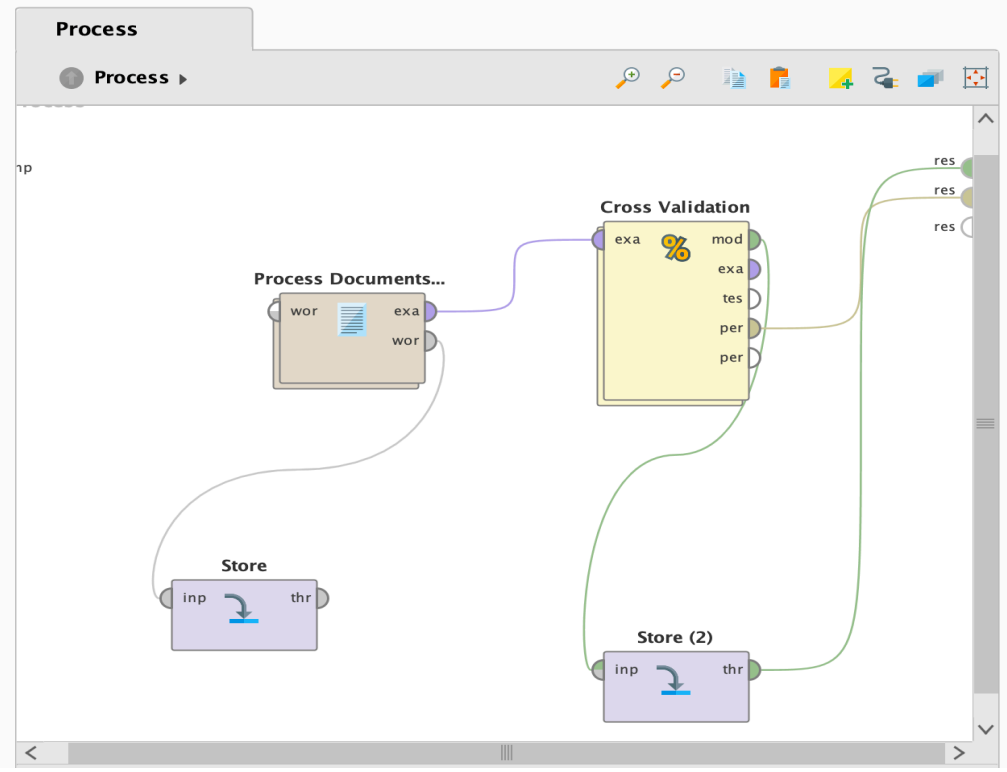


Figure 31. Lecture de documents textuels et application de la Cross-Validation

Dans notre expérimentation, nous avons exploré différentes valeurs pour le nombre de plis (folds) de la validation croisée (paramètre *k*). Il est important de souligner que la valeur du paramètre *k* dépend de la taille de l'ensemble de données, ainsi que l'objectif de la classification. Nous avons trouvé que *k=5* donne les meilleurs résultats.

La figure 32 suivante illustre la création du modèle de classification dans la phase d'apprentissage à gauche de la figure, puis l'application du modèle créé avec le calcul de performances sur les données de test à droite de la figure.

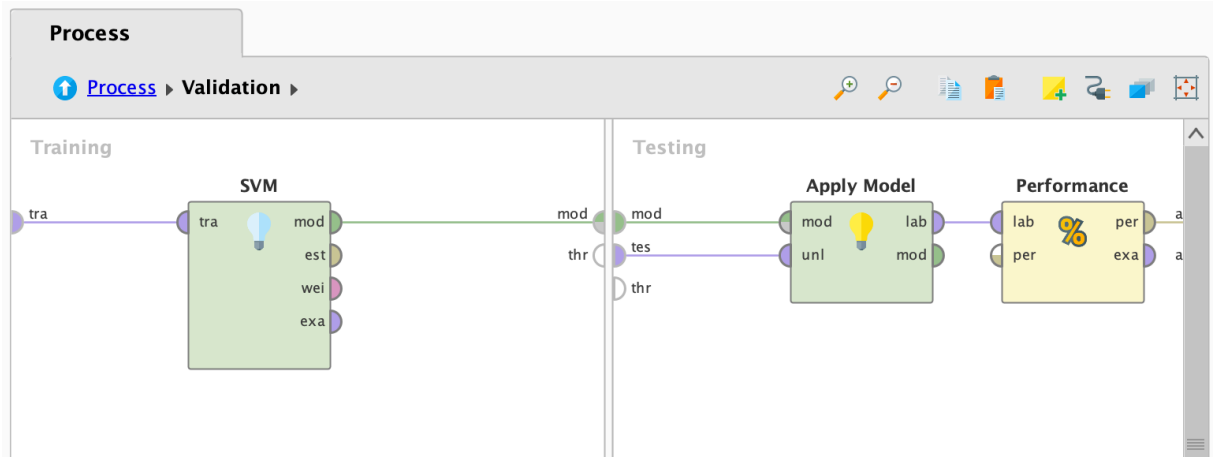


Figure 32. Exemple d'apprentissage Machine avec l'algorithme « SVM »

Pour ce qui est des performances du modèle, nous avons choisi les métriques : Accuracy, précision, rappel et F-measure. Les résultats obtenus sont présentés dans la section suivante.

3.4. Résultats obtenus

Nous présentons dans cette section les résultats obtenus après implémentation des deux scénarios d'évaluation. La figure 33 illustre la matrice de confusion obtenue avec l'implémentation du scénario 1 avec 4000 tweets en Anglais.

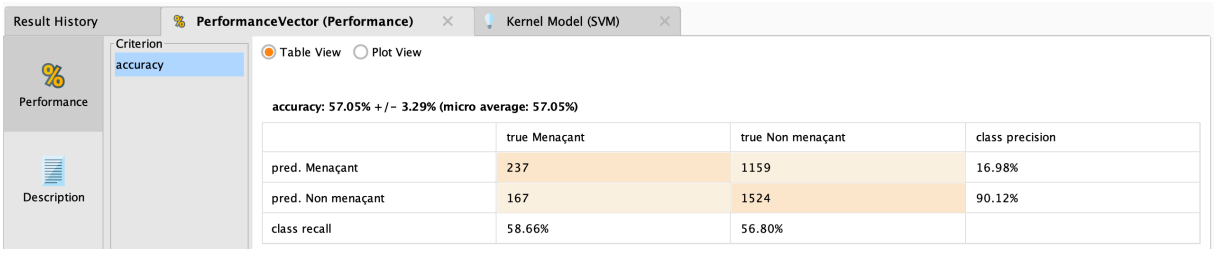


Figure 33. Matrice de confusion obtenue sur 4000 tweets en Anglais

Comme nous pouvons le voir sur la figure 33, le taux de rappel pour la classe Menaçant est de 58.66 %, ce qui veut dire que sur les 404 tweets Menaçants, notre modèle a bien classé 237 tweets (vrais positifs) et a mal classé 167 tweets (faux négatifs). Le taux de précision pour la classe Menaçant est de 16.98%, ce qui veut dire que notre modèle a bien classé 237 tweets et a mal classé 1159 tweets dans cette classe (faux positifs).

Les résultats de performances incluant la f-measure et l'accuracy sont présentés dans les tableaux 14, 15, 16 et 17 ci-après.

Tableau 14. Résultats obtenus sur 4000 tweets en Anglais

English	Rappel	Précision	F-measure	Accuracy
Classe menaçant	0.59	0.17	0.26	0.57
Classe non menaçant	0.57	0.9	0.70	

Tableau 15. Résultats obtenus sur 4000 tweets en Arabe

Arabe	Rappel	Précision	F-measure	Accuracy
Classe menaçant	0.75	0.29	0.42	0.67
Classe non menaçant	0.66	0.93	0.77	

Tableau 16. Résultats obtenus sur 4000 tweets bilingue

Bilingue	Rappel	Précision	F-measure	Accuracy
Classe menaçant	0.77	0.16	0.26	0.38
Classe non menaçant	0.32	0.89	0.47	

Tableau 17. Résultats obtenus sur 12000 tweets

Corpus complet	Rappel	Précision	F-measure	Accuracy
Classe menaçant	0.82	0.14	0.24	0.25
Classe non menaçant	0.16	0.84	0.27	

Nous remarquons d'abord d'une manière générale que les performances obtenues dans cette 2ème contribution sont largement supérieures à celles de la 1ère contribution. Les taux de rappels ont grimpé de 0.017 à 0.58 pour l'anglais, et de 0.018 à 0.70 pour l'arabe. Même chose pour les taux de la F-measure qui sont passés de 0.033 à 0.48 pour l'anglais et de 0.035 à 0.59 pour l'arabe. Ceci s'explique par le fait d'avoir utilisé l'apprentissage machine, en entrainant des données et en créant un modèle de classification binaire avec SVM.

