République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université M'hamed bougara de Boumerdès
Faculté des Sciences
Département des mathématiques

Mémoire Présenté

Pour l'obtention du Diplôme de Master

En Recherche Opérationnelle

Option : Recherche Opérationnelle, Optimisation et management stratégique

Par : Roza Chentoufi
Et : Djelouah Zineb

Optimisation intelligente de la réputation numérique, cas : SONATRACH

Soutenu à l'Université M'Hamed Bougara de Boumerdès,
Le ....., devant le jury composé de :

D^{r} B. Ferjellah	M.A. classe/ A	Présidente	à l'UMB-Boumerdès.
D^{r} F. Cheurfa	M.C. classe B	Encadreur	à l'UMB-Boumerdès
D^{r} M. Bezoui	M.C. classe B	Co-Encadreur	à l'UMB-Boumerdès.
D^{r} W. Drici	M.C. classe B	Examinatrice	à l'UMB-Boumerdès.

Année Universitaire 2021 – 2022

Remerciements

En tout premier lieu, nous remercions le bon Dieu, tout puissant, de nous avoir donné la force pour survivre, ainsi que l'audace pour dépasser toutes les difficultés. Nous tenons à

remercie messieurs**M. Bezoui** et **F. Cheurfa** pour l'encadrement qu'il nous assuré et leurs précieux et judicieux conseils qu'ils n'ont cessé de nous prodiguer tout au long de ce projet, leurs confiance témoignée, sans oublier leurs qualités humaines.
Ils trouveront ici notre gratitude et notre reconnaissance profondes.

Nous remercions les membres du jury, qui nous fait l'honneur de participer au jugement de ce travail.
notre remerciements vont également à tous nos enseignants du département mathématiques pour leurs conseils et surtout leur compréhension.

De peur d'en oublier certains, On ne nous aventurerais pas à tous les citer et, nul doute qu'ils sauront ici se reconnaître.

Dédicaces

Je dédie ce travail : *À la mémoire de ma mère, que Dieu l'accueille dans son vaste paradis.*

À ma grand mère ***Massaouda*** *qui a toujours été là pour moi.*

À ma chère soeur ***Tafath*** *que je trouve toujours derrière mon dos.*

À ma chère soeur ***Thaninna*** *et sa fille* ***Sadja*** *que j'aime infiniment.*

À mes enseignants de département recherche opérationnelle de l'université de béjaïa

À mes enseignants de département mathématiques de l'université de boumerdes

À mes chères enseignants ***Dr.Asli, Dr.Issaadi, Dr.Cheurfa*** *et* ***Dr.bezoui****. À ma moitié* ***Ryma heddouch*** *.*

À mes chères amies ***Nadjet, Baya et Mira****.*

Chentoufi Roza

Dédicaces

Je dédie ce travail : *À ma très chère mère.*

À mon cher père

À mes chères soeurs.

À mes chers frères

À mes enseignants de département mathématiques de l'université de boumerdes

Djelouah Zineb

Table des matières

Table des figures

Introduction générale

Avec la révolution numérique les internautes peuvent créer, consulter, publier, échanger et partager très facilement avec le reste du monde, créant ainsi une multitude de contenus déversés sur la toile et parcourus chaque seconde par les moteurs de recherche.

Désormais, la réputation d'une personne, d'une marque ou d'une entreprise ne se fait plus seulement par le bouche à oreille. Il faut donc compter avec les informations fournies par les internautes à leur égard et parler de " réputation en ligne ".

Internet vient comme un outil qui permet pour l'entreprise à créer et bâtir son capital image et valoriser sa réputation dont les informations se transmettent à un plus grand nombre de personnes bien plus rapidement, par les échanges entre internautes, et aussi par le référencement des nouveaux contenus.

Aujourd'hui l'apparition des entreprises sur Internet et sur les réseaux sociaux devient un facteur primordial.

