

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Polytechnique



Département Génie Industriel

Mémoire de projet de fin d'études

Pour l'obtention du diplôme d'ingénieur d'état en Génie Industriel

Contribution à l'optimisation du processus de Due Diligence réputationnelle par
l'automatisation.

Application au sein de KPMG

Présenté par Driss BENHADJI SERRADJ

Sous la direction de Mme Nadjwa Noual

Présenté et soutenu publiquement le (01/07/2019)

Composition du Jury :

Président	Mr Ali BOUKABOUS	MAA	ENP
Promotrice	Mme Nadjwa NOUAL	MAA	ENP
Examinateur	Mr Mabrouk AIB	Docteur	ENP
Invité	Mr Lotfi ABDI	Consultant Sénior	KPMG

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Polytechnique



Département Génie Industriel

Mémoire de projet de fin d'études

Pour l'obtention du diplôme d'ingénieur d'état en Génie Industriel

Contribution à l'optimisation du processus de Due Diligence réputationnelle par
l'automatisation.

Application au sein de KPMG

Présenté par Driss BENHADJI SERRADJ

Sous la direction de Mme Nadjwa Noual

Présenté et soutenu publiquement le (01/07/2019)

Composition du Jury :

Président	Mr Ali BOUKABOUS	MAA	ENP
Promotrice	Mme Nadjwa NOUAL	MAA	ENP
Examineur	Mr Mabrouk AIB	Docteur	ENP
Invité	Mr Lotfi ABDI	Consultant Sénior	KPMG

Dédicaces

*À La mémoire de mon grand-père Djamel, diplômé des mêmes bancs
un demi-siècle avant l'arrivée de son petit-fils,*

À La mémoire de ma grand-mère Nafissa qui a tant prié pour moi,

*À La mémoire de mon grand-père Mohamed qui nous a transmis le
sang de la persévérance,*

À Ma grand-mère Hafida qui m'a toujours hautement considéré,

*À Mes chers parents qui se sont sacrifiés pour ma réussite, m'ont
soutenu, fait confiance et cru en moi jusqu'au bout,*

*À Mes deux sœurs Yasmine et Ryma qui m'ont toujours encouragé
avec leurs petites intentions irremplaçables,*

À Mon frerot, Walid, qui aurait dû être mon binôme durant ce projet,

*À Mes Meilleurs : Fatchy, Sandra et Amir ainsi que l'ensemble des
collègues de la promo,*

*À mes Salim et Moha ainsi que tous mes amis, cousins et cousines de
Tlemcen,*

À Tous les membres du Lions club, du CAP et du PAC,

Je dédie ce modeste travail.

Remerciements

Il y a trois ans, J'arrivais de ma petite ville natale, Tlemcen, avec à la main, mon concours d'accès à ENP. Le choix de la spécialité était prédestiné par des rencontres d'anciens Indus. J'ai, donc, tracé ma trajectoire, contrôlé les paramètres, puis grâce au soutien divin, me voilà à la fin d'un palier que je n'aurai pas atteint sans la contribution de beaucoup de personnes.

Il me tarde de remercier, tout d'abord, Madame Nadjwa NOUAL, mon encadreur et promotrice pour son écoute, sa disponibilité et sa bienveillance tout au long de mon projet de fin d'étude. Tous les mots ne suffisent pour remercier la marraine qui m'a aiguillé dès mes premiers pas à l'ENP.

Je remercie vivement l'équipe Deal Advisory de KPMG Algérie en général et en particulier à l'équipe de « Recherche et Stratégie », les seniors Lotfi ABDI et Amina DAOUD qui m'ont énormément aidé pour ce travail; ainsi que les stagiaires Mehdi et Lyria avec lesquels règne une osmose parfaite, sans oublier notre précieuse Aida.

Je remercie également mon ainée au département Amina SADAoui, consultante au sein du cabinet, qui m'a épaulé dans la structure de mon travail, puis la responsable RH, Rym SEMGHOUNI, qui a veillé au bon déroulement de mon stage.

Mes chaleureux remerciements à Mr Mathieu BEAUCOURT (PDG), et Mehdi BETTAHAR (Manager), pour avoir rendu possible mon stage au sein de KPMG.

A mes parents et mes petites sœurs, un remerciement spécial pour leur soutien inconditionnel. Je ne saurai vous exprimer tout le respect que je vous dois, l'amour et la considération que je vous porte.

Je remercie les membres du jury de m'avoir fait honneur d'examiner mon humble travail.

Sans oublier l'ensemble du corps pédagogique du département Génie Industriel, et ce précieux CAP, grâce à qui j'ai beaucoup évolué et où j'ai rencontré des gens formidables.

Enfin, à tous ceux qui ont contribué de près ou de loin à l'accomplissement de ce travail, je dis: MERCI !

KPMG هي مؤسسة دولية من شركات التدقيق والمحاسبة والاستشارات العاملة في العديد من البلدان. قسم "استشارات الصفقات" هو المسؤول عن عمليات الاندماج والاستحواذ من خلال العمل كرفيق للصفقات. في هذا القسم، يهتم فريق "البحث والاستراتيجية" بسمعة الشركات عبر الإنترنت من خلال تحليل مؤشراتنا المختلفة بهدف جعلها أكثر قيمة ودمجها معلومة أساسية لأخذ القرار في للاستثمار في شركة. تسمى هذه العملية "العناية الواجبة للسمعة". يناسب مشروعنا في سياق تحسين هذه الأخيرة. للقيام بذلك، تم إضفاء الطابع الرسمي على العملية لأول مرة، وتم إجراء تشخيص، واكتشاف المهام الحرجة مما أدى بنا إلى تنفيذ التشغيل الآلي باستخدام برامج الكمبيوتر والتعلم الآلي للقدرة على معالجة عدد بيانات كبيرة في وقت قصير مع مراعاة الجانب التنظيمي وهذا يهدف الى تقديم تقارير سمعة عالية الجودة عبر الإنترنت بموارد متالية.

الكلمات المفتاحية: عمليات الدمج والاستحواذ، العناية الواجبة للسمعة، التحسين، التشغيل الآلي، التعلم الآلي.

Abstract :

KPMG is an international network of audit, accounting and consulting firms operating in several countries. The Deal Advisory department is in charge of mergers and acquisitions projects, acting as an advisor during the transaction. Within the latter, the "Research and Strategy" team is concerned by the online reputation of companies by analysing its various indicators in order to enhance it and integrate it as an essential parameter for investor decision-making. This process is called "Reputational Due Diligence". It is within the framework of the optimization of the latter that our project is part of. To do this, the process was first formalized, a diagnostic was made, critical tasks were detected, which led us to implement automation using computer programs and machine learning to be able to process a considerable amount of data in a reduced time while taking into account the organizational aspect to deliver quality online reputation reports with optimal resources.

Keywords: Mergers & Acquisition, Reputational Due Diligence, Optimization, Automation, Machine Learning.

Résumé :

KPMG est un réseau international de cabinets d'audit, d'expertise comptable et de conseil exerçant dans plusieurs pays. Le département Deal Advisory est chargé des missions de fusions et acquisitions en intervenant en tant qu'accompagnateur de la transaction. Au sein de ce dernier, l'équipe « Recherche et Stratégie » s'intéresse à la réputation en ligne des entreprises en faisant une analyse sur ses différents indicateurs dans le but de la valoriser et l'intégrer comme paramètre essentiel de décision de l'investisseur. Ce processus est appelé « Due Diligence réputationnelle ». C'est dans le cadre de l'optimisation de ce dernier que s'inscrit notre projet. Pour ce faire, le processus a, d'abord, été formalisé, un diagnostic a été fait, les tâches critiques détectées ce qui nous a mené à implémenter des automatisations en utilisant des programmes informatiques et de l'apprentissage machine pour pouvoir traiter un nombre de données considérable en un temps réduit tout en prenant en compte l'aspect organisationnel pour pouvoir délivrer des rapports de réputation en ligne de qualité avec des ressources optimales.

Mots Clés: Fusions & Acquisition, Due Diligence réputationnelle, Optimisation, Automatisation, Machine Learning.

Tables des matières :

Liste des figures	
Liste des tableaux	
Liste des Annexes	
Liste des abréviations	
Introduction Générale :.....	14
Partie 1 : Etat de l'art.....	19
Chapitre 1: Enjeux et leviers majeurs dans le monde des M&A	19
1. Mergers and Acquisitions (M&A) :.....	19
1.1. Définition des M&A :	19
1.2. Similitudes et différences entre les fusions, acquisitions et joint-ventures:	20
1.3. Les motifs des M&A :.....	20
1.4. Le processus M&A :	20
2. Due diligence:.....	21
2.1. Définition de la due diligence (DD) :.....	21
2.2. L'importance de la due diligence dans le processus M&A :	22
2.3. Types de due diligence :	22
3. Online reputation (OR):.....	24
3.1. Corporate reputation (CR):	24
3.1.1. Définition de la CR :.....	24
3.1.2. L'importance de la CR dans les M&A :	24
3.1.3. L'importance de la CR dans la due diligence :.....	25
3.1.4. Mesure de la CR:	25
3.2. Online reputation (OR) :	26
3.2.1. Définition de la OR :	26
3.2.2. La OR comme mesure essentielle de la CR :	27
4. Protection des données :	28
Chapitre 2: Le Machine Learning au service de l'optimisation.....	31
1. Machine Learning :.....	31
1.1. Définition de l'intelligence artificielle :.....	31
1.2. Définition du Machine Learning :.....	32
1.3. Objectif du Machine Learning :	32
1.4. Types d'apprentissages :	33

1.4.1.	Apprentissage Supervisé :	33
1.4.2.	Apprentissage Non-Supervisé :	33
1.5.	Fonctionnement de l'apprentissage supervisé :	33
1.6.	Types d'algorithmes utilisés :	35
1.6.1.	Régression linéaire:	35
1.6.2.	Régression Logistique :	36
1.6.3.	Arbres de décision :	36
1.6.4.	Nearest Neighbor learning (k-NN) :	36
1.7.	Natural Language Processing :	37
1.7.1.	Bag-of-Words Model :	37
1.7.2.	Character N-gram Model:	38
1.7.3.	Word2vec :	39
2.	Les outils de programmation :	39
2.1.	Python:	39
2.2.	R:	40
2.3.	C++:	40
Partie 2 : Etat des lieux.....		44
Chapitre 3: Présentation de KPMG.....		44
1.	KPMG International :	44
1.1.	Présentation de KPMG:	44
1.2.	Les activités de KPMG:	44
2.	KPMG en Algérie :	46
2.1.	Présentation de KPMG Algérie SPA :	46
2.2.	Chiffre clés de KPMG Algérie SPA :	46
2.3.	Structure organisationnelle de KPMG Algérie SPA :	47
2.4.	Présentation du Deal Advisory :	48
2.4.1.	Transaction services:	49
2.4.2.	Recherche & Stratégie :	50
Chapitre 4 : Diagnostic du processus de Due Diligence réputationnelle.....		53
1.	Description de l'approche OR pour KPMG:	53
1.1.	OR pour l'équipe R&S Alger :	53
1.2.	OR pour l'équipe Forensic India (Astrus):	54
2.	Définition du Processus OR Actuel :	54
2.1.	Scope of Work:	55
2.1.1.	Réception et compréhension du scope :	55

2.1.2.	Organisation des tâches:	56
2.2.	Data retrieving and exploration:	56
2.2.1.	Company presence analysis :	59
2.2.2.	Projects and brands analysis :	59
2.2.3.	Executive analysis :	59
2.2.4.	HR Network :	60
2.3.	Data cleaning :	60
2.4.	Databook Fulfillment :	60
2.5.	Report Construction :	62
3.	Diagnostic et constats :	64
3.1.	L'automatisation des tâches:	64
3.1.1.	Extensive Research :	64
3.1.2.	Le Sentiment Analysis :	64
3.2.	Le databook :	65
3.3.	L'organisation :	65
4.	Genèse de la problématique:	65
5.	Formalisation de la problématique :	67
Partie 3: Apports et Solutions Proposées		71
Chapitre 5 : Solutions, validations et apports		71
1.	Solutions :	71
1.1.	Automatisation des tâches :	71
1.1.1.	Automatisation du sous processus « Extensive Search » :	71
1.1.1.1.	First Google Page :	72
1.1.1.2.	Google Suggest :	73
1.1.1.3.	Crisis Tracking :	74
1.1.2.	Automatisation du sous processus « Sentiment Analysis » :	74
1.1.2.1.	Description du modèle ML :	75
1.1.2.2.	Validation du modèle :	78
1.2.	Le Databook :	80
1.3.	L'organisation :	83
2.	Application sur cas pratique (Projet Hermès):	84
2.1.	Contexte de la mission :	84
2.2.	Présentation de l'entreprise cible et ses concurrents :	84
2.3.	Présentation du rapport final de la mission :	85
3.	Apports :	85

3.1. La dimension qualitative :.....	85
3.2. La dimension temporelle :	86
3.3. La dimension organisationnelle :	88
4. Perspectives :	88
Conclusion Générale :	91
Bibliographie :.....	93
Annexes:.....	96

Liste des figures :

Figure 1: Processus M&A	21
Figure 2: Processus d'apprentissage	34
Figure 3: Exemple modèle BOW 2-gram	38
Figure 4: Présence de KPMG dans le monde.....	46
Figure 5: Evolution CA et EBE de 2014 à 2018	47
Figure 6: Evolution des charges du personnel et RN de 2014 à 2018	47
Figure 7: Organigramme KPMG Algérie SPA.	48
Figure 8: Structure Deal Advisory Alger.	49
Figure 9: Schéma du Processus OR	55
Figure 10: Schéma du sous processus « Scope of work ».....	55
Figure 11: Schéma représentatif des techniques de collecte et d'exploration des données.	56
Figure 12: Schéma représentatif de la technique « Extensive research »	57
Figure 13: Schéma du sous processus « Data retrieving and exploration »	58
Figure 14: Schéma du sous processus « Data cleaning »	60
Figure 15: Schéma du sous processus « Data Fulfillment »	60
Figure 16: Exemple Databook.....	61
Figure 17: Exemple Première version Databook	61
Figure 18: La méthode du risque exprimé par des feux tricolores.....	62
Figure 19: Les 4 rubriques d'un rapport d'OR.....	62
Figure 20: Schéma du sous processus « Report Construction »	63
Figure 21: Processus de Due Diligence réputationnel – (Input/Output)	64
Figure 22: Histogramme d'utilisation des ressources en heures.....	65
Figure 23: Les 3 dimensions des dysfonctionnements.....	66
Figure 24: Exemple de balise d'identification de la date	72
Figure 25: First Google Page de KPMG	73
Figure 26: Exemple Google Suggest pour KPMG.....	73
Figure 27: Description Modèle ML	75
Figure 28: Interface AWS « Amazon Comprehend »	78
Figure 29: Coûts des analyses “Amazon Comprehend”	79
Figure 30: Exemple avis négatif.....	79
Figure 31: Feuille YouTube databook	80
Figure 32: Feuille Sentiment Analysis - HR Network	81
Figure 33: Feuille KPIs - HR Network	81
Figure 34: Feuille Récapitulatif (Ancienne Version).....	81
Figure 35: Feuille Récapitulatif (Nouvelle Version).....	82
Figure 36: Rapport des coûts.....	86
Figure 37: Durée du projet avant l'optimisation	87
Figure 38: Durée du projet après l'optimisation	87

Liste des tableaux :

Tableau 1: Exemple des vecteurs Bag-of-Words	38
----------------------------------------------------	----

Liste des Annexes :

Annexe n°1: Processus Due Diligence sur MS Project.....	96
Annexe n°2: Programme informatique – Sentiment Analysis (SA)	97
Annexe n°3: Explication des ligne du programme SA	112
Annexe n°4: Programme Informatique – Automatisation FGP/GS	115
Annexe n°5: Programme Informatique - Crisis Tracking	121

Liste des abréviations :

AWS: Amazon Web Services.

BDD: Buyer Due Diligence.

BFR: Besoin en Fond de Roulement.

BOW: Bag Of Words.

BtoB: Business to Business.

BtoC: Business to Consumers.

CA: Chiffre d'Affaires.

CR: Corporate Reputation.

DD: Due Diligence.

EBE: Excédent Brut d'Exploitation.

EBIT: Earnings Before Interest, Taxes.

EBITDA: Earnings Before Interest, Taxes, Depreciation, and Amortization.

ESG: Environnemental, Social et Gouvernance.

FGP: First Google Page.

FO: Financial Organisation.

GS: Google Suggest.

HR: Humain Ressources.

HTML: HyperText Markup Language.

IA: Intelligence Artificielle.

k-NN: k-Nearest Neighbour.

KPI: Key Performance Indicators.

KPMG: Klynveld Peat Marwick Goerdeler.

M&A: Mergers & Acquisition.

ML: Machine Learning.

NLP: Natural Language Processing.

OR: Online Reputation.

PDG: Président-Directeur Général.

PIB: Produit Intérieur Brut.

PME: Petite et Moyenne Entreprise.

R&S: Recherche & Stratégie.

RGPD: Règlement Général de la Protection des Données.

RH: Ressources Humaines.

RN: Résultat Net.

SA: Sentiment Analysis.

SPA: Société Par Action.

TPE: Très Petite Entreprise.

TS: Transaction Services.

VDD: Vendor Due Diligence.

Introduction Générale

Introduction Générale :

Le monde qui nous entoure a connu des moments de gloire alternés et de moments moins honorifiques. La machine humaine est la principale responsable de ses deux états ; de par sa créativité engendrée par son intelligence et son savoir-faire pour se propulser vers le haut, et par sa fatuité, qui peut l'entraîner vers une chute incontrôlable et dégénérative.

Lorsque nous sommes dotés des bagages qu'il faut, nous nous mettons intuitivement à développer des idées, mettre en route des projets, puis à un certain sommet, nous n'arrivons plus à grimper vers un autre. L'énergie est là mais la propulsion ne suit pas. Curieusement ou jalousement, nous regardons chez le voisin. Qu'il soit sur un sommet plus haut ou plus bas que le nôtre, il peut nous être utile. Nous nous rendons compte aussi que nous ne pouvons applaudir d'une seule main. C'est ainsi que naîtra la fusion, cette union qui fait une force. Cette complicité stratégiquement étudiée permettra aux deux prétendants de se courtiser jusqu'à la collaboration optimale.

Ceci dit, nous brûlons tout de même certaines étapes essentielles qu'il est bon de rappeler. Avons-nous choisi le bon partenaire ? La progression financière positive est-elle une bonne et unique raison pour valider notre choix ?

C'est comme dans la vie sociale, nous ne pouvons-nous unir avec une personne qui n'a pas les mêmes valeurs que nous donc des renseignements sont nécessaires pour s'assurer de notre complicité et notre compatibilité.

Le « bouche à oreille » a un poids considérable. Il crée la bonne et la mauvaise réputation. Prenons exemple sur un restaurant où vous avez dîné: si vous aimez, vous allez sûrement le recommander, et si vous n'avez pas aimé, vous n'allez pas en parler jusqu'à l'oublier.

Encore plus simple, actuellement nous nous rendons sur la toile pour confirmer ou infirmer des informations, pour vérifier une réputation : en exemple concret, nous parcourons les commentaires concernant un hôtel qui définissent notre choix en général, même si ce n'est pas le moins cher, le moins centré ou plus étoilé. C'est ce que nous appelons dans un autre langage et un autre contexte: Le processus de Due Diligence Réputationnelle.

Est-ce une manière de regarder au-delà de l'aspect financier qui ne deviendrait plus la seule priorité ? Iriez-vous acheter un excellent pain même si ça vous prendra plus de temps et d'argent juste parce que le boulanger a une bonne réputation ?

Dans un langage plus technique, l'essor qu'a connu le marché des entreprises est intrinsèquement lié aux avantages que procurent les fusions acquisitions telles que la croissance, la préservation des emplois et la création de valeur pour les preneurs de risque.

De ce fait, les fusions acquisitions sont devenues un champ d'investigation et de recherche privilégié, de par leur importance et leur rôle dans l'économie mondiale ainsi de par leur complexité. Une opération de rapprochement entre entreprises n'est pas une simple tâche mais un long processus qui regorge d'obstacles et de risques inhérents.

La complexité à laquelle doit faire face un investisseur, lorsqu'il entame un processus de rapprochement, est multidimensionnelle et regroupe plusieurs facteurs de décision tels que le volet financier, fiscal, réputationnel, juridique, commercial et managérial.

Cette complexité justifie le recours aux prestataires de services spécialisés dans ce type de rapprochement tels que les banques d'affaires ou les cabinets d'audit et de conseil. KPMG en est, un leader dans le marché des grandes transactions grâce à l'expertise métier qui est reflétée par ses consultants au sein du département « Deal Advisory » spécialement dédié à ce type de mission. Dans son approche, le cabinet offre un accompagnement à son client tout au long du processus de fusion ou acquisition qu'il soit du côté du vendeur ou de l'acheteur en allant de la détection des entreprises en difficulté jusqu'à la restructuration de l'entreprise résultante de la fusion ou de l'acquisition en passant par les différentes investigations et vérifications faites avant la transaction (Vérification financière, juridique, réputationnelle, etc.). Ce type de vérification est appelée Due Diligence et c'est l'un des produits phares que propose KPMG à ses clients lors d'une mission de ce type.

De nos jours, la réputation de l'entreprise et son image de marque jouent un rôle important dans ce genre de transaction. Les investisseurs ne s'intéressent plus qu'à la santé financière de l'entreprise avant d'investir de l'argent mais aussi à sa notoriété et à la perception des consommateurs envers elle. A cet effet, la Due Diligence réputationnelle est devenue une étape primordiale dans un processus de fusion/acquisition qui pèse sur la décision.

La Due Diligence réputationnelle résume tous les indicateurs qui reflètent l'image et la notoriété d'une entreprise. KPMG fournit à son client un rapport qui balaye une vue d'ensemble sur la réputation de l'entreprise en passant par plusieurs étapes d'un processus.

Le présent travail s'inscrit dans le cadre d'une démarche d'amélioration du processus de Due Diligence réputationnelle qui représente un produit avec lequel le cabinet se différencie de ses concurrents. Cette démarche d'amélioration s'inscrit dans la stratégie du cabinet qui veut s'aligner avec les meilleures pratiques au monde en s'alignant aux standards internationaux dans ses processus opérationnels.

Le processus Due Diligence réputationnelle n'ayant jamais été formalisé a été sujet de notre travail. En effet, des questionnements ont été évoqués quant à l'identification des dysfonctionnements et à son amélioration. C'est pourquoi la question relative à son optimisation devient primordiale.

Afin de répondre aux besoins du cabinet, ce travail s'intéresse particulièrement à l'optimisation des tâches qui constituent ce processus en automatisant ses dernières grâce à des programmes informatiques et au Machine Learning. En vue de l'exposer de manière claire et explicite et de dérouler les résultats obtenus, nous l'avons structuré en trois grandes parties, réparties comme suit :

La première partie a pour objectif de rassembler les fondements théoriques et conceptuels du travail, et elle contient deux chapitres :

- Le premier comprend des généralités sur le monde des fusions et acquisitions;
- Le second aborde les outils techniques utilisés pour l'optimisation;

La seconde partie du document a pour objectif de synthétiser le contexte dans lequel s'effectue l'étude en décrivant un état des lieux. Elle comporte deux chapitres :

- Le troisième chapitre présente l'entreprise, ses activités et ses produits ;
- Le quatrième chapitre ambitionne de mieux cerner la problématique en s'appuyant sur un diagnostic du processus existant après sa formalisation;

Et enfin la troisième partie qui comprend le dernier chapitre présente notre contribution à la résolution de la problématique en proposant des programmes informatiques développés pour l'automatisation des tâches en utilisant entre autres le Machine Learning tout en tenant en compte de leurs conséquences sur l'aspect organisationnel.

Les résultats satisfaisants obtenus à l'issu du déroulement des programmes développés nous ont poussé à proposer des perspectives d'amélioration en guise de conclusion.

Partie 1 : Etat de l'art

Chapitre 1: Enjeux et leviers majeurs dans le monde des M&A

Partie 1 : Etat de l'art

Cette partie va traiter l'ensemble des aspects théoriques des différents concepts et terminologies utilisés dans le cadre de ce travail. Il y a une multitude de définitions et typologies qui peuvent être considérées afin de pouvoir cerner un concept, un processus ou même avant de définir une problématique liée à un thème en particulier.

« *La connaissance commence par la tension entre savoir et non-savoir : pas de problème sans savoir – pas de problème sans non-savoir.* » (Popper, 1979)

La revue de littérature permet, également, de voir ce qui a été fait dans le domaine investigué, de comprendre les défis auxquels sont confrontés les acteurs de ce même domaine et ceci afin d'apporter une originalité à l'apport qui découle de ce travail.

Chapitre 1: Enjeux et leviers majeurs dans le monde des M&A

Dans un environnement à forte intensité concurrentielle, les entreprises se retrouvent dans des situations difficiles où certains acteurs du marché prennent le dessus sur d'autres en ayant des parts de marché plus importantes ce qui se répercute sur le reste des acteurs du marché qui n'arrivent pas à assurer la pérennité de leurs entreprises. Dans une optique de croissance, ces dernières ont tendance à s'allier entre elles pour former une entité plus solide sous forme d'une fusion de deux entreprises ou plus. Elles procèdent, dans certains cas, à la cession de leurs actifs partiellement ou en totalité en faveur des entreprises dominant le marché ou alors, ce sont ces dernières qui engagent une procédure d'acquisition afin d'améliorer leur rentabilité à long terme.

1. Mergers and Acquisitions (M&A) :

1.1. Définition des M&A :

L'expression fusions & acquisitions (parfois aussi appelée « *Fusac* », ou en anglais M&A, un acronyme pour Mergers and Acquisitions) fait référence à une transaction d'ordre stratégique entre deux entreprises distinctes ou plus.

La fusion est une opération par laquelle deux sociétés décident de combiner leurs activités. À la suite d'une fusion deux sociétés individuelles cessent d'exister et une nouvelle société combinée est créée. L'acquisition est une opération par laquelle une société reprend l'exploitation d'une autre société. L'entreprise, qui est acquise, est dénommée "société cible" et la société qui l'acquiert s'appelle "acquéreuse".

Les fusions se produisent lorsque deux entreprises unissent leurs forces. De telles transactions ont généralement lieu entre deux entreprises de la même taille et qui reconnaissent les avantages que l'autre offre en termes d'augmentation des ventes, d'efficacité et de capacités. Les termes de la fusion sont souvent assez amicaux et mutuellement convenus et les deux sociétés deviennent des partenaires égaux dans la nouvelle entreprise.

Les acquisitions ont lieu lorsqu'une entreprise achète une autre entreprise et l'intègre à ses opérations. Parfois, l'achat est amical et parfois, il est hostile sans qu'il n'y ait d'accord fixé et

négocié auparavant avec le conseil d'administration et les dirigeants de la société cible, selon que la société acquise estime qu'elle est mieux placée en tant qu'unité d'exploitation d'une grande entreprise.

En parlant de M&A, il est nécessaire d'évoquer les joint-ventures qui sont formées par deux ou plusieurs sociétés mères en tant qu'entités distinctes en combinant une partie de leurs ressources pour atteindre un objectif stratégique. Une joint-venture est aussi décrite comme la réunion de deux ou plusieurs entreprises partenaires provenant de juridictions différentes pour échanger des ressources, partager les risques et partager les bénéfices d'une entreprise commune. Habituellement, mais pas toujours, l'un des partenaires est physiquement situé dans la juridiction de l'entreprise commune. (Thi Quynh Van, 2013)

1.2. Similitudes et différences entre les fusions, acquisitions et joint-ventures:

La littérature ne fait pas explicitement la distinction entre " fusion " et " acquisition ", et tous deux font référence à un processus de regroupement de deux sociétés auparavant distinctes en une seule entreprise. Toutefois, la littérature laisse entendre que les fusions se produisent entre des entreprises de taille de marché similaire, alors que dans le cas d'une acquisition, la société acquéreuse est généralement plus grande que l'entreprise cible en termes de taille de marché. Le résultat final des deux processus est relativement identique, mais la relation entre les deux sociétés diffère selon qu'il s'agisse d'une fusion ou d'une acquisition (Babanazarov, 2012).

McConnell et Nantell proposent deux facteurs qui distinguent les joint-ventures des fusions et acquisitions. Tout d'abord, *"une joint-venture implique la réunion d'un sous-ensemble des ressources de deux (ou plusieurs) sociétés mères"*. Deuxièmement, *"la direction initiale des entreprises fondatrices reste intacte dans le cadre de la joint-venture"* (McConnell et al. 1985).

1.3. Les motifs des M&A :

Les motifs des fusions sont complexes, diversifiés et évoluent au fil du temps, ce qui fait que chaque transaction est évaluée individuellement. Les fusions et acquisitions peuvent également être un moyen de créer de la valeur pour les actionnaires en exploitant les synergies, en augmentant la croissance, en remplaçant les gestionnaires inefficaces, en gagnant du pouvoir sur le marché et en tirant profit de la restructuration financière et opérationnelle. Les synergies opérationnelles et/ou financières sont une autre raison d'être des fusions et acquisitions, où au lieu d'exploiter deux sociétés séparément, on peut créer une plus grande valeur actionnariale combinée. La diversification, le pouvoir de marché, l'alignement stratégique, les ventes croisées d'avantages fiscaux, le transfert de ressources, l'embauche et l'accès aux actifs sont également des motifs importants derrière les fusions et acquisitions. Lorsque les entreprises fusionnent, elles cherchent à minimiser l'incertitude à un niveau donné de profit attendu, la minimisation de l'incertitude est donc également un motif de fusion (Brandtzæg, 2014).

1.4. Le processus M&A :

Le Processus de M&A se décline en 3 étapes essentielles résumées dans ce qui suit :

Pre-merger step: La première étape consiste en l'élaboration d'une stratégie d'acquisition qui exige que l'acquéreur ait une idée claire de ses motifs et ses objectifs vis à vis de la fusion ou

acquisition. Il procède ensuite à la définition des critères de recherche des entreprises cibles pour arriver à sélectionner des cibles potentielles ce qui lui permettra d'effectuer des simulations de transaction pour s'assurer de la cohérence de ses choix avec les objectifs de la stratégie M&A.

Transaction step: La transaction est faite qu'après avoir fait une analyse d'évaluation de tous les aspects des activités de l'entreprise cible ce qui est appelé dans le jargon de la M&A « Due Diligence » (Un processus exhaustif développé dans le titre qui suit). L'acquéreur, après avoir contacté l'entreprise cible, lui demande de fournir des informations substantielles (états financiers actuels, ses clients, ses ressources, etc.) qui lui permettront de l'évaluer davantage. Après l'évaluation, une première offre est transmise à la cible afin d'entamer les négociations dans le but de trouver un terrain d'entente sur le prix de cette dernière.

Post-merger step: Dans le cas d'un accord entre les deux parties de la transaction, un contrat d'acquisition/fusion est signé et un plan d'intégration, représentant une feuille de route conçue pour expliciter la façon de combiner les deux entreprises, est déployé et, dans certains cas, des processus de transformation structurelle sont entamés avec un audit à la fin pour une meilleure consolidation des comptes.

Le schéma ci-dessous illustre les 3 étapes expliquées et mets en valeur les éléments importants de chacune d'elles.

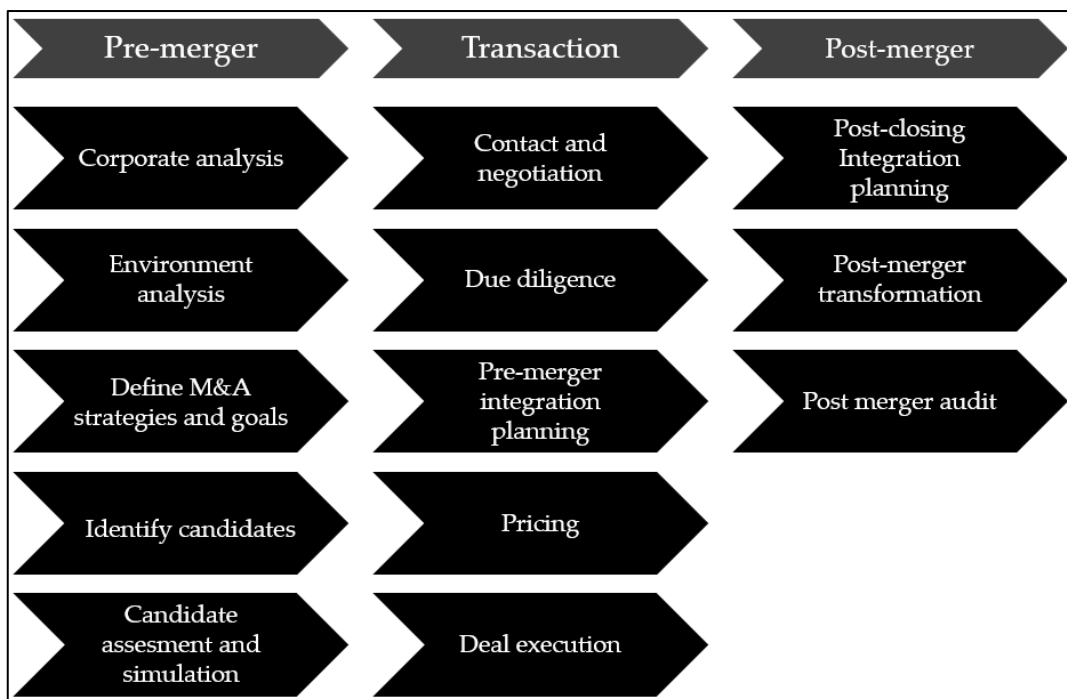


Figure 1: Processus M&A (Freitag, 2014)

2. Due diligence:

2.1. Définition de la due diligence (DD) :

La due diligence est définie en littérature comme " la diligence raisonnablement attendue d'une personne qui cherche à satisfaire à une exigence légale ou à s'acquitter d'une obligation et qui est habituellement exercée par cette personne ". Le processus de due diligence désigne un ensemble de vérifications qu'opère un investisseur en vue d'une transaction et fait référence

à l'examen d'une société cible pour permettre à un futur acquéreur de se faire une idée de la situation précise de cette dernière avant de se prononcer sur son investissement. Cela peut notamment permettre de vérifier la stratégie d'une entreprise, sa situation fiscale, comptable, sociale, environnementale et autre (Lajoux et al. 2010).