Bien que ces taux soient améliorés, ils restent tout de même non suffisants. Pour la classe *Menaçant*, les taux de rappels varient entre 0.59 à 0.77 seulement, ce qui signifie que près d'un tiers des données qui représentent de vraies menaces terroristes ne sont pas détectés.
Nous remarquons qu'il y un grand écart dans les taux de précisions obtenues entre les deux classes *Menaçant* et *Non Menaçant,* avec un écart de 0.73 pour l'anglais, et de 0.70 pour l'arabe. Ceci peut s'expliquer par le fait d'un déséquilibre de données entre les deux classes. La classes *Non Menaçant* est plus majoritaire que la classe *Menaçant*.

Nous constatons également que les résultats obtenus sur les données bilingues sont moins performants que ceux obtenus sur chaque langue séparée, et ce dans les deux scénarios. Sur les données bilingues, nous avons obtenus les taux les plus faibles, avec une accuracy de 0.38 seulement, un rappel de 0.32, une précision de 0.16 et une f-measure de 0.26 dans le 1er scénario. De même pour le 2ème scénario, en faisant un apprentissage automatique sur l'ensemble des tweets (arabe, anglais et bilingue), les taux de performances restent très faibles, avec une accuracy de 0.25 seulement, un rappel de 0.16, une précision de 0.14 et une f-measure de 0.24. Ceci s'explique par le fait que c'est très difficile de traiter des données écrites dans des langues différentes, surtout quand il s'agit de langues complexes comme l'arabe. Les techniques de traitement et d'analyse de texte sont différentes d'une langue à une autre. A titre d'exemple, l'algorithme de lemmatisation utilisé pour l'anglais ne peut pas être utilisé dans l'arabe, et vice versa. Nous constatons, que chaque langue devra être étudiée séparément.

4. Contribution 3 (Classification ternaire avec apprentissage par transfert)

Nous proposons dans cette troisième contribution un processus générique indépendant de la langue pour détecter et classer les apologies des terroristes sur Twitter en trois catégories : apologie, non apologie et neutre.

Notre objectif principal dans cette contribution est de construire un modèle de classification capable de détecter automatiquement les publications qui sont des apologies du terrorisme. Nous utilisons dans cette contribution une classification ternaire, en se basant sur trois classes très répandues dans l'analyse de sentiments à savoir : positive, négative et neutre. Un tweet est considéré comme positif (apologie) s'il représente une véritable apologie du terrorisme, comme l'incitation à des actes de terrorisme, la célébration, la présentation ou le commentaire favorable d'un acte, la justification ou la défense d'un acte ou de ses auteurs, etc. Si un tweet évoque le terrorisme mais ne représente pas une menace ou un danger, il est classé comme négatif (non apologie). Si un tweet n'a aucun rapport avec le terrorisme sous quelque forme que ce soit, il est classé comme neutre.

4.1. Méthodologie

Les éléments clés de la classification de texte consistent à élaborer une structure de données capable de représenter les documents (tel que les tweets) et à construire un classifieur qui pourra prédire avec précision l'étiquette de classe d'un document (positif, négatif ou neutre) comme souligné par AlGhamdi et Khan (AlGhamdi & Khan, 2020). En outre, il est primordial de souligner que la représentation du texte demeure l'un des éléments les plus importants dans le domaine de la classification de documents. La dimension extrêmement élevée des données textuelles est l'un des aspects clés du problème de la classification de textes, surtout pour les langues complexes comme l'arabe (Shang, et al., 2006).

Pour résoudre ces problèmes, et en tenant compte des difficultés de classification des données textuelles dans diverses langues, nous proposons un nouveau processus indépendant de la langue pour la détection automatique et immédiate des apologies des terroristes sur Twitter.

Pour valider notre démarche, nous avons opté pour deux langues d'entrée : l'arabe et l'anglais. Notre choix est justifié par 1) l'anglais est sans doute la langue la plus utilisée dans les réseaux sociaux, et 2) il y a peu de recherches menées en langue Arabe (une langue non latine) et nous aspirons à apporter notre contribution.

Notre approche repose sur l'application des techniques de normalisation du texte, dont la distance de Levenshtein (Konstantinidis, 2005), ainsi que sur des techniques du traitement naturel du langage pour le prétraitement des données. En outre, nous utilisons des algorithmes d'apprentissage automatique et profond pour la classification. L'objectif principal de notre étude est de créer un modèle de classification capable de déterminer si un tweet constitue une véritable apologie du terrorisme ou non.

Les cinq contributions principales de ce travail se résument comme suit :

1. Construction d'un ensemble de données de 12155 tweets bilingues (arabe et anglais) étiquetés manuellement comme apologie du terrorisme (classe positive), pas d'apologie (classe négative), et neutre (classe neutre).
2. Un processus indépendant de la langue pour détecter les apologies au terrorisme sur Twitter.
3. Comparaison des performances de quatre algorithmes d'apprentissage automatique et de cinq algorithmes d'apprentissage profond pour la tâche de détection des apologies du terrorisme.
4. Évaluation du récent modèle de représentation du langage BERT (Devlin, et al., 2018) sur la tâche de détection des apologies du terrorisme.
5. Comparaison des performances de classification entre les deux langues arabe et anglaise, et étude de l'influence de la langue sur la tâche de classification.

La figure 34 ci-après présente les différentes étapes de notre proposition.

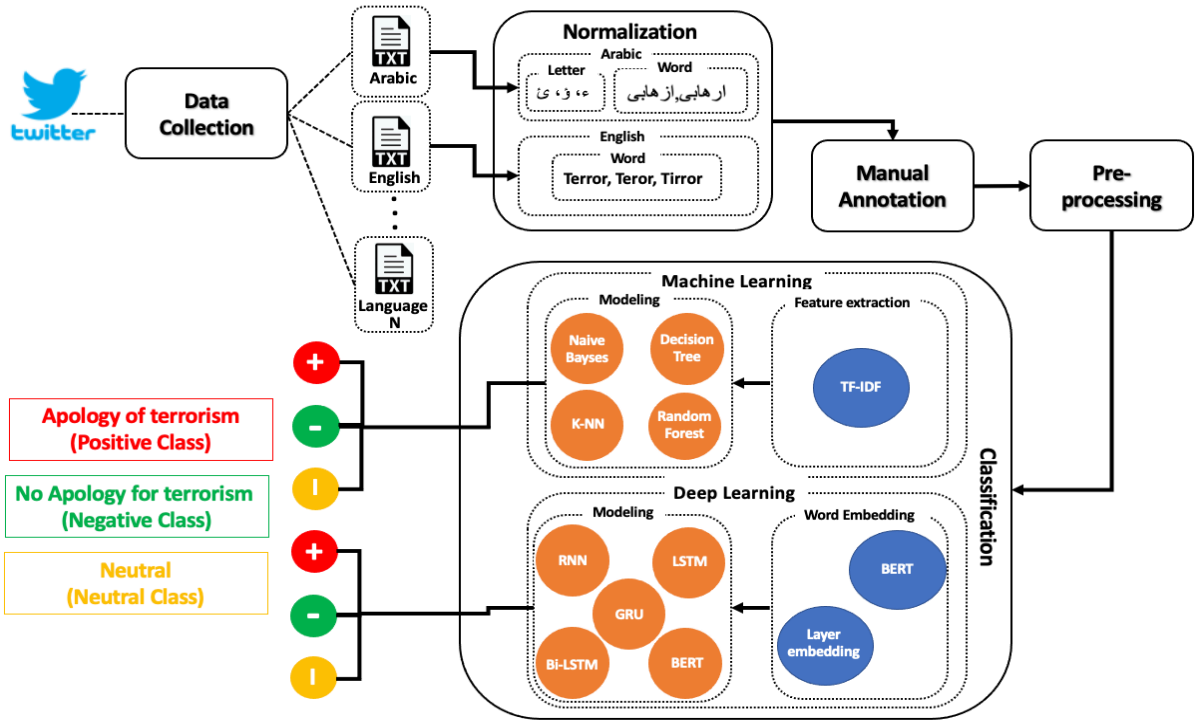


Figure 34. Processus de Langage-indépendant de détection des apologies du terrorisme

Pour mettre notre approche à l'épreuve, nous avons opté pour deux langues de départ, à savoir l'anglais et l'arabe. Ces choix ont été faits suite aux observations suivantes

(1) L'anglais est la langue la plus utilisée dans les réseaux sociaux,

(2) L'arabe a fait l'objet de peu de recherches en raison de sa complexité par rapport à d'autres langues (latines),

(3) Compte tenu de la croissance exponentielle des médias sociaux, les utilisateurs s'exprimant en langue arabe méritent plus d'attention que jamais.

Le processus que nous proposons se compose de 5 étapes que nous résumons dans ce qui suit :

1- Premièrement, en utilisant l'API de Twitter, nous extrayons les tweets qui contiennent des mots liés au terrorisme dans nos langues cibles. Nous organisons l'ensemble des tweets (Dataset) en fichiers texte où chaque fichier contient un seul tweet.

2- Ensuite, nous appliquons des techniques de normalisation et de standardisation sur les tweets pour avoir une seule forme d'écriture pour chaque langue adoptée, mais aussi pour vérifier que tous les mots qui composent un tweet sont dans le dictionnaire de cette langue, sinon nous appliquons une recherche hybride sur les mots inconnus, en utilisant le calcul de la distance de Levenshtein (Konstantinidis, 2005).