Avec l'avènement de la technologie numérique et des appareils intelligents, une grande quantité de données numériques est générée chaque jour. Cette forte augmentation des données, tant en taille qu'en forme, est principalement due aux réseaux sociaux qui permettent à des millions d'utilisateurs de partager des informations, d'exprimer et de diffuser leurs idées et leurs opinions sur un sujet, et montrer leurs attitudes envers un contenu.

Toutes ces actions stockées sur les médias sociaux génèrent un ensemble massif d'opinions qui offre une opportunité pour les systèmes automatiques de fouille et d'analyse de données pour déterminer les tendances des internautes. Plusieurs chercheurs ont montré un vif intérêt pour l'exploitation de ces informations afin de prédire les comportements humains des domaines aussi variés que la médecine, la politique, le marketing, etc. Cette exploitation est principalement basée sur l'analyse d'opinion.

L'analyse d'opinion vise à déterminer le sentiment des gens en analysant leurs messages et différentes actions sur les médias sociaux. Elle consiste à classer la polarité des messages en différents sentiments opposés tels que positif et négatif.

Problématique

La SONATRACH est une entreprise nationale d'hydrocarbures. Première entreprise en Afrique et surnommée la Major d'Afrique, est un acteur énergétique mondial. Elle est reconnue comme un partenaire incontournable dans l'industrie des hydrocarbures depuis plus de 58 ans, et est la locomotive de l'économie algérienne.

Pour pouvoir concrétiser ses ambitions de devenir l'une des cinq premières entreprises pétrolières nationales parmi les plus performantes et les plus rentables de l'industrie

énergétique mondiale, et, pour faciliter la gestion de plus de 200000 employés, et 154 filiales, SONATRACH a entamé la miseen œuvre d'un programme de transformation digitale. Parmi les axes de cette transformation, on trouve la gestion de l'image de marque de l'entreprise au niveau des réseaux sociaux.

En fait, SONATRACH ne se contente plus d'exister sur tous les réseaux sociaux, plus que ça, SONATRACH veut refléter l'image de son excellence et de son leadership et donner une image de marque forte et séduisante.

Pour cela, SONATRACH et ses équipes techniques, vont se lancer dans l'analyse des sentiments sur les réseaux sociaux. Le sentiment sur les médias sociaux représente la perception que les internautes ont de SONATRACH et apporte du contexte à chaque commentaire, partage, réaction ou publication dans lesquels cette dernière est mentionnée.

Notre thème est : **Optimisation intelligente de la réputation numérique, cas : SONATRACH.**

Afin de mieux cerner notre problématique, nous allons essayer de répondre aux interrogations suivantes :

? la SONATRACH dispose-t-elle d'une bonne E-réputation ?

? Comment aider l'entreprise SONATRACH pour soigner son image sur les réseau sociaux en analysant les sentiments qu'elle procure à ses abonnés sur les réseaux sociaux ?

Pour répondre aux interrogations nous avons formulé les hypothèses suivantes :

$H1$. L'entreprise SONATRACH dispose d'une bonne E-réputation lui permettant d'influencer le comportement de ses clients ;

$H2$. L'E-réputation de l'entreprise SONATRACH a un impact positif sur le comportement de ses consommateurs.

Objectifs :

L'objectif de ce mémoire est de détecter automatiquement les sentiments des internautes et leurs avis positifs, négatifs ou neutres sur un produit. En développant un système d'analyse des sentiments pour classer les opinions en trois catégories : positif, négatif et neutre, et représenter les différentes techniques et approches proposées et leurs résultats.

Schéma de la thèse

La présentation du manuscrit s'articule autour de 5 chapitres.

. Le premier chapitre on mettra une présentation générale pour le lieu d'organisme qui nous a accueillis pour effectuer notre stage à savoir SONATRACH ;

. Le deuxième chapitre concernera les la réputation numérique des entreprises sur les médias sociaux ;

. Le troisième chapitre se base sur les déffirente notions de l'analyse des sentients, facebook réaction et la machines learning ;

. Le quatrième chapitre expose la notion d'exploration de données et les déffirentes méthodes d'analyse exploratoire des données ;

. Le cinquième chapitre nous présentons notre approche pour résoudre la problématique et les différentes étapes qu'on a établie.