2.2.L'importance de la due diligence dans le processus M&A :

La due diligence intervient dans la deuxième étape de processus M&A. Elle aide à comprendre plus profondément les questions importantes sur l'entreprise cible. Cette phase permet de planifier et d'évaluer le processus d'intégration post-transaction et minimiser les risques qui pourraient survenir.

De nos jours, les entreprises sont confrontées à différents types de risques : risque financier, conformité juridique, instabilité politique et géopolitique, concurrence abusive, espionnage économique, protection de la réputation, etc., qui apparaissent souvent dans le cadre de grands contrats de M&A. Il convient de noter que la responsabilité des acquéreurs peut être mise en cause, y compris dans le cas de situations héritées de pratiques antérieures à l'opération. Les risques des pratiques non-conformes peuvent être considérables, qu'il s'agisse d'amendes, de pertes de profits mais aussi d'atteinte à l'image et la réputation de l'entreprise concernée. C'est pour cela qu'il est nécessaire de connaître en profondeur tous les tenants et aboutissants concernant la cible avant de conclure la transaction¹.

2.3.Types de due diligence :

Il existe plusieurs types de due diligence selon la littérature. Ces derniers peuvent être opérés du côté de l'acheteur ou l'acquéreur et cette due diligence est appelée « buyer due diligence (BDD) » ou bien du côté du vendeur/cible et cette due diligence est appelée « vendor due diligence (VDD) ».

La VDD est généralement faite par le vendeur afin de mieux valoriser l'entreprise dans le but de maximiser le prix de sa vente. Par contre la BDD est exercée par l'acheteur afin de d'investiguer l'entreprise cible dans le but de minimiser le prix de son achat.

Dans un processus M&A les due diligence majeures (applicables aussi bien pour la BDD et la VDD) opérées sont :

Due diligence administrative : La DD administrative est l'aspect qui consiste à vérifier les éléments liés à l'administration tels que les installations, le taux d'occupation, le nombre de postes de travail, etc. L'idée est de vérifier les diverses installations appartenant ou occupées par le vendeur et de déterminer si tous les coûts opérationnels sont pris en compte dans les états financiers ou non. Elle donne également une meilleure idée du type de coûts que l'acheteur est susceptible d'encourir au cas où il envisage de poursuivre l'expansion de l'entreprise cible.

Due diligence environnementale : La DD environnementale est l'évaluation et la gestion des responsabilités et des risques environnementaux. C'est un exercice à la fois juridique et technique qui peut prendre différentes formes et être accompli par plusieurs méthodes, mais l'objectif est toujours d'établir que l'entreprise cible est conforme à la réglementation environnementale et protégée contre les accidents environnementaux tels que la contamination du sol et/ou des eaux souterraines.

¹ Source: Netsources, N°124 Septembre / Octobre 2016

Due diligence financière : La DD financière est l'un des éléments les plus importants dans une transaction. Elle représente une analyse des éléments financiers relatifs à l'entreprise cible, ayant pour objectif d'identifier les points clés de la transaction. Elle comprend l'analyse des grands comptes clients, l'analyse des coûts fixes et variables, l'analyse des marges bénéficiaires et l'examen des procédures de contrôle interne. La DD financière examine en outre le carnet de commandes et les ventes de l'entreprise afin de créer des prévisions plus précises.

De nombreux acquéreurs ont une section distincte d'analyse financière axée sur la situation de la dette de la société cible, évaluant la dette à court et à long terme, les taux d'intérêt applicables, la capacité de la société à assurer le service de sa dette et à obtenir plus de financement au besoin, ainsi qu'un examen et une évaluation globale de la structure du capital de la société.

Au final, une due diligence financière permet à un acquéreur d'obtenir des bases financières pour valoriser une cible et des éléments de négociation du prix.

Due diligence légale : La DD légale examine tous les documents légaux qu'une société possède. Il est important de voir comment ces documents juridiques sont structurés et quelles sont les obligations qui existent pour un vendeur. Au cours d'un processus de fusion et acquisition, les aspects juridiques sont importants pour le succès de la transaction. Les conseillers juridiques voudront bien comprendre les risques juridiques possibles. Ces risques peuvent exister dans la structure de l'entreprise, les actifs, les contrats des clients ou des employés ou dans la propriété intellectuelle (Nouboussi et al. 2008).

Due diligence réputationnelle: La DD réputationnelle concerne l'évaluation minutieuse des risques de réputation attachés à un partenaire commercial ou à une entreprise cible, y compris les questions relatives à l'intégrité et à la crédibilité des personnes qui la gouvernent ainsi que la fiabilité et la prévisibilité de l'environnement politique.

Elle consiste à améliorer les chances d'un investisseur de trouver des partenaires d'affaires intègres et fiables et de lui permettre de protéger ses entreprises contre les enchevêtrements politiques nuisibles. Elle permet également d'être en concordance avec les " meilleures pratiques " en matière de conformité et de règles de gouvernance d'entreprise. En outre, elle accroît la transparence des réseaux économiques et politiques environnants. Le fait de fournir à un stade précoce des informations essentielles pour identifier les risques de réputation cachés et connexes contribue à élargir la base de prise de décision en ce qui concerne les investissements étrangers. Par conséquent, les ressources de gestion peuvent être utilisées efficacement et les coûts de transaction optimisés².

Selon une étude du cabinet de conseil Deloitte datée du 4 décembre 2012, 87% des dirigeants considèrent le risque de réputation comme le risque stratégique le plus important pour leur entreprise, et 41% de ceux-ci estiment qu'une atteinte à leur réputation personnelle peut avoir des répercussions directes sur le chiffre d'affaires de l'entreprise. Enfin, plus de la moitié des dirigeants interrogés souhaitent mettre en place des outils permettant d'analyser l'image de la marque afin d'anticiper tout risque de réputation.

Les médias sociaux à l'instar des blogs, des forums de discussions ou des wikis et le Web 2.0 (qui représente l'évolution du Web vers l'interactivité à travers une complexification interne de la technologie mais permettant plus de simplicité d'utilisation) ont renforcés le rôle de la

² Source: Berlin Risk Brief, No° 06, Octobre 2010.

réputation dans une grande variété de processus décisionnels économiques: Les consommateurs consultent la réputation fondée sur les médias sociaux dans leur choix de marques, les employés talentueux sont sensibles à la réputation fondée sur les médias sociaux lorsqu'ils décident de quitter une entreprise ou d'y rester et les investisseurs utilisent de plus en plus l'analyse de la réputation de l'entreprise cible sur les médias sociaux dans le cadre de leurs décisions de placement.

A cet effet, la due diligence réputationnelle s'effectue principalement par l'analyse de la réputation en ligne des entreprises ce que l'on appelle également «E-réputation » ou « Online Reputation (OR) ». Donc la OR est considérée comme l'outil principale de mesure de la réputation d'une entreprise utilisé lors d'une due diligence réputationnelle au sein de KPMG.

3. Online reputation (OR):

De nos jours, le web 2.0 est devenu une plateforme qui révolutionne la liberté d'expression des internautes. Il a complètement bouleversé le comportement des individus entre eux mais surtout face aux entreprises et aux institutions. Les nouvelles technologies accordent une large place à l'humain, à l'internaute qui peut réagir, contester, féliciter ou réprimander aussi facilement. Les avis des internautes ainsi que ceux des parties prenantes de l'entreprise peuvent influencer la réputation des entreprises. A cet effet, la mesure de la réputation est indispensable afin de préserver une image saine et pouvoir assurer la pérennité de l'activité de l'entreprise et de ses relations.

3.1. Corporate reputation (CR):

3.1.1. Définition de la CR :

A l'heure actuelle, une définition claire et universellement acceptée du terme « réputation de l'entreprise » ou « Corporate reputation » n'a pas encore été définie, en partie en raison de la récente apparition de ce terme. Fombrun (1996) définit la CR comme " *une représentation perceptuelle des actions passées et des perspectives d'avenir d'une entreprise qui décrit son attrait global pour toutes ses composantes clés, en comparaison des autres grands concurrents* ". Selon cette définition, " *la réputation est une perception globale de la mesure dans laquelle une organisation est tenue en haute considération ou en haute estime* ". Une bonne réputation est de prime abord une réputation qui conduit l'entreprise dans la direction souhaitée, tout en améliorant ses performances. Une bonne réputation est importante en raison de la capacité de créer de la valeur pour l'organisation, ainsi que de la possibilité de maintenir un profit supérieur. Le caractère immatériel de la réputation rend difficile aux concurrents de l'imiter, ce qui permet de créer un avantage concurrentiel durable. En outre, une bonne réputation peut accroître le rendement financier en empêchant d'autres acteurs d'entrer sur le marché de peur d'être absorbé (Brandtzæg, 2014).

3.1.2. L'importance de la CR dans les M&A :

Définir la réputation de l'entreprise comme l'actif le plus important de l'entreprise pourrait servir d'incitation à inclure une évaluation de la CR dans les fusions et acquisitions. La réputation de l'entreprise est en soi, constituée par plusieurs groupes d'intérêt à savoir: l'entreprise, les travailleurs, les gestionnaires, les concurrents, les collectivités et l'économie. Si les entreprises comparent la valeur financière d'avoir une bonne réputation avec l'effet financier de la perdre, il serait plus avantageux de dépenser de l'argent pour l'évaluer et la gérer, évitant ainsi d'éventuelles atteintes à la réputation.

Le fait de négliger la réputation d'une acquisition ou d'une fusion entraînera donc une surestimation ou une sous-estimation de la valeur de la transaction et pourrait servir à expliquer pourquoi tant de fusions/acquisitions échouent (Brandtzæg, 2014).

3.1.3. L'importance de la CR dans la due diligence :

Lajoux et Elson (2010) affirment que la diligence raisonnable classique est quelque peu limitée, ce qui pourrait faire référence au rôle et à la portée de la réputation de l'entreprise. L'élargissement de la pratique actuelle et l'adaptation de la due diligence à la fusion/acquisition concernée pourraient être bénéfiques pour l'acquéreur, la société visée, les actionnaires et les autres acteurs intéressés.

La raison principale d'inclure la réputation de l'entreprise dans la due diligence et, par conséquent, dans les fusions et acquisitions, dépend des sociétés qui tirent leur valeur des actifs incorporels. Il faut procéder à un examen attentif des forces importantes et des faiblesses de l'entreprise et, par conséquent, des fusions et acquisitions, principalement parce que les entreprises tirent leur valeur des actifs incorporels.

Epson (2005) soutient également qu'une due diligence incomplète est l'une des raisons pour lesquelles les fusions et acquisitions échouent, ce qui appuie l'hypothèse selon laquelle il faut étendre la due diligence classique (financière) et par conséquent la pratique des fusions et acquisitions. De plus, l'objectif premier du processus de due diligence est de rassurer les deux parties sur le fait que la fusion ou l'acquisition ne présente pas de risques en matière de concurrence et qu'elle peut avoir un impact sur les résultats de l'entreprise.

Toutefois, si l'enquête n'est pas faite sur tous les aspects importants, comme la réputation, la valorisation n'est pas fiable. Par exemple, si l'on découvrait une réputation négative dans le processus de due diligence réputationnelle, cela pourrait affecter le prix payé à la société cible, si ce n'est la décision d'annulation de la transaction (Brandtzæg, 2014).

3.1.4. Mesure de la CR:

La réputation de l'entreprise n'est pas une variable facile à mesurer avec précision. De multiples disciplines ont étudié les aspects clés de la définition de la réputation de l'entreprise et diverses approches hétérogènes ont été utilisées pour mesurer la construction de la réputation de l'entreprise. Parmi les méthodologies les plus utilisées pour évaluer la CR:

Les indicateurs de fortune: AMAC (America's Most Admired Companies) et GMAC (Global Most Admired Companies). Ces deux indices classent les organisations en fonction des résultats financiers, de la meilleure performance et du chiffre d'affaires.

Le quotient de réputation : C'est un indice fondé sur l'enquête auprès de la population générale et de l'ensemble de la population qui vise à savoir quelles entreprises sont appréciées et respectées par les individus, et pour quelles raisons. Il est basé sur 6 facteurs de réputation : Attractivité émotionnelle, produits et services, vision et leadership, environnement du lieu de travail, performance financière et la responsabilité sociale.

La méthodologie RepTrak : Elle a vu le jour en 2006 et a été créée pour fournir aux cadres supérieurs un instrument d'analyse qui pourrait non seulement suivre et évaluer les perceptions des parties prenantes à l'égard des entreprises, mais aussi pour qui permettrait

également une compréhension plus complète de l'information sous-jacente des moteurs de réputation qui suscitent l'attachement émotionnel.

Les critiques de ces méthodes estiment que les indices ne prennent en compte que certaines parties prenantes, la mesure évalue peu au-delà de la performance financière et que les évaluations peuvent ne pas correspondre à la réalité et ne pas intégrer une vision multi-stakeholder en incluant toutes les parties prenantes.

Outre les méthodologies présentées précédemment, il existe des approches quantitatives qui tendent à surmonter les faiblesses des méthodologies qualitatives et qui sont :

Approche du capital intellectuel : L'approche du capital intellectuel évalue la différence entre le prix du marché et la valeur comptable par l'estimation appropriée de 5 dimensions : marque de commerce, marques de service, droits d'auteur, autorisations et droits exclusifs. La limite évidente de cette approche réside dans l'hétérogénéité des différences dans les bilans, qui ne permettent pas une comparaison entre plusieurs entreprises. De plus, elle ne couvre pas les événements soudains qui peuvent sérieusement affecter la réputation d'une organisation.

Approche comptable : Une approche qui est davantage axée sur l'évaluation de la performance des actifs incorporels de l'entreprise c'est à dire que l'on ne peut pas toucher. Ce sont les brevets, frais d'établissement, fonds de commerce, licences et tous autres bien immatérielles que l'entreprise peut posséder. Le problème est qu'il n'existe pas de critères clairs pour l'évaluation de la juste valeur.

Approche marketing : C'est une approche qui mesure la réputation de l'entreprise à travers le concept de capital de marque qui représente la différence provoquée par la connaissance de la marque dans la manière dont les consommateurs réagissent au produit et à son marketing. Il y a 5 facteurs contribuant à la constitution du capital de marque : la fidélité, la notoriété, la qualité perçue, les associations de marques et les autres atouts liés à la marque. Cependant, la marque ne représente qu'une dimension et ne peut donc pas expliquer tous les aspects liés au concept de réputation (Barbato, 2016).

3.2. Online reputation (OR) :

3.2.1. Définition de la OR :

L'apparition du Web 2.0 et l'émergence des médias sociaux ont responsabilisé les consommateurs et entraîné un changement de comportement, puisqu'ils sont devenus les principaux contributeurs au contenu du Web en partageant informations et expériences, influençant ainsi leurs pairs dans leur processus décisionnel de produits et services et rendant les décisions d'achat dépendantes de cette information sans que les entreprises soient en mesure d'influencer de tels processus.

La réputation en ligne, aussi appelée E-Réputation ou Online reputation (OR) désigne l'image perçue par les internautes pour une marque, une entreprise ou un site web. Elle représente la réputation des entreprises établies sur Internet et se construit principalement par la participation de la communauté et par le biais des " agrégateurs de réputation ", tels que les moteurs de recherche comme Google, qui permettent aux gens de trouver du contenu en ligne.

La OR représente, de nos jours, un enjeu majeur qui constitue la Corporate Réputation (CR) des entreprises. Elle reflète une grande partie de la CR mais les deux termes ne sont pas interchangeables.

La participation de la communauté s'obtient par une communication cohérente avec les parties prenantes des entreprises, mais la réputation peut même se forger par des expériences indirectes avec l'entreprise, déclenchées par le bouche-à-oreille, les médias ou d'autres plateformes. Il est important que les communicateurs des entreprises sachent ce qui se dit sur leur entreprise sur Internet et qu'ils participent à la communication de leurs positions sur les questions clés dans les forums, les blogs et autres moyens qui atteignent leurs parties prenantes. Les blogues et les courriels des employés sont de plus en plus un moyen d'améliorer la réputation de l'entreprise mais peuvent aussi générer des problèmes pour l'entreprise.

De nos jours, les entreprises commencent à apprendre comment gérer ces nouveaux canaux de communication et essayent de gérer l'influence sur l'opinion des parties prenantes sachant que toute personne ayant un ordinateur relié à Internet peut générer du contenu pouvant être positif, neutre ou négatif.

3.2.2. La OR comme mesure essentielle de la CR :

L'attitude à l'égard des entreprises est de plus en plus influencée par les opinions qui circulent dans les réseaux numériques. Traditionnellement, les médias d'information étaient les principaux pour permettre aux intervenants d'acquérir des connaissances sur la CR qui était toutefois difficile à observer directement. Aujourd'hui, de plus en plus de personnes acquièrent des connaissances sur une entreprise en recherchant et en interprétant des mentions en ligne de cette dernière.

Les internautes ne se basent plus seulement sur l'éventail des comparaisons entre entreprises ayant des offres similaires, mais aussi sur la façon dont un réseau social perçoit la performance et la qualité d'une entreprise. Une fois que les gens ont construit une image, ils partagent leurs opinions et leurs sentiments avec les autres et " la vérité subjective devient une vérité collective sur ce qu'est une organisation et ce qu'elle devrait être ".

L'évaluation des opinions et des sentiments des utilisateurs dans les médias sociaux en ligne représente un bon indicateur de la CR, il est donc possible, à partir des données produites dans les sites web ainsi que les médias sociaux, d'extraire, de surveiller et même de prédire les tendances de la réputation de l'entreprise en agrégeant les opinions subjectives à l'aide de techniques de data mining. Parmi les différentes techniques d'analyse des grandes données, l'analyse de sentiment représente le meilleur outil pour interpréter la grande quantité de données disponibles sur les médias sociaux (Barbato, 2016).

Pour effectuer une due diligence réputationnelle, les entreprises utilisent des outils adaptés pour scruter le web et les réseaux sociaux en permanence afin de récolter, traiter puis classer l'ensemble des informations, mentions et dialogues qui se diffusent en ligne. Ce mode opératoire nécessite la prise en compte de l'aspect « confidentialité des données » afin de pouvoir exercer une analyse tout en respectant la vie privée des internautes ainsi que les données à caractère personnel même si l'ensemble des sources utilisées sont des sources ouvertes au grand public.

D'autre part, les entreprises et particuliers exigent de plus en plus de transparence de la part des fournisseurs de solutions utilisant des données personnelles quant aux moyens employés pour exécuter leurs prestations. Le recours à des sous-traitants et le transfert des données à l'étranger peuvent constituer un risque quant à la garantie de confidentialité des données et c'est pour cette raison que la protection des données va être traitée dans la partie qui suit.

4. Protection des données :

Aujourd'hui, toute entreprise ou organisme public détient des informations sur les citoyens, dès lors qu'ils utilisent leurs services. La question est d'autant plus sensible sur internet où chaque service en ligne (gratuit ou non), comme les réseaux sociaux, les moteurs de recherche, les plateformes de vente en ligne, requièrent et enregistrent des données sur ses utilisateurs. Il faut bien comprendre que la collecte de données n'est pas un problème en soi, mais c'est l'utilisation qui peut en être faite qui est controversée lorsqu'elle se fait à des fins commerciales.

L'importance de la protection des données augmente à mesure que la quantité de données créées et stockées continue de croître à un rythme sans précédent. La protection des données contre les compromis et la garantie de la confidentialité des données sont d'autres éléments clés de la protection des données.

Depuis les années 1960 et l'expansion des capacités des technologies de l'information, les entreprises et les gouvernements ont stocké les données personnelles dans des bases de données. Les bases de données peuvent être recherchées, éditées, croisées et partagées avec d'autres organisations à travers le monde. Une fois que la collecte et le traitement des données se sont généralisés, les gens ont commencé à poser des questions sur ce qui arrivait à leurs données une fois qu'ils les avaient fournies. Qui avait le droit d'accéder aux données ? A-t-il été conservé avec précision ? Était-elle recueillie et diffusée à leur insu ? Pourrait-il être utilisé pour discriminer ou violer d'autres droits fondamentaux ?

Après toutes ces questions et plusieurs études, il a été convenu qu'un cadre solide de protection des données peut responsabiliser les individus, limiter les pratiques nuisibles en matière de données et limiter l'exploitation des données. Il est essentiel de mettre en place les cadres de gouvernance nécessaires à l'échelle nationale et mondiale pour garantir que les individus ont de solides droits sur leurs données, des obligations strictes sont imposées à ceux qui traitent des données personnelles (dans les secteurs public et privé), et des pouvoirs d'exécution puissants peuvent être utilisés contre ceux qui violent ces obligations et protections. La région allemande de Hesse a adopté la première loi pour la protection des données en 1970. Aux États-Unis, la confidentialité des données n'a pas été très réglementée aussi tôt, de sorte que, par extension, il n'existe pas de lois strictes en matière de protection des données qui s'appliquent, bien que cela change rapidement à mesure que les gens prennent conscience de la valeur de la confidentialité et de la protection des données. Au Royaume-Uni, toutefois, l'organe législatif a adopté la loi sur la protection des données de 1998, une révision de la loi fondamentale de 1984 qui énonçait des règles pour les utilisateurs de données et définissait les droits des individus en ce qui concerne les données qui leur sont directement liées. La Loi est entrée en vigueur le 1er mars 2000. La loi elle-même s'efforce d'établir un équilibre entre les droits individuels à la vie privée et la capacité d'un plus grand nombre d'organismes publics d'utiliser ces données dans le cadre de leurs activités.

L'Algérie a mis en place un cadre juridique de protection des données à caractère personnel avec la promulgation de la loi n°18-07 du 10 juin 2018. Mais l'activité du département Deal Advisory de KPMG est soumise au règlement européen de la protection des données car ses missions sont dédiées exclusivement à l'activité off-shore pour le compte de KPMG France. Le Règlement Général sur la Protection des Données (RGPD) est le nouveau règlement européen sur la protection des données à caractère personnel. Il a été mis en application le 25 mai 2018 dans tous les pays européens et il fait sentir ses effets à travers le monde, y compris aux États-Unis et en Chine, puisque toute entreprise traitant des données personnelles d'Européens est obligée de l'appliquer. L'adoption de ce règlement permettra à l'Europe de s'adapter aux nouvelles réalités du numérique et de mieux protéger les droits et libertés des personnes.

Le règlement définit les droits des personnes physiques et fixe les obligations des organisations qui effectuent le traitement des données à caractère personnel et de celles qui sont responsables de ces traitements. Il définit également les méthodes visant à assurer le respect des dispositions prévues, ainsi que l'étendue des sanctions imposées à ceux qui enfreignent les règles.

Les données sont considérées "à caractère personnel" dès lors qu'elles concernent des personnes physiques identifiées directement ou indirectement. Par exemple : un nom, une photo, une empreinte, une adresse postale, une adresse mail, un numéro de téléphone, un numéro de sécurité sociale, un matricule interne, une adresse IP, un identifiant de connexion informatique, un enregistrement vocal, etc. Peu importe que ces informations soient confidentielles ou publiques.

Nous vivons aujourd'hui dans un monde connecté et KPMG évolue dans ce monde connecté. Nous utilisons chaque jour les technologies de l'information et de la communication pour gérer le patrimoine informationnel. Des renseignements à caractères personnels des clients, des prospects, des fournisseurs ou prestataires font l'objet d'une exploitation informatique. Le RGPD définit les principes à respecter lors de la collecte et de la conservation des données mais aussi le respect à l'égard des finalités et cela à propos des objectifs du traitement des données (KPMG intranet)

Conclusion :

Dans le cadre de ce travail, il a été nécessaire de définir le processus M&A, le rôle de la due diligence dans ce dernier ainsi que l'importance de la OR dans la mesure de la réputation des entreprises qui demeure une étape majeure en amont d'une transaction de fusion/acquisition et représente un indicateur incontournable pour des décisions stratégiques de grande envergure. Enfin, il nous a semblé important de traiter le volet de la protection des données qui demeure une pratique non négligeable dans l'exercice de toutes activités suscitant la manipulation de données.

Chapitre 2: Le Machine Learning au service de l'optimisation

Chapitre 2: Le Machine Learning au service de l'optimisation

Le modèle économique mondial est en pleine mutation suite au développement du numérique et des technologies qui en découlent : l'intelligence artificielle, le big data, l'impression 3D, les biotechnologies, la robotique ou encore l'internet des objets. Cette révolution a des conséquences gigantesques sur l'organisation du travail, l'emploi, l'environnement et des pans entiers de notre économie.

1. Machine Learning :

Définir le Machine Learning (ML) ou l'apprentissage machine n'est pas chose facile. Le champ est si vaste qu'il est impossible de la restreindre à un domaine de recherche spécifique; c'est plutôt un programme multidisciplinaire. Si son ambition initiale était d'imiter les processus cognitifs de l'être humain, plus précisément celui du raisonnement, ses objectifs actuels visent plutôt à mettre au point des automates qui résolvent certains problèmes bien mieux que les humains.

Au fond, le ML est simplement un moyen d'atteindre l'intelligence artificielle. Donc il est nécessaire de la définir avant d'aller vers d'autres détails.

1.1.Définition de l'intelligence artificielle :

Le dictionnaire Larousse 2018 définit l'IA comme un « ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine ».

Marvin Lee Minsky, un des créateurs de l'IA, l'a défini comme : "la construction de programmes informatiques qui s'adonnent à des tâches qui sont pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critiquée."

En envoyant une lettre aux actionnaires d'Amazon dont il est fondateur, Jeff Bezos a écrit: «Au cours des dernières décennies, les ordinateurs ont largement contribué à l'automatisation des tâches que les développeurs pouvaient transposer en règles et algorithmes clairs. Les nouvelles techniques de Machine Learning nous permettent désormais de faire de même pour des tâches bien plus complexes.»³

En d'autres termes, une intelligence artificielle est un programme informatique visant à effectuer, au moins aussi bien que des humains, des tâches nécessitant un certain niveau d'intelligence. L'horizon à atteindre concerne donc potentiellement l'ensemble des champs de l'activité humaine : déplacement, apprentissage, raisonnement, socialisation, créativité, etc. (Bertrand, 2017).

³ Business Insider France, Kif Leswing, 12 Apr 2017.

1.2.Définition du Machine Learning :

Le Machine Learning (ML) ou " apprentissage machine " est un sous-domaine de l'informatique. Il constitue le fondement d'un ensemble d'outils statistiques qui permettent d'estimer des fonctions complexes en tirant des leçons des données.

Arthur Samuel a inventé l'expression peu de temps après AI, en 1959, en la définissant comme " la capacité d'apprendre sans être explicitement programmé " en soulignant qu'il est possible d'obtenir l'IA sans utiliser le ML, mais cela nécessiterait de construire des millions de lignes de code avec des règles complexes.

Dans cette optique, Tom Mitchell donne une définition utile de l'apprentissage machine dans son livre *Machine Learning* :

"On dit qu'un programme d'ordinateur apprend de l'expérience E par rapport à une certaine classe de tâches T et à la mesure de performance P, si sa performance aux tâches en T, telle que mesurée par P, s'améliore avec l'expérience E." Et ceci en notant que :

La tâche T est le but d'apprentissage de l'algorithme. Il est défini par l'utilisateur et est généralement communiqué à l'algorithme en fournissant un exemple de la façon dont un événement doit être traité.

La mesure de performance P est la mesure par laquelle la compétence de l'algorithme est évaluée. La métrique peut être choisie par l'utilisateur et doit généralement être adaptée à la tâche T.

L'expérience E englobe l'ensemble des données fournies et toute information supplémentaire que l'algorithme de ML peut utiliser pour apprendre. C'est là que les algorithmes supervisés et non supervisés peuvent différer (Borovkov, 2017).

1.3.Objectif du Machine Learning :

Une différence majeure entre l'homme et la machine est depuis longtemps qu'un être humain a tendance à améliorer automatiquement sa façon d'aborder un problème. Les humains tirent des leçons de leurs erreurs passées et tentent de les corriger ou de trouver de nouvelles approches pour résoudre le problème. Les programmes informatiques traditionnels ne tiennent pas compte du résultat de leurs tâches et sont donc incapables d'améliorer leur comportement. Le domaine du ML s'attaque à ce problème et implique la création de programmes informatiques capables d'apprendre et donc d'améliorer leurs performances en recueillant plus de données et d'expérience.

Certaines tâches sont plus facilement "programmées" en utilisant la programmation conventionnelle, alors que d'autres se prêtent mieux au ML. Pour illustrer cela, un parallèle peut être établi avec le cerveau humain. Par exemple, il est possible d'expliquer comment faire ses lacets de chaussures, mais pas simple d'expliquer pourquoi l'on reconnaît le visage de quelqu'un qui nécessite une réflexion associée à une mémoire. Par conséquent, il est plus compliqué d'écrire un programme d'ordinateur traditionnel pour identifier un visage. Néanmoins, il est possible d'abreuver l'ordinateur avec des milliers d'images de visages et une image du visage que l'on souhaite qu'il reconnaisse jusqu'à ce qu'il puisse le déchiffrer et c'est ainsi que la reconnaissance faciale est née (Luckert et al. 2015).

Par ailleurs, une étude de PWC, faite en 2017, estime que le PIB mondial pourrait croître de 14% d'ici 2030 grâce à l'IA et le ML. Celle-ci devrait contribuer à hauteur de 15 700 milliards de dollars à l'économie mondiale en 2030, soit plus que le PIB cumulé actuel de la Chine et de l'Inde.

1.4.Types d'apprentissages :

Le ML consiste à concevoir des algorithmes qui permettent à un ordinateur d'apprendre. L'apprentissage n'implique pas nécessairement la conscience, mais l'apprentissage consiste à trouver des régularités statistiques ou d'autres modèles dans les données. Ainsi, de nombreux algorithmes de ML ressembleront à peine à la façon dont l'homme pourrait aborder une tâche d'apprentissage. Cependant, les algorithmes d'apprentissage peuvent donner un aperçu de la difficulté relative de l'apprentissage dans différents environnements. Pour cela il y a deux types d'apprentissages majeurs: Supervisé et non Supervisé.

1.4.1. Apprentissage Supervisé :

L'apprentissage supervisé signifie généralement que le programme reçoit à la fois l'entrée et la sortie souhaitée, par exemple, des images d'objets avec des étiquettes correspondantes de ce qui est représenté. Le but de l'apprentissage est de construire une carte entre ces deux éléments. Il est souvent assimilé à la classification.

Les valeurs d'entrée sont définies comme les informations externes que l'algorithme est autorisé à utiliser, telles que les valeurs d'attribut et les métadonnées, tandis que les valeurs de sortie sont les étiquettes spécifiques de l'attribut de classe. Cela signifie que la structure des données est déjà connue et que le but de ces programmes est d'affecter de nouvelles données aux classes correctes.

1.4.2. Apprentissage Non-Supervisé :

Contrairement à l'apprentissage supervisé, l'approche d'apprentissage non supervisé ne fournit pas au programme le résultat correct. Le but de l'apprentissage est de trouver la structure à l'intérieur de l'entrée de donnée. Il inclut, donc, toutes les tâches qui n'ont pas accès aux valeurs de sortie et tentent donc de trouver des structures dans les données en créant leurs propres classes.