3- Ensuite, nous annotons manuellement les tweets en 3 catégories : positif (i.e., le tweet représente une véritable apologie du terrorisme), négatif (i.e., le tweet ne représente pas une apologie du terrorisme), neutre (i.e., le tweet n'a aucun rapport avec le terrorisme).
4- Une fois l'annotation terminée, nous utilisons des techniques NLP (tokenisation, lemmatisation et suppression des mots vides) pour prétraiter les tweets afin de filtrer et normaliser le contenu du jeu de données.
5- Après les étapes de normalisation, d'annotation et de prétraitement décrites ci-dessus, nos tweets sont maintenant prêts à être utilisés pour le benchmarking de différentes approches de classification. Plus précisément, nous avons comparé 4 algorithmes ML (RF, DT, NB et KNN) utilisant TF-IDF comme technique d'extraction de caractéristiques, et 5 algorithmes DL (GRU, SimpleRNN, LSTM, BiLSTM et BERT) utilisant la technique d'incorporation de couches pour cette extraction, à l'exception de BERT qui utilise ses propres techniques pour l'initialisation des neurones d'entrées de la phase d'extraction des de caractéristiques.

Nous avons testé notre proposition sur un ensemble de données bilingue (arabe et anglais) de 12155 tweets annotés manuellement. Nous avons mené deux séries d'expériences, l'une avec des données déséquilibrées et l'autre avec des données équilibrées (sur échantillonnées). Nous avons comparé les performances de classification des 9 classifieurs cités précédemment. Les résultats obtenus sont donnés dans les sections suivantes.

4.2. Évaluation

Pour évaluer notre proposition, nous avons appliqué les 5 premières étapes de notre processus décrites précédemment sur notre jeu de données. Après l'étape de collecte, nous avons obtenu un total de 12155 tweets (en arabe et en anglais), qui ont ensuite été annotés manuellement en 3 classes (positif, négatif, ou neutre). La figure 35 montre la répartition de ces tweets entre les 3 classes. Les détails de l'annotation (nombre de tweet par classe par mot clé) sont déjà présentés dans les tableaux 10 et 11 (Pages 52 & 53).

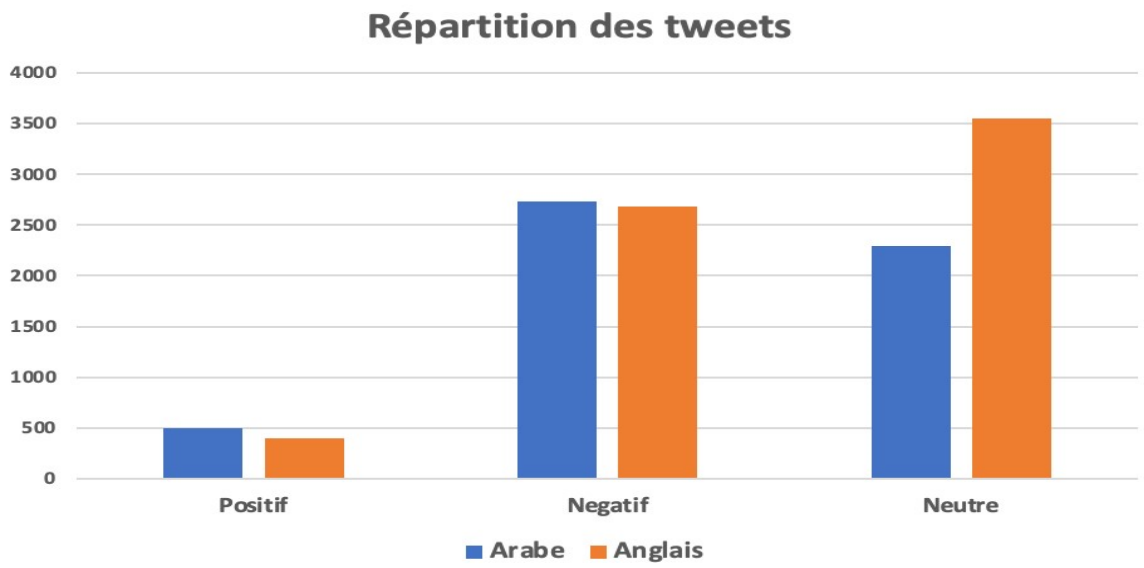


Figure 35. Répartition des tweets par classe

Comme on peut le voir sur la figure 35, la distribution des tweets est déséquilibrée entre les trois classes, tant en arabe qu'en anglais. En effet, les tweets annotés comme positifs sont environ 5 fois moins nombreux que ceux annotés comme négatifs. Cette disparité peut avoir un impact sur les performances de la classification. Par conséquent, nous avons choisi de mener deux séries d'expériences : l'une avec les données déséquilibrées, et l'autre avec les données équilibrées en utilisant le sur échantillonnage sur la classe positive.

Étant donné que notre jeu de données est de taille réduite, nous avons opté à exploiter et tester des modèles préalablement construits et entraînés sur de gros ensembles de données. Nous avons donc choisi d'utiliser BERT avec un modèle déjà entraîné sur des millions de données. L'un des principaux avantages de BERT est qu'il peut être adapté à des domaines spécifiques et entraîné sur un certain nombre de tâches différentes. Cela signifie qu'étant donné que BERT a été si bien entraîné au préalable, il peut être appliqué à de petits ensembles de données tout en ayant de bonnes performances. Nous avons utilisé la bibliothèque Transformers de la plateforme Hugging-Face[12] pour y parvenir. Ainsi, nous avons effectué le processus de réglage fin sur nos données (fine-tunning).

Il faut noter qu'il y a des dizaines de modèles déjà entraînés avec BERT sur la plateforme Hugging-Face. Nous avons opté pour le modèle ArabicBert d'ASafaya 'asafaya/bert-base-arabic' (Safaya, et al., 2020) pour l'arabe comme il a été entraîné sur un corpus de 8,2 milliards

[12] https://huggingface.co

de mots qui a été proposé pour identifier les discours offensifs dans les médias sociaux, ce qui est proche de notre domaine d'étude. Quant à l'anglais, nous avons opté pour le modèle de base BERT 'bert-base-uncased' proposé par (Devlin, et al., 2018), formé et entrainé sur un ensemble de données comprenant le contenu de l'encyclopédie Wikipedia en anglais, accompagné de plus de 11000 ouvrages anglais.

Les expériences menées ainsi que les résultats obtenus sont présentés dans ce qui suit.

a. Protocole d'expérimentation

Pour chaque série d'expériences, nous avons utilisé le même jeu de données, et comparé les résultats de classification entre les 4 classifieurs d'apprentissage automatique et les 5 classifieurs d'apprentissage profond.
Pour comparer les performances de ces classifieurs, nous avons utilisé la validation croisée (avec k=5) pour les algorithmes d'apprentissage automatique et la division Train-Test de Scikit-learn pour les algorithmes d'apprentissage profond.

b. Métriques de performances

Dans les deux scénarios (déséquilibré et sur-échantillonné), les performances des 9 classifieurs ont été comparées en utilisant les métriques communément adoptées pour ce type de tâches, à savoir : accuracy, f-measure, rappel, et précision. Ces métriques de performances sont décrites dans le chapitre 2. Les résultats de nos expériences sont résumés dans le tableau 18 pour les données déséquilibrées, le tableau 21 pour les données sur échantillonnées, et dans la figure 36 (page 85) pour l'accuracy de tous les classifieurs dans les deux expériences.

4.3. Résultats obtenus

Nous présentons dans cette section les résultats obtenus dans les deux séries d'expérimentations. Nous commençons par les données déséquilibrées, puis les données sur échantillonnées.

4.3.1. Résultats obtenus sur les données déséquilibrées

Les résultats de performances de la classification en termes de rappel, précision et f-measure de chaque classifieur sont présentés dans le tableau 18. Les meilleures performances sont indiquées en caractères gras souligné.