1

Présentation de l'entreprise

Introduction

Dans ce chapitre, nous présenterons en premier lieu l'organisme qui nous a accueillis pour effectuer notre stage sous l'engaderement de monsieur **Benmokhtar chorahbil** à savoir SONATRACH, ses missions et fonctionnalités, ses directions et son organigramme.

FIGURE 1.1 – Logo de SONATRACH

Sonatrach est une compagnie étatique algérienne et un acteur international majeur dans l'industrie des hydrocarbures, le groupe pétrolier et gazier est classé 1ère en Afrique et 12ème dans le monde en 2013, toutes activités confondues, avec un chiffre d'affaires à l'exportation plus de 60 milliards de US$.

Née le 31 décembre 1963, la compagnie intervient dans l'exploration, la production, le transport par canalisations, la transformation et la commercialisation des hydrocarbures et de leurs dérivés. Elle est 4ème exportateur mondial de GNL, 3ème exportateur mondial de GPL et 5ème exportateur de Gaz Naturel.
Adoptant une stratégie de diversification, Sonatrach se développe aussi bien dans les activités de génération électrique, d'énergies nouvelles et renouvelables, de dessalement d'eau de mer, de recherche et d'exploitation minière.
Poursuivant sa stratégie d'internationalisation, SONATRACH opère en Algérie et dans plusieurs régions du monde : Afrique (Mali, Niger, Libye, Egypte), Europe (Espagne, Italie, Portugal, Grande Bretagne), Amérique Latine (Pérou) et USA.
Sonatrach est une Société Nationale pour la Recherche, la Production, le Transport, la Transformation, et la Commercialisation des Hydrocarbures, " S.P.A " est une entreprise publique algérienne créée le 31 décembre 1963, un acteur majeur de l'pétrolière surnommé la major africaine. SONATRACH est classée la première entreprise d'Afrique.
Aujourd'hui, les activités de SONATRACH ne se limitent pas seulement sur la production pétrolière, elle se développe également dans les activités de pétrochimie, de génération électrique, d'énergies nouvelles et renouvelables, de dessalement d'eau de mer et d'exploitation minière[27].

1.1 Historique de la SONATRACH :

Une Algérie prospère, une Algérie portée par la volonté d'un état qu'après l'indépendance, a très tôt, compris que l'accès à l'énergie est une voie essentielle menant au développement économique, social et politique. C'est dans cette perspective qu'au lendemain de son indépendance, l'Algérie a créé, le 31.12.1963, Sonatrach " **SO**ciété **NA**ationale pour la recherche,la production,le **TRA**nsport et de la **C**ommercialisation des **H**ydrocarbures ".

□ **Année 1964 :**

1. SONATRACH, pour confirmer son acte de naissance, a lancé la construction du premier oléoduc algérien, l'OZ1, d'une longueur de 805 KM, reliant Haoud El Hamra à Arzew.
2. L'Algérie décide de lancer la grande aventure du gaz, en mettant en service le premier complexe de liquéfaction de gaz naturel, dénommé GL4Z (CAMEL - Compagnie Algérienne du Méthane Liquéfié), d'une capacité de traitement de 1.8 milliards m^3 gaz/an.
3. Mise en service de la raffinerie d'Alger. La réalisation de ces infrastructures a permis à l'Algérie d'entrer de plain−pied dans l'industrie des hydrocarbures.

□ **Année 1965 :**

1. Les négociations algéro−françaises relatives au règlement des questions touchant les hydrocarbures et le développement industriel de l'Algérie, ont abouti à la création d'une association coopérative " ASCOOP " entre SOPEFAL, représentant

l'Etat français, et l'Etat Algérien. Cette étape a permis à l'Etat algérien d'élargir considérablement son champ d'activités dans la gestion des hydrocarbures du pays.

2. Lancement de la première campagne sismique de recherche d'hydrocarbures par Sonatrach avec l'implantation de 3 forages.