L'apprentissage non supervisé est souvent assimilé au clustering. Le processus d'apprentissage n'est pas supervisé puisque les exemples d'entrée ne sont pas étiquetés par classe. Typiquement, il peut être utilisé pour découvrir des classes dans les données. Par exemple, une méthode d'apprentissage non supervisée peut prendre, en entrée, un ensemble d'images de chiffres manuscrits. Supposons qu'il trouve 10 groupes de données. Ces grappes peuvent correspondre aux 10 chiffres distincts de 0 à 9, respectivement. Cependant, comme les données de formation ne sont pas étiquetées, le modèle appris ne peut pas nous dire la signification sémantique des grappes trouvées (Ayodele, 2010).

1.5.Fonctionnement de l'apprentissage supervisé :

La classification est une tâche très répandue en Machine Learning. Dans ce genre de problématique, on cherche à mettre une étiquette (un label) sur une observation par exemple: une tumeur est-elle maligne ou non, une transaction est-elle frauduleuse ou non. Quand on a deux choix d'étiquettes possibles (tumeur maligne ou non), on parle de « Binary

Classification » (classification binaire). Par ailleurs, l'étiquette Y aura deux valeurs possibles 0 ou 1.

Le processus d'apprentissage est décrit dans la figure suivante :

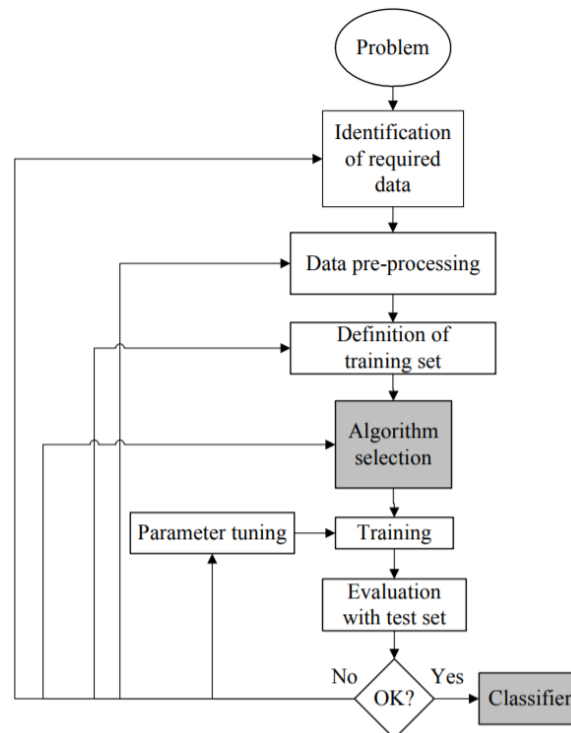


Figure 2: Processus d'apprentissage (Kotsiantis, 2007)

La première étape consiste à recueillir l'ensemble de données. Un expert du domaine peut suggérer quels champs (attributs, caractéristiques) sont les plus informatifs. Dans le cas contraire, la méthode la plus simple est celle de la "force brute", qui consiste à mesurer tout ce qui est disponible dans l'espoir de pouvoir isoler les bonnes caractéristiques (informatives, pertinentes). Cependant, un ensemble de données collectées par la méthode "brute-force" ne convient pas directement à l'induction. Il contient dans la plupart des cas du bruit et des valeurs de caractéristiques manquantes, et nécessite donc un prétraitement important.

La deuxième étape est la préparation et la préprogrammation des données. Selon les circonstances, les chercheurs ont le choix entre plusieurs méthodes pour traiter les données manquantes ou aberrantes. La sélection de caractéristiques n'est pas seulement utilisée pour gérer le bruit, mais aussi pour faire face à l'impossibilité d'apprendre à partir de très grands ensembles de données. Il existe une variété de procédures pour échantillonner des instances à partir d'un vaste ensemble de données. La sélection de sous-ensembles de caractéristiques est le processus d'identification et de suppression d'autant de caractéristiques non pertinentes et redondantes que possible. Cela réduit la dimensionnalité des données et permet aux algorithmes d'exploration de données de fonctionner plus rapidement et plus efficacement. Ce problème peut être résolu en construisant de nouvelles fonctionnalités à partir de l'ensemble de fonctionnalités de base. Cette technique s'appelle la construction/transformation d'entités. Ces nouvelles fonctionnalités peuvent conduire à la création de classificateurs plus concis et plus précis. De plus, la découverte de caractéristiques significatives contribue à une meilleure compréhension du classificateur produit et à une meilleure compréhension du concept appris.

La quatrième étape consiste au choix du modèle d'apprentissage qui va permettre de passer aux dernières étapes qui sont l'apprentissage du modèle et l'évaluation de son efficacité.

Le choix de l'algorithme d'apprentissage spécifique que nous devrions utiliser est une étape critique. Une fois que les tests préliminaires sont jugés satisfaisants, le classificateur (mappage des instances non étiquetées aux classes) est disponible pour une utilisation de routine. L'évaluation du classificateur est le plus souvent basée sur la précision de la prédiction (le pourcentage de la prédiction correcte divisé par le nombre total de prédictions). Il existe au moins trois techniques utilisées pour calculer la précision d'un classificateur. Une technique consiste à diviser l'ensemble d'entraînement en utilisant les deux tiers pour l'entraînement et l'autre tiers pour l'estimation de la performance. Dans une autre technique, appelé validation croisée, l'ensemble de formation est divisé en sous-ensembles de taille égale et mutuellement exclusifs et, pour chaque sous-ensemble, le classificateur est formé sur l'union de tous les autres sous-ensembles.

La technique de validation croisée permet de séparer les données d'une base de données en k partitions. Les $k-1$ partitions servent à l'apprentissage. La partition restante sert à la validation du modèle. L'opération est répétée k fois jusqu'à ce que toutes les partitions aient servi à la validation. Cela permet d'utiliser l'ensemble des données pour l'apprentissage et pour la validation. La moyenne du taux d'erreur de chaque sous-ensemble est donc une estimation du taux d'erreur du classificateur.

Si l'évaluation du taux d'erreur n'est pas satisfaisante, nous devons revenir à une étape précédente du processus (Figure 2). Divers facteurs doivent être examinés : les caractéristiques pertinentes pour le problème ne sont peut-être pas utilisées, un ensemble d'apprentissage plus large est nécessaire, la dimensionnalité du problème est trop élevée, l'algorithme sélectionné est inapproprié ou un réglage des paramètres est nécessaire (Kotsiantis, 2007).

1.6.Types d'algorithmes utilisés :

Il existe plusieurs types d'algorithmes qui sont utilisés en Machine Learning chacun à des fins précises et basé sur des méthodes mathématiques différentes. Parmi ces derniers :

1.6.1. Régression linéaire:

Le but de la classification dans les classificateurs linéaires est de regrouper en groupes les éléments qui ont des valeurs de caractéristiques similaires. Un classificateur linéaire est souvent utilisé dans des situations où la vitesse de classification constitue un problème, puisqu'il est considéré comme le classificateur le plus rapide. Le taux de convergence entre les variables de l'ensemble de données dépend toutefois de la marge. En gros, la marge quantifie à quel point un ensemble de données est séparable linéairement, et donc à quel point il est facile de résoudre un problème de classification donné.

Le modèle de régression linéaire décrit la variable de sortie y (un scalaire) comme une combinaison affine des variables d'entrée x_1, x_2, \dots, x_p (chacun un scalaire) plus un terme de bruit ε :

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}_{f(\mathbf{x}; \boldsymbol{\beta})} + \varepsilon.$$

Nous nous référons aux coefficients $\beta_0, \beta_1, \dots, \beta_p$ comme paramètres dans le modèle, et nous nous référons parfois à β_0 spécifiquement comme terme d'interception. Le terme de bruit ε tient compte des erreurs non systématiques, c'est-à-dire aléatoires, entre les données et le modèle. Le bruit est indépendant de x et tend vers la valeur zéro quand les β_i (paramètres du modèle) sont précis (Lindholm et al. 2019).

1.6.2. Régression Logistique :

La régression logistique indique habituellement où se trouve la frontière entre les classes, et indique également que les probabilités de classe dépendent de la distance de la frontière, selon une approche spécifique. La régression logistique est une approche de prédiction. Cependant, avec la régression logistique, la prédiction aboutit à un résultat dichotomique. La régression logistique est l'un des outils les plus couramment utilisés pour les statistiques appliquées et l'analyse de données discrètes.

Le modèle de régression logistique est défini par la relation suivante :

$$P = \frac{1}{1 + e^{-y^w x}}$$

Où x est le vecteur de la donnée où $x_i \in \mathbb{R}^n$, y est le vecteur de l'étiquette de la classe où $y_i \in \{1, -1\}$ et $w \in \mathbb{R}^n$ est le vecteur des poids (Lindholm et al. 2019).

1.6.3. Arbres de décision :

Un arbre de décision est une technique de classification qui met l'accent sur une forme de représentation facile à comprendre et qui est l'une des méthodes d'apprentissage les plus courantes. Les arbres de décision utilisent des ensembles de données constitués de vecteurs d'attributs, qui contiennent à leur tour un ensemble d'attributs de classification décrivant le vecteur et un attribut de classe attribuant l'entrée de données à une certaine classe. Un arbre décisionnel est construit en divisant itérativement l'ensemble de données sur l'attribut qui sépare le mieux possible les données en différentes classes existantes jusqu'à ce qu'un certain critère d'arrêt soit atteint. Le formulaire de représentation permet aux utilisateurs d'obtenir une vue d'ensemble rapide des données, car les arbres de décision peuvent être facilement visualisés dans un format structuré en arbre, facile à comprendre pour les humains (Luckert et al. 2015).

1.6.4. Nearest Neighbor learning (k-NN) :

Nearest Neighbor learning aussi connu sous le nom *Instance-Based Learning* décrit le processus de résolution de problèmes basé sur des solutions à des problèmes similaires déjà connus. Chaque système k-NN nécessite un ensemble de paramètres :

- Une fonction de distance qui mesure la similitude entre les problèmes ou les entrées de données. Ceci est nécessaire pour mesurer quels sont les voisins les plus proches du nouveau problème.
- Un certain nombre de voisins dont il faut tenir compte lorsqu'on s'attaque au nouveau problème
- Une fonction de pondération qui permet de quantifier davantage les voisins trouvés afin d'augmenter la prédiction et la qualité de l'apprentissage.

- Une méthode d'évaluation qui décrit une fonction sur la façon d'utiliser les voisins trouvés pour résoudre le problème donné (Luckert et al. 2015).

- ❖ Après avoir vu ces différentes méthodes qui regorgent de différents algorithmes du Machine Learning il est nécessaire de définir les principes de traitement de texte pour aboutir aux classifications citées précédemment.

1.7.Natural Language Processing :

Le Natural Language Processing (NLP) est un domaine multidisciplinaire qui emprunte à l'informatique, à la linguistique et à la psychologie cognitive. Il combine leur théorie avec le calcul pour traiter les textes en langage naturel (humain). En d'autres termes, la NLP implique la représentation et l'analyse informatique, c'est-à-dire la compréhension et la génération de texte.

Le NLP traite les textes à différents niveaux d'analyse linguistique :

- Phonologie : Il s'agit de l'étude du fonctionnement des sons de la parole et de leur organisation dans un langage naturel particulier.
- Morphologie : Ce niveau effectue ensuite la décomposition morphologique des mots en racines et affixes pour en déduire leur structure interne.
- La lexicologie : L'analyse lexicale détermine la signification ou le sens sous-jacent des mots individuels, généralement par la recherche dans un dictionnaire appelé lexique.
- Syntaxe : Ce niveau déduit la structure grammaticale de la phrase, c'est-à-dire les dépendances structurelles entre les mots constitutifs.
- Sémantique : En général, il s'agit de l'étude du sens des expressions linguistiques. Plus étroitement définie, il s'agit de l'étude du sens des mots au niveau de la phrase, sans tenir compte du discours et des facteurs pragmatiques.
- Discours : Alors que la syntaxe et la sémantique sont donc des analyses au niveau des phrases, ce niveau d'analyse fonctionne sur l'ensemble du document ou du discours, reliant le sens entre les phrases.
- Pragmatisme : Il s'agit de l'étude de la signification dans le contexte au-delà de ce qui peut être saisi par le texte, par exemple l'intention, le plan et/ou le but de l'orateur, le statut des parties impliquées et d'autres connaissances mondiales.

L'un des aspects les plus importants du ML dans le NLP est de décider quelles fonctions utiliser. Pour divers types de problèmes de classification de texte semblables à celui de la présente thèse, l'approche de base la plus courante est le modèle Bag of Words mais ce n'est pas la seule (Schlünz, 2014).

1.7.1. Bag-of-Words Model :

a. BOW model :

Le modèle du BOW (BOW) considère chaque message comme un ensemble de mots qui se répètent un certain nombre de fois. La représentation du document est totalement désordonnée, car chaque mot est traité indépendamment du mot précédent et du mot suivant. Par exemple, nous avons un ensemble de données composé de seulement deux messages : « Le chat vaut mieux que le chien » et « Le temps est meilleur qu'hier ».

Le tableau ci-dessous montre la représentation des deux vecteurs :

	The	Cat	Is	Better	Than	Dog	Weather	Yesterday
Vector 1	2	1	1	1	1	1	0	0
Vector 2	1	0	1	1	1	0	1	1

Tableau 1: Exemple des vecteurs Bag-of-Words (Lundborg, 2017.)

Au fur et à mesure que le nombre d'échantillons augmente, le nombre de mots uniques augmente. Puisque chaque mot unique est représenté par une position spécifique dans le vecteur, ces vecteurs vont naturellement s'agrandir également. Le vecteur aura la longueur du nombre total de mots uniques existant dans l'ensemble de données, mais peut aussi être limité au nombre X des mots les plus courants de l'ensemble de données.

b. BOW N-gram model:

Le modèle N-gram, de plus que le BOW model classique, tient compte de l'ordre des mots en comptant les séquences de mots, où N est le nombre de mots à inclure dans une séquence. En incluant des séquences de mots, nous pouvons rendre compte d'un sens plus profond dans les phrases et capturer plus de nuances dans le texte.

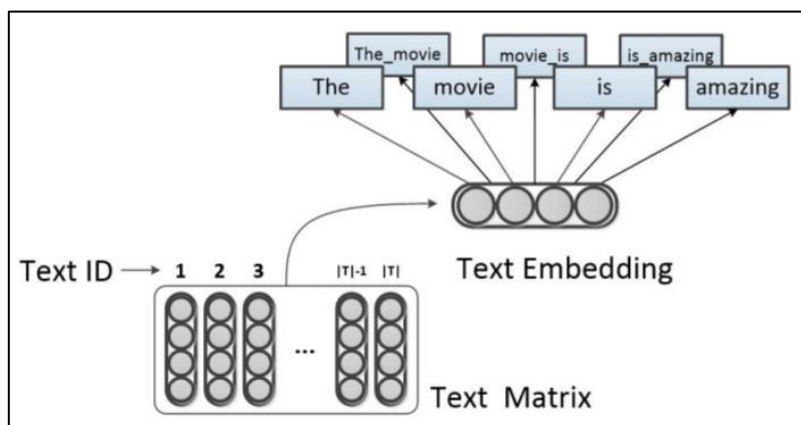


Figure 3: Exemple modèle BOW 2-gram

La figure ci-dessus décrit un exemple de modèle BOW à 2-gram pour la phrase « The movie is amazing » où chaque deux (2) mots consécutifs de la phrase sont pris en compte ensemble pour le traitement.

Les modèles de BOW (N-gram) traitent un texte comme un ensemble de mots (N-gram), qui ignorent le fait que les textes sont essentiellement des données séquentielles bien que les informations d'ordre figurant dans les séquences de mots soient rejetées, les modèles de BOW (N-gram) sont d'une efficacité surprenante, et bénéficient également des avantages d'être plus performants et solides. Ils ont été largement utilisés dans divers types de tâches de PNL telles que la recherche d'information, les réponses aux questions et la classification des textes. Habituellement, les fonctions BOW éparées nécessitent des techniques de pondération pour obtenir de meilleures performances, où les mots importants ont plus de poids alors que les mots sans importance ont moins de poids.

1.7.2. Character N-gram Model:

Le Character N-gram est similaire au BOW N-gram, mais au lieu de créer des représentations vectorielles de mots et de combinaisons de mots, nous créons une représentation vectorielle basée sur les caractères et combinaisons de caractères. De ce fait, on peut plus facilement

expliquer les fautes d'orthographe. A titre d'exemple, nous utilisons le texte suivant : one dog, one cat. En prenant $N=2$, c'est-à-dire 2-gram, et en représentant les caractères espace avec « _ », on obtiendrait le vecteur de caractéristique suivant:

on	ne	e_	_d	do	og	g,	_o	_c	ca	at
2	2	2	1	1	1	1	1	1	1	1

1.7.3. Word2vec :

Word2vec est une technique d'intégration de mots capable de saisir les degrés de similitude entre les mots. Dans le traitement du langage naturel, l'incorporation de mots est le nom des techniques utilisées pour associer des mots ou des phrases à des vecteurs de nombres réels et word2vec est une technique capable de produire ces vecteurs, où des mots ayant une signification ou un contexte similaire se produisent à proximité les uns des autres dans l'espace vectoriel.

Les vecteurs de mots qui en résultent ont des propriétés intéressantes et quelque peu surprenantes. Par exemple, Si l'on considère un modèle word2vec formé avec des messages suédois sur Twitter², les 5 vecteurs les plus proches du mot Stockholm sont : sthlm, stockholm, Göteborg et Malmö et Uppsala, ce qui signifie que la distance en espace vectoriel capture quelque peu les similitudes entre les villes (grandes villes en Suède). Les mots qui se trouvent souvent à proximité l'un de l'autre dans les phrases auront des vecteurs situés à proximité l'un de l'autre dans l'espace vectoriel (Lundborg, 2017).

2. Les outils de programmation :

Le ML aide à remodeler la vision de l'informatique de façon plus générale. En déplaçant la question de "comment programmer des ordinateurs" à "comment leur permettre de se programmer eux-mêmes", il met l'accent sur la conception de systèmes d'autosurveillance qui s'auto-diagnostiquent et s'autoréparent, et sur des approches qui modélisent leurs utilisateurs, et qui tirent profit du flux constant de données qui circulent dans le programme plutôt que de simplement le traiter.

De même, le ML aidera à remodeler le domaine de la statistique, en mettant l'accent sur une perspective informatique et en soulevant des questions comme l'apprentissage sans fin. Mais, avant tout, la programmation, l'informatique en général et les statistiques contribueront à façonner le ML au fur et à mesure qu'ils progressent et fourniront de nouvelles idées pour changer la façon dont nous envisageons l'apprentissage.

Il existe de nombreuses et excellentes boîtes à outils qui fournissent un support pour le développement de logiciels de Machine Learning sous Python, C++, R, Matlab, et des environnements similaires.

2.1. Python:

Le langage de programmation Python s'impose comme l'un des langages les plus populaires pour le calcul scientifique. Grâce à son caractère interactif de haut niveau et à son écosystème de bibliothèques scientifiques en pleine maturité, c'est un choix intéressant pour le développement algorithmique et l'analyse exploratoire de données.

Scikit-learn est une boîte à outils qui exploite ce riche environnement pour fournir des implémentations de pointe de nombreux algorithmes de ML bien connus, tout en maintenant une interface facile à utiliser étroitement intégrée avec le langage Python. Cela répond au

besoin croissant d'analyse statistique de données par des non-spécialistes dans les industries du logiciel et du web, ainsi que dans des domaines autres que l'informatique, comme la biologie ou la physique.

Scikit-learn expose une grande variété d'algorithmes d'apprentissage machine, supervisés ou non, à l'aide d'une interface cohérente et orientée tâche, permettant ainsi une comparaison facile des méthodes pour une application donnée. Puisqu'il s'appuie sur l'écosystème scientifique de Python, il peut facilement être intégré dans des applications en dehors du cadre traditionnel de l'analyse de données statistiques.⁴

2.2.R:

R est l'un des systèmes logiciels les plus populaires et les plus utilisés pour les statistiques, l'exploration de données et l'apprentissage machine. Toutefois, elle ne définit pas d'interface normalisée pour, par exemple, la modélisation prédictive supervisée. Pour toute expérience non triviale, il faut écrire un code long, fastidieux et sujet aux erreurs pour unifier les méthodes d'appel et gérer les résultats. Le progiciel MLR offre un langage propre, facile à utiliser et flexible pour les expériences d'apprentissage machine en R. Il prend en charge la classification, la régression, le clustering et l'analyse de survie avec plus de 160 techniques de modélisation. Définir les tâches d'apprentissage, les modèles de formation, faire des prédictions et évaluer les résumés de performance de la mise en œuvre de l'apprenant sous-jacent via une interface orientée objet.

MLR va bien au-delà de la simple fourniture d'une interface unifiée. Il met en œuvre une architecture générique qui permet d'évaluer les performances de généralisation, de comparer différents algorithmes d'une manière scientifiquement rigoureuse, de sélectionner des caractéristiques et d'accorder des hyperparamètres pour toute méthode, ainsi que d'étendre les fonctionnalités des apprenants grâce à un mécanisme de wrapper. Les propriétés interrogeables fournissent un mécanisme de réflexion pour les objets d'apprentissage machine. Enfin, MLR fournit des méthodes de visualisation sophistiquées qui permettent de montrer les effets de la dépendance partielle des modèles. L'objectif à long terme de MLR est de fournir un langage de haut niveau spécifique au domaine pour exprimer autant d'aspects que possible des expériences d'apprentissage machine.⁵

2.3. C++:

Le C++ est un langage très populaire qui allie puissance et rapidité. Il est dit "orienté objet (OO)" conçu à partir du langage C. De ce fait la plupart des fonctionnalités du C restent utilisables en C++.

Dlib-ml est une bibliothèque open source, destinée à la fois aux ingénieurs et aux chercheurs, qui vise à fournir un environnement aussi riche pour le développement de logiciels d'apprentissage machine en langage C++.

Sa conception est fortement influencée par les idées issues de la conception par contrat et du génie logiciel basé sur les composants. Cela signifie qu'il s'agit avant tout d'une collection de composants logiciels indépendants, accompagnés chacun d'une documentation complète et de modes de débogage complets. De plus, la bibliothèque est conçue pour être utile à la fois dans

⁴ Journal of Machine Learning Research 12 (2011) 2825-2830

⁵ Journal of Machine Learning Research 17 (2016) 1-5

les projets de recherche et dans les projets commerciaux du monde réel et a été soigneusement conçue pour faciliter son intégration dans l'application C++ d'un utilisateur.

L'un des principaux objectifs de conception de cette partie de la bibliothèque est de fournir une architecture très modulaire et simple pour traiter les algorithmes du noyau. En particulier, chaque algorithme est paramétré pour permettre à l'utilisateur de fournir l'un ou l'autre des éléments suivants des noyaux dlib-ml prédéfinis, ou un nouveau noyau défini par l'utilisateur. De plus, les implémentations des algorithmes sont totalement séparées des données sur lesquelles ils opèrent.

Cela rend l'implémentation dlib-ml suffisamment générique pour fonctionner sur n'importe quel type de données, qu'il s'agisse de colonnes ou de vecteurs, des images ou toute autre forme de données structurées. Tout ce qu'il faut, c'est un noyau approprié.

C'est une fonctionnalité unique à dlib-ml. De nombreuses bibliothèques permettent l'utilisation de noyaux pré-calculés arbitraires et d'éléments certains permettent même des noyaux définis par l'utilisateur mais ont des interfaces qui les limitent au fonctionnement sur colonne vecteurs. Cependant, aucune ne permet la flexibilité d'opérer directement sur des objets arbitraires, ce qui en fait un outil très utile plus facile d'appliquer des noyaux personnalisés dans le cas où les noyaux opèrent sur des objets autres que des vecteurs de longueurs fixes.⁶

Conclusion :

Lors de ce chapitre, les outils utilisés pour procéder à l'optimisation du processus de Due Diligence réputationnelle ont été explicités et ceci en commençant par la définition du Machine Learning qui représente une des techniques de l'intelligence artificielle, en citant les types d'apprentissage, les types d'algorithmes utilisés puis le Natural Language Processing qui représente un outils important dans l'analyse d'un texte ainsi que les différents langages de programmations qui sont utilisés dans le domaine.

La partie qui suit traitera un état des lieux qui étudiera le contexte dans lequel ce projet est fait, le diagnostic du processus actuel pour définir l'étendue du travail ainsi que les pistes d'améliorations.

⁶ Journal of Machine Learning Research 10 (2009) 1755-1758

Partie 2 : Etat des lieux

Chapitre 3: Présentation de KPMG

Partie 2 : Etat des lieux

Chapitre 3: Présentation de KPMG

Introduction :

Au cours des dernières années, le marché mondial du conseil ou « consulting » a connu une croissance importante, bien que les taux de croissance diffèrent entre les marchés plus matures et les économies émergentes.

L'industrie du conseil ou « consulting » se développe de plus en plus et offre des prestations dans des domaines divers procédant par des actions de nature très variée allant du diagnostic et de la simple recommandation à la mise en place de solutions complètes, engageant le court, le moyen et le long terme de l'entreprise cliente sur des aspects opérationnels, tactiques et stratégiques.

KPMG est l'un des cabinets de conseil les plus réputés en audit et en fiscalité, ainsi que dans les services-conseils. Les cabinets de KPMG encouragent les associés et le personnel à s'impliquer dans leurs collectivités pour bâtir des liens de confiance et favoriser le changement.

1. KPMG International :

1.1. Présentation de KPMG:

KPMG est un réseau international de cabinets d'audit, d'expertise comptable et de conseil exerçant dans 154 pays (Figure 4). C'est une société de services professionnels et l'un des 4 grands cabinets, appelés « Big Four », avec Deloitte, Ernst & Young (EY) et PricewaterhouseCoopers (PwC).

Le nom "KPMG" signifie "Klynveld Peat Marwick Goerdeler." Il a été choisi lors de la fusion de KMG (Klynveld Main Goerdeler) avec Peat Marwick en 1987.

Basée à Amstelveen, aux Pays-Bas, ce cabinet intervient auprès des petites, moyennes et grandes entreprises. Il comptait 207 050 employés (dont 47% de femmes et 53% d'hommes) en 2018 après avoir enregistré un record pour son chiffre d'affaire en 2017 d'une valeur de 26,4 milliards de dollars avec un taux de croissance de 4,8% par rapport à 2016.

En 2018, KPMG confirme son attractivité auprès des étudiants et jeunes diplômés en se plaçant 11ème du classement Universum des employeurs préférés.⁷

1.2. Les activités de KPMG:

KPMG a élargi ses domaines d'activités afin de répondre aux besoins sans cesse changeants de ses clients. Ils s'investissent dans de nouveaux services et de nouvelles technologies pour les secteurs où les défis et les perturbations sont les plus importants. Les activités de KPMG tournent autour de 3 pôles:

⁷ Source : Documents Internes KPMG

Audit – Commissariat aux comptes: KPMG audite, c'est à dire vérifie la sincérité et l'exactitude des comptes de ses clients afin de les certifier. La démarche d'audit s'appuie sur une connaissance approfondie des organisations et de leur environnement pour émettre une opinion sur les comptes de l'entreprise cliente.

Advisory (Conseil) : KPMG aide les entreprises à relever les défis auxquels elles doivent faire face dans un monde en pleine mutation, qu'il s'agisse de ruptures technologiques ou d'évolutions réglementaires quel que soit leur secteur. Le cabinet les conseille également dans la gestion de leurs opérations de restructuration, acquisition ou cession, et dans des situations de fraude ou litiges.

Cette activité compte 3 principales divisions :

- **Management consulting:** KPMG accompagne ses clients dans la définition et la conduite de leurs projets de transformation et d'amélioration de la performance dans les domaines opérationnels, l'organisation, la fonction finance et les systèmes d'information.
- **Risk consulting:** KPMG apporte des solutions pour évaluer et optimiser le dispositif de contrôle interne et de management des risques en conformité avec les orientations stratégiques et les obligations légales et réglementaires.
- **Deal advisory:** KPMG accompagne ses clients dans le cadre de leurs transactions de fusions/acquisitions en mettant à leur disposition des équipes spécialisées qui permettent de maximiser la valeur ajoutée créée pendant ces transactions.

ESC (Expertise, Services et Conseil): KPMG propose un accompagnement durable aux entrepreneurs dans tous leurs projets. Cet accompagnement concerne l'ensemble des métiers de l'ESC, à savoir : Expertise, gestion sociale, juridique et fiscale qui contribuent aux succès des clients et ceci à tous les stades de développement grâce à trois métiers conjoints: Expertise-Comptable, Gestion Sociale, Tax & Legal.

- **Expertise comptable :** KPMG intervient en tant qu'expert-comptable auprès des PME, groupes familiaux, TPE, Artisans, commerçant et professions libérales pour les accompagner et les conseiller à chaque étape de leur développement, aider à la création, évaluation, et la gestion sociale.
- **Gestion sociale :** KPMG s'engage même socialement. Cette offre est portée en région par des professionnels dédiés qui accompagnent les entreprises dans l'établissement de la paie et des déclarations sociales, le suivi des indicateurs sociaux et l'administration du personnel, etc.
- **Tax & Legal :** KPMG assure pour ses clients des prestations fiscales telles que la revue des déclarations fiscales, l'audit fiscal, la revue fiscale dans le cadre de l'exécution des contrats et l'assistance à la mise en place de procédures de conformité. Elle accompagne également ses clients dans le cadre des prestations juridiques comme la création de sociétés, les études sur les formes d'implantation et la réalisation d'une opération, la revue juridique de conformité, l'assistance en matière contractuelle et le secrétariat juridique.

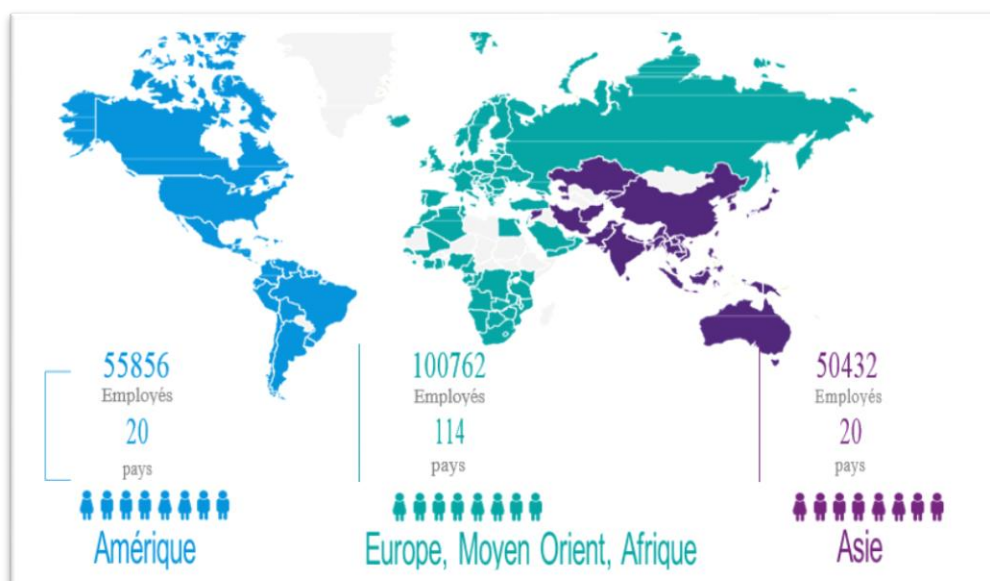


Figure 4: Présence de KPMG dans le monde (KPMG Intranet, 2018)

2. KPMG en Algérie :

2.1. Présentation de KPMG Algérie SPA :

KPMG Algérie est membre du réseau KPMG International constitué de cabinets indépendants adhérents de KPMG International Cooperative, une entité de droits Suisse. Le cabinet est présent à Alger et Oran sous forme de Société Par Actions (SPA). C'est une filiale de KPMG France avec laquelle une activité Off-shore a vu le jour récemment. C'est-à-dire que l'équipe française sous-traite des missions avec l'équipe Algérienne ce qui lui permet de gagner en coûts de revient sur les livrables et donc d'avoir un avantage concurrentiel sur le marché français.

Dans son approche de proximité et de disponibilité à travers le globe, et conscient du mouvement de libéralisation qui s'amplifie en Algérie, générant de nouveaux besoins pour les entreprises, KPMG Algérie SPA a été le premier des « Big Four » à s'implanter en Algérie en mars 2002 et compte parmi ses clients les plus prestigieuses références locales et internationales.

KPMG Algérie exerce la majorité des activités de KPMG International pour des clients locaux et internationaux. Elle offre, donc, des prestations d'audits, de conseil et d'expertise comptable.

2.2. Chiffre clés de KPMG Algérie SPA :

En 2018, KPMG Algérie SPA comptait plus de 120 employés pour servir plus de 200 clients enregistrant un chiffre d'affaires (CA) de 795 millions de DZD.