Tableau 18. Résultats sur des données déséquilibrées

			Precision		**Recall**		**F-measure**	
		Class	**Arabic**	**English**	**Arabic**	**English**	**Arabic**	**English**
Machine Learning	**RF**	Positive	0.43	**0.60**	0.02	**0.74**	0.04	**0.66**
		Negative	0.55	**0.80**	**0.98**	0.38	0.71	0.52
		Neutral	**0.93**	0.64	0.25	**0.97**	0.40	0.77
	DT	Positive	0.53	0,46	0.13	0,03	0.21	0,06
		Negative	0.60	0,79	0.61	0,32	0.61	0,45
		Neutral	0.55	0,62	0.63	**0,97**	0.58	0,76
	NB	Positive	0.41	0,30	0.47	0,10	0.44	0,15
		Negative	0.78	0,74	0.63	0,65	0.70	0,69
		Neutral	0.70	0,76	0.84	0,88	0.76	0,81
	KNN	Positive	0.16	0,24	0.53	0,13	0.24	0,17
		Negative	0.61	0,58	0.65	0,63	0.63	0,6
		Neutral	0.67	0,69	0.27	0,68	0.38	0,68
Deep Learning	**GRU**	Positive	0.39	0.30	0.38	0.29	0.38	0.29
		Negative	0.71	0.71	0.76	0.72	0.74	0.72
		Neutral	0.78	0.83	0.72	0.77	0.75	0.80
	BERT	Positive	**0.58**	0,47	**0.56**	0,32	**0.57**	0,38
		Negative	**0.84**	0,76	0.86	**0,81**	**0.85**	**0,78**
		Neutral	0.91	**0,88**	**0.88**	0,87	**0.90**	**0,87**
	SimpleRNN	Positive	0.36	0,22	0.34	0,32	0.35	0,26
		Negative	0.75	0,64	0.64	0,6	0.69	0,62
		Neutral	0.72	0,77	0.79	0,73	0.76	0,75
	LSTM (1 layer)	Positive	0.44	0,57	0.43	0,23	0.44	0,33
		Negative	0.79	0,76	0.76	0,79	0.77	0,77
		Neutral	0.84	0,85	0.81	0,83	0.82	0,84
	BiLSTM	Positive	0.56	0,48	0.39	0,26	0.46	0,34
		Negative	0.76	0,77	0.84	0,77	0.80	0,77
		Neutral	0.87	0,85	0.77	0,84	0.82	0,84

Comme on peut le voir dans le tableau 18, les mesures de performance de la classe positive en arabe et en anglais sont relativement faibles par rapport aux autres classes ; la meilleure F-measure est seulement de 0,57 pour les données arabes obtenues par BERT, et de 0,66 pour les données anglaises obtenues par Random Forest. Alors que les classes négatives et neutres atteignent des F-measures allant jusqu'à 0,78 et 0,90. Cela peut s'expliquer par le fait que les

données sont fortement déséquilibrées (peu de tweets annotés comme positifs), ce qui a eu un impact sur les résultats de performance.

On constate que BERT obtient les meilleurs résultats en arabe. Cela est dû au fait que BERT apprend le contexte d'un mot à partir de tout son environnement (à gauche et à droite du mot), ainsi qu'au fait que nous avons utilisé un modèle qui avait déjà été entraîné avec BERT sur un ensemble de données volumineux.

La performance la plus faible est obtenue par DT, avec une f-measure de 0.21 en arabe et seulement 0.06 en anglais, pour la classe positive. Cela peut s'expliquer par le fait que cet algorithme n'est pas adapté aux données déséquilibrées, car les points de séparation de l'arbre sont conçus pour séparer au mieux les échantillons en groupes équilibrés avec le moins de mélange possible (Haixiang, et al., 2017). Lorsque les données sont déséquilibrées, les instances de la classe minoritaire (classe positive dans notre étude) sont ignorées.

4.3.2. Résultats obtenus sur des données équilibrées (sur-échantillonnées)

Dans cette expérience, nous avons appliqué un suréchantillonnage aux données annotées positives pour équilibrer les classes. Nous avons donc augmenté les données de la classe positive en dupliquant les tweets pour chaque collection afin que le nombre de tweets de la classe positive corresponde à celui de la classe négative. Le tableau 19 montre le nombre de tweets, par mot-clé, de la classe positive après le suréchantillonnage.

Tableau 19. Nombre de tweets de la classe positive après sur-échantillonnage

Arabic			English	
Collection	***Meaning in English***	**Positive Class**	**Collection**	**Positive Class**
البغدادي	Albaghdadi	243	Albaghdadi	218
القاعدة	Alqaida	108	Attack	272
شهيد	Martyr	212	Bomb	168
داعش	Daech	774	Daesh	270
حماس	Hamas	127	Hamas	126
حزب الله	Hizbollah	496	Hizbollah	98
ارهاب	Terrorism	236	#IS	116
مفخخ	Booby-trapped	64	Isis	398
نصرالله	Nasrallah	245	IslamicState	272
قصف	Bombing	208	Jihad	312
Total		**2713**	Nasrallah	42
			Terror	380
			Total	**2672**

Nous avons augmenté le nombre de tweets annotés comme positifs afin qu'ils soient en équilibre avec ceux classés comme négatifs. Nous avons réalisé cette étape, en équilibrant chaque collection par mot clé, pour garder l'ensemble de données diversifié, et minimiser ainsi le risque d'un surajustement (overfitting). Pour la classe neutre, comme elle ne représente pas de grand écart des données par rapport à la classe négative, nous avons choisi de ne pas l'augmenter.
Le nombre de tweets par classe dans ces données suréchantillonnées est donnée dans le tableau 20 suivant.

Tableau 20. Nombre de tweets par classe après sur-échantillonnage

	Positive	***Negative***	***Neutral***
Arabic	*2713*	*2729*	*2290*
English	*2672*	*2679*	*3553*

Les 9 algorithmes de classification étudiés ont été exécutés à nouveau sur cet ensemble de données sur échantillonnées. Les résultats de ces expériences sont donnés dans le tableau 21. Les meilleures performances sont indiquées en caractères gras soulignés.

Tableau 21. Résultats sur les données sur-échantillonnées

			Precision		Recall		F-measure	
		Class	Arabic	English	Arabic	English	Arabic	English
Machine Learning	RF	Positive	0.82	0.71	0.49	0.2	0.61	0.31
		Negative	0.45	0.81	0.90	0.09	0.60	0.16
		Neutral	0.90	0.44	0.26	0.97	0.40	0.61
	DT	Positive	0.89	0,93	0.23	0,07	0.37	0,13
		Negative	0.40	**0,85**	**0.97**	0,04	0.57	0,08
		Neutral	**0.92**	0,41	0.13	**0,99**	0.23	0,58
	NB	Positive	0.69	0,73	0.74	0,62	0.72	0,67
		Negative	0.67	0,66	0.52	0,52	0.59	0,58
		Neutral	0.66	0,69	0.80	0,86	0.72	0,76
	KNN	Positive	0.50	0,67	0.98	0,95	0.66	0,79
		Negative	0.78	0,69	0.47	0,49	0.59	0,57
		Neutral	0.69	0,7	0.22	0,63	0.33	0,66
Deep Learning	GRU	Positive	0.88	0.90	**0.99**	0.98	0.93	0.94
		Negative	0.86	0.77	0.76	0.64	0.81	0.70
		Neutral	0.79	0.82	0.79	0.84	0.79	0.83
	BERT	Positive	**0.94**	0,96	**0.99**	**0,99**	**0.97**	**0,97**
		Negative	**0.87**	0,8	0.83	**0,87**	**0.85**	**0,83**
		Neutral	0.84	**0,95**	**0.82**	0,87	**0.83**	**0,91**
	SimpleRNN	Positive	0.92	0,95	0.96	0,98	0.94	0,96
		Negative	0.81	0,74	0.75	0,62	0.78	0,68
		Neutral	0.77	0,77	0.75	0,82	0.76	0,8
	LSTM (1 layer)	Positive	0.91	0,96	0.96	0,92	0.94	0,94
		Negative	0.86	0,78	0.79	0,78	0.83	0,78
		Neutral	0.81	0,86	0.81	0,86	0.81	0,86
	BiLSTM	Positive	0.90	**0,97**	0.96	0,93	0.93	0,95
		Negative	**0.87**	0,76	0.78	0,79	0.82	0,78
		Neutral	0.82	0,86	0.81	0,85	0.81	0,86

Dans le tableau 21, nous constatons une augmentation significative des performances pour la plupart des algorithmes de classification. Cela est dû au fait que nous avons augmenté le nombre de tweets dans la classe positive, ce qui a finalement influencé les différentes métriques de performances.

Avec une différence significative, les algorithmes d'apprentissage profond surpassent les algorithmes d'apprentissage automatique en arabe et en anglais. Comme indiqué

précédemment, contrairement aux algorithmes d'apprentissage automatique, les algorithmes d'apprentissage profond ne nécessitent pas de méthodes d'extraction de caractéristiques. Ils s'entraînent plutôt à extraire les aspects pertinents de la détection que nous voulons faire.

En ce qui concerne la classe positive, une amélioration remarquable de toutes les mesures (supérieure à 0,90 pour tous les algorithmes d'apprentissage profond), comme le montre le tableau 21. La meilleure F-measure de cette expérience a été atteinte par BERT (0,97 en arabe et en anglais). En d'autres termes, ce classifieur est capable de détecter 94% des tweets menaçants avec seulement 6% de faux positifs. Cette augmentation peut cependant s'expliquer par le sur-échantillonnage des données en augmentant la classe positive, ce qui a eu un impact sur les résultats de performance.

En outre, BERT a également obtenu le meilleur taux de rappel avec 0,99 en arabe et en anglais, suivi de GRU avec respectivement 0,99 et 0,98. Cela signifie que 99 % des tweets d'apologie du terrorisme ont été repérés et détectés. Il s'agit d'un excellent résultat, proche de la détection parfaite ou complète, qui est conforme à notre objectif de détecter tous les tweets qui peuvent être des apologies au terrorisme.

Certains des résultats obtenus par DT sont nettement améliorés par rapport à la première expérience. Cela est dû au fait que la classe positive, qui était minoritaire dans la première expérience, a presque la même taille que la classe négative dans l'expérience suréchantillonnée (c'est-à-dire qu'il n'y a plus de classe minoritaire), ce qui permet à l'algorithme d'obtenir de meilleures performances.