□ **Année 1966 :**

1. La mise en service de l'Oléoduc OZ1, un ouvrage d'une grande portée stratégique, a permis d'augmenter les capacités de production et d'acheminement de près de 30%.
2. Augmentation du capital de SONATRACH qui passe de 40 à 400 millions de Dinars.
3. Les missions de SONATRACH, qui étaient limitées à la gestion des pipelines et à la commercialisation, sont élargies à la recherche, à la production et à la transformation des hydrocarbures.
4. SONATRACH devient la société nationale de recherche, production, transport, transformation et commercialisation des hydrocarbures et de leurs dérivés.

□ **Année 1967 :**

1. L'Algérie se lance dans un processus de nationalisation des activités de raffinage et de distribution, au terme duquel SONATRACH est à la tête de la distribution des produits pétroliers sur le marché national et inaugure la première station−service aux couleurs de l'entreprise.
2. Première découverte de pétrole à El Borma (Hassi Messaoud Est).
3. Lancement de la construction du nouvel oléoduc Mesdar Skikda.
4. SONATRACH devient majoritaire (à plus de 50%) dans le transport terrestre des hydrocarbures en Algérie, elle crée ses sociétés de services et détient le monopole dans la commercialisation du gaz.
5. SONATRACH se lance aussi dans la réalisation d'une usine d'ammoniac et prévoit la construction d'un complexe de produits pétrochimiques à Skikda et l'aménagement d'un port méthanier.

□ **Année 1968 :**

1. Découverte de gaz à Gassi EL Adem, au sud Est de Hassi Messaoud.
2. SONATRACH est autorisée à transporter des hydrocarbures gazeux en provenance du gisement de Hassi R'mel et des zones productrices algériennes, à travers le gazoduc Hassi R'Mel−Skikda.

SONATRACH évolue comme une société intégrée à la faveur de ses découvertes de pétrole, et devient une société qui détient des réserves en hydrocarbures.

□ **Année 1969 :**

1. L'Algérie devient membre de l'OPEP.

2. Le projet de transport de gaz de pétrole liquéfié (GPL) et de condensat " Hassi Messaoud Arzew ", présenté par SONATRACH, est approuvé par l'Etat. SONATRACH est autorisée à exploiter l'ouvrage.
3. SONATRACH débute les premières opérations d'exploitation pétrolière par ses propres moyens sur le champ d'El BORMA.

□ **Année 1971 :**

24 Février 1971 : Nationalisation des hydrocarbures Une nouvelle ère pour le développement économique du pays.

La nationalisation des hydrocarbures décidée par l'Algérie en Février 1971 place la compagnie nationale des hydrocarbures dans une nouvelle dynamique. Une planification de plus en plus rigoureuse est mise en place, les objectifs de SONATRACH étaient alors l'extension de toutes ses activités à l'ensemble des installations gazières et pétrolières et l'atteinte de la maitrise de toute la chaine des hydrocarbures.

Cette année a été marquée aussi par l'acquisition du premier méthanier baptisé au nom du gisement gazier Hassi R'Mel.

□ **Année 1972 :**

1. Mise en service du complexe de liquéfaction de gaz naturel (GL1K) à Skikda, d'une capacité de production de 6.5 millions m^3/an de GNL, 170000 tonnes/an d'Ethane, 108400 tonnes/an de Propane, 92600 tonnes / an de Butane, 60250 tonnes /an de Gazoline et des postes de chargement de 2 méthaniers d'une capacité de 50000 à 70000 m^3.
2. Mise en service de la raffinerie d'ARZEW, d'une capacité de production de 2400000 tonnes/ an de carburants, 70000 tonnes/an de bitumes, 55000 tonnes/an de lubrifiants et 110000 tonnes/ an de GPL.

□ **Année 1973 :**

Mise en service du complexe de séparation de GPL (GP2Z), d'une capacité de production de 600000 tonnes/ an de GPL.

□ **Année 1974 :**

La capacité de production du gisement de Hassi R'mel a été portée à 14 milliards de m^3 de gaz naturel et 2400000 tonnes de condensat stabilisé.

□ **Année 1975 :**

Découverte du gisement de pétrole de Mereksen.