La figure, ci-dessous représente l'évolution du chiffre d'affaire de KPMG Algérie SPA durant les 5 dernières années ainsi que l'excédent brut d'exploitation (EBE) en pourcentage du chiffre d'affaires. L'évolution du chiffre d'affaires est relativement constante en allant de 635 en 2014 à 795 millions de DZD en 2018. Pour l'EBE, il représente 20% du chiffre d'affaires en 2018, ce qui reste considérable pour le cabinet.

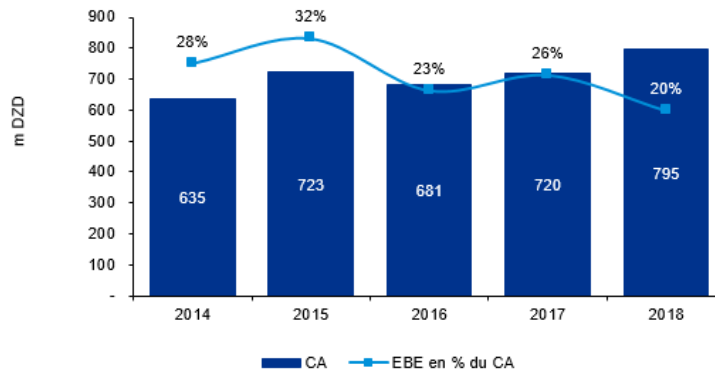


Figure 5: Evolution CA et EBE de 2014 à 2018

En 2018, KPMG Algérie SPA a généré un résultat net (RN) de 6 millions de DZD. Ce faible résultat est dû à l'amortissement de l'immeuble KPMG construit en 2015 et aux charges du personnel qui ne cessent d'augmenter d'année en année.

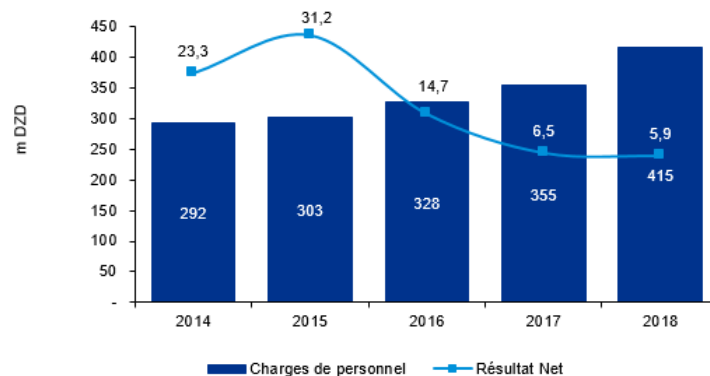


Figure 6: Evolution des charges du personnel et RN de 2014 à 2018

La figure ci-dessus décrit l'évolution des charges du personnel de KPMG Algérie SPA qui varie de 292 en 2014 à 415 millions de DZD en 2018 ce qui représente une augmentation considérable de plus de 42% vu que c'est une société de service. Le résultat net est en plein déclin depuis 2015 suite aux dotations aux amortissements de l'immeuble KPMG qui sont de l'ordre d'approximativement 100 millions de DZD chaque année.⁸

2.3. Structure organisationnelle de KPMG Algérie SPA :

Le cabinet est structuré en 5 départements en plus de la division d'Oran qui est constituée d'une petite équipe œuvrant à acquérir des marchés dans la région Ouest du pays pour le compte de KPMG Alger du fait de son manque d'effectif mais cette dernière participe à des petites prestations fiscales et d'expertise comptable.

⁸ Source : Documents internes KPMG

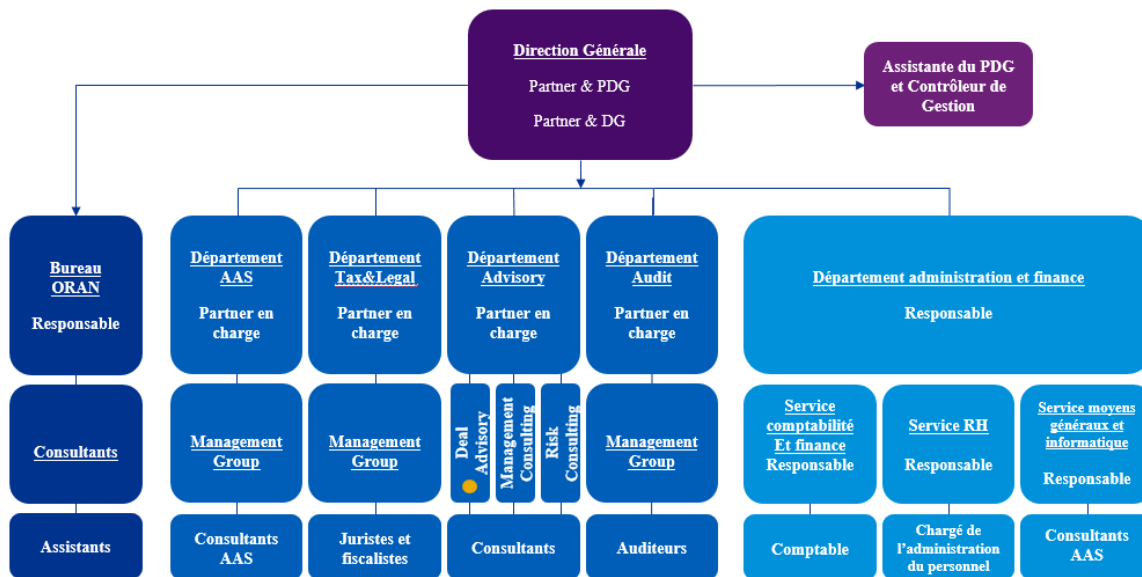


Figure 7: Organigramme KPMG Algérie SPA (Document Interne KPMG).

L'organigramme ci-dessus est fait en sorte de prendre en charge les activités de KPMG Algérie SPA décrites dans la première partie du chapitre, les départements ainsi que leurs missions sont résumés ci-après:

- Département administration et finance : C'est le département qui s'occupe des procédures administratives, des moyens généraux, de la comptabilité du cabinet, de la gestion des ressources humaines ainsi que du réseau informatique.
- Département Accounting Advisory Services (AAS) : C'est le département chargé des missions d'expertise comptable.
- Département Tax&Legal : C'est le département qui s'occupe des prestations fiscales et juridiques.
- Département Audit : C'est le département chargé des missions d'audits dans le domaine de la finance.
- Département Advisory : C'est le département qui englobe Risk Management, Management Consulting et Deal Advisory comme expliqué dans la partie « Activités de KPMG »

2.4.Présentation du Deal Advisory :

C'est le département qui se charge des transactions de fusions ou acquisitions pour l'activité Off-shore de KPMG Algérie SPA pour le compte de KPMG France. C'est dans ce dernier que nous avons effectué notre stage dans le cadre de notre projet de fin d'étude.

L'équipe Deal Advisory est entièrement dédiée à l'accompagnement de sociétés de toutes tailles et de fonds d'investissement durant toutes les étapes de leurs opérations de recherche de cible, évaluation financière, due diligence, business plan, revue du contrat d'acquisition ou de cession, ainsi que dans la gestion des situations de fraudes et de litiges qui peuvent affecter le bon déroulement d'une transaction et ceci en se positionnant du côté acheteur (l'acquéreur ou Buy Side) ou bien du côté vendeur (l'entreprise cible ou Sell Side).

Dans leur activité, les consultants du Deal Advisory font face à 3 types de clientèles :

- **Les entreprises Business to Consumer** : Les entreprises BtoC sont des entreprises qui opèrent dans les échanges commerciaux avec une clientèle de particuliers. La cible de client est plus large, mais moins experte en BtoC.
- **Les entreprises Business to Business** : Les entreprises BtoB sont des entreprises qui opèrent dans les échanges commerciaux réalisés avec une autre entreprise. La cible de clientèle est restreinte et se compose d'entreprises ayant, généralement, plus d'expertise métier qu'un simple consommateur.
- **Les fonds d'investissement** : ou « Financial organisation » (FO) sont des sociétés financières dont l'objectif consiste à investir dans des sociétés cibles pour leurs opportunités d'évolution, d'expansion et de développement.

Les consultants du Deal Advisory Alger combinent à la fois, ambition, jeunesse et dynamisme. Lancé en Avril 2017, ce service offshore compte aujourd'hui une équipe de 32 collaborateurs, dont des collaborateurs spécialisés, issus pour la plupart de grandes écoles et rassemblant des profils d'exception. Cette équipe est répartie en deux divisions Transaction Services (TS) et Recherche & Stratégie (R&S) représentées dans la figure ci-dessous.

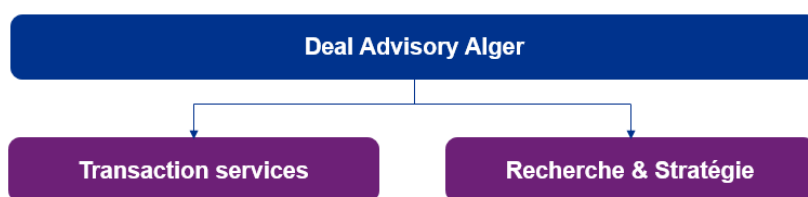


Figure 8: Structure Deal Advisory Alger (Document Interne KPMG).

2.4.1. Transaction services:

L'équipe « TS » d'Alger a été opérationnelle dès le lancement de l'activité offshore du Deal Advisory et à tout de suite su porter les projets au sein du cabinet. Les consultants TS opèrent dans 3 activités majeures :

- **L'élaboration des rapports de due diligence financière** : C'est un rapport contenant une analyse détaillée des éléments financiers relatifs à l'entreprise cible, ayant pour objectif d'identifier les points clés de la transaction et d'obtenir des bases financières pour valoriser une cible (EBE/EBITDA, EBIT, BFR) et des éléments de négociation du prix (ajustements de dette nette, garanties d'actifs et de passifs).
- **L'élaboration des Business Plan financiers** : Les Business Plan financiers sont des projections, sur les années à venir, des différents états financiers afin d'avoir une prévision sur l'évolution des agrégats financiers à moyen et long terme de l'entreprise cible généralement dans le but d'approcher les sociétés de financement pour acquérir des crédits d'investissement.
- **L'élaboration des rapports de valorisation** : La valorisation c'est l'estimation de la valeur de l'entreprise dans le cadre de la transaction. L'une des méthodes les plus utilisées est la valorisation patrimoniale (en prenant en compte les passifs de l'entreprise : ressources matérielles, immobilisations, etc.) mais cette dernière ne prend pas en compte les éléments qui ne sont pas tangibles comme le savoir-faire ou la

réputation. La deuxième méthode, qui reflète mieux la valeur de l'entreprise, est la valorisation par les multiples où chaque industrie a un coefficient de valorisation qui est multiplié par un indicateur financier. Par exemple, dans le secteur pharmaceutique africain, la valeur moyenne de l'entreprise est 1,8 fois la valeur de son chiffre d'affaire.

Afin de réaliser ces activités plusieurs tâches élémentaires sont prise ne compte comme l'analyse du chiffre d'affaires (saisonnalité, par activité, par client), l'analyse des charges de personnel, l'analyse du besoin en fonds de roulement, l'élaboration de bilans, P&L, trésorerie d'exploitation à partir d'une balance générale, la préparation de databooks dans lesquels sont présentés les états financiers et les différents tableaux de détails et d'analyses effectués, etc.

2.4.2. Recherche & Stratégie :

L'équipe « R&S » est une jeune équipe créée après l'équipe TS en fin 2017 et travaille en off-shore avec plusieurs équipes de KPMG France dans le cadre d'activités diverses à savoir :

- **L'équipe Pré-Deal** : L'équipe travaille sur des rapports d'opportunités contenant les informations clés pour un acheteur afin d'avoir les informations nécessaires sur l'entreprise qu'il veut acquérir telles que: L'aperçu de l'activité (Business Overview), les chiffres clés, l'historique de ses transactions, la présence géographique, la description des produits et marques, l'étude de la tendance du marché sur l'industrie en question, analyse de la concurrence et des parts de marché, etc. Des missions de valorisation sont également faites afin d'inclure l'ordre de grandeur de la valeur de l'entreprise dans le rapport d'opportunité.
- **L'équipe Distressed M&A** : C'est l'activité relative à la reprise d'entreprise en difficulté comme les entreprises placées dans le cadre d'une procédure collective (judiciaire) et qui sont en voie de liquidation. L'équipe d'Alger intervient dans l'établissement des teasers (Un présentation brève) pour des entreprises en difficultés afin de présenter leur activité et donner des chiffres clés à d'éventuels repreneurs. Ces derniers sont aussi recherchés par l'équipe R&S sous forme de liste de repreneurs potentiels.
- **L'équipe Environnement, Social et Gouvernance (ESG)** : Les investisseurs s'intéressent de plus en plus à la performance environnementale, sociale et de gouvernance (ESG) des entreprises dans lesquelles ils investissent. L'attention ne vient pas seulement du marché croissant des investisseurs responsables, mais aussi des investisseurs traditionnels qui s'intéressent de plus près à l'ESG alors qu'ils commencent à regarder au-delà des horizons d'investissement à court terme pour créer de la valeur à plus long terme pour les actionnaires.

C'est donc cette équipe qui est chargée de la réalisation de la Due Diligence réputationnelle, objet de travail de notre projet de fin d'étude.

Conclusion :

La présentation du cabinet a été faite afin de situer le champ d'application à l'issue de ce travail. La présentation a commencé par KPMG International passant par KPMG Algérie SPA puis le département Deal Advisory où intervient l'équipe « Recherche et Stratégie » (R&S) sur des missions avec l'équipe ESG de Paris.

La demande croissante des investisseurs qui sont en quête d'investissements dans des entreprises durables, conjuguée aux attentes croissantes du public en matière de responsabilité sociale des entreprises, mettent davantage l'accent sur les questions environnementales, sociales et de gouvernance (ESG) relatives à leurs métiers.

Lors de son activité ESG, l'équipe R&S se concentre essentiellement sur l'aspect réputationnel, plus précisément sur la Online Reputation (OR), des entreprises en procédant à une revue digitale pour capter un maximum d'informations du Web et les réseaux sociaux relatives à l'étude ESG.

Chapitre 4 : Diagnostic du processus de Due Diligence réputationnelle

Chapitre 4 : Diagnostic du processus de Due Diligence réputationnelle

Introduction :

De nos jours, dans un processus M&A, les investisseurs requièrent de plus en plus d'informations avant de s'engager dans des contrats dispendieux qui relèvent de décision stratégique. Il a été démontré que la Due Diligence financière ne suffit plus pour trancher sur des décisions de cette ampleur. A cet effet, ces investisseurs accordent plus d'importance à leur image de marque et aucun d'eux ne souhaite être associé à un scandale financier, environnemental ou social nuisant à sa réputation de sérieux, d'honnêteté et d'exemplarité. Contre le risque de réputation, le recours à l'intégration de critères ESG offre une bonne protection.

C'est ainsi que le besoin du recours à la due diligence réputationnelle est né. Cette dernière permet, grâce à l'analyse de la OR, de mettre en valeur un aspect immatériel qui peut nuire à l'entreprise ou lui être valorisant.

Dans ce qui suit, le processus OR va être formalisé en prenant en compte tous les intervenants durant ce dernier.

1. Description de l'approche OR pour KPMG:

L'équipe « R&S » du Deal Advisory Alger réalise l'étude qui résume les indicateurs clés de la réputation en ligne de l'entreprise et ses dirigeants sans investiguer en profondeur les détails des sujets évoqués.

Une étude supplémentaire, relative à l'intégrité de l'entreprise et son comité de direction, est faite par l'équipe Forensic de KPMG India basée sur des technologies qui analyse le contenu profond d'internet auquel l'équipe d'Alger n'a pas accès. Les deux études, étant indépendantes dans leur exécution, sont complémentaires afin de pouvoir traiter tous les aspects de la OR.

Notre travail portera sur la première partie de l'étude qui est réalisée par l'équipe Algérienne du fait que notre stage s'est déroulé au niveau du département Deal Advisory de KPMG Algérie.

1.1. OR pour l'équipe R&S Alger :

N'ayant jamais été formalisé auparavant, le processus de due diligence réputationnelle a été un long sujet de discussion avec les responsables du département Recherche & Stratégie. Afin de structurer ce dernier, nous avons travaillé sur la transposition d'une mission OR sur un projet qui a été explicité sur le fichier MS Project en Annexe 1 (Les missions étant identiques, le choix d'une seule mission s'avère suffisant pour la transposition). Ce dernier a fait l'objet d'une structuration préliminaire du processus. Il a été tenu compte, dans le cadre de cette structuration, des principes de base annoncés par l'approche processus notamment celle relative à la planification et l'identification de ces derniers. Il est à rappeler que l'approche

processus désigne une méthode de modélisation des activités d'une entreprise et ceci en décrivant son fonctionnement en comprenant les mécanismes et les interactions préalablement nécessaires à toute démarche qui génère un Output et ceci en passant par les étapes suivantes : Identification/classification des processus, formalisation/description de chaque processus, adaptation des processus à l'organisation et enfin le pilotage de ces processus.

Il est à mentionner que seule la première étape du processus a été balayée servant de base pour sa modélisation.

Ceci nous a permis de cerner ses activités ainsi que : les différentes tâches, les ressources utilisées et les résultats de ce dernier (Inputs et Outputs). Les résultats du travail sont présentés dans les parties à venir.

1.2.OR pour l'équipe Forensic India (Astrus):

L'équipe Forensic India est une équipe de KPMG International localisée en Inde qui collabore avec l'équipe « R&S » dans le cadre de la OR et qui travaille plus précisément sur la prévention, la détection et l'atténuation des risques de fraude, de non-conformité et de mauvaise conduite des entreprises et ses dirigeants afin d'enrichir le rapport d'OR. Pour ce faire, ils utilisent Astrus, une solution logicielle, qui permet d'extraire, transformer et visualiser l'information de n'importe quelle source dans n'importe quel format, y compris les ordinateurs portables, les téléphones mobiles et autres appareils électroniques afin de juger l'intégrité des parties prenantes de l'entreprise.

L'approche de revue de l'intégrité consiste à vérifier le contenu profond d'Internet en utilisant les technologies propres à KPMG (dont Astrus présenté en annexe).

La revue de l'intégrité couvre :

- Les fonds de l'entreprise et les membres du conseil d'administration ;
 - Les actionnaires/bénéficiaires propriétaires ;
 - Les articles de presse défavorables ;
 - Les litiges/sanctions ;
 - La liste des tiers politiquement exposés ;
 - Les entités à risque élevé.
- ❖ La différence entre des deux parties est que la première assure une vue d'ensemble sur la réputation en ligne d'une entreprise cible en résumant les différents indicateurs et la deuxième est beaucoup plus concentrée vers l'intégrité de l'entreprise et ses dirigeants mais permet également de creuser en profondeur les sujets clés résultant de la première partie à travers des analyses plus approfondies.

2. Définition du Processus OR Actuel :

Comme mentionnée plus avant et après plusieurs séances de travail et en se basant sur le premier fichier MS Project (Annexe 1) établi et sur les éléments clés apportés par l'approche

processus, nous avons identifié les activités du processus et nous les avons réparties sur 5 étapes majeures qui se résument dans le schéma suivant et les descriptions qui suivent.



Figure 9: Schéma du Processus OR

2.1. Scope of Work:

Durant la première étape, l'étendue du travail est envoyée par mail par nos collègues de KPMG France. C'est la première étape du processus qui est déclinée en deux points représentés dans la figure ci-dessous :



Figure 10: Schéma du sous processus « Scope of work »

2.1.1. Réception et compréhension du scope :

Dans cette étape, l'équipe R&S réceptionne le périmètre de l'étude et repère les différentes parties à traiter. Ce dernier comprend les éléments qui doivent être pris en compte pour constituer le rapport et se résument dans:

- Le type de client:

En accord avec les 3 approches citées précédemment, il s'agit de déterminer quel est le type d'entreprise pour laquelle l'étude sera faite : « Business to Consumer », « Business to Business » ou « Financial Organisations »

- La liste des concurrents à Benchmarker:

Une fois le type d'entreprise défini, il s'agit d'effectuer un Benchmark des concurrents, c'est-à-dire analyser l'activité de ses principaux concurrents de façon à montrer en quoi ils sont différents ou en quoi ils sont meilleurs en termes de présence en ligne. Ce benchmark concurrentiel est censé aller plus loin qu'une simple analyse de la concurrence il aide à comparer l'entreprise cible aux best practices sur le même segment de marché, et donc de connaître ses forces et ses faiblesses.

- Les noms des dirigeants de la cible :

Les dirigeants sont des acteurs clés de l'entreprise qu'ils pilotent au jour le jour. En effet, pour 81%⁹ des cadres, la visibilité et l'image du PDG impactent directement la réputation de l'entité qu'ils gouvernent. Il ne s'agit pas d'avoir juste de la visibilité, mais aussi de la crédibilité sur les canaux pertinents.

⁹ Source : KRC Research, 2015.

- Les implantations de l'entreprise :

Il est important de recenser tous les pays où l'entreprise est présente afin d'avoir une vision globale sur l'entreprise. La réputation d'une filiale peut influencer celle du groupe et nuire à celle-ci en cas de scandale. En sachant où l'entreprise est implantée, cela permet d'agrandir le champ de collecte des informations et de n'omettre aucun détail important. Mais cela permet aussi de limiter le bruit qui viendrait ajouter des données inutiles qui ne sont pas en rapport avec l'entreprise.

- L'étendue temporelle de l'étude :

Il s'agit d'une donnée essentielle pour la OR. Grâce à l'étude, l'entreprise pourra observer l'évolution de ses mentions au fil des mois et des années, et d'identifier les périodes de pics ainsi que ce qui est dit sur les différents réseaux sociaux, les médias mais aussi les forums et blogs durant cette période.

- Identification des marques de l'entreprise cible:

Ceci permettra de clarifier les recherches en utilisant les marques, filiales ou participations de l'entreprise comme mot clés lors de la collecte des informations pour assurer l'exhaustivité de cette dernière.

2.1.2. Organisation des tâches:

Une fois l'étendue reçue, une réunion d'équipe est faite pour lancer la mission en priorisant les étapes et les parties du rapport. Lors de cette dernière la répartition des tâches est discutée également afin d'organiser le travail d'une manière optimale.

2.2. Data retrieving and exploration:

Une fois tout le périmètre défini, il s'agit de l'étape de collecte des informations. Celle-ci se fait à l'aide d'outils de collecte. Ces outils, qui seront définis dans la partie « Extensive Research », ne font pas que surfer sur le web, ils permettent une analyse plus en profondeur et avec plus d'efficacité des mentions faites de la marque, ses dirigeants et ses projets (filiales aussi par moment) en prenant en compte ce qui se dit sur les réseaux sociaux, les médias en ligne ainsi que les blogs.

La collecte de données est faite à l'aide de 4 techniques, qui seront répétées dans les rubriques du rapport OR, représentées dans la figure ci-dessous :



Figure 11: Schéma représentatif des techniques de collecte et d'exploration des données.

Les 3 premières techniques sont des techniques de collecte et le sentiment analysis est plutôt une technique de traitement. Elles sont décrites comme suit :

➤ **Primary Research :**

Les premières recherches sur Google sont effectuées en intégrant des mots clé (le nom de l'entreprise, le nom d'un dirigeant, etc.) pour des recherches superficielles et pour avoir de premières informations sur la cible, ses dirigeants, ses concurrents ou autres.

➤ **Extensive Research :**

Une recherche manuelle approfondie sur le web est faite en utilisant 2 techniques essentielles qui interviendront dans les autres rubriques du rapport et représentés dans la figure qui suit:

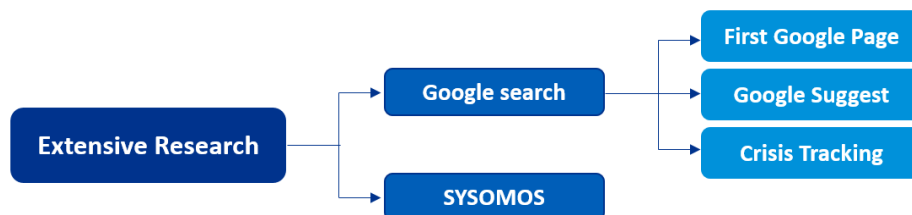


Figure 12: Schéma représentatif de la technique « Extensive research »

○ **Google Search :**

- La « First Google Page » est analysée manuellement pour identifier, par exemple, si l'entreprise a un bon référencement par mots clés et si les liens présents sur la première page sont de type « controlled » (l'entreprise gère à elle seule le contenu publié), « semi controlled » (l'entreprise peut agir sur le contenu) ou « uncontrolled » (l'entreprise n'a aucune main sur le contenu) et cela peut avoir un impact positif, négatif ou neutre surtout en prenant en compte les liens non contrôlés sur lesquels l'entreprise n'a pas d'ascendant.
- Depuis 2004, Google offre un service de suggestion automatique dans son moteur de recherche. Ce service suggère les mots les plus recherchés avec un mot clés de référence. Cela permettra d'identifier les recherches les plus fréquentes ainsi que les mots auxquels l'image de l'entreprise est liée. Les 10 mots les plus associés à celui de la cible sont extraits et analysés en identifiant les négatifs d'entre eux.
- Le Crisis Tracking est également fait. C'est une technique qui consiste à l'association du nom de la cible, ses dirigeants ou ses projets à 10 termes de crises (Justice, arnaque, sécurité, fraude, accusation, controverse, escroc, détournement et corruption) sur des recherches Google. Le résultat des dix (10) premières pages (10 Mots X 10 Pages X 10 Liens/Page = 1000 liens/entreprise) est analysé manuellement pour identifier d'éventuels éléments négatifs en procédant à la lecture des articles pour porter un jugement.

○ **SYSOMOS :** C'est une plateforme qui permet aux marques et aux organismes de transformer les informations fondées sur des données en opportunités d'engagement exploitables.

En automatisant de nombreux efforts manuels au sein d'un moteur d'analyse alimenté par l'intelligence artificielle, les spécialistes du marketing peuvent transformer des milliards de données détenues et acquises en perspectives et prévisions contextuelles qui permettent de prendre de meilleures décisions commerciales. Grâce à des alertes en temps réel, les

spécialistes du marketing peuvent également prendre connaissance des changements et des tendances liés à leur image de marque et agir immédiatement.

À l'aide SYMOS, l'équipe R&S analyse les mentions de l'entreprise, ses dirigeants ou ses projets sur le web durant les 12 derniers mois en intégrant le nom de ces derniers en mot clé et en ajoutant des mots relatifs au domaine d'activité pour cadrer la recherche. L'outil permet également de filtrer les mentions selon la langue, le pays, la source, etc.

➤ **Social media analysis :**

La présence dans les réseaux sociaux de l'entreprise est identifiée à l'aide de recherche Google sur les différentes plateformes Facebook, Tweeter, LinkedIn pour identifier si la cible et ses concurrents ont des pages actives. Une page active compte au minimum une publication par semaine.

Pour une analyse plus approfondie sur chaque réseau social, un split par mois est effectué sur le nombre de mention en identifiant les pics et en commentant leurs causes. Le nuage des mots les plus récurrents est extrait pour twitter ainsi que d'autres informations supplémentaires sur LinkedIn comme le nombre des followers, le nombre d'employés ou la durée moyenne de carrière.

➤ **Sentiment Analysis :**

Il s'agit du processus qui permet de déterminer la tonalité émotionnelle qui se cache derrière une série de mots. Les données analysées quantifient les sentiments ou les réactions du grand public à l'égard de certains produits, personnes ou idées et révèlent la nature de l'information dans son contexte respectif.

Cette analyse est utilisée pour mieux comprendre la perception, les opinions et les émotions exprimées dans une mention en ligne et cela en attribuant un adjectif « Neutre, positif ou négatif » en ce qui concerne le contenu de chaque mention pour pouvoir faire une analyse sur la réputation globale de la cible. La lecture manuelle de chaque mention est faite afin de juger la tonalité du contenu.

❖ Précédemment nous avons parlé des techniques utilisées et dans la partie qui suit ce sont les activités qui seront décrites.

Le processus de collecte de donnée alimente les 4 rubriques du rapport d'OR qui seront présentées dans la partie « Report Construction » considéré comme Output du processus. Cependant il est à signaler que les activités de collecte des données décrites dans ce qui suit sont structurées de la même manière que le rapport final, elles sont ventilées en quatre rubriques comme représentées dans la figure ci-dessous :



Figure 13: Schéma du sous processus « Data retrieving and exploration »

2.2.1. Company presence analysis :

L'analyse de la présence de l'entreprise cible sur le web concerne toutes les plateformes indépendamment des réseaux sociaux (Média en ligne, Wikipédia, YouTube, Forum, etc.) et ceci passe par l'utilisation des outils précédemment cités à savoir:

a. Extensive Research:

Une recherche approfondie sur le web est faite pour la cible et ses concurrents en utilisant Google search ainsi que Sysomos.

b. Social media analysis :

L'analyse de la présence de la cible sur les réseaux sociaux ainsi qu'un comparatif avec ses concurrents sont établis lors de cette étape.

c. Sentiment analysis :

Le sentiment analysis est fait après la lecture des mentions de l'entreprise ainsi que ceux de ses concurrents, extraites avec Sysomos, en attribuant l'adjectif « Positif, Négatif ou Neutre » à chaque mention manuellement.

2.2.2. Projects and brands analysis :

L'analyse de la présence des projets/marques de l'entreprise sur le web passe par l'utilisation des outils précédemment cités à savoir:

a. Primary research:

De premières recherches Google sont faite pour identifier les marques ainsi que les projets de l'entreprise qui ne sont pas déjà définies par le scope ou pour détecter de nouveaux éléments.

b. Extensive Research:

Une recherche approfondie sur le web est faite pour les projets ainsi que les marques de l'entreprise en utilisant Google search ainsi que Sysomos.

c. Social media analysis :

L'analyse de la présence des projets/marques de la cible sur les réseaux sociaux ainsi qu'un comparatif avec ses concurrents sont faits pour estimer leur visibilité et détecter des éléments à investiguer.

d. Sentiment analysis :

Le sentiment analysis est fait après la lecture des mentions des projets/marques de l'entreprise cible, extraites avec Sysomos, en attribuant l'adjectif « Positif, Négatif ou Neutre » à chaque mention manuellement.

2.2.3. Executive analysis :

La réputation des chefs d'entreprises, des différents actionnaires ainsi que des personnalités les plus influentes dans l'entreprise est analysée vu son impact direct sur la réputation de l'entité dirigée. Ceci passe par l'utilisation des outils précédemment cités à savoir:

a. Primary research:

De premières recherches sont faites pour identifier le PDG et les responsables de la cible et ses concurrents si ces derniers ne sont pas déjà définis par le scope qui est envoyé par e-mail.

b. Extensive Research:

Une recherche approfondie sur le web est faite sur le PDG et les responsables de la cible et ses concurrents en utilisant Google search ainsi que Sysomos dans certains cas.

2.2.4. HR Network :

La présence en ligne sur les différentes plateformes RH comme Glassdoor, Indeed ou Viadeo est analysée à travers les avis des employés et des notations attribuées. Ceci passe par l'utilisation des outils précédemment cités à savoir:

a. Primary research :

De premières recherches sont faites pour évaluer la présence de l'entreprise ainsi que ses concurrents sur les plateformes RH citées précédemment.

b. Sentiment analysis :

Le nombre d'avis ainsi que la note globale des avis est prise en compte afin de donner une appréciation globale sur ces derniers. Une lecture diagonale des premiers avis est faite pour déterminer s'il n'y a pas de sujet important à investiguer afin d'attribuer une tonalité positive.

2.3. Data cleaning :

Cette étape concerne les extractions faite à l'aide de l'outil SYSOMOS qui sont téléchargés sous format Excel et elle vient juste après la collecte des données comme mentionné sur la figure ci-dessous :

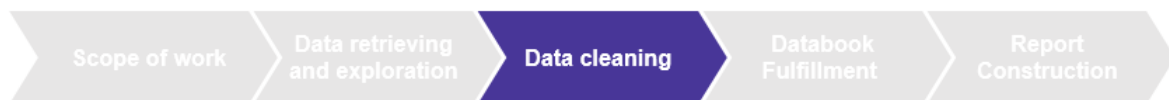


Figure 14: Schéma du sous processus « Data cleaning »

Après la collecte des données, le nettoyage de données est l'opération de détection et suppression du bruit (Mentions qui n'ont pas de rapport avec notre recherche) et qui engendre une analyse biaisée.

Ce nettoyage dépend du nombre de mentions qui ressortent après l'analyse sur SYSOMOS qui varie d'aucune à des millions quand le nom de la cible est un mot commun, ce qui engendre un temps de traitement manuel des mentions énorme et cela pour la cible, ses concurrents, ses projets, ses marques ainsi que ses dirigeants.