4.3.3. Accuracy des données déséquilibrées et sur-échantillonnées

L’exactitude (accuracy) est la mesure de performance qui calcule le rapport entre les observations correctement prédites et toutes les observations. Un modèle est optimal s'il est précis (proche de 1). La figure 36 montre les résultats de performance de l’exactitude pour les deux expérimentations que nous avons menées, couvrant les deux langues (Arabe et Anglais), et ce, pour chaque classifieur.

Figure 36. Accuracy sur les données déséquilibrées et sur-échantillonnées

Nous constatons sur la figure 36 que la meilleure exactitude est obtenue avec BERT : 0,84 pour l'arabe et 0,81 pour l'anglais sur des données déséquilibrées, et 0,88 pour l'arabe et 0,91 pour l'anglais sur des données sur échantillonnées.

Les quatre autres méthodes d'apprentissage profond (GRU, Simple RNN, LSTM et BiLSTM) obtiennent des résultats assez similaires en arabe avec une précision de 0,76 à 0,79 pour les données déséquilibrées, et de 0,86 à 0,87 pour les données équilibrées. Un comportement similaire peut être observé sur les données en anglais avec une précision de 0,79 à 0,80 pour les données déséquilibrées, et de 0,84 à 0,86 pour les données équilibrées.

De plus, nous observons que les performances des classifieurs d'apprentissage automatique (RF, DT, NB et KNN) sont significativement inférieures à celles des classifieurs d'apprentissage

profond. En effet, pour l'arabe, les mesures de précision vont de 0,48 à 0,70 pour les données déséquilibrées et de 0,46 à 0,68 pour les données équilibrées. En ce qui concerne l'anglais, les mesures de précision vont de 0,63 à 0,74 pour les données déséquilibrées et de 0,43 à 0,69 pour les données équilibrées. Cela s'explique par les mêmes raisons que celles évoquées précédemment sur l'extraction de caractéristiques.

Nous constatons que tous les classifieurs ont atteint une meilleure précision avec des données sur échantillonnées qu'avec des données déséquilibrées, à l'exception des trois classifieurs RF, NB et DT. La figure 36 montre qu'en général tous les classifieurs ont des performances assez similaires pour les deux langues (arabe et anglais). En effet, malgré une très légère différence, ils semblent ne pas être affectés par la langue d'entrée dans les deux scénarios (déséquilibré et sur échantillonné).

Nos expériences relèvent également que BERT surpasse presque tous les autres algorithmes en termes de performances. Ceci est dû au fait que BERT est non séquentiel, c'est-à-dire que les phrases sont traitées dans leur ensemble plutôt que mot par mot ; il a une meilleure compréhension du contexte (plus fine et sur une fenêtre bidirectionnelle beaucoup plus grande) ainsi qu'une meilleure compréhension de l'importance de l'ordre des mots (utilisation de poids fixes ou appris qui encodent des informations liées à une position spécifique d'un jeton dans une phrase pour remplacer la récurrence).
Les résultats obtenus par BERT peuvent également être expliqués par le fait que nous avons utilisé un réglage fin (fine-tunning). En effet, nous avons pris un modèle qui avait été entraîné avec BERT sur un grand corpus de données non étiquetées (ce qui nous a permis d'acquérir une meilleure compréhension de la langue) et nous l'avons affiné avec une seule couche de sortie supplémentaire pour créer un modèle adapté à notre étude de cas " apologies du terrorisme ".

Avec toutes ces observations, nous pouvons résumer nos constats en deux points essentiels :
a. Pour les jeux de données de petite taille, il est préférable d'utiliser un modèle déjà entraîné sur jeu de données volumineux plutôt que de construire un nouveau modèle à partir de zéro.
b. Avoir plus des documents annotés (tweets dans notre cas) dans la classe positive, en l'équilibrant avec la classe négative, améliore significativement la précision de la détection.

5. Synthèse et comparaison

En guise de synthèse de nos trois contributions, nous procédons à une comparaison des résultats des **meilleures** performances entre elles, ainsi qu'en les mettant en perspectives par rapport à d'autres travaux de recherche similaires. Il est important de rappeler qu'il est difficile de faire une comparaison directe avec les travaux de la littérature en raison de l'absence des jeux de données standardisés. Dans le tableau 22 suivant, nous présentons une synthèse de nos contributions et une comparaison avec quelques travaux similaires.

Tableau 22. Synthèse et comparaison entre travaux de recherche

	Travail	Jeu de données	Modèle	Accuracy	F-measure
Notre travail	Contribution1 (Recherche lexicale) **(Bedjou, et al., 2018)**	4000 tweets	Apache Lucene	-	0.035
	Contribution 2 (Classification binaire) **(Bedjou, et al., 2019)**	4000 tweets arabes 4000 tweets anglais 4000 tweets bilingue	TF-IDF avec SVM	0.67	0.77
	Contribution 3 (Classification ternaire) **(Bedjou & Azouaou, 2023)**	12155 tweets arabes et anglais	BERT	0.91	0.97
(Mazari & Kheddar, 2023)	Détection des messages haineux en dialecte algérien	14150 commentaires	FastText avec Bi-GRU	0.74	0.76
(Al-Hassan & Al-Dossari, 2022)	Détection des messages haineux en arabe avec Deep learning	11000 tweets arabes	LSTM avec CNN	-	0.73
(b-Aldera, et al., 2021)	Détection de l'extrémisme dans les réseaux sociaux	89,816 Arabic Tweets labelled as extremist or non-extremist	BERT	0.9749	-
(Alshalan & Al-Khalifa, 2020)	Détection des messages haineux sur Twitter à l'Arabie saoudite	Arabic dataset that contained 9316 annotated tweets	CNN	-	0,79

Comme on peut le constater à partir du tableau 22, les résultats de performances de chacun des travaux présentés dépendent des ensembles de données utilisés. Globalement, nous pouvons

dire que notre troisième contribution est alignée avec les autres travaux de recherche, en apportant une valeur significative grâce à la création du jeu de données créé comprenant plus de 12.000 tweets annotés manuellement dans deux langues différentes. De plus, nous avons réalisé une comparaison de performances de 9 classifieurs, en mettant en évidence l'efficacité particulière de l'apprentissage par transfert (transfer learning) en utilisant un modèle déjà pré-trainé avec BERT sur des millions de données. Nous espérons que la publication de notre ensemble de données, accompagnée de nos résultats de performance, servira de point de départ pour les futures recherches dans le domaine de la détection d'apologies ou de menaces terroristes. De plus, cela offrira aux chercheurs la possibilité d'utiliser nos jeux de données et de réaliser des comparaisons de résultats de manière aisée.

6. Conclusion

Dans ce chapitre, nous avons présenté les trois contributions essentielles de notre thèse. Dans chaque contribution, nous avons détaillé l'approche utilisée, les ensembles de données (data sets), les différentes techniques, algorithmes et classifieurs.

De plus, nous avons présenté les expérimentations que nous avons menés pour chacune des contributions, en fournissant une détaillé et discuté les résultats obtenus dans chacune d'entre elles.

Nous avons remarqué que l'utilisation des techniques de représentation de texte basées sur le contexte, combinés aux algorithmes de l'apprentissage profond, améliore significativement les résultats de performances de la classification.

Enfin, nous avons constaté que l'utilisation de BERT avec l'apprentissage par transfert (transfer learning) avec un modèle déjà pré-trainé sur un ensemble de données volumineux, en ajoutant une couche de nos propres données (BERT fine-tuning) s'est avéré le meilleur moyen pour détecter efficacement les menaces terroristes sur les réseaux sociaux.

Conclusion générale

Les réseaux sociaux sont des plateformes web rassemblant des identités sociales telles que des individus, des entreprises et des organisations, favorisant les échanges d'informations par le biais d'interactions sociales. Ces plateformes sont utilisées par des millions de personnes dans le monde, pour exprimer leurs opinions, rechercher des informations, partager leur quotidien, des actualités, etc. Ainsi, des masses gigantesques de données sont partagées et échangées chaque jour, ce qui rend leur gestion et contrôle particulièrement complexes. En effet, ces plateformes contiennent souvent des contenus violents, racistes, sexistes, des cas d'harcèlement, voir même des menaces, qui se propagent sans qu'ils soient détectés automatiquement. Cela pose un véritable danger, et une lacune en matière de sécurité, particulièrement lorsque ces contenus sont accessibles et visibles par des individus mineures ou vulnérables. Cette thèse aborde spécifiquement la problématique des menaces terroristes sur ces plateformes, se concentrant sur les menaces textuelles telles que les publications, commentaires, réponses, tweets, etc.

Afin d'apporter une solution à la problématique posée, nous avons entamé une étude approfondie des différentes techniques de représentation de texte. Nous avons minutieusement comparé leurs avantages et inconvénients pour choisir les approches les plus pertinentes à intégrer dans nos propositions. Ensuite, nous avons réalisé une revue des travaux de recherche portant sur la détection des menaces terroristes, l'analyse de sentiment, ainsi que la détection des discours haineux sur les réseaux sociaux. Cette revue a exploré les techniques, les approches, les algorithmes de classification et les diverses méthodes d'extraction de caractéristiques utilisées. Une analyse attentive des résultats obtenus ainsi que des limites de ces études a été conduite afin d'élaborer des solutions plus performantes. Ainsi, nous avons proposé 3 contributions de recherche comme suit.