□ **Année 1976 :**

Mise en service de deux (02) unités de transformation des matières plastiques, une à Sétif et l'autre à Chlef.

□ **Année 1977 :**

Avec la diversification de ses activités (de la recherche à la pétrochimie), la nécessité d'un plan directeur s'est imposée à l'Algérie. Le plan " Valhyd " (Valorisation des Hydrocarbures) est lancé. Il a pour objectif, l'accroissement des taux de production de pétrole et de gaz, la récupération des gaz associés au pétrole pour les réinjecter dans le cadre de la récupération secondaire, la production maximale de GPL et de condensat, la commercialisation du gaz naturel sous ses formes gazeuses et liquides, la substitution de produits finis au brut à l'exportation, la satisfaction des besoins du marché national en produits raffinés, pétrochimiques, engrais et matières plastiques.

Grâce à des investissements massifs, l'Algérie est devenue un grand pays pétrolier exportateur.

□ **Année 1977 :**

1. Mise en service du Module 1 de Hassi R'Mel, avec une capacité de production de 18 milliards m^3/an de gaz et 3 millions de tonnes/ an de condensat.
2. Mise en service du complexe de liquéfaction (GL1Z) à Arzew, d'une capacité de production de $17,5$ millions de m^3/ an de GNL.

□ **Année 1979 :**

1. Mise en service du Module 2 de Hassi R'Mel, avec une capacité de production de 20 milliards m^3/an de gaz, 4 millions de tonnes/ an de condensat et 880000 tonnes/an de GPL.
2. Achèvement des travaux du Module 4 de Hassi R'Mel, avec une capacité de production de 20 milliards m^3/an de gaz, 4 millions de tonnes/ an de condensat et 880000 tonnes/an de GPL.

□ **Année 1980−1985 :**

Durant cette période, l'Algérie a lancé de grands projets économiques qui ont permis la mise en place d'une assise industrielle dense. Ce qui lui a permis de tirer profit de la rente pétrolière dont une bonne partie a été réinvesti dans les projets de développement économique.

SONATRACH s'est engagée selon un plan quinquennal dans un nouveau processus de restructuration étendue, qui a abouti à la création de 17 entreprises.

◇ 4 entreprises industrielles :
- NAFTAL (raffinage et distribution des hydrocarbures).
- ENIP (l'industrie pétrochimique).
- ENPC (industrie du plastique et du caoutchouc).
- ASMIDAL (engrais).

◇ 3 entreprises de réalisation :
- ENGTP (Grands travaux pétroliers).
- ENGCB (Génie−civil et bâtiment).
- ENAC(Canalisation).

◇ 6 entreprises de services pétroliers :
- — ENAGEO (Géophysique).
- — ENAFOR & ENTP (Forage).
- — ENSP (Service aux puits).
- — ENEP (Engineering pétrolier).
- — CERHYD (Centre de recherche en hydrocarbures).

◇ 4 entreprises de gestion des zones industrielles à Arzew, Skikda, Hassi R'mel et HassiMessaoud. Cette restructuration a permis à SONATRACH de se consacrer essentiellement à ses métiers de base. D'une entreprise de 33 personnes en 1963 avec pour objectif principal le transport et la commercialisation des hydrocarbures, à une entreprise de plus de 103300 travailleurs en 1981 avec un domaine d'activité englobant la maitrise de toute la chaine des hydrocarbures.

En 1981, mise en service du complexe de liquéfaction (GL2Z) à Bethioua, d'une capacité de traitement de 13 milliards de m^3/AN.

En 1983, le gazoduc " Enrico Mattei " a été mis en fonction pour alimenter l'Italie et la Slovénie via la Tunisie voisine, avec une capacité dépassant aujourd'hui les 32 milliards de m^3 par an.

□ **Année 1986−1990 :**

Ouverture au partenariat La loi de $86-14$ du 19 août 1986 définissait les nouvelles formes juridiques des activités de prospection, d'exploration, de recherche et de transport d'hydrocarbures permettant à SONATRACH de s'ouvrir au partenariat.