2.4. Databook Fulfillment :

Le remplissage du Databook est fait après le nettoyage des données comme représenté sur la figure suivante :

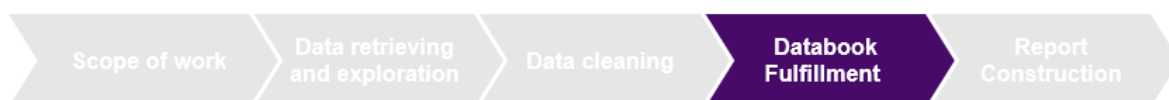


Figure 15: Schéma du sous processus « Data Fulfillment »

Le databook est un fichier Excel comportant toutes les données recueillies et traitées, qui sont mises, par la suite, sous forme de schémas ou de graphiques comme l'exemple représenté dans la figure suivante :

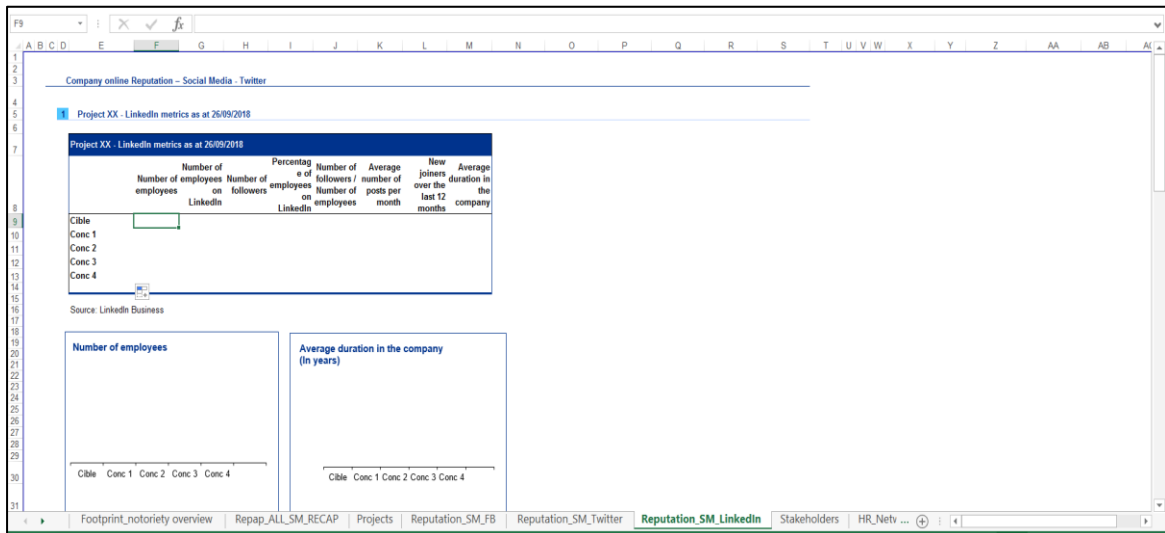


Figure 16: Exemple Databook

C'est un fichier qui est utilisé comme intermédiaire entre les plateformes de collectes de données (Google, Sysomos, etc.) et le rapport OR final délivré au client sous forme de fichier PowerPoint. Il sert à stocker les données dans des tableaux, les schématiser sous formes de graphes (Histogrammes, radars, camemberts, etc.) ainsi qu'à conserver les sources de provenances des données (Sites web, liens, etc.)

La conception du databook avait été effectuée au début pour répondre à un besoin client lors d'une première mission OR en intégrant les éléments nécessaires pour subvenir à ce dernier. Le databook intègre plusieurs feuilles relatives aux différentes parties du rapport schématisées dans la figure 19 plus bas. Un exemple de la première version du Databook est représenté dans la figure suivant avec les onglets des feuilles citées précédemment.

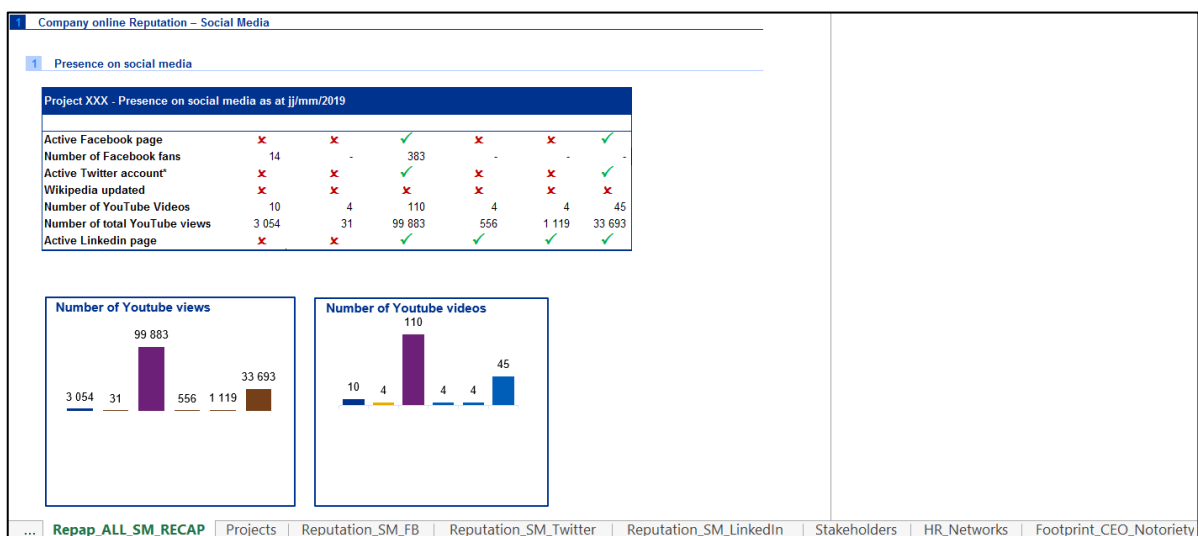


Figure 17: Exemple Première version Databook

Après la collecte de données, le renseignement du databook se fait manuellement pour chaque rubrique citée précédemment. Un template est prédéfini pour la construction des graphiques et des schémas en parallèle après le renseignement des données.

Cependant plusieurs ajustements ou changements de format sont effectués pour ses derniers afin de répondre aux besoins de l'analyse et aux exigences de la charte graphique de KPMG.

2.5. Report Construction :

Le rapport d'OR est construit sur la base de la méthode du risque exprimé par des feux tricolores (des extraits de rapport seront présentés lors de l'application) qui a été développée par les équipes de KPMG où un niveau de risque élevé est exprimé en rouge, moyen en orange et bas en vert, comme représenté sur la figure suivante :

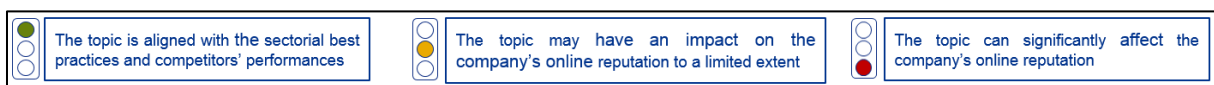


Figure 18: La méthode du risque exprimé par des feux tricolores.

Le Rapport d'OR (Output du processus) comporte 4 rubriques principales, représentées dans la figure 19, qui traitent l'ensemble des indicateurs reflétant la réputation en ligne d'une entreprise.

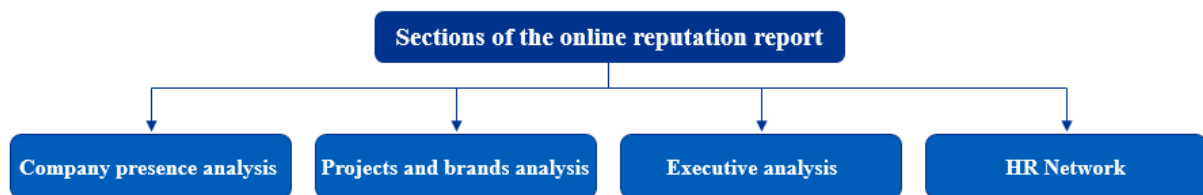


Figure 19: Les 4 rubriques d'un rapport d'OR

- **Company presence analysis** : Elle englobe la présence sur Internet de l'entreprise, sa notoriété, le benchmark avec ses concurrents ainsi que sa réputation vis-à-vis de ses parties prenantes (couverture presse, opinions des consommateurs, les avis sur les réseaux sociaux, etc.) ;
- **Projects and brands analysis** : La visibilité des projets et marques de l'entreprise est analysée dans cette partie (Nombre des mentions, analyses des mentions, détection de sujet à investiguer, notoriété des projets/marques, etc.)
- **Executive analysis** : elle concerne la présence sur Internet du PDG et des membres du conseil d'administration de l'entreprise (les points de vue sur Internet, forums, blogs, le nombre de mentions, les pics d'activité, etc.) ;
- **HR Network** : il résume l'ensemble des avis des employés/dirigeants (présence des employés sur Internet, leurs opinions, vue d'ensemble des RH, revue des forums spécialisés).

Le rapport OR contient les données qui ont été renseignées sur le databook en ajoutant des analyses et des commentaires sur le contenu en un format PowerPoint.

La construction du rapport passe par les trois étapes représentées sur la figure qui suit:

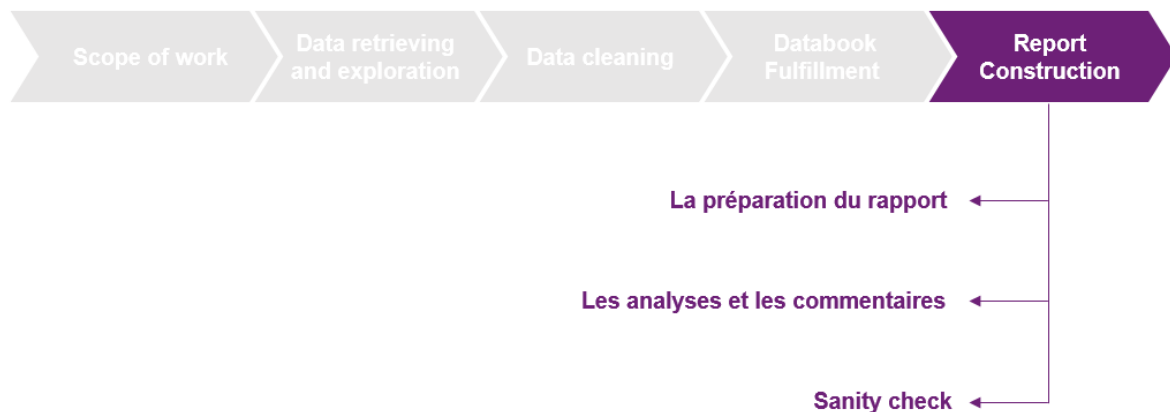


Figure 20: Schéma du sous processus « Report Construction »

- La préparation du rapport :

Sur l'aspect forme, les données relatives au projet (nom de code, date, etc.) sont mises sur le masque du PowerPoint, les sources sont citées et les périodes de collecte et de traitement de données sont précisées pour situer la période de l'analyse dans une période de temps bien déterminée.

- Les analyses et les commentaires :

L'exportation des données à partir du databook vers le rapport PPT est faite à l'aide d'Upslide.

Upslide est un outil facilitant la transition des tableaux et graphiques à partir d'un fichier Excel vers un fichier PowerPoint. Il permet également de normaliser les couleurs et les formats suivant la charte KPMG en un clic. Il contient aussi des bibliothèques riches en pictogrammes et formes permettant d'illustrer les idées d'une façon claire et précise.

Après cela, les analyses concernant les différentes courbes sont faites et les commentaires à propos des résultats du traitement des données sont émis après une discussion entre les membres qui ont participé à la mission.

Un résumé sur les différentes parties du rapport est mis en valeur dans les « Findings » pour donner une idée générale sur la réputation en ligne de l'entreprise par rapport à ses concurrents.

- Sanity check :

Le Sanity Check est la dernière étape du processus. Il représente une vérification qui est faite sur deux aspects :

a. L'aspect fond :

La révision du contenu est nécessaire pour vérifier si les chiffres ont été soigneusement repris, les textes correctement conçus et les analyses bien présentées.

b. L'aspect forme :

La charte graphique des couleurs et formes, les ajustements, les alignements ainsi que la mise en forme globale sont vérifiés lors de cette étape.

Après cette vérification, une revue générale est effectuée puis le rapport est envoyé à l'équipe

de KPMG France pour relecture avant de le présenter au client.

Le diagnostic précédemment explicité nous a permis de définir les inputs du processus (Demandes du client et les données collectées) pour aboutir aux outputs (Rapport OR Final et Databook) en passant par les étapes cités ci-dessus. Ce qui nous a permis d'illustrer le processus comme suit :

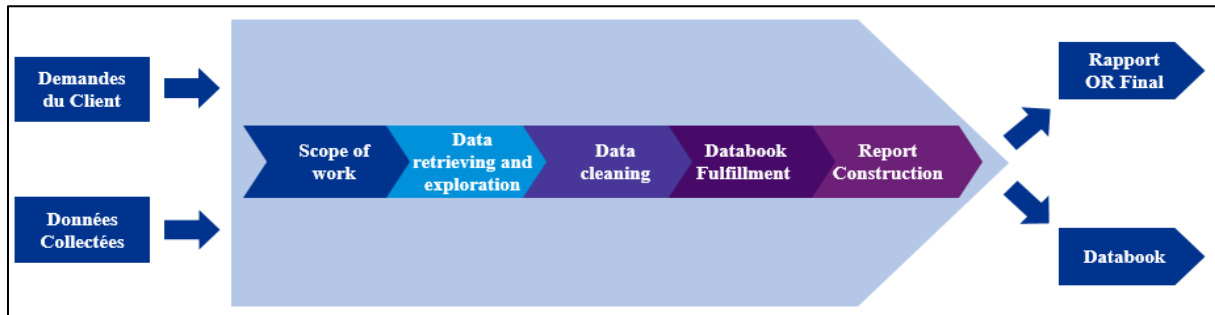


Figure 21: Processus de Due Diligence réputationnel – (Input/Output)

3. Diagnostic et constats :

Suite à l'identification des tâches critiques, des ressources utilisées ainsi que les résultats du processus OR et tenant compte de la modélisation de la mission effectuée sur MS Project, nous avons détecté un certain nombre de constats que nous structurons autour des points suivants :

3.1. L'automatisation des tâches:

Les tâches sont lourdes, répétitives et la quasi-totalité d'entre elles sont manuelles, notamment celles relatives aux sous processus Extensive Research et Sentiment Analysis.

Nous avons également constaté que les tâches relatives à ces deux sous processus peuvent être optimisées contrairement aux tâches afférentes aux autres processus où l'intervention de l'humain est nécessaire pour le jugement et la validation. D'où notre intérêt à se focaliser sur l'optimisation des tâches les constituant. Un diagnostic plus approfondi de ces derniers nous a permis de porter les constats suivants :

3.1.1. Extensive Research :

En plus d'être lourdes et répétitives, la prise en charge des tâches des éléments figurant dans la « Extensive Research » représentés dans la figure 12 prend énormément de temps (environ 28% de la durée d'une mission). Aucune réflexion quant à l'optimisation de ces dernières n'a été développée. Les automatiser nous semble une solution qui va permettre un gain de temps considérable et offre beaucoup plus d'efficacité en termes d'exécution.

3.1.2. Le Sentiment Analysis :

Etant fait manuellement pour un nombre de donnée énorme variant entre des dizaines et des millions de données, l'attribution des tonalités (positives, négatives ou neutres) est jugée comme la tâche la plus pénible du processus en terme d'effort et en temps consommé (environ 36% de la durée d'une mission). A cet effet, il est nécessaire de réfléchir à une

méthode fiable d'attribution automatique des tonalités pour pouvoir se concentrer uniquement sur les tonalités négatives et investiguer leurs raisons.

3.2. Le databook :

Le databook a été fait une première fois pour répondre à un besoin spécifique d'un client lors de la toute première mission OR et depuis, il a été adopté pour toutes les autres missions sans forcément remettre en question les éléments qui le constituent, les types d'analyses et de graphiques ni les rubriques qu'il faudra intégrer ou enlever spécifiquement pour chaque cas. Il est donc utile de revoir sa structure ainsi que ses éléments pour juger la pertinence des analyses qui sont faites et leurs interprétations.

3.3. L'organisation :

Après avoir analysé les ressources sur le fichier MS Project, il a été remarqué que la répartition des tâches au sein de l'équipe « Recherche & Stratégie » n'est pas optimale où nous remarquons un déséquilibre lors de l'utilisation des ressources comme représenté sur la figure suivante :

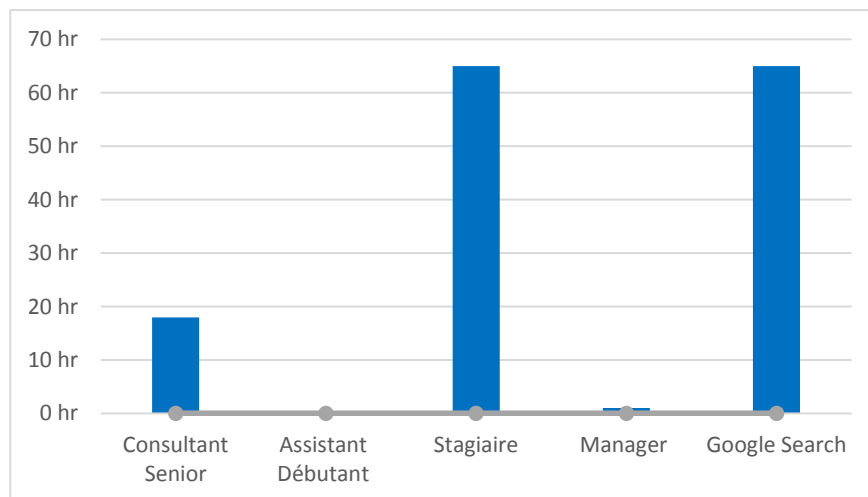


Figure 22: Histogramme d'utilisation des ressources en heures.

Le chevauchement entre les différentes missions pose également problème pour la planification des ressources et leur affectation.

Une meilleure répartition des tâches et synchronisation entre les membres de l'équipe peut contribuer à délivrer un rapport OR de qualité en un temps plus optimal.

4. Genèse de la problématique:

À l'issue de ce diagnostic, l'état des lieux a été clôturé par un grand débat à travers l'organisation de divers entretiens avec les responsables autour des dysfonctionnements remarqués. Il a été convenu de catégoriser ces derniers en trois dimensions et ce pour mieux identifier les apports des améliorations attendues issues de leurs prises en charge, ces dimensions sont représentées dans la figure qui suit:



Figure 23: Les 3 dimensions des dysfonctionnements

4.1.La dimension Qualitative:

L'internaute joue un rôle important dans le partage de contenu, il est libre et n'est pas limité. Ceci engendre une quantité énorme de contenu sur le web. De ce fait le traitement de ces données (contenu), étant manuel, ne peut être exhaustif vu la quantité énorme de données. Toutes les tâches du processus, étant manuelles, il est très difficile de faire une analyse en prenant en compte toutes les informations.

Il a été remarqué que le « sentiment analysis » est une tâche qui se répète dans toutes les rubriques du rapport mis à part dans « executive analysis ». Etant faite manuellement, il est difficile et pénible de lire avec une concentration totale le nombre considérable (qui peut être en nombre de millions) de contenu pour lui attribuer une tonalité positive ou négative. De ce fait l'analyse des sentiments est biaisée par des lectures transversales et affecte donc la fiabilité et la qualité du résultat attendu. Par exemple, l'avis individuel des employés sur les plateformes RH n'est que rarement pris en compte quand le nombre d'avis n'excède pas la cinquantaine mais lorsqu'il y a des milliers d'avis une lecture transversale est faite des premiers avis et la note globale de ces derniers est prise en compte pour juger leur tonalité. D'autre part, la subjectivité dans l'attribution des tonalités est un sujet à étudier car chacun à sa propre vision et interprétation des choses donc cela peut également affecter la fiabilité des résultats et leur pertinence.

L'aspect qualitatif est affecté également par la qualité du databook où le contenu est aussi important que le contenant. La qualité de la présentation des données, leur structure ainsi que l'exhaustivité des analyses apportent une grande valeur ajoutée à l'analyse OR. Il est donc nécessaire d'améliorer le databook sur l'aspect fond (contenu) et forme (graphiques, représentations, structure des données, etc.)

4.2.La dimension temporelle :

Le temps est un phénomène inflexible, intangible, non négociable. Il s'écoule inlassablement et personne n'a le pouvoir de l'arrêter ou de revenir en arrière. Dans le domaine des sociétés de service, tel que le consulting, il représente une ressource essentielle et précieuse qu'il va falloir optimiser.

Il est, donc, nécessaire d'évoquer l'aspect chronophage des activités qui engendre un temps de traitement très important.

Pour illustrer cela, un constat est fait quand à la répétition du processus « Extensive research » dans les différentes rubriques du rapport sauf pour « HR Network ». Ce processus ou, plus exactement, la partie « Google search » du processus regroupe des tâches opérationnelles répétitives chronophages telles que la « First Google Page » qui représente une reproduction manuelle de des éléments de la 1^{ère} page de Google (Titres, liens, etc.) sur un fichier Excel, le « Google Suggest » qui est de même fait manuellement en reprenant les mots les plus recherchés ainsi que le « Crisis Tracking » qui nécessite la consultation manuelle de 1000 liens pour chaque cible donnée.

4.3.La dimension organisationnelle :

Pour assurer une analyse complète des dysfonctionnements, il est indispensable de traiter l'aspect organisationnel du travail au sein de l'équipe « R&S ». Pour ce faire, il est fondamental de dire que le processus n'a jamais été formalisé pour cadrer les différentes tâches et sous processus et leur interdépendance. La planification du projet chronologiquement en prenant en compte une répartition rigoureuse des tâches n'est pas faite systématiquement pour chaque membre de l'équipe. Enfin, le chevauchement entre les projets (missions) rend la gestion des deadlines compliquée et peut engendrer des retards de traitement ou une surexploitation des ressources qui peut être nuisible à long terme.

5. Formalisation de la problématique :

Dans le cadre des fusions-acquisitions, les missions de due diligence sont d'une importance majeure dans la prise de décision d'un investisseur. Chez KPMG, le département Deal Advisory s'occupe de ce type de mission en ayant une expertise métier assez développée dans le monde des grandes transactions.

Au sein du Deal Advisory, l'équipe « Recherche & Stratégie » de KPMG Algérie est déterminée à évoluer vers de nouveaux horizons pour s'aligner avec les meilleures pratiques dans le monde. C'est dans cette perspective que notre intervention a eu lieu dans ce service.

Le diagnostic nous a permis de définir plusieurs chantiers sur lesquels nous pouvons agir afin d'améliorer le processus.

Dans un premier lieu, nous avons évoqué des dysfonctionnements temporels où les tâches sont macrophages qui nous poussent à réfléchir à : Comment peut-on passer moins de temps dans la réalisation d'un rapport et minimiser notre temps d'exécution tout en répondant aux exigences du client ?

En second lieu, nous avons parlé des imperfections de qualité relatives à l'exhaustivité des données traitées qui nous ont menées à poser la question : Comment peut-on traiter un nombre important de donnée pour assurer leur exhaustivité dans le but d'améliorer le livrable final ?

Enfin, les manquements organisationnels ont été évoqués pour aboutir au questionnement suivant : Comment peut-on améliorer l'organisation de l'équipe afin d'être plus productifs et efficaces dans la réalisation des missions ?

Ces différents questionnements convergent vers une réflexion plus approfondie sur la façon avec laquelle nous pouvons améliorer le processus de due diligence en général, ce qui nous mène à poser la question principale suivante :

« Comment peut-on optimiser le processus de due diligence réputationnelle ? »

Plaçant ainsi l'optimisation du processus comme étant la problématique majeure à laquelle s'intéresse notre travail et ceci afin de gagner en temps, en qualité tout en générant un gain en coût représentatif.

Conclusion :

Le besoin au sein du cabinet pour l'amélioration des différents processus a été exprimé et ceci en accord avec la stratégie de développement des différentes business units de KPMG. De ce fait, un diagnostic a été fait pour le processus de due diligence réputationnelle afin de faire un état des lieux, détecter les pistes d'amélioration et d'agir dans l'optique d'optimiser ce processus pour un gain en temps, en qualité et donc en argent.

La partie qui va suivre va expliciter en détail les actions concrètes qui ont été mises en place dans le but de répondre au besoin du cabinet.

Partie 3 : Apports et solutions proposées

Chapitre 5 : Solutions, validations et apports.

Partie 3: Apports et Solutions Proposées

Chapitre 5 : Solutions, validations et apports

Introduction :

Dans une optique d'amélioration de la performance de ses missions, KPMG engage une démarche d'analyse des processus afin de les optimiser. Le diagnostic a été fait pour le processus de OR afin de l'analyser et de penser aux différentes améliorations qui pourraient apporter de la valeur ajoutée. Les constats ont été structurés en 3 dimensions. La première étant temporelle où nous avons évoqué l'aspect macrophage des tâches faites, le second étant qualitatif où l'exhaustivité des informations impactait la qualité des analyses faites et le dernier étant l'aspect organisationnel où la répartition des tâches et le chevauchement entre les missions ne facilitaient pas l'affectation des ressources.

Dans ce chapitre, nous allons aborder les solutions concrètes qui ont été mises en place afin de répondre au besoin d'optimisation et d'amélioration de la performance. Pour cela, les solutions vont être explicitées en détails, la validation des modèles va être assurée et l'apport synthétisé à la fin pour valoriser l'impact des actions implémentées.

Pour une meilleure visibilité, les solutions seront présentées en respectant le listing des constats présentés dans la partie diagnostic, les apports quant à eux seront présentés en respectant la catégorisation proposée plus haut.

1. Solutions :

En accord avec la synthèse du diagnostic citée dans le troisième point du chapitre précédent, les solutions ont été apportées sur les 3 éléments majeurs cités :

1.1. Automatisation des tâches :

1.1.1. Automatisation du sous processus « Extensive Search » :

Suite à la transposition du processus sur MS Project (Annexe 1) il a été remarqué que le sous processus « Extensive Search », schématisé dans la Figure 12, se répète dans les 3 rubriques Company Presence Analysis, Project and Brands Analysis et dans Executive Analysis. Ce dernier, dans sa partie Google, incluant la First Google Page, le Google Suggest ainsi que le Crisis Tracking, constitue une tâche macrophage qui peut être optimisée avec de la programmation afin d'extraire les informations à partir des pages Web et de les inscrire sur le Databook.

Pour cela, nous avons décidé de concevoir un programme informatique qui permet d'extraire des données clés en scrapant (Le scraping est une technique informatique permettant d'extraire des textes ou des informations à partir d'un site web existant) le Web et qui intègre ces données dans le databook directement.

Pour ce faire, il existe plusieurs langages de programmation qui ont été cités dans la partie état de l'art. Chacun avec ses avantages et ses inconvénients. Il faut bien en choisir un. Nous avons opté pour le langage Python (Version Python 3.5) pour les plusieurs raisons. Parmi ces dernières, selon le cabinet myTectra, un spécialiste des solutions d'apprentissage, Python est

le langage favori d'un très grand nombre de développeurs comparé à d'autres langages comme Java, PHP, C++, etc.¹⁰. Pour un ingénieur qui n'a pas eu des notions poussées en programmation, la syntaxe de Python est très simple et, combinée à des types de données évoluées (listes, dictionnaires...), conduit à des programmes à la fois très compacts et très lisibles. Les développeurs attestent, aujourd'hui, que Python est leur langage de programmation privilégié dans presque tous les domaines de l'informatique, y compris le développement Web, le cloud computing (AWS, OpenStack, VMware, Google Cloud, etc.), l'automatisation, les tests de logiciels, Big Data, Hadoop, etc. D'autre part, les bibliothèques standard de Python et les paquetages contribuéés donnent accès à une grande variété de services : chaînes de caractères et expressions régulières, protocoles Internet (Web, News, FTP, CGI, HTML...), interfaces graphiques, etc.

Dans cette phase d'automatisation, python a été utilisé afin d'exploiter des données à partir d'une source HTML (HyperText Markup Language : langage de balisage conçu pour représenter les pages web) et de les extraire sur le databook (format Excel) pour les exploiter dans l'analyses de la OR. Afin de repérer les informations dans un script HTML, les balises spécifiques pour chaque information sont identifiées. Ces dernières sont caractérisées dans le code par des chevrons (<, >) encadrant les noms des balises. Un exemple de la balise qui permet d'identifier une date dans un code HTML est représenté dans la figure suivante :

```
id="am-b0" aria-label="Options relatives au résultat" aria-exp
"2ahUKEwjny_WJj6_iAhVLbBoKHSaIDUQQ7B0wAHoECAMQAg"><span class=
"keydown:m.hdke;mouseover:m.hdhne;mouseout:m.hdhue" data-ved="
uitem"><a class="fl" href="
ebcache.googleusercontent.com/search?q=cache:ufBRSrVTRHMJ:http
-le-banc-des+&cd=1&hl=fr&ct=clnk&gl=dz" ping="
ebcache.googleusercontent.com/search%3Fq%3Dcache:ufBRSrVTRHMJ:
-sur-le-banc-des%2B%26cd%3D1%26hl%3Dfr%26ct%3Dclnk%26gl%3Ddz&a
><div><span class="st"><span class="f">9 janv. 2018 - </span>I
devant la <em>justice</em>. Le 9 janvier&nbsp;...</span></div>
"2ahUKEwjny_WJj6_iAhVLbBoKHSaIDUQQFSgAMAF6BAGAEAA"><div class=
u.pvc.com/cx/en/industries/government-public-services/public
```

Figure 24: Exemple de balise d'identification de la date

Dans l'exemple cité la balise qui nous a permis de reconnaître une date dans un code de milliers de ligne est : ``

Pour une meilleure efficacité et une rapidité d'exécution, nous avons opté pour cette technique pour le sous processus « Extensive Research » qui était répétitif et qui nécessitait pas l'intelligence humaine. L'intervention concerne les 3 points du sous processus :

1.1.1.1. First Google Page :

La First Google Page, comme son nom l'indique, est une synthèse de la première page Google d'une entreprise. Après l'introduction du nom de la société comme mot clé dans la barre de recherche Google, les titres de la première page Google sont copiés manuellement sur un tableau Excel (figurant dans le databook) en faisant également l'extraction manuelle de la source (lien) et la date d'apparition si elle est mentionnée (généralement mentionnée quand il s'agit d'un article de presse). Le niveau de control est jugé à partir de la source qui a publié l'article (Controlled, Semi-Controlled ou Uncontrolled) pour juger si l'entreprise à un bon

¹⁰ <https://www.developpez.com/actu/133538/Programmation-decouvrez-les-sept-raisons-pour-lesquelles-vous-devez-apprendre-le-langage-Python-selon-myTetra/>

référencement par mot clé (Si elle maîtrise sa première page Google) et la Tonalité de chaque article est également attribuée en lisant le contenu des titres représentés. Un exemple de ces informations est illustré dans la figure qui suit (Les dates n'apparaissent pas dans ce cas car il n'y a pas d'article de presse qui apparait dans la première page) :

Project Consulting- First Google page for KPMG				
Order	Headers	Source	Level of control	Tone
1	KPMG — Wikipédia	fr.wikipedia.org	semi-controlled	Neutral
2	Algeria - KPMG Global	home.kpmg.com	controlled	Neutral
3	kpmg algerie spa	www.kpmg.dz	controlled	Neutral
4	KPMG Algérie Spa ALGER KPMG est le réseau ... - Formation DZ	www.formation-dz.com	uncontrolled	Neutral
5	Kpmg Algérie,spa, Immeuble Kpmg Lot 94 Zone... - Kompass	dz.kompass.com	semi-controlled	Neutral
6	KPMG France - KPMG France - KPMG International	home.kpmg	controlled	Neutral
7	KPMG ALGERIE Conseils et études financières Alger	www.lespagesmaghreb.com	semi-controlled	Neutral
8	Guide Investir en Algérie - FCE	www.fce.dz	uncontrolled	Positive
9	KPMG - Comptabilité	www.wiki-compta.com	uncontrolled	Neutral
10	KPMG Algérie LinkedIn	www.linkedin.com	uncontrolled	Neutral

Figure 25: First Google Page de KPMG

Au lieu de faire toutes ses tâches manuellement et répétitivement pour les 3 rubriques du rapport, un programme informatique (Annexe 4) a été développé afin de d'automatiser l'extraction des données à partir des pages Web et l'écriture des données sur le databook.

Le titre, la date et la source sont extraits directement grâce aux balises de recherche respectives comme expliqué précédemment (Figure 24). Le niveau de control est jugé automatiquement en vérifiant la liste des sites Controlled (Ceux qui contiennent le nom de l'entreprise) et la liste des sites Semi-controlled que nous avons recensé (Les annuaires, Wikipédia, etc.). Pour la tonalité, elle dépend du contenu des titres donc reste manuelle vu le temps infinitésimal qu'elle prend mais fait l'objet de perspective d'amélioration après le développement d'un modèle plus rigoureux de traitement de texte qui pourra affecter la tonalité automatiquement. Le résultat est le même mais l'exécution prend moins de temps et s'avère plus simple pour les exécutant.