Contribution 1 : nous avons proposé LexD3T, un processus de recherche lexicale des menaces terroristes sur le réseau social Twitter. Ce processus est utilisé pour analyser et détecter les contenus associés au terrorisme. Le processus débute avec en analysant les publications (tweets) à l'aide les techniques NLP (Tokenisation, Lemmatisation et Suppression de mots vides). Nous

avons ensuite indexé le corpus traité en utilisant l'API Apache Lucene, pour finalement détecter les tweets susceptibles de représenter des menaces terroristes. Nous avons expérimenté cette proposition sur un corpus de 4000 tweets. Les résultats obtenus en termes de f-measure varient entre 0.033 à 0.035. Ces résultats, très faibles, soulignent que la recherche lexicale manque de précision et génère un bruit trop important (tweets détectés lors de la recherche, mais qui ne représentent pas de vraies menaces).
Nous avons constaté que, bien que la recherche lexicale puisse repérer tous les contenant un vocabulaire lié au terrorisme, elle ne suffit pas pour distinguer les vraies menaces.

Contribution 2 : nous avons proposé un processus utilisant SVM pour détecter et classifier les menaces terroristes sur le réseau social Twitter. Ce processus analyse et détecte les contenus liés au terrorisme en traversant 12 étapes. Les 5 premières relèvent du domaine du traitement automatique de la langue (NLP), tant dis que les 7 suivantes relèvent du domaine l'apprentissage machine. Cette classification est binaire, elle permet de classer les tweets en deux catégories : *Menaçant* ou *Non menaçant*.
Nous avons développé deux scénarios pour l'évaluation de ce processus. Le premier concerne l'apprentissage de 4.000 tweets, séparés en ensembles anglais, arabe, et bilingue. Pour chaque ensemble de tweets, un modèle de classification est créé, et des taux de rappel et précision sont calculés. Par comparaison avec les résultats de notre première contribution, où une approche lexicale a été utilisée, toutes les performances montrent une nette amélioration avec des taux de f-measure qui varient entre 0.26 à 0.77. Cette amélioration est justifiée par le fait d'utiliser l'apprentissage automatique en entrainant des données et en créant un modèle de classification binaire avec SVM. Quant au 2ème scénario, l'apprentissage se fait sur les 12000 tweets à la fois, ce qui a permis d'obtenir un modèle différent de ceux du 1er scénario. Cette comparaison a permis d'évaluer l'impact de la taille du corpus sur l'apprentissage utilisant les SVM et a donné un aperçu du comportement de ces SVM lorsqu'ils traitent un corpus multilingue. En effet, les résultats obtenus dans ce scénario sont relativement faibles, avec une f-measure de 0.24 à 0.27. Cette diminution de performances être à la fusion de données provenant de deux langues complétement différentes.

Contribution 3 : nous avons étudié le problème de la détection des apologies du terrorisme sur les médias sociaux en utilisant Twitter comme source. Pour ce faire, nous avons proposé un

processus indépendant de la langue, s'appuyant sur des techniques de traitement automatique de la langue du langage naturel (NLP). Ce processus débute par la collecte de tweets à l'aide de mots-clés liés au terrorisme, suivi d'une phase de normalisation. Une fois les tweets normalisés, nous les avons manuellement annotés en trois classes : Positif, Négatif et Neutre. Ensuite, et afin de rendre les tweets annotés facilement traitables par la machine, nous avons appliqué 3 techniques NLP : la tokenisation, la lemmatisation et la suppression de mots vides. La dernière étape de notre processus implique l'utilisation et la comparaison des performances de 4 algorithmes d'apprentissage automatique (RF, DT, NB et KNN) avec TF-IDF comme technique d'extraction de caractéristiques, et de 5 algorithmes d'apprentissage profond (GRU, Simple RNN, LSTM, BiLSTM et BERT) avec l'intégration de couches (layer embedding) et BERT comme techniques de représentation du texte.

Pour évaluer notre proposition, une première série d'expériences est menée impliquant les 9 classifieurs mentionnés sur un ensemble de 12155 tweets annotés manuellement. Cependant, comme cet ensemble de données est déséquilibré (seul un cinquième des tweets est annoté comme positif), nous avons mené une deuxième série d'expériences sur un ensemble de données sur échantillonné (équilibré).

Les résultats de ces expériences montrent que BERT est clairement le plus performant parmi les classifieurs étudiés. En effet, avec les tweets en arabe, BERT atteint une précision de 0,84 et un f-score de 0,90 sur des données déséquilibrées, et une précision de 0,88 et un f-score de 0,97 sur des données sur échantillonnées. Des performances similaires ont été observées avec BERT sur les tweets en anglais également (précision : 0,81, f-score : 0,87 sur les données déséquilibrées, et précision : 0,91, f-score : 0,97 sur les données sur échantillonnées).

Nos recherches ont validé certains résultats existants précédemment publiés, tout en apportant de nouvelles observations, comparaisons et résultats. Dans l'ensemble, nos études dégagent au moins trois conclusions essentielles :

(1) Nous avons noté que l'augmentation des tweets classés comme positifs (c'est-à-dire les tweets représentant une apologie du terrorisme) améliore considérablement les performances de détection en particulier en ce qui concerne le rappel.

(2) De meilleures performances peuvent être obtenues en utilisant un modèle pré-entraîné, issus d'un grand corpus, affiné avec des données spécifiques à un domaine (détection d'apologies du terrorisme dans notre cas) plutôt qu'en créant un nouveau modèle à partir de zéro.

(3) L'arabe est sans aucun doute une langue complexe, mais sa manipulation peut être améliorée en standardisant ses lettres et ses mots et en utilisant les techniques NLP appropriées. En conséquence, les performances obtenues sont très satisfaisantes et se rapproche étroitement de de celles obtenues pour l'anglais.

Nous estimons que nos résultats et le jeu de données construit, peuvent servir de point de départ pour des recherches poussées ou approfondies dans le domaine des menaces terroristes sur les médias sociaux. En perspectives de notre travail, nous avons l'attention d'explorer et de comparer d'autres modèles de représentation de texte comme FastText de Facebook (Bojanowski, et al., 2017), avec d'autres classifieurs tels que CNN ou SVM-multi classes. De plus, il serait intéressant d'explorer les techniques des grands modèles linguistiques LLM (Larage Language Models) comme Jais, Llama2 ou GPT (Radford, et al., 2018). Nous envisageons également d'étudier des aspects spécifiques à la langue en considérant les publications écrites dans des dialectes (Ouchene & Bessou, 2023) , ou dans le langage des SMS et tchat comme l'Arabizi (Guellil, et al., 2017). En outre, nous projetons d'appliquer notre processus sur des documents écrits dans d'autres langues comme le français, l'espagnol, l'italien... en développant des modules de normalisation plus raffinés et spécifiques à chaque langue.

Références Bibliographiques

a-Aldera, S. et al., 2021. Online Extremism Detection in Textual Content: A Systematic Literature Review. *IEEE Access,* pp. vol. 9, p. 42384-42396.

Al-Azani, S. & El-Alfy, E. S. M., 2017. Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text.. *Procedia Computer Science,* pp. 109, (pp. 359-366).

AFIA, 2023. https://afia.asso.fr/domaines-de-lia/, consulté le 10.03.2023

AlGhamdi, M. A. & Khan, M. A., 2020. *Intelligent Analysis of Arabic Tweets for Detection of Suspicious Messages.* s.l., Springer.

Al-Hassan, A. & Al-Dossari, H., 2022. Detection of hate speech in arabic tweets using deep learning,. *Multimedia Systems,* Volume 28(6), pp. pp. 1963-1974.

Al-Khafaji, D. H. K. & Habeeb, A. T., 2017. Efficient algorithms for preprocessing and stemming of tweets in a sentiment analysis system. *IOSR J. Comput. Eng., Vol. 19, No. 3,* p. pp.44–50.

Al-Sanabani, M. & Al-Hagree, S., 2015. Improved an algorithm for Arabic name matching. *Open Transactions on Information Processing,* pp. 2374-3778.

Alshalan, R. & Al-Khalifa, H., 2020. A deep learning approach for automatic hate speech detection in the saudi twittersphere.. *Applied Sciences, 10(23),,* p. 8614.

Al-Smadi, M., Jaradat, Z., Mahmoud, A. A. & Jararweh, Y., 2017. Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features.. *Information Processing & Management,* pp. 53(3), 640-652.

Areej Al-Hassan, H. A.-D., 2019.. Detection of hate speech in social networks: a survey on multilingual corpus.. *COSIT, AIAPP, DMA, SEC,* pp. © CS & IT-CSCP pp. 83–100, 2019.

b-Aldera, S. et al., 2021. Exploratory data analysis and classification of a new Arabic online extremism dataset. *IEEE Access, 9, 161613-161626.*

Batrinca, B. & Treleaven, P. C., 2015. Social media analytics: a survey of techniques, tools and platforms.. *Ai & Society,* Volume 30, pp. 89-116.