Quatre formes d'associations étaient possibles tout en accordant à SONATRACH le privilège de détenir une participation minimum de 51% :
- — Association " Production Sharing Contracte " (PSC) : contrat de partage de production.
- — Association de " contrat de service ".
- — Association en participation sans personnalité juridique dans laquelle l'associé étranger constitue une société commerciale de droit algérien ayant son siège en Algérie
- — Association en forme de société Commerciale par actions, de droit algérien, ayant son siège social en Algérie.

□ **Année 1986−1990 :**

SONATRACH, Un groupe pétrolier et gazier de renommée international.

Les amendements introduits par la loi 91/01 en décembre 1991, ont permis aux sociétés étrangères activant notamment dans le domaine gazier, la récupération des fonds investis et leur ont accordé une rémunération équitable des efforts consentis. Plus de 130 compagnies pétrolières dont les majors, ont noué contact avec SONATRACH et 26 contrats de recherche et de prospection ont été signés durant les 2 années qui ont suivi le nouveau cadre institutionnel.

Mise en service en 1996 du gazoduc Maghreb Europe appelé " Pedro Duran Farell " qui approvisionne l'Espagne et le Portugal via le Maroc. Sa capacité est de plus

de 11 milliards de m3 de gaz par an.

□ **Année 2000−Aujourd'hui :**

SONATRACH a consenti des efforts considérables : en exploration, développement et exploitation de gisements, en infrastructures d'acheminement des hydrocarbures (gazoducs et stations de compression), en usines de liquéfaction de gaz naturel et en méthaniers. Depuis l'an 2000, plusieurs projets ont été lancés, dans le processus de développement des performances, l'internationalisation, le développement de la pétrochimie et la diversification des activités du groupe SONATRACH, ainsi l'objectif de production primaire fixé pour la période $1999 - 2007$ a été largement dépassé.

Les gisements mis en production durant la période $(99 - 2009)$ par SONATRACH seule ou en association ont assuré la croissance de la production primaire des hydrocarbures qui est passée de 8 millions de tep à 233 millions de tep.

SONATRACH est aujourd'hui devenu un puissant élément d'intégration nationale, de stabilité et de développement économique et social.

1.2 Organisation de la SONATRACH

La Direction Générale du Groupe SONATRACH est assurée par
Monsieur *Toufik HAKKAR*, Président Directeur Général, depuis le 05 février 2020.

— **La Direction Transformation SH2030 (TRF)** est chargée de la coordination et du suivi de la mise en œuvre du plan de transformation de SONATRACH SH2030.
— **La Direction Communication (CMN)** est chargée de l'élaboration et de la mise en œuvre de la stratégie de communication de SONATRACH.
— **La Direction Corporate Stratégie, Planification et Economie (SPE)** est chargée de l'élaboration et le développement à moyen et long terme et d'évaluer leur mise en œuvre.
— **La Direction Corporate Finances (FIN)** est chargée d'élaborer les politiques et stratégies dans le domaine de la Finance. Elle évalue leur mise en œuvre et veille à la qualité de l'information financière.
— **La Direction Corporate Business Development et Marketing (BDM)** est chargée de formuler la stratégie de croissance et de recherche des opportunités d'investissement pour la Société.
— **La Direction Corporate Ressources humaines (RHU)** est chargée de l'élaboration des politiques et stratégies en matière de ressources humaines et du contrôle de leur mise en œuvre.
— **La Direction Centrale Procurement & Logistique (P&L)** a pour mission de piloter les processus d'Achats et la Logistique pour le Groupe.
— **La Direction Centrale Ressources Nouvelles (R&N)** est chargée de piloter et d'exploiter, depuis le centre, les projets de ressources non conventionnelles et l'Offshore.
— **La Direction Centrale Engineering & Project Management (EPM)** assure le pilotage et l'exécution des grands projets industriels du Groupe.