1.1.1.2. Google Suggest :

Il est de même pour le Google Suggest, qui résume les mots les plus associés au nom de la société lors des recherches Google. Cette dernière se faisait manuellement en copiant puis collant les termes suggérés sur ubersuggest (site : neilpatel.com/fr/ubersuggest/) après avoir inscrit le nom de l'entreprise cible en intégrant le nom de la cible en mot clé. La figure suivante décrit le résultat de cette recherche :

Project Consulting- First Google page for KPMG as keyword - Top 10	
kpmg	kpmg algerie pdf
kpmg algerie	kpmg dz
kpmg recrutement	kpmg wikipedia
kpmg algerie 2019	kpmg assurance algerie 2017
kpmg oran	kpmg logo

Figure 26: Exemple Google Suggest pour KPMG

Pour automatiser cette tâche répétitive, le programme informatique (Annexe 4 – suite du programme de la First Google Page) a été conçu afin de minimiser l'intervention de l'humaine sur des actions aussi élémentaires. Le programme va permettre, donc, l'extraction des données de la page web vers le databook directement en un clic comme c'est le cas de la First Google Page. Le résultat est le même mais l'exécution prend moins de temps et s'avère plus simple pour les exécutant.

1.1.1.3. Crisis Tracking :

Le Crisis Tracking est la tâche la plus pénible et celle qui prend énormément de temps dans le sous processus « Extensive Research ». En effet, la consultation de mille (1000) liens par entreprise pour identifier des éléments négatifs prend en moyenne entre 2 et 3 heures où il faut intégrer le nom de l'entreprise cible en mot clés dans la recherche Google en l'associant à un des 10 termes de crises et défiler manuellement les 10 premières page Google pour chaque termes de crises et en vérifiant le contenu des 10 liens présents sur chaque page.

Pour éviter de faire tout ce travail manuellement, un programme informatique (Annexe 5) a été développé afin d'extraire les 1000 liens avec leur contenu dans une même feuille Excel en un clic et de permettre une lecture plus rapide des contenus pour identifier les éléments à investiguer dans le cadre de la OR. Ce programme va permettre de scraper les pages Web, d'extraire les données nécessaires et de les classer dans un fichier Excel pour une meilleure fluidité de lecture et une rapidité d'exécution. Le résultat est le même mais l'exécution prend moins de temps et s'avère plus simple pour les exécutant.

Il serait également intéressant de pouvoir identifier les éléments négatifs (Eléments de tonalité négative) automatiquement pour minimiser le temps de lecture et, donc, d'optimiser cette tâche. Mais cette fonctionnalité nécessite le développement d'un modèle robuste utilisant de l'intelligence artificielle (Un modèle va être développé dans le point suivant) afin de pouvoir assurer la fiabilité de cette automatisation.

1.1.2. Automatisation du sous processus « Sentiment Analysis » :

Le Sentiment Analysis (SA) comporte des tâches chronophages (qui consomment beaucoup de temps) et qui sont faites manuellement pour juger la tonalité d'une mention. En d'autres termes, le SA permet de juger si une mention (Un texte, une phrase, un mot, un commentaire, un avis, etc.) est positive ou négative. Dans notre cas, il est à préciser que les mentions à caractère neutre (qui ne sont ni positives ni négatives) sont considérées comme positives car le but de l'étude de la OR est d'identifier les éléments négatifs (Fraude, scandale, etc.) et de les investiguer pour juger la réputation de l'entreprise.

Le SA est fait d'une manière répétitive lors des 3 rubriques du rapport : Company Presence Analysis, Projects and Brands Analysis, HR Network (représentés dans la figure 19). Le jugement d'une tonalité nécessite la lecture de la mention et de la réflexion ce qui engendre un temps de traitement manuel énorme, variant de quelques minutes à plusieurs heures, qui dépend du nombre de mentions (si l'entreprise à un Nom qui ressemble à un nom commun, des millions de mentions sont en résultat et le traitement est quasiment impossible manuellement). C'est ainsi que le besoin de recourir à l'intelligence artificielle est né pour automatiser ce traitement.

Un tel traitement nécessite un modèle d'apprentissage machine (Machine Learning) assez développé pour pouvoir lire un nombre de mention puis leur attribuer une tonalité ce qui n'est pas simple à réaliser. De ce fait nous avons pu bénéficier d'un premier programme qui a été fait par l'équipe D&A (Data & Analytics) de KPMG Paris pour développer plus rapidement le modèle de Machine Learning qui répond à nos besoins.

Dans un premier lieu, nous nous sommes intéressés au SA relatif à la partie HR Network. Durant cette dernière, les avis des employés concernant leur employeur sont traités par le modèle Machine Learning pour pouvoir attribuer une tonalité positive ou négative afin de juger la réputation de l'entreprise chez ses employés ou ses anciens employés/dirigeants.

Le modèle a été développé sous Python grâce à la librairie « Scikit-learn » qui représente une des librairies informatiques les plus riches dans le monde du Machine Learning.

1.1.2.1. Description du modèle ML :

Le modèle d'apprentissage machine supervisé (supervised ML) est un modèle à qui une base d'apprentissage (Training Set) est fournie afin de lui permettre de développer une certaine logique mathématique pour pouvoir, par la suite, interpréter d'autres résultats.

Les lignes de codes du programme développé (Annexe 2) sont expliquées dans le tableau en annexe 3 afin de comprendre le lien entre la description ci-dessous et le développement.

En accord avec le processus d'apprentissage expliqué dans l'état de l'art (Figure 2), le modèle d'apprentissage machine (Annexe 2) pour la partie HR Network a été développé selon les étapes représentées dans la figure ci-dessous et explicitées par la suite.

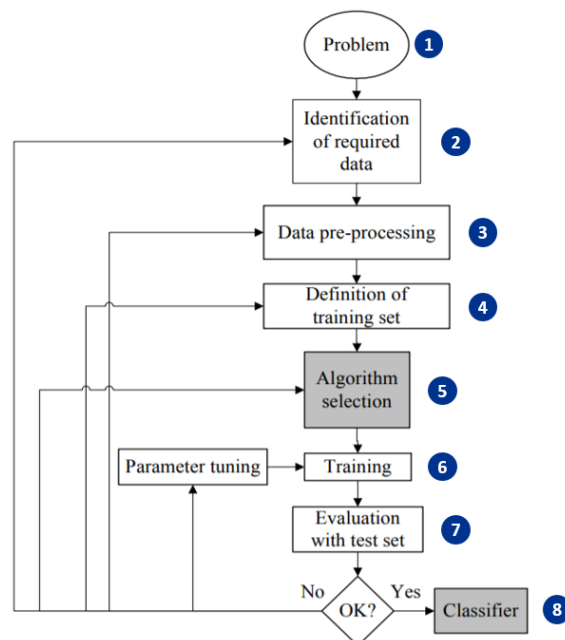


Figure 27: Description Modèle ML

① Notre problème réside dans le fait de savoir si un avis d'un employé est positif ou négatif afin de pouvoir généraliser le traitement sur un nombre considérable de d'avis pour en tirer des conclusions générales sur la réputation de l'entreprise auprès des réseaux et plateformes

dont bénéficie les ressources humaines (RH) pour exprimer leur opinion vis-à-vis de leur travail.

② Les données d'entrée du modèle d'apprentissage dont nous avons besoin sont des avis ou commentaires des utilisateurs des plateformes RH, il est donc nécessaire d'avoir une base d'apprentissage diversifiée et exhaustive. Pour cela, un programme informatique a été développé (Annexe 2) pour extraire la base d'apprentissage en scrapant les pages Web de ses plateformes afin d'extraire des avis dans un fichier Excel. Pour les besoins de l'apprentissage, nous avons opté pour 70 entreprises (5 entreprises de chacun des 14 secteurs différents) de chaîne de valeur distinctes allant des industries aux hôtels en passant par différentes entreprises de services pour pouvoir entraîner le modèle sur divers avis et lui permettre d'apprendre un maximum.

③ Le prétraitement des données se fait par le nettoyage de bruit (données qui peuvent biaiser notre modèle). Dans notre cas, nous avons retenu, dans un premier lieu, les commentaires en anglais seulement, pour ne pas chevaucher les langues ce qui rend l'apprentissage plus compliqué. Le nombre de commentaires retenu après le nettoyage excédait les six mille (6000) avis et ceci pour assurer un apprentissage complet et donc pour cibler une grande précision lors des prédictions du modèle.

Dans cette étape il est nécessaire d'éliminer également les caractères usuels qui se répètent beaucoup comme les déterminants, les « S » des pluriels, les caractères spéciaux (/, @, \, #, etc.) et ceci grâce au Tokenizer qui nous permet d'éliminer les termes indésirables. Ce dernier est défini dans le programme comme suit :

def my_tokenizer(s, stoplist=stop_words):

```
special = ['\.', '\*', '<br />', '\", '^-', '#', '@', ':', 'RT']  
s = s.lower()  
tokens = nltk.tokenize.word_tokenize(s)  
tokens = [wnl.lemmatize(t, pos='v') for t in tokens]  
tokens = [t for t in tokens if not t in stoplist]  
tokens = [t for t in tokens if not all(i.isdigit() for i in t)]  
tokens = [t for t in tokens if not re.match("^\d", t)]  
tokens = [remove_plural(t) for t in tokens]  
tokens = [remove_special(t) for t in tokens]  
tokens = [t for t in tokens if len(t)>3]  
return tokens
```

④ Après le nettoyage des données, la base d'apprentissage (Training Set) doit être définie. Comme expliqué dans l'état de l'art, un modèle d'apprentissage supervisé nécessite des données en Input et leur résultat en Output pour pouvoir les prendre comme exemple et apprendre en développant une logique. Dans notre cas, il est donc impératif d'alimenter notre modèle avec les 6000 avis en plus de leurs tonalités respectives qui ont été faites manuellement.

5 Le choix de l’algorithme de traitement dépend du besoin qui est exprimé. Dans notre cas, l’objectif est de traiter du texte en utilisant le Natural Language Processing (NLP), expliqué dans l’état de l’art, pour aboutir à des prédictions grâce à un modèle de prévision mathématique qui compte le nombre d’occurrence des mots en Input avec leur tonalité en Output pour définir une matrice de fréquence qui sera utilisée pour la prévision.

Le traitement de texte, dans notre modèle, est fait grâce à la technique du Bag-of-Word (BOW) N-gram model. Le (BOW) N-gram model est l’un des modèles de langage les plus simples utilisés en NLP. Dans notre cas, nous avons essayé un modèle Unigramme (N=1) du texte en gardant la trace du nombre d’occurrences de chaque mot. Dans ce modèle, les mots sont pris en compte individuellement en donnant à chaque mot une tonalité spécifique. Si la note totale est négative, le texte sera classé comme négatif et si elle est positive, le texte sera classé comme positif. Le modèle Unigramme est simple à faire, mais moins précis parce qu’il ne tient pas compte de l’ordre des mots ou de la grammaire. Donc il faudrait penser à augmenter le N (N=2 ou N=3) si la précision du modèle Unigramme ne donne pas des résultats satisfaisant. La définition de ce dernier dans le modèle est comme suit :

```
def init_tfidf(X_train, ngram_range=(1, 1), min_df=5, max_df=.9):
```

```
    tfidf = TfidfVectorizer(min_df=min_df, max_df=max_df, ngram_range=ngram_range,
    stop_words=stop_words, tokenizer=my_tokenizer)

    tfidf.fit(X_train)

    bag_of_words = tfidf.get_feature_names()

    return tfidf, bag_of_words
```

Le modèle de régression choisis pour les prédictions est le modèle de régression logistique car c’est le plus utilisé dans les techniques de classifications de texte et peut être facilement généralisée à plusieurs classes. La régression logistique estime $P(y|x)$ en extrayant certains ensembles de caractéristiques de l’Input, en les combinant de façon linéaire (en multipliant chaque caractéristique par un poids et en les additionnant). La définition de ce dernier dans le modèle est comme suit :

```
from sklearn.linear_model import LogisticRegression
```

```
    clf = LogisticRegression(penalty= 'l2', C=1.0)
```

où *penalty*= 'l2' et *C*=1.0 sont choisis comme paramètre par défaut du modèle.¹¹

6 Une fois l’algorithme sélectionné, la base d’apprentissage est introduite dans le modèle pour lui permettre de s’auto-former et de développer une logique grâce au modèle de régression logistique. Il est à noter qu’une partie de la base d’apprentissage est faite pour l’apprentissage et la deuxième partie est faite pour les tests (étape suivante). Dans notre cas les deux tiers (2/3) de la base ont été utilisés pour l’apprentissage et le dernier tiers pour les tests.

7 Comme mentionné, ci-dessus, la deuxième partie de la base d’apprentissage (33% - dernier tiers de la taille de la base) utilisé pour que le modèle se teste par lui-même et évalue

¹¹ Source : https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

sa propre efficacité afin de revenir vers une des étapes précédentes pour s'améliorer ou réajuster ses paramètres. Le choix de diviser la base en 2/3 pour l'apprentissage et 1/3 pour les tests a été fait pour mieux tester notre modèle car le pourcentage de test varie généralement de 1/5 à 1/3 de la taille de la base. Ce dernier est défini dans le modèle comme suit :

```
y_test = train_test_split(corpus, df.score, stratify = df.score, test_size = .33)
```

8 Le résultat final est la classification des avis en avis positif ou négatif en affectant chaque avis à la classe à laquelle il appartient selon les l'apprentissage cité ci-dessus.

1.1.2.2. Validation du modèle :

Afin de tester la robustesse du modèle et de le valider, le degré de précision du modèle Unigramme a été comparé au modèle utilisé par Amazon dans la classification de texte. La particularité des modèles développés sur AWS (Amazon Web Services) est la puissance de leurs modèles en termes de rapidité et précisions de traitement car ils ont été entraînés sur des bases d'apprentissages très exhaustives.

Le service « Amazon Comprehend » de AWS est un service en ligne qui peut découvrir le sens et les relations dans un texte à partir d'incidents de support client, de critiques de produits, de flux de médias sociaux, d'articles de presse, de documents et autres sources. Par exemple, nous pouvons identifier la caractéristique qui est le plus souvent mentionnée lorsque les clients sont satisfaits ou mécontents d'un service ou produit. Dans le contexte de notre travail, ce service a été exploité dans le traitement de texte afin d'analyser sa tonalité.

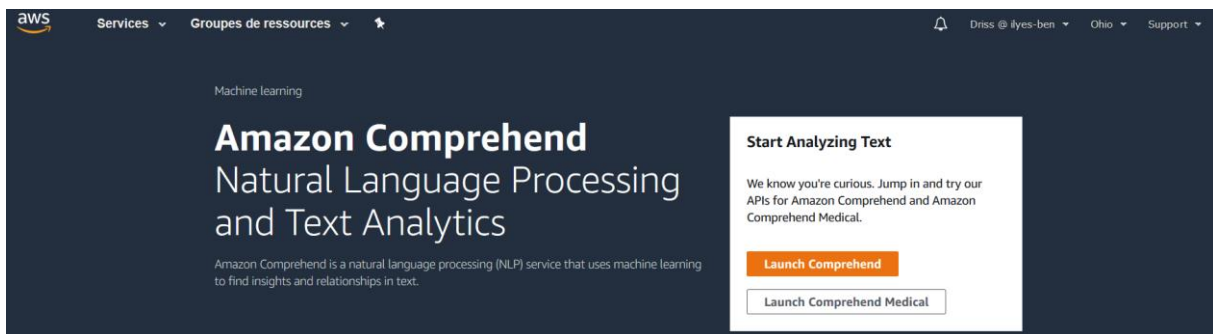


Figure 28: Interface AWS « Amazon Comprehend »

Amazon Comprehend utilise le traitement du langage naturel (NLP) pour extraire des informations sur le contenu des documents. Il traite tout fichier texte au format UTF-8 (Universal Character Set Transformation Format - 8 bits). Il développe des idées en reconnaissant les entités, les phrases clés, le langage, les sentiments et d'autres éléments communs dans un document.

Il utilise un modèle préformé pour examiner et analyser un document ou un ensemble de documents afin de recueillir des informations à son sujet. Ce modèle est formé en permanence sur un grand nombre de textes, de sorte qu'il n'est pas nécessaire de fournir une base d'apprentissage et il peut examiner et analyser des documents en 6 langues.


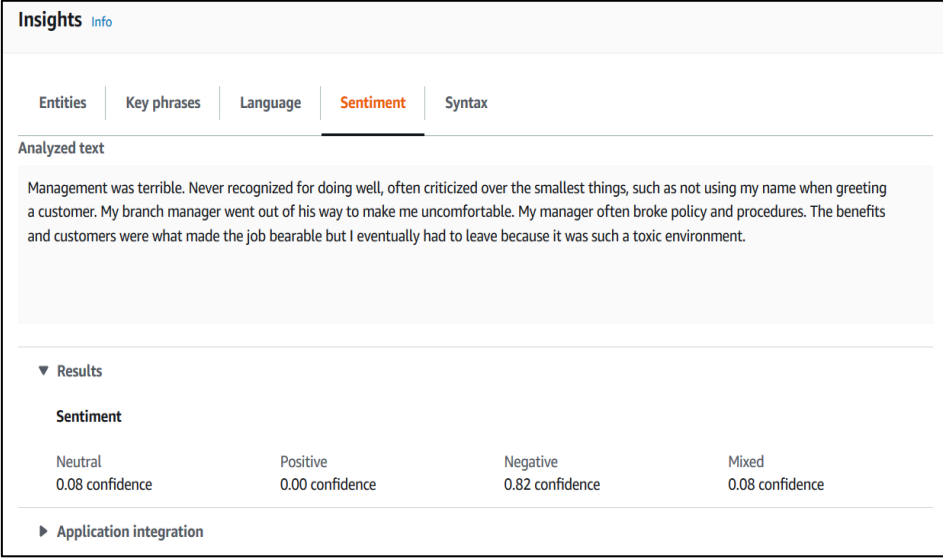
Topic modeling	
The first 5 jobs (up to 1MB each) in each monthly cycle will be free for the first year upon signup. Usage beyond the first 5 jobs will be billed at \$1.00 per job (flat rate) for the first 100MB.	
5 jobs/month	Free
1 job	\$1.00
Cost calculator 	

Figure 29: Coûts des analyses “Amazon Comprehend”

Afin de valider le modèle de nombreux essais ont été effectués en utilisant ce service en incorporant des exemples d’avis en parallèle avec notre modèle. Le service étant gratuit que pour les premières utilisations (Figure 29), nous avons pu tester 500 avis et ça a donné la même tonalité pour 481 avis en comparant le modèle développé et le modèle en ligne d’AWS qui est très performant et donne un résultat très précis. Ce qui nous mène à dire que notre modèle est d’une précision de 481/500 donc plus de 96% par rapport au modèle quasi parfait de AWS. Cette précision, étant très satisfaisante, est relative à l’unicité de la langue d’apprentissage qui est « l’anglais » et au caractère de similitude entre les avis des employés envers leur employeur.

Dans la figure ci-dessous un exemple de test sur l’interface de AWS est représenté.



The screenshot shows the AWS Insights interface with the 'Sentiment' tab selected. The analyzed text is a negative review about a manager. The results section shows a sentiment distribution where 'Negative' is the dominant category with a confidence of 0.82.

Insights <small>Info</small>				
Entities	Key phrases	Language	Sentiment	Syntax
Analyzed text				
Management was terrible. Never recognized for doing well, often criticized over the smallest things, such as not using my name when greeting a customer. My branch manager went out of his way to make me uncomfortable. My manager often broke policy and procedures. The benefits and customers were what made the job bearable but I eventually had to leave because it was such a toxic environment.				
▼ Results				
Sentiment				
Neutral 0.08 confidence	Positive 0.00 confidence	Negative 0.82 confidence	Mixed 0.08 confidence	
▶ Application integration				

Figure 30: Exemple avis négatif

Ceci dit, le modèle d’apprentissage Unigramme choisi donne des résultats satisfaisants dans un premier lieu qui répondent à notre besoin primaire qui est d’identifier les avis négatifs pour pouvoir investiguer leur cause.

1.2. Le Databook :

Comme expliqué, le databook, étant établi pour des besoins spécifiques à un client lors d'une première mission, n'a jamais été remis en question. Il a été adapté tel qu'il était la première fois à toutes les missions.

Lors du diagnostic, il a été remarqué que certaines améliorations pourraient être intéressantes et donc, nous avons pu intervenir sur les points suivants :

- Une analyse plus approfondie sur YouTube qui enrichie le Databook :

Dans l'ancienne version du Databook représentée dans la figure 17, l'apparition sur YouTube était mentionnée dans le récapitulatif sans donner de détails sur l'activité sur cette plateforme. Une nouvelle feuille a été créée sur le Databook dédiée uniquement à YouTube en mettant en avant la présence d'une chaîne YouTube de la cible, le nombre de vidéos partagées ainsi que le nombre de vue pour refléter l'activité de l'entreprise sur cette plateforme. Tout cela illustré par des histogrammes pour une meilleure comparaison entre la cible et ses concurrents comme cela est représenté sur la figure ci-dessous :

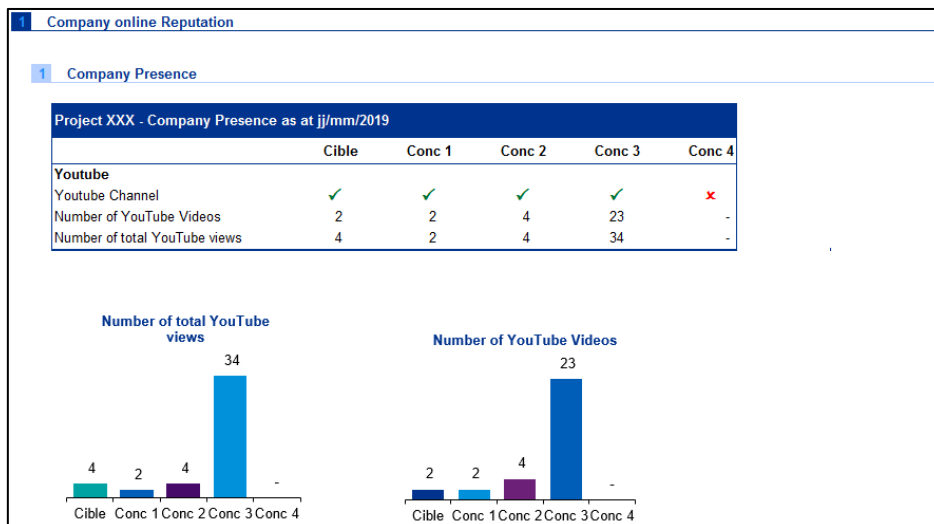


Figure 31: Feuille YouTube databook

- Une meilleure exploitation des données HR Network :

Le HR Network ou les plateformes pour les ressources humaines sont de plus en plus utilisées de nos jours. Les employés donnent des avis sur leur employeur, les anciens dirigeants laissent leurs feedbacks sur l'entreprise également et ce genre d'information permettent de faire sortir des éléments négatifs clés qui peuvent nuire à la réputation de l'entreprise et qui doivent être investigués de plus près. En développant le modèle du Sentiment Analysis, il a été plus simple de trier les avis négatifs et de les lire afin d'enquêter sur les plus redondants d'entre eux. La partie HR Network a donc été enrichie par une analyse des sentiments des avis qui ne se faisait pas avant compte tenu de l'importance du nombre d'avis qui rendait la tâche quasiment impossible manuellement. Les résultats de l'analyse des sentiments qui a été faite grâce au programme développé (annexe 2) ont été valorisés sous forme de tableau et anneau tel que représenté sur la figure ci-dessous.

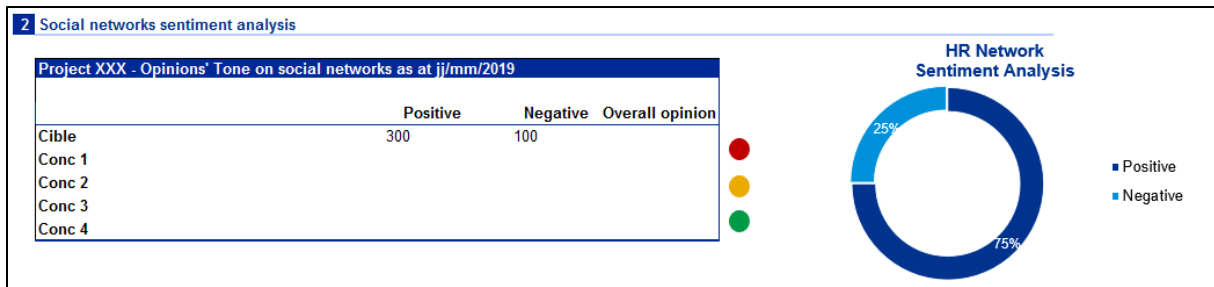


Figure 32: Feuille Sentiment Analysis - HR Network

Il a été également remarqué que certaines informations disponibles sur les plateformes ne sont pas exploitées comme les notations de certains indicateurs clés à savoir : L'équilibre vie social/vie professionnelle, le salaire/avantages sociaux, l'évolution de carrière, le management et la culture de l'entreprise. Ces derniers sont importants à mentionner afin d'effectuer la comparaison avec les concurrents et de détecter des pistes de recherche sur les 5 indicateurs.

Ces informations ont été extraites puis résumées dans un tableau et illustrées par des radars qui servent à superposer les différents indicateurs des concurrents avec ceux de la cible et ceci afin de faire la comparaison entre eux et détecter les points de faiblesses comme l'exemple représenté sur la figure ci-dessous :

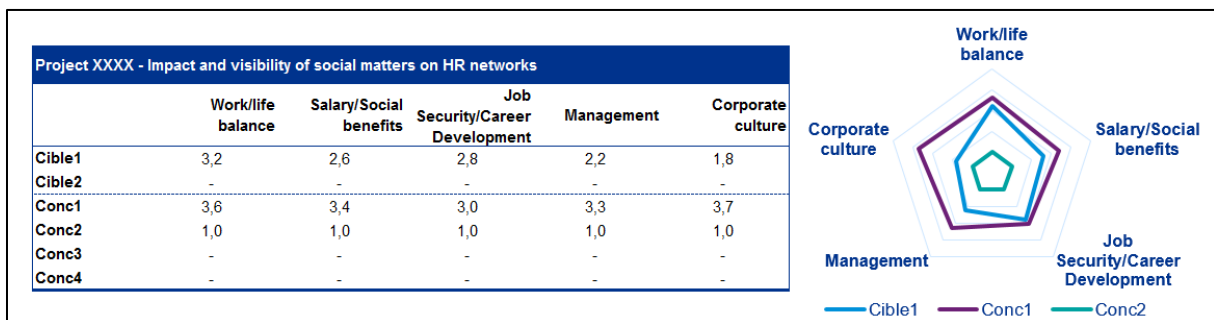


Figure 33: Feuille KPIs - HR Network

- Un récapitulatif plus riche :

Lors de notre revue du databook, nous avons constaté que l'ancienne version du récapitulatif ne résumait que quelques points des données collectées lors de l'analyse (Figure 34) et ne résumait pas l'intégralité des parties du rapport.

1 Company online Reputation – Social Media

1 Presence on social media

Project xxx - Presence on social media as at jj/mm/2018						
Active Facebook page	×	×	✓	×	×	✓
Number of Facebook fans	14	-	383	-	-	-
Active Twitter account*	×	×	✓	×	×	✓
Wikipedia updated	×	×	×	×	×	×
Number of YouTube Videos	10	4	110	4	4	45
Number of total YouTube views	3 054	31	99 883	556	1 119	33 693
Active LinkedIn page	×	×	✓	✓	✓	✓

Figure 34: Feuille Récapitulatif (Ancienne Version)

Pour assurer l'exhaustivité des informations sur le récapitulatif et sa cohérence avec le rapport, il a été structuré de la même manière que le rapport (Structure représentée dans la figure 19) en illustrant les 3 grandes parties à savoir : Company Presence Analysis, Project and Brands Analysis et dans Executive Analysis. Dans chaque partie, les éléments clés qui ressortent sont mentionnés comme représenté dans la figure 35 ci-après.

1 Company Presence					
Project XXX - Company Presence as at jj/mm/2019					
	Cible	Conc 1	Conc 2	Conc 3	Conc 4
Social Media					
Facebook page					
Active Facebook page					
Twitter account					
Active Twitter account*					
Linkedin Page					
Active Linkedin page					
HR Network					
Indeed					
Viadeo					
Glassdor					
Wikipedia					
Wikipedia page					
Updated Wikipedia page					
Youtube					
Youtube Channel					
Active youtube Channel					
2 Company's Projects Presence					
Project XXX - Company Presence as at jj/mm/2019					
	Cible	Conc 1	Conc 2	Conc 3	Conc 4
Projects					
Visibility					
3 Company Executive Presence					
Project XXX - Company Executive Presence as at jj/mm/2019					
	Cible	Conc 1	Conc 2	Conc 3	Conc 4
CEO Notoriety					
Twitter account					
Wikipedia page					
Linkedin profile					
Website					

Figure 35: Feuille Récapitulatif (Nouvelle Version)

- Un nettoyage des données non exploitées :

Sur l'ancienne version du Databook, il y avait des tableaux et des graphiques non exploités vu l'absence de l'information comme, par exemple, les données LinkedIn auxquels nous n'avions plus accès depuis que le compte professionnel premium LinkedIn utilisé n'est plus actif ou encore depuis la nouvelle politique de confidentialité de Facebook qui a rendu les données disponibles sur ce dernier très restreintes. Ces tableaux et graphiques non exploités ne faisaient qu'encombrer le fichier Excel et rendaient sa manipulation difficile (Enregistrement, extraction, ouverture, etc.) donc nous avons opté pour leur suppression en attendant d'aller vers des outils de data visualisation plus performant.

1.3. L'organisation :

Après la formalisation du processus, les différentes tâches qui concernent chaque rubrique du rapport ont été distinguées ce qui a permis de les classer dans un enchaînement logique comme cela est représenté dans le fichier MS Project (Annexe 1) mais aide également à estimer le temps nécessaire pour le traitement de chaque partie du rapport.

- L'organisation générale de l'équipe :

Pour pallier au chevauchement des missions, l'idée de l'établissement d'un planning prévisionnel a été suggérée malgré qu'il n'y ait pas de visibilité sur le court et moyen terme par rapport aux missions OR. Pour cela, une proposition, qui est en train d'être discuté, a été émise comme suit :

Un planning prévisionnel n'est pas forcément fixe mais peut être flexible selon le volume des missions. Nous pouvons donc, établir un premier planning sur 3 mois disant que, par exemple, nous allons avoir 5 missions OR durant cette période, que les ressources humaines soient allouées suivant le planning à chaque mission équitablement (sachant que les opérations dans le processus OR ont été chronométrés) et que ce planning doit être mis à jour d'une façon hebdomadaire (Mise à jour continue) pour réaffecter le personnel qui n'a pas été sollicité à d'autres types de missions ou d'autres projets de développement interne.

D'un point de vue plus stratégique, il a été proposé d'aller conquérir le marché local pour compenser l'incertitude de la demande du marché Français afin de signer des missions OR plus régulièrement avec les entreprises locales ce qui permettra d'assurer une régularité dans le travail et un minimum de visibilité pour la charge de l'équipe R&S.

Il a été suggéré également d'organiser, durant les périodes creuses, de petites formations sur le processus OR et les outils utilisés pour que l'ensemble de l'équipe soit opérationnel dans le cas où nous aurons besoin d'un effectif plus volumineux (Par exemple : plusieurs missions OR à la fois) car ce n'est pas l'ensemble des éléments de l'équipe qui maîtrisent ce type de mission.

- La répartition des tâches relative à l'exécution d'une mission OR :

Dans les missions OR, généralement, l'effort de deux personnes, un junior et un sénior, parmi les membres de l'équipe suffit afin de délivrer le rapport dans les délais et la répartition des tâches n'est pas souvent un long sujet de discussion. Il est, donc, nécessaire de bien catégoriser les tâches faites par l'un et l'autre afin d'exploiter le temps de façon optimal. A cet effet, une proposition de répartition des tâches a été émise comme suit :

Une fois le scope de travail reçu par mail, une séance de brainstorming entre junior et senior doit être entamée afin creuser un peu plus les éléments envoyés par mail pour définir un scope de travail personnalisé pour chaque personne où chacun à ses propres limites d'extension du travail à ne pas dépasser. Après cela la répartition des tâches doit être de sorte que le junior s'occupe de l'extraction des données pour alimenter les parties « Company Presence Analysis » et « HR Network » du rapport pendant que le senior investigue les projets/marques de la société pour en sélectionner les plus significatifs ainsi que les dirigeants qui sont le plus présent sur le Web si ces derniers ne sont pas déjà définis dans le scope. Une fois terminé, le junior passe aux deux parties « Projects and Brands Analysis » et « Executif Analysis » en ayant comme feuille de route les commentaires déjà fait par le senior afin d'avancer

efficacement pendant que le sénior s'occupe de la vérification de ce qui a été fait et supervise la cohérence des informations utilisées. Après cela, la consolidation des informations est faite par le junior en exportant avec UpSlide sur le PowerPoint et en faisant une première analyse sous forme de commentaires. Le senior, par la suite, porte un jugement sur le contenu et la pertinence des commentaires selon son expérience en prenant l'avis du junior pour ne pas diverger et sortir du scope. La mise en forme et les ajustements sont faits par les deux collègues avant de revoir le rapport projeté en entier avec le reste de l'équipe (Un regard extérieur) pour porter une appréciation globale et modifier les éléments jugés incomplets.