Bedjou, K., Aloui, A. & Azouaou, F., 2018. *LexD3T : A Lexical Detection Process of Terrorist Threats on Twitter.*. Béjaia, 7th. International Symposium ISKO-Maghreb'2018 Knowledge Organization in the perspective of Digital Humanities.

Bedjou, K. & Azouaou, F., 2023. Detection of terrorism's apologies on Twitter using a new bi-lingual dataset. *Int. J. Data Mining, Modelling and Management,* Volume Vol. 15, No. 04, pp.331–334.

Bedjou, K., Azouaou, F. & Aloui, A., 2019,. *Detection of terrorist threats on Twitter using SVM.*. Paris, France. July 1–2, , In Proceedings of the 3rd International Conference on Future Networks and Distributed Systems (pp. 1-5).

Benali, A., Maaloul, M. H. & Belguith, L. H., 2023. Automatic Processing of Algerian Dialect: Corpus Construction and Segmentation. *SN Computer Science,* Volume 4(5), 597.

Bourgonje, P., Moreno-Schneider, J., Srivastava, A. & Rehm, G., 2017. Automatic Classification of Abusive Language and Personal Attacks in Various Forms of Online Communication.. *In International Conference of the German Society for Computational Linguistics and Language Technology,* pp. (pp. 180– 191). Springer, Cham.

Burnap, P. et al., 2014. Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining, 4, 1-14.*

Chau, E. C., Lin, L. H. & Smith, N. A., 2020. Parsing with multilingual BERT, a small corpus, and a small treebank.. *arXiv preprint arXiv:2009.14124.*

Cheong, M. & Lee, V. C., 2011. "A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter".. *Information Systems Frontiers,,* pp. 13(1), 45-59.

Dang, N. C., Moreno-García, M. N. & De la Prieta, F., 2020. Sentiment analysis based on deep learning: A comparative study.. *Electronics,,* pp. 9 (3), 483.

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805,,* Issue https://huggingface.co/bert-base-uncased.

El-Shishtawy, T., 2013. A Hybrid Algorithm for Matching Arabic Names. *arXiv preprint arXiv:1309.5657.*

Farzindar, A. &. R. M., 2013. Les défis du traitement automatique du langage pour l'analyse des réseaux sociaux. Revue TAL–Traitement Automatique des langues, , Volume 54(3), , pp. 7-16.

Fernández, A., Garcia, S., Herrera, F. & Chawla, N. V., 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary.. *Journal of artificial intelligence research, 61,* pp. 863-905.

Fifth Tribe, 2018. «—How ISIS Uses Twitter,‖ Kaggle,» consulted on March 12, 2018.. [En ligne]. Available: https://www.kaggle.com/fifthtribe/how-isis-uses-twitter.

Founta, A. M. et al., 2019. A unified deep learning architecture for abuse detection.. *Proceedings of the 10th ACM conference on web science,,* pp. (pp. 105-114).

Fsih, E., Kchaou, S., Boujelbane, R. & Belguith, L. H., 2022. Benchmarking transfer learning approaches for sentiment analysis of Arabic dialect. *In Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pp. (pp. 431-435).

Gospodnetic, O., Hatcher, E. & McCandless, M., 2010. Lucene in Action‖, Second Edition: Covers Apache Lucene 3.0. *pp. Manning Publications Co., Greenwich, CT, USA.*

Guellil, I. et al., 2020. Detecting hate speech against politicians in Arabic community on social media.. *International Journal of Web Information Systems. Vol. 16 No. 3,,* pp. pp. 295-313..

Gupta, S. & Kumari, N., 2019. Security mechanism for twitter data using Cassandra in cloud. *Int. J. Distrib. Cloud Comput., Vol. 7, No. 2,* p. p. pp.1–6.

Gu, Q., Tian, J., Li, X. & Jiang, S., 2022. A novel Random Forest integrated model for imbalanced data classification problem. *Knowledge-Based Systems, 250, 109050.*

Haddad, H. et al., 2023. TunBERT: Pretraining BERT for Tunisian Dialect Understanding. *SN Computer Science,* Volume 4(2), 194.

Haixiang, G. et al., 2017. Learning from class-imbalanced data: review of methods and applications. *Expert Systems with Applications,,* pp. Vol. 73, , p. pp.220– 239,.

HALIM, A. H. B. A., MANSHOR, N. B., AZURA, N. & HUSIN, B., 2023. CLASS IMBALANCE LEARNING WITH COSTSENSITIVE-ACGAN. *Journal of Theoretical and Applied Information Technology,* Volume 101(12).

Harris, Z. S., 1954. *Distributional structure.* s.l., Word, 10(2-3), 146-162.

Hasib, K. M., Showrov, M. I. H., Al Mahmud, J. & Mithu, K., 2022. Imbalanced data classification using hybrid under-sampling with cost-sensitive learning method. *In Edge Analytics: Select Proceedings of 26th International Conference—ADCOM 2020,* pp. (pp. 423-435). Singapore: Springer Singapore.

Huang, F., Kwak, H. & & An, J., 2023. s chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736.*.

Iskandar, B., 2017. Terrorism detection based on sentiment analysis using machine learning".. *Journal of Engineering and Applied Sciences,12(3), ,* pp. 691-698.

JABER, M. M., AL-GHURIBI, S. M. & ABD, D. H., 2023. Arabic Text Detection Using Rough Set Theory: Designing a Novel Approach.. *IEEE.*

Jiang, J. & Conrath, D., 1997. *Semantic similarity based on corpus statistics and lexical taxonomy.* Taiwan, s.n.

Johnston, A. H. & Weiss, G. M., 2017. *Identifying Sunni Extremist Propaganda with Deep Learning.* s.l., In 2017 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1-6). IEEE.

Kchaou, S., Boujelbane, R. & Hadrich, L., 2023. Hybrid pipeline for building Arabic Tunisian Dialect-Standard Arabic Neural machine translation model from scratch.. *ACM Transactions on Asian and Low-Resource Language Information Processing,* Volume 22(3), pp. 1-21.

Khan, A., Baharudin, B., Lee, L. H. & khan, K., 2010. A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of advances in information technology,* Volume VOL. 1, NO. 1, pp. 4-20.

Klausen, J., 2015. Tweeting the Jihad: Social Media Networks of Western Foreign Fighters in Syria and Iraq". *Studies in Conflict & Terrorism,* Volume 38(1), pp. 1-22.

Konstantinidis, S., 2005.. Computing the Levenshtein distance of a regular language. *IEEE Information Theory Workshop, IEEE,* pp. pp. 4-pp.

Kumari, R. & Srivastava, S. K., 2017. Machine learning: A review on binary classification. International Journal of Computer Applications, , Volume 160(7).

Last, M., Markov, A. & Kandel, A., 2006. *Multi-lingual Detection of Terrorist Content on the web.* s.l., H. Chen et al. (Eds.) : WISI 2006, LNCS 3917, pp 16-30.

Leenuse, M. L. & Pankaj, D. S., 2023. Detection and Prediction of Terrorist Activities and Threatening Events in Twitter-A Survey.. *In 2023 International Conference on Control, Communication and Computing (ICCC). IEEE.,* pp. (pp. 1-6).

Le, Q. & Mikolov, T., 2014. *Distributed Representations of Sentences and Documents.* s.l., International conference on machine learning (pp. 1188-1196). PMLR.

Madabushi, H. T., Kochkina, E. & Castelle, M., 2020. Cost-sensitive BERT for generalisable sentence classification with imbalanced data.. *arXiv preprint arXiv:2003.11563.*

Malek, N. H. A. et al., 2023. Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data.. *Indones. J. Elec. Eng. Comput. Sci, 29, 598-608.*

Malmasi, S. & Zampieri, M., 2017. Detecting Hate Speech in Social Media. *arXiv preprint arXiv:1712.06427.*

Mansour, M., Tohamy, M., Ezzat, Z. & Torki, M., 2020. Arabic dialect identification using BERT fine-tuning.. *In Proceedings of the Fifth Arabic Natural Language Processing Workshop*, December, pp. pp. 308-312.

Matrane, Y., Benabbou, F. & Sael, N., 2023. A Systematic Literature Review of Arabic Dialect Sentiment Analysis. *Journal of King Saud University-Computer and Information Sciences, 101570.*

Mazari, A. C. & Kheddar, H., 2023. Deep Learning-based Analysis of Algerian Dialect Dataset Targeted Hate Speech, Offensive Language and Cyberbullying. *International Journal of Computing and Digital Systems.*

Mierswa, I. & Klinkenberg, R., 2018. RapidMiner Studio (9.1) [Data Science, Machine Learning, Predictive Analytics],. *[online] https://rapidminer.com/ (accessed 14 November 2021).*

Mikolov, T., Chen, K., Corrado, G. & Dean, J., 2013. *Efficient estimation of word representations in vector space.* s.l., arXiv preprint arXiv:1301.3781.