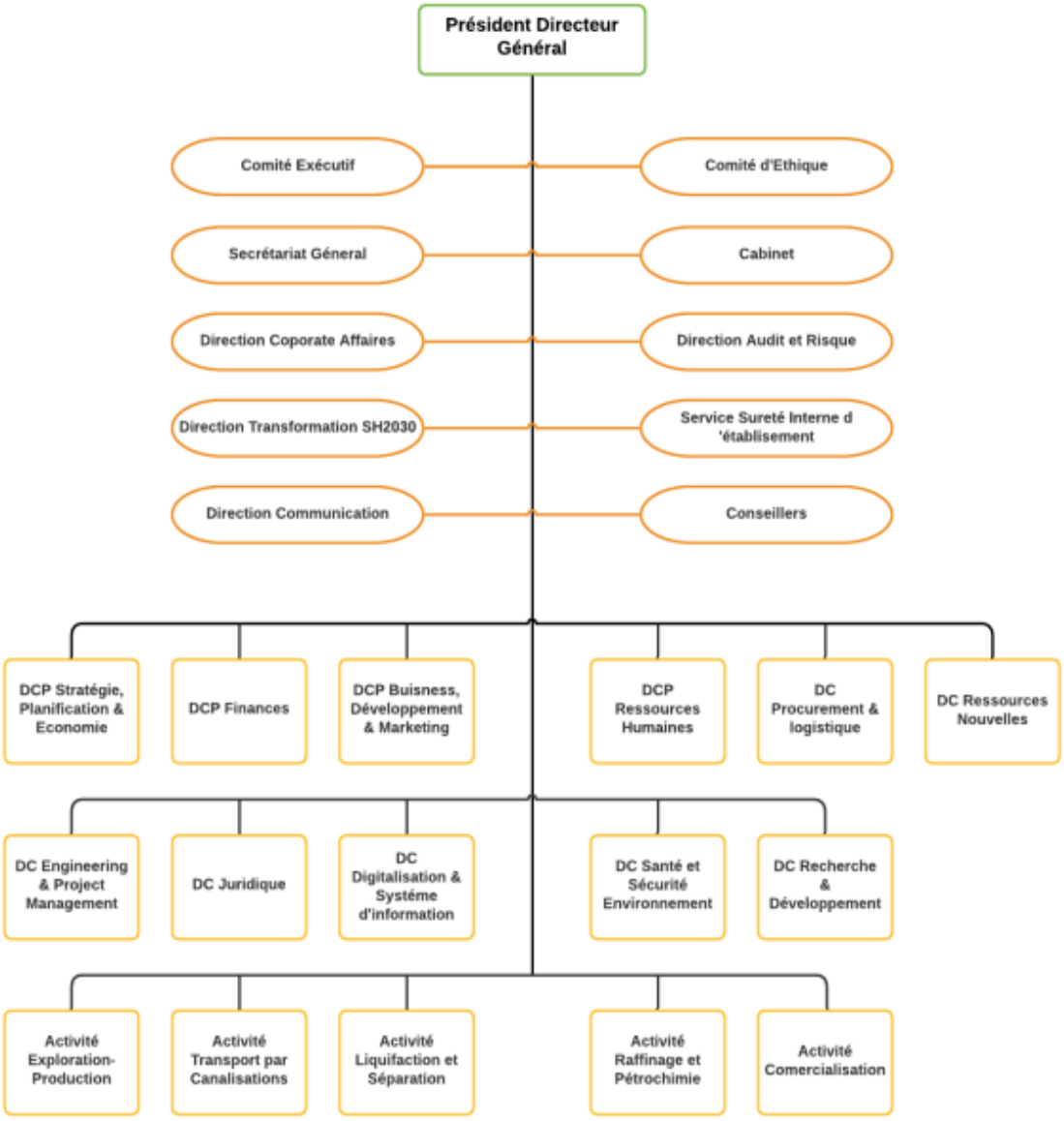


FIGURE 1.2 – Organigramme de SONATRACH

— **La Direction Centrale juridique (JUR)** est en charge de l'élaboration et de l'harmonisation des instruments juridiques et du contrôle de leurs applications.
— **La Direction Centrale Digitalisation et Système d'information (DSI)** est chargée de la définition et du contrôle de la politique informatique et de la digitalisation de la Société.
— **La Direction Centrale Santé, Sécurité et Environnement (HSE)** a en charge l'élaboration des politiques en matière d'environnement, de sécurité et de qualité de vie au travail. Elle assure le contrôle de leur application.
— **La Direction Centrale de la Recherche et du Développement (R&D)** est chargée de promouvoir et de mettre en uvre la politique de la recherche appliquée et développement des technologies dans les métiers de base de la Société.