Cette organisation est faite dans l'objectif de responsabiliser le junior et de l'impliquer un maximum dans la prise de décision ce qui permet de faciliter le travail de contrôle de son aîné.

2. Application sur cas pratique (Projet Hermès):

Afin de valider les solutions mises en place tout au long du processus de Due Diligence Réputationnelle et de soutenir le travail fait dans le cadre de ce projet de fin d'étude, une application sur une mission réelle a été faite et présentée dans ce qui suit.

Le choix du projet a été fait en prenant en compte le nombre de mentions dans ce dernier pour pouvoir le considérer comme un projet pilote représentatif des résultats obtenus. Selon l'historique, le nombre de données dans une mission OR varie dans un intervalle de [100, 2×10^6] et dans cette mission un peu moins d'un million de mentions ont été extraites ce qui représente un nombre moyen de données par rapport au minimum et au maximum.

2.1. Contexte de la mission :

Pour des raisons de confidentialité, l'équipe « Recherche & Stratégie » donne un nom de code pour chaque mission afin de préserver le secret client. La mission d'Online Réputation réalisée par l'équipe dans notre cas a été nommée « Projet Hermès ».

Pour cette mission, le client de KPMG est un Fond d'Investissement qui voulait acquérir un groupe qui opère dans le domaine de la production immobilière. Le client a sollicité KPMG afin de faire une Due Diligence Réputationnelle sur l'entreprise qu'il veut acquérir afin de valoriser sa notoriété, son image de marque et détecter d'éventuels sujets conflictuels à investiguer avant d'investir des millions d'euros.

2.2. Présentation de l'entreprise cible et ses concurrents :

Pour des raisons de confidentialité, les dénominations des entreprises vont être anonymisées dans ce qui suit pour préserver le secret client.

Nous avons dit que l'entreprise acquéreuse (acheteuse) est un fond d'investissement dans le paragraphe précédent et l'entreprise cible qu'il veut acquérir est un groupe français qui réunit deux filiales en France. Le groupe, nommé « Groupe Cible1 », active dans le domaine des constructions immobilières, de la conception et réalisation des plateformes logistiques. Il porte également une attention particulière à mettre à profit les techniques les plus modernes de modélisation 3D, de reconstitution des bâtiments existants et de simulation de processus industriels.

Le groupe « Groupe Cible1 » détient deux filiales : Un contractant général qui porte la même dénomination que le groupe qui va être nommé « Cible1 » (La différence avec le groupe est l'association du préfixe « Groupe » à ce dernier) et un promoteur immobilier qui va être nommé « Cible2 ». Ces 2 entités opèrent dans des domaines similaires et ont pour ambition de proposer à chaque client une approche globale de leur projet allant de la recherche de foncier à la livraison “prêt à piloter” et ce, dans une démarche intégrant responsabilités économique, sociale et environnementale.

La filiale « Cible1 » est un producteur d'immobilier d'entreprise qui repose sur un management qui privilégie l'expérience terrain des collaborateurs, la spécialisation “métiers“, une très forte implication personnelle et le partage des valeurs de qualité, d'esprit de service, de transparence et d'innovation de l'entreprise.

La filiale « Cible2 » est un créateur bâtisseur d'opération immobilière qui propose une démarche orientée « utilisateur » c'est-à-dire de la « sélection de foncier » au « prêt à utiliser ». Anticipant les attentes et parlant le même langage que les utilisateurs, ses chefs de projet intègrent l'ensemble des attentes pour être force de propositions et de solutions créatives directement inspirées des besoins opérationnels auxquels le bâtiment devra répondre.

Le benchmark avec quatre (4) concurrents a été transmis par le client. Les concurrents ont été nommés : Conc1, Conc2, Conc3 et Conc4. Les deux premiers sont les concurrents directs de la filiale Cible1 et les deux derniers ceux de la filiale Cible2 mais les quatre opèrent dans le même domaine d'activité avec quelques petites différences de spécialisation.

Le groupe « Groupe Cible1 », étant en pleine croissance, a suscité l'intérêt du fond d'investissement afin de réaliser une importante croissance bénéfique pour les deux parties.

2.3. Présentation du rapport final de la mission :

Le rapport final de la mission a été présenté au client en décortiquant chaque partie. Il est à noter que les résultats obtenus sont le fruit de l'exécution des programmes développés (En annexe).

3. Apports :

Après l'implémentation des solutions citées ci-avant et l'application sur la mission OR (Projet Hermès), de nettes améliorations ont été remarquées durant les différentes étapes du processus de Due Diligence Réputationnelle. Dans cette partie, les apports de ce travail dans l'optimisation du processus ont été structurés en 3 dimensions, en accord avec le dernier point du chapitre précédent.

3.1.La dimension qualitative :

Il est souvent difficile de quantifier une amélioration qualitative. Cependant, nous avons remarqué, qu'avec l'automatisation des tâches et le modèle de Sentiment Analysis développé, le nombre de mentions ne dérangeait plus. C'est-à-dire qu'avoir des milliers/millions de mentions à traiter n'était plus un obstacle. Ce qui permet d'élargir l'étendue du travail et donc, d'assurer une exhaustivité des données permettant de refléter réellement la réputation de

l'entreprise et de s'approcher beaucoup plus de la réalité en traitant un maximum de données accessibles.

Par ailleurs, la restructuration du Databook, autant sur l'aspect fond que l'aspect forme, a permis une meilleure représentation des données et une meilleure cohérence dans l'enchaînement des rubriques. Cela assure une fluidité de lecture lors de la présentation du rapport en donnant du sens à chaque partie et en le valorisant avec un contenu exhaustif. Ce qui a impacté positivement la qualité de l'analyse qui a été revalorisée avec de nouveaux indicateurs et de nouvelles représentations. Ces derniers ont permis d'enrichir le rapport avec plus de détails sur les mentions et donc plus de précision lors des différentes analyses.

3.2. La dimension temporelle :

Les écarts de productivité entre les entreprises de services, en général, s'expliquent principalement par le temps investi en exerçant leurs activités. Dans les cabinets de conseil en particulier, une productivité horaire plus élevée se traduirait par une diminution du nombre total d'heures travaillées, pour le même nombre d'employés.¹² Les répercussions de la dimension temporelle se remarquent également sur l'aspect organisationnel où les durées des tâches sont considérées comme éléments majeurs lors de la planification et l'organisation.

La gestion du temps est l'une des compétences les plus importantes à posséder dans le domaine du Consulting car le temps représente la ressource clé qui est utilisée pour créer de la valeur. En optimisant le temps dédié pour chaque mission, l'effort et les coûts sont optimisés.

Le coût de la main d'œuvre et le coût variable le plus important dans le processus de Due Diligence Réputationnelle (Ce qui est représenté dans la figure 57 issue du rapport des coûts de MS Project en Annexe) en estimant que le coût fixe de l'outil Sysomos, qui est de l'ordre de 30000€/an, est totalement amorti par la multitude des missions OR ainsi que les autres utilisations de ce dernier pour d'autres types de missions donc ce coût fixe est considéré optimal.

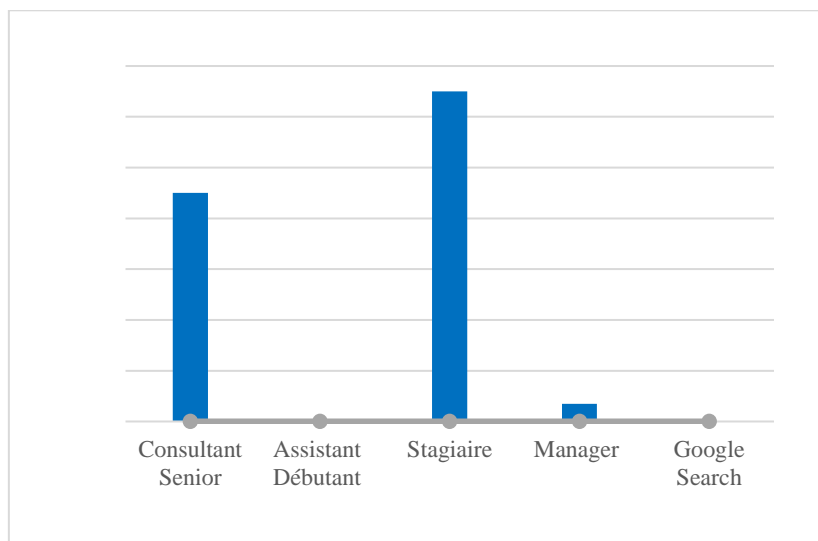


Figure 36: Rapport des coûts (MS Project)

¹² Kremp, Élisabeth : Économie et Statistique N° 270, pp. 63-76

A cet effet, plus le temps investi dans une mission est long (Calculé en terme d'heures chargeables) plus le coût de la main d'œuvre est élevé et donc le coût d'une mission revient plus cher et par conséquent son prix l'est également.

Le gain de temps a été calculé sur la même mission afin de comparer le temps dédié à la mission avant et après l'optimisation du processus en gardant le même scope de travail.

Avant l'optimisation du processus le temps moyen passé sur le projet Hermès a été calculé en élaborant le diagramme de Gant sur le fichier MS Project (Figure 58) et a été estimé à 72 heures chargeables (Heures facturables pour le client).

OR	Online Réputation Project	72 h	Dim 17/03/19	Dim 31/03/19		
OR.SW	Scope of Work	2 h	Dim 17/03/19	Dim 17/03/19		
OR.DR	Data Retrieving and exploration	70 h	Dim 17/03/19	Dim 31/03/19		
OR.DR.1	Company presence analysis	70 h	Dim 17/03/19	Dim 31/03/19		
OR.DR.2	Projects and Brands Analysis	44 h	Lun 18/03/19	Mar 26/03/19		
OR.DR.2.1	Primary Research	3 h	Lun 18/03/19	Lun 18/03/19	4	Google Search;Sta
OR.DR.2.2	Extensive Research (Google, Sysomos)	6 h	Mar 19/03/19	Mer 20/03/19	11	Google Search;Sta
OR.DR.2.5	Social media analysis	6 h	Dim 24/03/19	Mar 26/03/19	11	Google Search;Sta
OR.DR.2.4	Sentiment analysis	12 h	Jeu 21/03/19	Mar 26/03/19	11	Comptes Réseaux
OR.DR.3	Executive Analysis	9 h	Lun 18/03/19	Mar 19/03/19		
OR.DR.3.1	Primary Research	2 h	Lun 18/03/19	Lun 18/03/19	4	Comptes Réseaux
OR.DR.3.2	Extensive Research (Google, Sysomos)	6 h	Lun 18/03/19	Mar 19/03/19	16	Google Search;Sta
OR.DR.4	HR Network	5 h	Dim 17/03/19	Dim 17/03/19		
OR.DR.4.3	Primary Research	1 h	Dim 17/03/19	Dim 17/03/19	4	
OR.DR.4.4	Sentiment analysis	4 h	Dim 17/03/19	Dim 17/03/19	19	Comptes Réseaux
OR.DC	Data Cleaning	18 h	Mar 19/03/19	Dim 24/03/19		
OR.DF	Databook Fulfillment	6 h	Dim 24/03/19	Lun 25/03/19		
OR.DF.1	Renseignement du Databook	5 h	Dim 24/03/19	Dim 24/03/19	22;23;24	Pack Microsoft Of
OR.DF.2	Ajustement des graphiques	1 h	Lun 25/03/19	Lun 25/03/19	26	Pack Microsoft Of
OR.RC	Report Construction	18 h	Lun 25/03/19	Mer 27/03/19		
OR.RC.1	Préparation du rapport	2 h	Lun 25/03/19	Lun 25/03/19	27	Pack Microsoft Of
OR.RC.2	Analyses et commentaires	12 h	Lun 25/03/19	Mer 27/03/19	29	Consultant Senior;
OR.RC.3	Sanity Check	4 h	Mer 27/03/19	Mer 27/03/19	30	Consultant Senior;

Figure 37: Durée du projet avant l'optimisation (MS Project)

Après l'optimisation du processus, la même mission a été refaite en utilisant les automatisations et le modèle développé. La durée de cette dernière a été calculée en élaborant le diagramme de Gant sur le fichier MS Project (Figure 59) et a été estimé à 50 heures chargeables (Heures facturables).

OR	Online Réputation Project	50 h	Dim 17/03/19	Mar 26/03/19		
OR.SW	Scope of Work	2 h	Dim 17/03/19	Dim 17/03/19		
OR.DR	Data Retrieving and exploration	45 h	Dim 17/03/19	Lun 25/03/19		
OR.DR.1	Company presence analysis	39 h	Dim 17/03/19	Dim 24/03/19		
OR.DR.1.1	Extensive Research (Google, Sysomos)	4 h	Dim 17/03/19	Lun 18/03/19	4	Google Search;Sys
OR.DR.1.2	Social media analysis	2 h	Lun 18/03/19	Dim 24/03/19	7	Comptes Réseaux
OR.DR.1.5	Sentiment analysis	6 h	Lun 18/03/19	Dim 24/03/19	7	Comptes Réseaux
OR.DR.2	Projects and Brands Analysis	41 h	Dim 17/03/19	Lun 25/03/19		
OR.DR.2.1	Primary Research	2 h	Dim 17/03/19	Lun 18/03/19	4	Google Search;Sta
OR.DR.2.2	Extensive Research (Google, Sysomos)	4 h	Lun 18/03/19	Lun 18/03/19	11	Google Search;Sta
OR.DR.2.5	Social media analysis	5 h	Jeu 21/03/19	Lun 25/03/19	11	Google Search;Sta
OR.DR.2.4	Sentiment analysis	10 h	Mer 20/03/19	Dim 24/03/19	11	Comptes Réseaux
OR.DR.3	Executive Analysis	5 h	Dim 17/03/19	Lun 18/03/19		
OR.DR.3.1	Primary Research	1 h	Dim 17/03/19	Dim 17/03/19	4	Comptes Réseaux
OR.DR.3.2	Extensive Research (Google, Sysomos)	3 h	Lun 18/03/19	Lun 18/03/19	16	Google Search;Sta
OR.DR.4	HR Network	5 h	Dim 17/03/19	Dim 17/03/19		
OR.DR.4.3	Primary Research	1 h	Dim 17/03/19	Dim 17/03/19	4	
OR.DR.4.4	Sentiment analysis	0 h	Dim 17/03/19	Dim 17/03/19	19	Comptes Réseaux
OR.DC	Data Cleaning	17 h	Lun 18/03/19	Mer 20/03/19		
OR.DF	Databook Fulfillment	5 h	Mer 20/03/19	Jeu 21/03/19		
OR.DF.1	Renseignement du Databook	4 h	Mer 20/03/19	Jeu 21/03/19	22;23;24	Pack Microsoft Of
OR.DF.2	Ajustement des graphiques	1 h	Jeu 21/03/19	Jeu 21/03/19	26	Pack Microsoft Of
OR.RC	Report Construction	18 h	Jeu 21/03/19	Mar 26/03/19		

Figure 38: Durée du projet après l'optimisation (MS Project)

Le gain de temps est estimé donc à une réduction de $\frac{(72-50)}{72} = 30,55\%$ du temps dédié à une mission habituellement. Ce gain de temps de plus de 30% est directement traduit par un gain sur le coût de revient de la mission. Ce qui permet à KPMG d'opter pour une des deux solutions stratégiques proposées :

- Baisser le prix d'une mission, qui variait de 7000 à 12000€ selon l'étendue du travail, ce qui permet d'avoir un avantage plus compétitif sur le marché et de proposer un prix imbattable (vu que le prix était déjà abordable grâce à la délocalisation de l'activité en Algérie) et donc d'acquérir de nouveaux marchés et de croître en terme de volume;
- Garder le même prix de la mission et faire plus de marge sur chaque mission, ce qui permet de financer le développement d'autres activités dans le cabinet ou de compenser d'autres types de missions qui n'ont pas été très profitable (Par exemple : Des missions négociées à un bas prix face à un concurrent)

3.3. La dimension organisationnelle :

Une bonne organisation du travail au sein d'une équipe favorise l'adhésion, aide à bénéficier des potentialités dans le but d'atteindre les objectifs fixés dans les meilleurs délais. La formalisation du processus a joué un rôle important dans l'aspect organisationnel. En effet, elle nous a permis de structurer les différentes étapes, de les chronométrer, d'optimiser quelques-unes d'entre elles et de faire sortir une répartition des tâches équitable en valorisant chaque contribution dans le rapport final.

Une équipe formée, structurée, soudée et organisée est une équipe qui a la capacité d'absorber une charge de travail importante et d'œuvrer en groupe pour atteindre les objectifs dans les deadlines malgré les difficultés et les heures de travail illimitées dans ce métier.

4. Perspectives :

Bien qu'elle soit limitée et parfois crainte, l'Intelligence Artificielle continue à faire son chemin, influençant les domaines auxquels elle est appliquée. Des Chatbots du service client aux automates extrêmement sophistiqués, L'IA a sans aucun doute un impact sur tout ce qui nous entoure. Faisant déjà partie de notre monde, profiter de certains de ses avantages pourrait avoir une incidence très positive sur notre quotidien. Les experts du domaine ont le choix quant à la mesure dans laquelle ils vont permettre à l'IA de jouer un rôle dans leurs stratégies et faire évoluer le processus du présent vers un futur plus ambitieux.

Après l'entame de ce travail, plusieurs pistes d'amélioration ont été remarquées et une multitude de procédures ont été remises en question. Ce travail a été l'exemple de ce que nous pouvons améliorer au quotidien en gagnant quelques minutes ou en fournissant moins d'effort grâce aux avancés technologiques. N'ayant pas encore achevé ce stage chez KPMG, plusieurs perspectives peuvent s'offrir dans le but d'instaurer une amélioration continue dans le département, à savoir :

- Le développement du modèle de Sentiment Analysis pour lui permettre de traiter, pas que de simples avis, mais des articles en entier et en tirer des conclusions (Il peut être utilisé pour identifier les articles négatifs du Crisis Tracking directement par exemple) ;
- La généralisation du traitement des commentaires en Anglais du modèle de SA à toutes les langues fréquemment rencontrées (Français, Italien, Arabe, etc.)
- Le développement d'un modèle qui permettra d'automatiser la partie Data Cleaning qui prend beaucoup de temps dans le processus en intégrant des mots clés associé à notre recherche ce qui permettra de faire sortir uniquement les articles et mentions souhaités ;
- La remise en question des limites de taille de donnée à extraire (10 pages Google pour le Crisis Tracking, 10 termes de crises, etc.) ce qui ne va plus poser problème car les tâches ont été automatisées donc le nombre n'est plus une contrainte ;
- La mise en place d'une plateforme dynamique de Data Visualisation qui peut énormément aidé lors de la présentation des résultats à un client sachant que des premiers travaux ont été entamé sur Power BI pour la présentation de certains KPI.

De manière plus générale, ce travail a ouvert la porte à plusieurs réflexions concernant l'automatisation des autres processus du cabinet ainsi qu'à l'utilisation des programmes informatiques et du Machine Learning dans chaque sous-processus qui peut être facilement programmable.

Conclusion :

La formalisation du processus était une étape primordiale dans la structuration des étapes du processus, l'identification des pistes d'amélioration et la contribution à l'optimisation du processus à travers les solutions explicités dans ce chapitre. Il est clair que ces solutions ne sont pas définitives mais font l'objet d'amélioration continue en accord avec le développement des technologies et en cohésion avec la stratégie de KPMG qui privilégie ses clients en premier lieu et donne une importance incontestable au développement des projets en interne en second lieu.

Conclusion Générale

Conclusion Générale :

Le présent travail s'inscrit dans le cadre de l'optimisation du processus de Due Diligence réputationnelle. Pour ce faire, nous avons opté pour une solution axée sur l'automatisation car c'est la mieux adaptée à ce type de processus tout en prenant en compte ses conséquences sur l'aspect organisationnel. L'étude a été effectuée au sein du cabinet d'audit, d'expertise comptable et de conseil KPMG dans la division Deal Advisory qui intervient lors des missions de rapprochement entre les entreprises telles les fusions et acquisitions.

Le processus de Due Diligence réputationnelle génère comme Output un rapport qui synthétise une vue d'ensemble sur la réputation de l'entreprise, sa notoriété et son image de marque qui deviennent de nos jours des paramètres de décision essentiels pour les investisseurs avant de s'engager avec un partenaire.

Afin d'épouser la stratégie de développement et d'amélioration continue de KPMG, l'optimisation de ce processus a été le cœur de ce travail. Pour ce faire, il a été convenu de passer par les étapes suivantes :

- La formalisation du processus afin de le structurer et mieux le visualiser qui a été faite sur MS Project en tenant compte des principes de l'approche processus ;
- Le diagnostic détaillé du processus afin d'identifier les pistes d'amélioration ;
- Le développement d'un programme mathématique pour automatiser l'extraction des données à partir des pages Web et les classer ;
- Le développement d'un modèle de Machine Learning pour automatiser l'analyse des sentiments (Tonalités) des avis des employés sur l'entreprise ;
- La remise en question de l'aspect qualitatif du rapport final délivré au client sur le fond et la forme ;
- La planification des tâches ainsi que l'organisation de l'équipe.

Le résultat de ce travail a été remarqué dans ce qui suit :

- Un gain de temps de plus de 30% sur la durée de la mission avec en conséquence la proposition de deux stratégies de pricing;
- Un rapport final délivré au client mieux valorisé avec de nouvelles analyses et interprétations ;
- Une proposition d'organisation plus rigoureuse qui génère une meilleure harmonie de travail.

Les solutions implémentées ont été validées avec succès par les responsables au sein du cabinet et ont été considérées comme une ouverture vers une approche d'amélioration continue plus régulière en concluant qu'avec quelques changements incrémentaux nous pouvons être plus efficaces. Dans le même esprit, des perspectives ont été proposées afin de s'inscrire dans cette démarche d'amélioration continue.

Pour finir, nous tenons à souligner que ce travail nous a permis de capitaliser les connaissances acquises durant notre cursus universitaire, d'avoir une idée claire sur le métier que nous ambitionnons qui est le consulting, d'apprendre que le monde professionnel c'est

avant tout une question de relationnel et de loyauté mais aussi, plus en relation avec le thème de notre projet, que pour acheter du bon pain il ne suffit pas de choisir le moins cher ou le plus proche mais plutôt le boulanger le mieux réputé si vous voulez manger une saveur unique.

Bibliographie :

Ouvrages : _____

[Borovkov, 2017] BOROVKOV Andriy. *Image Classification with Deep Learning*. April 2017. 49p. ISBN: 9780128104088.

[McConnell et al. 1985] McCONNELL John J. and NANTELL Timothy J. *Corporate Combinations and Common Stock Returns: The Case of Joint Ventures*. June 1985. 536p.

[Lajoux et al. 2010] LAJOUX, A. R., & ELSON, C. *The art of M&A due diligence*. Editeur: McGraw-Hill Professional, 2010. 561p. ISBN: 007162936X

[Lindholm et al. 2019] LINDLHOLM Andreas, WAHLSTORM Niklas, LINDSTEIN Fredrik, Thomas B. Schön *Supervised Machine Learning*. March, 2019. 112p.

[Schlünz, 2014] SCHLUNZ G. I. *Advanced natural language processing* May 2014 73p ISBN : 1787121429.

Thèses: _____

[Ayodele, 2010] AYODELE Taiwo Oladipupo. *Types of Machine Learning Algorithms*. 49p. Master Thesis University of Portsmouth. 2010.

[Babanazarov, 2012] BABANAZAROV Bahtiyar, M.S. *Agricultural and applied economic* 100p.

Thesis Faculty of Texas Tech University August, 2012.

[Barbato, 2016] BARBATO Massimo-Maria. *Measuring corporate reputation through online social media*. 51P.

Master Thesis University of Bologna. December 2016.

[Brandtzæg, 2014] BRANDTZAEG Emilie. *Corporate Reputation in Mergers and Acquisitions*. 85p.

Master Thesis university of stavanger 2014.

[Freitag, 2014] FREITAG Andreas. *Applying Business Capabilities in a Corporate Buyer M&A Process*. 248p.

Doctoral dissertation, University of Technology München. 2014. ISBN 978-3-658-07281-0.

[Kotsiantis, 2007] KOTSIANTIS, S. B. *Supervised Machine Learning*. 30p.

Thesis, Department of Computer Science and Technology University of Peloponnese, Greece, July, 2007.

[Luckert et al. 2015] LUCKERT Michael, SCHAEFER-KEHNERT Moritz. *Using Machine Learning Methods for Evaluating the Quality of Technical Documents*. 95p.

Master Thesis, 2015.

[Lundborg, 2017] LUNDBORG Anton. *Text classification of short messages* 56p.

Master's thesis lund university, faculty of engineering LTH 2017.

[Nouboussi et al. 2008] NOUBOUSSI Josiane, BEUKE Ndeye Diene. *Due diligence: learn from the past, but look toward the future.* 63p.

Master Thesis. 2008.

[Thi Quynh Van, 2013] THI QUYNH VAN, Nguyen. *Impact of Mergers and Acquisitions announcement on shareholder value.* 56p.

Master Thesis University of Nottingham. September 2013.

Articles: _____

[Bertrand, 2017] BERTRAND Antoine. *Émergence de l'intelligence artificielle.* 30p. éditeur: A.DEPLAE, SG UCM National. Janvier 2017.

[Fombrun, 1996] FOMBRUN. *Corporate reputation: definition and data.* December, 1996. 371p.

[Journal of Machine Learning Research 10 (2009) 1755-1758] JOURNAL OF MACHINE LEARNING RESEARCH 10 *Dlib-ml: A Machine Learning Toolkit* éditeur: Soeren Sonnenburg, july, 2009.

[Journal of Machine Learning Research 12 (2011) 2825-2830] JOURNAL OF MACHINE LEARNING RESEARCH 12 *Scikit-learn: Machine Learning in Python* éditeur: Mikio Braun, october, 2010.

[Journal of Machine Learning Research 17 (2016) 1-5] JOURNAL OF MACHINE LEARNING RESEARCH 17 *mlr: Machine Learning in R* éditeur : Antti Honkela, September, 2016.

Webographie: _____

LIBMANN Anne-Marie. *La Compliance et la Due Diligence dans le spectre de l'intelligence économique.* [consulté le 28 mars 2019].

Disponible sur <<https://www.fla-consultants.com/fr/blog-actualites/la-compliance-et-la-due-diligence-dans-le-spectre-de-l-intelligence-economique>>

MALICK mytectra. *Découvrez les sept raisons pour lesquelles vous devez apprendre le langage Python.* [consulté le 17 mars 2019].

Disponible sur <<https://www.developpez.com/actu/133538/Programmation-decouvrez-les-sept-raisons-pour-lesquelles-vous-devez-apprendre-le-langage-Python-selon-myTectra/>>

KEVORK Djansezian, what you need to know about artificial intelligence. [consulté le 08 avril 2019].

Disponible sur <<https://www.businessinsider.fr/us/jeff-bezos-shareholder-letter-on-ai-and-machine-learning-2017-4/>>

SHANDWICK Weber. *Une nouvelle ère dans l'engagement*. [consulté le 03 mars 2019].

Disponible sur <<http://webershandwick.fr/etude-ceo-reputation-premium-une-nouvelle-ere-dans-lengagement/>>

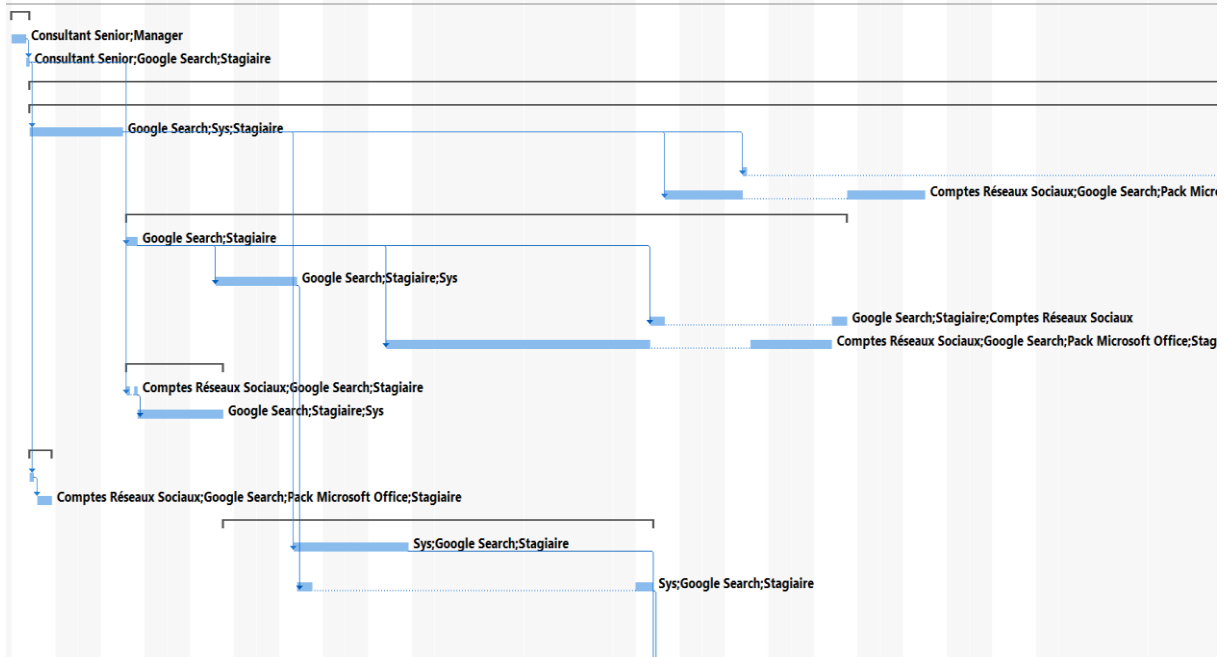
KPMG Intranet. 2018. [consulté le 21 avril 2019].