Mulki, H., Haddad, H., Ali, C. B. & Alshabani, H., 2019. L-hsab: A levantine twitter dataset for hate speech and abusive language.. *In Proceedings of the third workshop on abusive language online., August 2019,* pp. pp. 111-118.

Noguchi, Y. & Kholmann, E., 2006. *Tracking terrorists online,* s.l.: Washingtonpost. com video report. vol. 19.

Oh, O., Agrawal, M. & Rao, H. R., 2010. Tracking the Mumbai terrorist attack through twitter. *Information Systems Frontiers,* Volume 13, pp. 33-43.

On-line, 2016. https://www.lecourrierdelatlas.com/algerie-un-journaliste-d-echourouk-arrete-pour-apologie-au-terrorisme-6733/, consulté le 17.02.2017.

Ouchene, L. & Bessou, S., 2023. FastText Embedding and LSTM for Sentiment Analysis: An Empirical Study on Algerian Tweets. *In 2023 International Conference on Information Technology (ICIT).IEEE.*, pp. (pp. 51-55).

Park, J. H. & Fung, P., 2017. *One-step and Two-step Classification for Abusive Language Detection on Twitter.* s.l., arXiv preprint arXiv:1706.01206.

Peng, C. Y. & Park, Y. J., 2022. A new hybrid under-sampling approach to imbalanced classification problems.. *Applied Artificial Intelligence, 36(1), 1975393.*

Pennington, J., Socher, R. & Manning, C. D., 2014. *Glove: Global vectors for word representation.* s.l., Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) pp. 1532-1543 .

Pitsilis, G. K., Ramampiaro, H. & Langseth, H., 2018. Detecting Offensive Language in Tweets Using Deep Learning. *arXiv:1801.04433v1 [cs.CL]*.

Porter, M., 2001. *Snowball: A Language for Stemming [online]*. [En ligne] Available at: http://snowball.tartarus.org/ texts/introduction.html [Accès le 12 January 2022].

Rada, R., Bicknell, E. & Blettner, M., 1989. Development and Application of a Metric on Semantic Nets. *IEEE Trans. on Systems, Man, and Cybernetics,* Volume 19(1), pp. 17-30.

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I., 2018. Improving Language Understanding by Generative Pre-Training. *OpenAI,* https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf, consulté le 23.04.2022.

Resnik, P., 1999. Semantic similarity in a taxonomy: an information based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research,* Volume 11, pp. 95-130.

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F. & Cuenca-Jiménez, P. M., 2023. A review on sentiment analysis from social media platforms. *Expert Systems with Applications, 119862.*

Rout, N., Mishra, D. & Mallick, M. K., 2018. Handling imbalanced data: a survey. s.l., In International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications: ASISA 2016 (pp. 431-443). Springer Singapore.

Saad, M., 2016. *Arabic Light Stemmer. GitHub Repository.* [En ligne] https://github.com/motazsaad/arabic-light-stemming-py, consulté le 5.02.2022.

Safaya, A., Abdullatif, M. & Yuret, D., 2020. Bert-cnn for offensive speech identification in social media. s.l., Kuisail at semeval-2020 task 12. arXiv preprint arXiv:2007.13184.

Saleh, H., Alhothali, A. & Moria, K., 2023. Detection of hate speech using BERT and hate speech word embedding with deep model. *Applied Artificial Intelligence, 37(1), 2166719.*

Salem, F., 2017. *Social media and the internet of things towards data-driven policymaking in the Arab world: potential, limits and concerns.* s.l., The Arab Social Media Report, Dubai: MBR School of Government 7.

Salton, G. & Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management,* Volume 24, Issue 5(ISSN 0306-4573, https://doi.org/10.1016/0306-4573(88)90021-0.), pp. 513-523.

Santos, M. S. et al., 2018. Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier]. *IEEE Computational Intelligence Magazine,* Volume Vol. 13, No. 4, p. p. pp.59–76.

Sarnovský, M. & Paralic, M., 2008. *Text Mining Workflows Con- struction with Support of Ontologies.* s.l., In 2008 6th International Symposium on Applied Machine Intelligence and Informatics (pp. 173-177). IEEE.

Selva Birunda, S. & Kanniga Devi, R., 2021. *A review on word embedding techniques for text classification..* s.l., Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020, 267-281.

Shang, W. et al., 2006. A noval feature selection algorithm for text categorization. *Expert System with Application, Vol. 33, No. 1,,* p. pp.1–5.

Soliman, A. B., Eissa, K. & El-Beltagy, S. R., 2017. Aravec: A set of arabic word embedding models for use in arabic nlp.. *Procedia Computer Science,* Volume 117,, pp. 256-265.

Tanimoto, A. et al., 2022. Improving imbalanced classification using near-miss instances.. *Expert Systems with Applications, 201, 117130.*

Theodosiadou, O. et al., 2021. Change Point Detection in Terrorism-Related Online Content Using Deep Learning Derived Indicators. *Information. https://doi.org/10.3390/ info12070274,* pp. 12,274.

Varelas, G. et al., 2005. *Semantic similarity methods in wordNet and their application to information retrieval on the web.* s.l., In Proceedings of the 7th annual ACM international workshop on Web information and data management (pp. 10-16).

Vishwamitra, N. et al., 2017. *MCDefender: Toward effective cyberbullying defense in mobile online social networks.* s.l., In Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics (pp. 37-42).

Waseem, Z., Davidson, T., Warmsley, D. & Weber, I., 2017. *Understanding Abuse: A Typology of Abusive Language Detection Subtasks.* s.l., arXiv preprint arXiv:1705.09899.

We-are-social, 2023. https://wearesocial.com/uk/blog/2022/01/digital-2022-another-year- of-bumper-growth-2/, consulté le 11.10.2023.

Whang, S. E., Roh, Y., Song, H. & Lee, J. G., 2023. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal,* Volume 32(4), pp. 791-813.

Zhang, Z. & Luo, L., 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web,,* Volume 10(5), pp. 925-945.

Résumé

Nous avons traité dans cette thèse la problématique des menaces terroristes sur les réseaux sociaux. Nous nous sommes intéressés aux menaces écrites (publications textuelles) afin de les détecter automatiquement. Pour cela, nous avons proposé 3 contributions, la première est une approche lexicale basé sur la recherche des publications contenant des mots en relation avec le terrorisme même s'ils sont mal orthographiés ; la seconde est basée sur une classification binaire (menaçant ou non menaçant) en utilisant un apprentissage automatique avec SVM, et la troisième est une approche de classification ternaire (positif, négatif ou neutre) en utilisant l'apprentissage par transfert avec un modèles déjà entrainés sur des millions de données auquel nous rajoutons une couche avec nos propres données. Nous avons évalué notre proposition avec un dataset de plus de 12000 tweets manuellement annotés, dans les deux langues Arabe et Anglais. Les résultats obtenus sont très significatifs avec des précisions proches de 90%.

Mots clès : Analyse des réseaux sociaux, Détection des menaces terroristes, NLP, Analyse de sentiments, Machine learning, Deep learning, Transfer learning.

Abstract

In this thesis, we addressed the issue of terrorist threats on social networks. We focused on written threats (textual publications) in order to detect them automatically. To this end, we have proposed 3 contributions: the first is a lexical approach based on searching for publications containing words related to terrorism, even if they are misspelled; the second is based on binary classification (threatening or non-threatening) using machine learning with SVM, and the third is a ternary classification approach (positive, negative or neutral) using transfer learning with a model already trained on millions of data to which we add a layer with our own data. We evaluated our proposal with a dataset of over 12,000 manually annotated tweets, in both languages Arabic and English. The results obtained are highly significant, with accuracies close to 90%.

Keywords: Social network analysis, Terrorist's threat detection, NLP, Sentiment analysis, Machine learning, Deep learning, Transfer learning

ملخص

في هذه الأطروحة تناولنا مشكلة التهديدات الإرهابية على شبكات التواصل الاجتماعي. اهتممنا خاصة بالتهديدات المكتوبة (المنشورات النصية) من أجل اكتشافها آليا. لهذا، اقترحنا 3 مساهمات، الأولى نهج معجمي بالبحث عن المنشورات التي تحتوي على كلمات ذات صلة بالإرهاب حتى لو كانت بها أخطاء إملائية، والثانية اعتمدنا على التصنيف الثنائي (تهديد أو غير تهديد) باستخدام خوارزمية التعلم الآلي SVM، والثالثة انتهجنا تصنيفا ثلاثيا (إيجابي أو سلبي أو محايد) باستخدام التعلم التحوّلي مع نموذج تم تدريبه مسبقا على ملايين البيانات التي نضيف إليها طبقة ببياناتنا الخاصة. قمنا بتقييم اقتراحنا بمجموعة بيانات تضم أكثر من 12000 تغريدة مشروحة يدويًا باللغتين العربية والإنجليزية. النتائج التي تم الحصول عليها مرضية للغاية وبدقة تقترب من 90٪.

كلمات دلالية: تحليل الشبكات الاجتماعية، الكشف عن التهديد الإرهابي، البرمجة اللغوية العصبية ، تحليل المشاعر ، التعلم الآلي ، التعلم العميق ، نقل التعلم.