Disponible sur <<https://intra.ema.kpmg.com/sites/FR/Pages/Bienvenue.aspx> >

Annexes:

Annexe n°1: Processus Due Diligence sur MS Project

OR	Online Réputation Project	72 h	Dim 17/03/19	Dim 31/03/19		
OR.SW	Scope of Work	2 h	Dim 17/03/19	Dim 17/03/19		
OR.DR	Data Retrieving and exploration	70 h	Dim 17/03/19	Dim 31/03/19		
OR.DR.1	Company presence analysis	70 h	Dim 17/03/19	Dim 31/03/19		
OR.DR.2	Projects and Brands Analysis	44 h	Lun 18/03/19	Mar 26/03/19		
OR.DR.2.1	Primary Research	3 h	Lun 18/03/19	Lun 18/03/19	4	Google Search;Sta
OR.DR.2.2	Extensive Research (Google, Sysomos)	6 h	Mar 19/03/19	Mer 20/03/19	11	Google Search;Sta
OR.DR.2.5	Social media analysis	6 h	Dim 24/03/19	Mar 26/03/19	11	Google Search;Sta
OR.DR.2.4	Sentiment analysis	12 h	Jeu 21/03/19	Mar 26/03/19	11	Comptes Réseaux
OR.DR.3	Executive Analysis	9 h	Lun 18/03/19	Mar 19/03/19		
OR.DR.3.1	Primary Research	2 h	Lun 18/03/19	Lun 18/03/19	4	Comptes Réseaux
OR.DR.3.2	Extensive Research (Google, Sysomos)	6 h	Lun 18/03/19	Mar 19/03/19	16	Google Search;Sta
OR.DR.4	HR Network	5 h	Dim 17/03/19	Dim 17/03/19		
OR.DR.4.3	Primary Research	1 h	Dim 17/03/19	Dim 17/03/19	4	
OR.DR.4.4	Sentiment analysis	4 h	Dim 17/03/19	Dim 17/03/19	19	Comptes Réseaux
OR.DC	Data Cleaning	18 h	Mar 19/03/19	Dim 24/03/19		
OR.DF	Databook Fulfillment	6 h	Dim 24/03/19	Lun 25/03/19		
OR.DF.1	Renseignement du Databook	5 h	Dim 24/03/19	Dim 24/03/19	22;23;24	Pack Microsoft Of
OR.DF.2	Ajustement des graphiques	1 h	Lun 25/03/19	Lun 25/03/19	26	Pack Microsoft Of
OR.RC	Report Construction	18 h	Lun 25/03/19	Mer 27/03/19		
OR.RC.1	Préparation du rapport	2 h	Lun 25/03/19	Lun 25/03/19	27	Pack Microsoft Of
OR.RC.2	Analyses et commentaires	12 h	Lun 25/03/19	Mer 27/03/19	29	Consultant Senior;
OR.RC.3	Sanity Check	4 h	Mer 27/03/19	Mer 27/03/19	30	Consultant Senior;



Annexe n°2: Programme informatique – Sentiment Analysis (SA)

```
#-*- coding: utf-8 -*-
```

```
"""
```

```
Created on Tue Apr 23 16:57:55 2019
```

```
@author: dbenhadji
```

```
"""
```

```
# _____ Modules _____
```

```
import openpyxl as xl
```

```
import re
```

```
import pandas as pd
```

```
import json
```

```
import requests
```

```
from selenium import webdriver
```

```
import numpy as np
```

```
import matplotlib.pyplot as plt
```

```
import langdetect as ld
```

```
import codecs
```

```
import nltk
```

```
from nltk.stem import WordNetLemmatizer
```

```
from nltk.corpus import stopwords
```

```
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
from sklearn.model_selection import train_test_split
```

```
from sklearn.utils import shuffle
```

```
from wordcloud import WordCloud
```

```
# In[3]:
```

```
wnl = WordNetLemmatizer()
```

```
# In[4]:
```

```
stopwords_custom = set(w.rstrip() for w in open("C:\\Users\\dbenhadji\\Desktop\\PFE Recherche  
OR\\4. Sentiment analysis\\stopwords.txt"))
```

```
stopwords_nltk = set(stopwords.words('english'))
```

```
stop_words = stopwords_custom.union(stopwords_nltk)
```

```
# In[5]:
```

```

def extract_excel(all_):
    file_res= 'BDD_Lang.xlsx'
    writer = pd.ExcelWriter(file_res , engine='xlsxwriter')
    df_res = pd.DataFrame(all_ , index = ['Com', 'language'])
    df_res = df_res.transpose()
    df_res.to_excel(writer,sheet_name='All_Data', startrow=0, startcol=0, header=['Com', 'language'],
index= False)
    writer.save()
    return 'Done'

def get_file(path):
    df = pd.read_csv(path, encoding="iso-8859-1", dtype=str)
    df = shuffle(df)
    return df

def correctChar(df_old):
    """Remplace les caractères non reconnus (problèmes d'encodage)"""
    df_new = df_old.copy()
    df_new = df_new.apply(strip_char, axis=0)
    df_new.columns = map(str.lower, df_new.columns)
    df_new.columns = map(str.upper, strip_char(pd.Series(df_new.columns)))
    return df_new

def strip_char(df_series):
    """Corrige les erreurs d'encodage"""
    if not np.issubdtype(df_series.dtype, np.number):
        try:
            df_series = df_series.str.normalize('NFKD').str.encode('ascii', errors='ignore').str.decode('utf-8')
            df_series = [" ".join(re.findall("\w+[\']?\w+|[A-Za-z]", s)) for s in df_series]
        except:
            print("%s -- FAILED" % df_series.name)
    return df_series

def remove_plural(s):
    """Supprime le s final si la lettre precedente est une consonne"""

```

```

pattern = re.compile('[b-df-hj-np-tv-xz]s$')
if pattern.search(s):
    return re.sub('s$', '', s)
pattern = re.compile('ies$')
if pattern.search(s):
    return re.sub('ies$', 'y', s)
pattern = re.compile('es$')
if pattern.search(s):
    return re.sub('s$', '', s)
return s

def remove_special(s, special = [\., \*, <br />, \", ^-, #, @, :, 'RT']):
    """Supprime les caracteres speciaux specifiques dans la liste"""
    for sp in special:
        s = re.sub(sp, "", s)
    return s

def my_tokenizer(s, stoplist=stop_words):
    """Notre tokenizer"""
    s = s.lower()
    tokens = nltk.tokenize.word_tokenize(s)
    tokens = [wnl.lemmatize(t, pos='v') for t in tokens]
    tokens = [t for t in tokens if not t in stoplist]
    tokens = [t for t in tokens if not all(i.isdigit() for i in t)]
    tokens = [t for t in tokens if not re.match("^\d", t)]
    tokens = [remove_plural(t) for t in tokens]
    tokens = [remove_special(t) for t in tokens]
    tokens = [t for t in tokens if len(t)>3]
    return tokens

def init_tfidf(X_train, ngram_range=(1, 1), min_df=5, max_df=.9):
    """Initialise un tf-idf vectorizer"""
    tfidf = TfidfVectorizer(min_df=min_df, max_df=max_df, ngram_range=ngram_range,

```

```

        stop_words=stop_words, tokenizer=my_tokenizer)

tfidf.fit(X_train)

bag_of_words = tfidf.get_feature_names()

print("\t%d relevant features (words)!\n" % len(bag_of_words))

return tfidf, bag_of_words

def doc_to_term(corpus, tfidf):
    """Vectorise un corpus"""
    vectorized = tfidf.transform(corpus)
    return vectorized.toarray()

def contingency_table(y, fitted):
    """Construit une table de contingence"""
    df = pd.DataFrame({'Observed': y, 'Predicted': fitted})
    ct = pd.crosstab(df.Observed, df.Predicted, margins = False)
    ct.columns = [s[:6] for s in ct.columns]
    return ct

def print_score(model, Xtest, Ytest):
    """Renvoie la performance du model et la table de contingence"""
    fitted = model.predict(Xtest)
    score = model.score(Xtest, Ytest)
    print("\nClassification rate: %1.3f\n" % score)
    print(contingency_table(Ytest, fitted))

def print_predict(model, Xtest):
    """Prediction d'un nouvel ensemble de test """
    print ( "prediction ")
    fitted = model.predict(Xtest)
    return fitted

def make_wordcloud(Series, max_words = 1000):
    d = {w: int(np.abs(v)*100) for w, v in zip(Series.index, Series.values)}
    wordcloud = WordCloud(width=900, height=500, max_words=max_words,
        relative_scaling=1,normalize_plurals=False).generate_from_frequencies(d)
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis("off")

```

```
plt.show()
# In[6]:
# _____ definition _____
res_list = []
titles= []
coms= []
notes= []
positifs= []
negatifs= []
dates= []
biglist= []
nb_avis=[]
note_cats1=[]
NOTECAT1=[]
note_cats1=[]
NOTECAT2=[]
note_cats2=[]
NOTECAT3=[]
note_cats3=[]
NOTECAT4=[]
note_cats4=[]
NOTECAT5=[]
note_cats5=[]
LISTE=[]
L=[]
rates=[]
COMS=[]
NBAVIS=[]
NOTES=[]
RATES=[]
# In[6]:
# _____ Fonctions _____
```

```

def scrap(url):
    chromeOptions = webdriver.ChromeOptions()
    chromeOptions.add_experimental_option('useAutomationExtension', False)
    browser = webdriver.Chrome()
    browser.get(url)
    innerHTML = browser.execute_script("return document.body.innerHTML")
    browser.quit()
    return innerHTML

def extract_titles (index, text):
    title = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            title+= text[i]
            i+=1
    return title

def extract_coms (index, text):
    com = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<' or text[i+1] != '/' or text[i+2] != 's' or text[i+3] != 'p' or text[i+4] != 'a' or
text[i+5] != 'n' or text[i+6] != '>'):
            com+= text[i]
            i+=1
    return com

def extract_notes (index, text):
    note = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            note+= text[i]
            i+=1
    return note

```



```

def extract_positifs (index, text):
    positif = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            positif+= text[i]
            i+=1
    return positif

```

```

def extract_negatifs (index, text):
    negatif = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            negatif+= text[i]
            i+=1
    return negatif

```

```

def extract_dates (index, text):
    date = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            date+= text[i]
            i+=1
    date=date.replace('&nbsp;',' ')
    return date

```

```

def extract_note_cats(index,text):
    notescat = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            notescat+= text[i]

```

```

        i+=1
    return notescat
def extract_note_cats1(index,text):
    notescat1 = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            notescat1+= text[i]
            i+=1
    return notescat1
def extract_note_cats2(index,text):
    notescat2 = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            notescat2+= text[i]
            i+=1
    return notescat2
def extract_note_cats3(index,text):
    notescat3 = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            notescat3+= text[i]
            i+=1
    return notescat3
def extract_note_cats4(index,text):
    notescat4 = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            notescat4+= text[i]

```

```

        i+=1
    return notescat4
def extract_note_cats5(index,text):
    notescat5 = "
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            notescat5+= text[i]
            i+=1
    return notescat5
def extract_nb_avis (index,text):
    nb_avi = "
    if ( index != None):
        i= index.end()
        while ( text[i] != "" ):
            nb_avi+= text[i]
            i+=1
    return nb_avi
#_____Extraction des données vers le databook_____
databook = 'C:\\Users\\dbenhadji\\Desktop\\PFE Recherche OR\\Coding\\HRNET.xlsx'
def load_report (databook) :
    return xl.load_workbook(databook)
def get_sheet ( workbook, sheetname):
    return workbook.get_sheet_by_name (sheetname)
def edit_value (sheet, cell, value):
    sheet[cell].value= value
    return 'done'
def print_value (sheet, cell):
    print(sheet[cell].value)
    return 'done'
def get_value (sheet, cell):
    return sheet[cell].value

```

```

# _____ Appel au Fonctions _____
for company_name in companies :
    res_list= []
    hr=""
    nb_page=10
    hr
    scrap('https://www.indeed.fr/cmp/'+company_name+'/reviews?fcountry=ALL&start='+str(nb_page))
    nb_page+=10
    while (nb_page <70):
        print (nb_page)
        hr+=
        scrap('https://www.indeed.fr/cmp/'+company_name+'/reviews?fcountry=ALL&start='+str(nb_page))
        nb_page+=20
        print ( company_name, ' __', len(hr))
        res_list = hr.split("cmp-review-container")
        res_list.append(company_name)
        L.append(res_list)
dataset= {}
idf= 0
HRN=get_sheet(workbook,'HR_Network')
for elm in L:
    com=""
    coms=[]
    nb_avis=[]
    notes=[]
    for res in elm :
        idf+=1
        re_titles = re.search(r"<div class=\"cmp-review-title\"><span>", res)
        title= extract_titles(re_titles, res)
        titles.append(title)
        re_coms = re.search(r"<span class=\"cmp-review-text\" itemprop=\"reviewBody\">", res)
        com= extract_coms(re_coms, res)
        com = com.replace('<br>', ' ')

```

```

coms.append(com)

re_notes = re.search(r"<div class=\"cmp-ratingNumber\">", res)
note= extract_notes(re_notes, res)
notes.append(note)

re_dates = re.search(r"<span class=\"cmp-review-date-created\">", res)
date= extract_dates(re_dates, res)
dates.append(date)

re_nb_avis = re.search(r"<meta itemprop=\"reviewCount\" content=\"\", res)
nb_avi= extract_nb_avis(re_nb_avis, res)
nb_avis.append(nb_avi)

re_rates = re.search(r"<span class=\"cmp-header-rating-average\" itemprop=\"ratingValue\">",
elm[0])
rate= extract_negatifs(re_rates, elm[0])
rate = rate.replace(',', '.')

re_note_cats1 = re.search(r"<span class=\"cmp-ReviewCategories-rating\">", elm[len(elm)-2])
note_cat1= extract_note_cats1(re_note_cats1, elm[len(elm)-2])

re_note_cats2 = re.search(r"personnelle </span></div><div class=\"cmp-ReviewCategories-
category\"><span class=\"cmp-ReviewCategories-rating\">", elm[len(elm)-2])
note_cat2= extract_note_cats2(re_note_cats2, elm[len(elm)-2])

re_note_cats3 = re.search(r"Avantages sociaux </span></div><div class=\"cmp-ReviewCategories-
category\"><span class=\"cmp-ReviewCategories-rating\">", elm[len(elm)-2])
note_cat3= extract_note_cats3(re_note_cats3, elm[len(elm)-2])

re_note_cats4 = re.search(r"Évolution de carrière </span></div><div class=\"cmp-
ReviewCategories-category\"><span class=\"cmp-ReviewCategories-rating\">", elm[len(elm)-2])
note_cat4= extract_note_cats4(re_note_cats4, elm[len(elm)-2])

re_note_cats5 = re.search(r"Management </span></div><div class=\"cmp-ReviewCategories-
category\"><span class=\"cmp-ReviewCategories-rating\">", elm[len(elm)-2])
note_cat5= extract_note_cats5(re_note_cats5, elm[len(elm)-2])

COMS.append(coms)
NBAVIS.append(nb_avis)
NOTECAT1.append(note_cat1)
NOTECAT2.append(note_cat2)
NOTECAT3.append(note_cat3)
NOTECAT4.append(note_cat4)

```

```

NOTECAT5.append(note_cat5)

NOTES.append(notes)

RATES.append(rate)

#_____Remplissage DB_____

# In[7]:

df = pd.read_excel('C:\\Users\\dbenhadji\\Desktop\\PFE Recherche OR\\Coding\\DA6003F.xlsx',
encoding='iso-8859-1')

print("Here are the abstract categories...")

categories = df['score'].value_counts()

category_table = pd.DataFrame(categories).transpose()

print("\nCounts:\n")

print(category_table)

# In[8]:

corpus = np.array(df['text'])

X_train, X_test, y_train, y_test = train_test_split(corpus, df.score, stratify = df.score, test_size = .33)

tfidf, bag_of_words = init_tfidf(X_train, min_df=10)

# In[9]:

dtc_train = doc_to_term(X_train, tfidf)

dtc_test = doc_to_term(X_test, tfidf)

# In[10]:

print("%d examples for training\n%d samples for testing" % (dtc_train.shape[0], dtc_test.shape[0]))

# In[11]:

clf = LogisticRegression(penalty= 'l2', C=10)

print(clf.fit(dtc_train, y_train))

# In[12]:

print_score(clf, dtc_test, y_test)

# In[13]:

coefs={ }

Results = pd.DataFrame({'coefficient': clf.coef_[0]}, index=bag_of_words)

coefs = Results['coefficient'].to_dict()

ROW=25

f=-1

for elt in COMS :

```

```

f+=1
x=0
y=0
g=0
for com in elt:
    try:
#       Detecter les commentaires en Anglais
        langue=ld.detect(com)
    except:
        langue = "error"
    if (langue=='en'):
        g+=1
        tokens_ = my_tokenizer(com)
        poids = 0
        for t in tokens_ :
            if t in coefs.keys():
                poids += coefs[t]
        if (poids >= 0):
            x+=1
        else:
            y+=1

    print('le nombre de commentaire pour ' + companies[f] + ' est:',len(COMS[0]))
    print('Ce programme traite uniquement les commentaires en Anglais pour ' + companies[f] + ' dont
le pourcentage est de',g/len(COMS[0]),'%')
    print('Le nombre de commentaires en anglais pour ' + companies[f] + ' est:',g)
    print ('Le nombre de commentaires positifs pour ' + companies[f] + ' est:',x)
    print ('Le nombre de commentaires negatifs pour ' + companies[f] + ' est:',y)
    HRN.cell(row=ROW,column=6).value=x
    HRN.cell(row=ROW,column=7).value=y
    ROW+=1
i=0
ROW=11
for elt in NBAVIS:

```

```

HRN.cell(row=ROW,column=6).value=NBAVIS[i][0]
i+=1
ROW+=1
i=0
ROW=11
for elt in RATES:
    HRN.cell(row=ROW,column=7).value=RATES[i]
    i+=1
    ROW+=1
i=0
ROW=72
for elt in NOTECAT1:
    HRN.cell(row=ROW,column=12).value=NOTECAT1[i]
    i+=1
    ROW+=1
i=0
ROW=72
for elt in NOTECAT2:
    HRN.cell(row=ROW,column=13).value=NOTECAT2[i]
    i+=1
    ROW+=1
i=0
ROW=72
for elt in NOTECAT3:
    HRN.cell(row=ROW,column=14).value=NOTECAT3[i]
    i+=1
    ROW+=1
i=0
ROW=72
for elt in NOTECAT4:
    HRN.cell(row=ROW,column=15).value=NOTECAT4[i]
    i+=1

```



```
    ROW+=1
i=0
ROW=72
for elt in NOTECAT5:
    HRN.cell(row=ROW,column=16).value=NOTECAT5[i]
    i+=1
    ROW+=1
# In[14]:
plt.figure(figsize = (10, 12))
pos = Results['coefficient'][Results['coefficient'] > 3]
print ( pos )
make_wordcloud(pos)
plt.figure(figsize = (10, 12))
neg = Results['coefficient'][Results['coefficient'] < -3]
print ( neg )
make_wordcloud(neg)
# In[15]:
workbook.save(databook)
```

Annexe n°3: Explication des ligne du programme SA

Step	Comment	command line
1	Définition de la liste des Stopwords de deux sources différentes (Fichier & nltk corpus)	stopwords.words
2	Lecture du fichier d'apprentissage	read_csv
3	Comptage du nombre de commentaires pour chaque catégorie de l'attribut score	categories = df['score'].value_counts()
4	Création d'un tableau horizontal	category_table = pd.DataFrame(categories).transpose()
5	Récupération du text des commentaires	corpus = np.array(df['text'])
6	Définition de la répartition des données d'apprentissage et de test	X_train, X_test, y_train, y_test = train_test_split(corpus, df.score, stratify = df.score, test_size = .33)
7	Répartition des ensembles d'apprentissage où les données de test représentent 33% de l'ensemble.	X_train, X_test, y_train, y_test = train_test_split(corpus, df.score, stratify = df.score, test_size = .33)
8	Sélection des données et des classes à considérer dans le training	X_train, X_test, y_train, y_test = train_test_split(corpus, df.score, stratify = df.score, test_size = .33)
9	Définition de la méthode de la variable contenant les classes. Ici, le Shuffle est par défaut, à True. Les données sont mélangées.	X_train, X_test, y_train, y_test = train_test_split(corpus, df.score, stratify = df.score, test_size = .33)
10	Fonction qui retourne le token et le score tf*idf	tfidf, bag_of_words = init_tfidf(X_train, ngram_range=(1, 2), min_df=10)
11	Conversion des commentaires en matrice de fréquences selon la formule tf*idf	tfidf = TfidfVectorizer(min_df=min_df, max_df=max_df, ngram_range=ngram_range, stop_words=stop_words, tokenizer=my_tokenizer)
12	Permettre de ne prendre en considération que les tokens ayant un nombre d'occurrences dans un même commentaire se situant entre le min et le max définis.	tfidf = TfidfVectorizer(min_df=min_df, max_df=max_df, ngram_range=ngram_range, stop_words=stop_words, tokenizer=my_tokenizer)
13	Permettre de définir le N des N-grams à prendre en considération de manière à ce que min_n <= n <= max_n	tfidf = TfidfVectorizer(min_df=min_df, max_df=max_df, ngram_range=ngram_range, stop_words=stop_words, tokenizer=my_tokenizer)
14	Permettre d'éliminer les termes contenus dans la stop liste de l'ensemble des tokens	tfidf = TfidfVectorizer(min_df=min_df, max_df=max_df, ngram_range=ngram_range, stop_words=stop_words, tokenizer=my_tokenizer)
15	Permettre d'effectuer une tokenization autre que celle par défaut	tfidf = TfidfVectorizer(min_df=min_df, max_df=max_df, ngram_range=ngram_range, stop_words=stop_words, tokenizer=my_tokenizer)
16	Construction du vocabulaire	tfidf.fit(X_train)
17	Permet de récupérer la liste des tokens / N-grams	bag_of_words = tfidf.get_feature_names()
18	Fonction qui retourne la matrice de	doc_to_term(corpus, tfidf)

	fréquences	
19	Construction de la matrice Term-frequency en considérant les valeurs du tf*idf. Cette matrice est creuse.	vectorized = tfidf.transform (corpus)
20	Transformation en tableau	vectorized. toarray ()
21	Affichage du nombre de lignes de la matrice	dtc_train. shape [0], dtc_test. shape [0]
22	Déclaration d'une instance du modèle Logistic Regression	clf = LogisticRegression (penalty= 'l2', C=10)
23	Spécification des paramètres du modèle	clf = LogisticRegression (penalty= 'l2', C=10)
24	Apprentissage des données selon le modèle linéaire instancié (L'input étant la matrice de fréquences)	clf. fit (dtc_train, y_train)
25	Fonction qui prend en input l'instance du modèle et les données de test pour y appliquer la prédiction. Retourne la table de contingence	print_score (clf, dtc_test, y_test)
26	Application du modèle appris aux données de test. Renvoi des valeurs prédites	fitted = model. predict (Xtest)
27	Calcul de la précision du modèle appris en fonction des données de test	score = model. score (Xtest, Ytest)
28	Fonction qui crée la table de contingence à partir des résultats de validation et des valeurs prédites par le modèle appris	contingency_table (y, fitted)
29	Fonction qui permet la création d'un tokenizer	my_tokenizer (s, stoplist=stop_words)
30	Normalisation du corpus en lowercase	s = s.lower()
31	Récupération des tokens	tokens = nlk.tokenize.word_tokenize (s)
32	Instanciation de l'ontologie WordNet pour la lemmatisation	wnl = WordNetLemmatizer ()
33	Lemmatisation. Transformation des flexions de l'ensemble des tokens en lemmes.	tokens = [wnl.lemmatize (t, pos='v') for t in tokens]
34	Définition de la classe de mots de la méthode de lemmatisation POS (Part of speech). Ici, les verbes sont considérés.	tokens = [wnl.lemmatize (t, pos='v') for t in tokens]
35	Elimination des termes de la stoplist	tokens = [t for t in tokens if not t in stoplist]
36	Elimination des nombres	tokens = [t for t in tokens if not all(i. isdigit () for i in t)]
37	Elimination des termes commençant par un nombre	tokens = [t for t in tokens if not re. match ("^\d", t)]
38	Fonction qui élimine la marque du pluriel	tokens = [remove_plural (t) for t in tokens]
39	Fonction qui élimine les caractères spéciaux	tokens = [remove_special (t) for t in tokens]
40	Ne garde que les tokens dont la taille est supérieure à 3	tokens = [t for t in tokens if len (t)>3]
41	Fonction qui permet d'instancier le nuage de mots avec comme input les tokens à afficher selon leur polarité.	make_wordcloud (Series, max_words = 1000)
42	Définition des données du nuage de mots	d = {w: int(np.abs(v)*100) for w, v in zip (Series. index , Series. values)}
43	Instanciation du nuage de mot sur la base des fréquences en input	wordcloud = WordCloud (width=900, height=500, max_words=max_words, relative_scaling=1,normalize_plurals=False).

		generate_from_frequencies(d)
44	Récupération des tokens prédits positifs	pos Results['coefficient'][Results['coefficient'] > 3]
45	Récupération des tokens prédits négatifs	neg Results['coefficient'][Results['coefficient'] < -3]

Annexe n°4: Programme Informatique – Automatisation FGP/GS

```
# -*- coding: utf-8 -*-
```

```
"""
```

```
Created on Mon Apr 22 07:52:30 2019
```

```
@author: dbenhadji
```

```
"""
```

```
import openpyxl as xl
```

```
import re
```

```
import pandas as pd
```

```
import json
```

```
import requests
```

```
from selenium import webdriver
```

```
from datetime import datetime, timedelta
```

```
# FIRST GOOGLE PAGE #
```

```
nom_code_projet='Consulting'
```

```
company_name = 'KPMG'
```

```
url='https://www.google.com/search?q='+company_name
```

```
# Google suggest URL #
```

```
URL="http://suggestqueries.google.com/complete/search?client=firefox&q="+company_name
```

```
# _____ Fonctions _____
```

```
sugg=[]
```

```
headers = {'User-agent':'Mozilla/5.0'}
```

```
response = requests.get(URL, headers=headers)
```

```
result = json.loads(response.content.decode('utf-8'))
```

```
sugg=result[1]
```

```
print(sugg)
```

```
#Level of control
```

```
locf = pd.read_excel('C:\\Users\\dbenhadji\\Desktop\\PFE Recherche OR\\Coding\\listing  
Controlled.xlsx')
```

```
def dic_loc(df_lloc):
```

```
    lloc = { }
```

```
    for i in df_lloc.itertuples():
```

```
        lloc[i[1]] = i[2]
```

```

    return lloc
def scrap(url):
    chromeOptions = webdriver.ChromeOptions()
    chromeOptions.add_experimental_option('useAutomationExtension', False)
    browser = webdriver.Chrome()
    browser.get(url)
    innerHTML = browser.execute_script("return document.body.innerHTML")
    browser.quit()
    return innerHTML
def extract_title (index, text):
    title = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            title+= text[i]
            i+=1
    return title
def extract_sites (index, text):
    site = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '>'):
            site+= text[i]
            i+=1
    return site
def extract_dates (index, text):
    date = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != ','):
            date+= text[i]
            i+=1

```

```

date=date.replace('&nbsp;',' ')
if 'heures' in date:
    date= datetime.today().strftime('%d-%m-%Y')
    return date
elif 'heure' in date:
    date= datetime.today().strftime('%d-%m-%Y')
    return date
elif 'minutes' in date:
    date= datetime.today().strftime('%d-%m-%Y')
    return date
elif 'jour' in date:
    date= datetime.today().strftime('%d-%m-%Y')
    return date
elif 'jours'in date:
    nb_jr= date.split(' ')[3]
    date = datetime.today() - timedelta(days=int(nb_jr))
    date= date.strftime('%d-%m-%Y')
    return date
else:
    return date
def load_report (databook) :
    return xl.load_workbook(databook)
def get_sheet ( workbook, sheetname):
    return workbook.get_sheet_by_name (sheetname)
def edit_value (sheet, cell, value):
    sheet[cell].value= value
    return 'done'
def print_value (sheet, cell):
    print(sheet[cell].value)
    return 'done'
def get_value (sheet, cell):
    return sheet[cell].value

```

```
# _____Appel aux Fonctions_____
```

```
fgp = scrap(url)
dic_loc= dic_loc(locf)
res_list = []
titles= []
biglist=[]
sites= []
dates=[]
loc=[]
sources=[]
liste_site_seul=[]
#Diviser l'ensemble de la page par résultat
res_list = fgp.split("<div class=\"r\">")
res_list= res_list[1:]
#Stop list pour exclure les termes
stop_list= ['org', 'fr', 'com', 'www','dz', 'https', 'http']
for res in res_list :
    niv_con='uncontrolled'
    re_title = re.search(r"<h3 class=\"LC20lb\">", res)
    title= extract_title(re_title, res)
    titles.append(title)
    re_sites = re.search(r"<a href=\"", res)
    site= extract_sites(re_sites, res)
    site___ = site.split('/')
    SITES = site___[2]
    sources.append(SITES)
    site_seul = SITES.split('.')
    for s in site_seul:
        if s in stop_list:
            site_seul.remove(s)
    ss_ = ''.join(site_seul)
    liste_site_seul.append(ss_)
```



```

cn = company_name.replace(' ','')
print( cn, '____', ss_)
print (ss_.find(cn) )
if ( ss_.lower().find(cn.lower()) != -1 ):
    print ('here')
    niv_con='controlled'
else :
    if (ss_ in dic_loc.keys()):
        niv_con=dic_loc[ss_]
sites.append(site)
loc.append(niv_con)
re_dates = re.search(r"<span class=\"st\"><span class=\"f\">", res)
date= extract_dates(re_dates, res)
dates.append(date)
biglist.append(titles)
biglist.append(sources)
biglist.append(dates)
biglist.append(loc)
#Extraction des donnée vers le databook
databook = 'C:\\Users\\dbenhadji\\Desktop\\PFE Recherche OR\\Coding\\DATABOOK Coding.xlsx'
#_____First _____ Google
page_____
workbook=load_report(databook)
fgp=get_sheet(workbook,'FGP')
j=-1
for col_ in range (3,7):
    j+=1
    row_=6
    for i in range(0,len(titles)):
        fgp.cell(row=2,column=2).value='Project '+nom_code_projet + '- First Google page for '+company_name
        fgp.cell(row=4,column=2).value='Project '+nom_code_projet + '- First Google page for '+company_name

```

```

fgp.cell(row=row_,column=col_).value=biglist[j][i]
row_+=1
#_____GoogleSuggest_____
gs=get_sheet(workbook,'GS')
row_=7
for i in range(0,5):
    gs.cell(row=row_,column=2).value=sugg[i]
    gs.cell(row=6,column=2).value='Project ' +nom_code_projet + '- First Google page for
'+company_name +' as keyword - Top 10'
    gs.cell(row=2,column=2).value='Project ' +nom_code_projet + '- First Google page for
'+company_name +' as keyword - Top 10'
    row_+=1
row_=7
for i in range(5,len(sugg)):
    gs.cell(row=row_,column=3).value=sugg[i]
    row_+=1
workbook.save(databook)

```

Annexe n°5: Programme Informatique - Crisis Tracking

```
# -*- coding: utf-8 -*-
```

```
"""
```

```
Created on Mon Apr 29 09:46:37 2019
```

```
@author: dbenhadji
```

```
"""
```

```
# _____ Modules _____
```

```
import openpyxl as xl
```

```
import re
```

```
import pandas as pd
```

```
import json
```

```
import requests
```

```
from selenium import webdriver
```

```
import html2text
```

```
# _____ Zone de Recherche _____
```

```
company_name = 'pwc'
```

```
# _____ Définition _____
```

```
res_list = []
```

```
titles = []
```

```
coms = []
```

```
notes = []
```

```
positifs = []
```

```
negatifs = []
```

```
dates = []
```

```
biglist = []
```

```
# _____ Fonctions _____
```

```
def scrap(url):
```

```
    chromeOptions = webdriver.ChromeOptions()
```

```
    chromeOptions.add_experimental_option('useAutomationExtension', False)
```

```
    browser = webdriver.Chrome()
```

```
    browser.get(url)
```

```
    innerHTML = browser.execute_script("return document.body.innerHTML")
```

```
    browser.quit()
```

```
    return innerHTML
```

```

def extract_titles (index, text):
    title = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            title+= text[i]
            i+=1
        return title

def extract_content (index, text):
    text = text.replace("<em>", "")
    text = text.replace("&nbsp;", "")
    text = text.replace("</em>", "")
    text = text.replace("<span class=\"f\">", ' ')
    text = text.replace("- </span>", ' ')
    content = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            content+= text[i]
            i+=1
        return content

def extract_sources (index, text):
    source = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):
            source+= text[i]
            i+=1
        return source

def extract_dates (index, text):
    date = ""
    if ( index != None):
        i= index.end()
        while ( text[i] != '<'):

```

```

        date+= text[i]

        i+=1

    date=date.replace('&nbsp;',' ')

    return date

#_____Extraction des données vers le dataset_____

def load_report (databook) :

    return xl.load_workbook(databook)

def get_sheet ( workbook, sheetname):

    return workbook.get_sheet_by_name (sheetname)

def edit_value (sheet, cell, value):

    sheet[cell].value= value

    return 'done'

def print_value (sheet, cell):

    print(sheet[cell].value)

    return 'done'

def get_value (sheet, cell):

    return sheet[cell].value

def extract_excel(all_):

    file_res= 'BD_content.xlsx'

    writer = pd.ExcelWriter(file_res , engine='xlsxwriter')

    df_res = pd.DataFrame(all_ , index = ['title', 'site','contenu'])

    df_res = df_res.transpose()

    df_res.to_excel(writer,sheet_name='All_Data', startrow=0, startcol=0, header=['title',' site','
contenu'], index= False)

    writer.save()

    return 'Done'

#_____ Appel au Fonctions _____

f= 'C:\\Users\\dbenhadji\\Desktop\\PFE Recherche OR\\Coding\\crisiswords.txt'

crisiswords = open (f, "r", encoding='ISO-8859-1')

c = crisiswords.read()

words = c.split(",")

print (words)

print('_____Le nombre de mots est: ',len(words))

```

```

dataset= {}
contenus=[]
sites=[]
contenu=""
content=""
ct=""
for word in words :
    nb_page=00
    while (nb_page <10):
        print('compagny:',company_name)
        print('nombre de page',nb_page)
        ct +=
scrap('https://www.google.com/search?q='+company_name+'+'+word+'&start='+str(nb_page))
    res_list = ct.split("<div class=\"r\">")
    res_list = res_list[1:]
    idf=0
    i=0
    for res in res_list :
        i+=1
        idf+=1
        contenu=""
        re_titles = re.search(r"<h3 class=\"LC20lb\">", res)
        title= extract_titles(re_titles, res)
        titles.append(title)
        re_sites = re.search(r"<cite class=\"iUh30\">", res)
        site= extract_sources(re_sites, res)
        sites.append(site)

        re_content = re.search(r"<div class=\"s\"><div><span class=\"st\">", res)
        contenu= extract_content(re_content, res)
        contenus.append(contenu)
        dataset[idf]= [title, site, contenu]
    nb_page+=10
extract_excel(dataset)

```