1.3 Les activités et les missions de SONATRACH :

SONATRACH joue un rôle de leader
dans le développement de l'économie du pays. Ses missions et ses activités sont :

1. **Exploration et production :**
   Pour répondre à la demande d'une population mondiale et la fournir une énergie meilleure, plus abordable et propre, SONATRACH se concentre beaucoup sur les opérations d'exploration, de développement et de production de pétrole et de gaz naturel.
2. **Transport par canalisation :**
   Le TRC est une méthode de transport de matières gazeuses, liquides ou solides constituant généralement un système de transport par canalisation. SONATRACH cherche toujours à développer et améliorer ce système pour garantir le transport des hydrocarbures depuis les sources au sud vers le nord afin d'être transformer ou utiliser.
3. **Liquéfaction et séparation :**
   La liquéfaction et la séparation consistent à liquéfier et séparer les hydrocarbures liquides et gazeux de l'eau et des sédiments afin de produire des nouveaux produits tels que l'Ethan le Propane Butane et la Gazoline ou bien pour purifier un seul produit par exemple le gaz naturel GNL.
4. **Raffinage et pétrochimie :**
   A travers les activités de raffinage et la pétrochimie, le pétrole brut et le gaz naturel sont transformés en produits finis et intermédiaires tels que les carburants, plastiques, résines, etc. Et c'est pour répondre principalement à la demande du marché national en produits pétroliers.
5. **Commercialisation :** La commercialisation consiste à gérer les opérations de livraisons et de vente des produits et des hydrocarbures extraits et traités sur les marchés nationaux et internationaux de SONATRACH.

1.4 Missions et objectifs de la SONATRACH

L'entreprise SONATRACH a pour missions tant en Algérie qu'à l'étranger différentes tâches qu'on peut résumer en :

- La protection, la recherche et l'exploitation d'hydrocarbures solides, liquides et gazeux, ainsi que les substances dérivées, et la maintenance des installations pétrolières.
- Le développement, l'exploitation et la gestion des réseaux de transport, de stockage et de chargement des hydrocarbures.
- Le développement, l'exploitation et la gestion des réseaux de transport, de stockage et de chargement des hydrocarbures.
- La diversification des marchés et des produits à l'exportation.
- Le développement des techniques modernes de gestion par la formation continue de ses cadres.

L'approvisionnement de l'Algérie en hydrocarbures à court, moyen et long terme. Et d'un autre coté Elle a pour objectifs ce qui suit :

- Le renforcement de ses capacités technologiques.
- Le développement international et le partenariat.
- La diversification de son portefeuille d'activité.
- La maitrise continue de ses métiers de base.
- L'approvisionnement du pays en hydrocarbures à moyen et long terme.
- Les prises de participation et autres valeurs mobilières dans toutes société existante ou à créer en Algérie.
- L'étude, la promotion et de la valorisation de toute autre forme et source d'énergie.

Conclusion

L'entreprise SONATRACH que nous venons de présentée est l'entreprise publique par excellence, grâce à ces différentes activités, elle arrive à procurer la plus grande marge en terme de recettes pour l'économie algérienne. Il est nécessaire d'introduire quelques notions de base qui vont nous aider a proposé une solution à la problématique.

2
Réputation numérique

Introduction

La réputation classique était ce qu'on appelle le bouche-à-oreille, ce qui été dit par un proche, un collègue ou un voisin sur une personne ou sur une entreprise. Avec l'avènement du digital, de plus en plus les entreprises essayent de s'adapter à ces nouveautés et changements. On trouve l'apparition de l'E-réputation, c'est la réputation à l'ère du numérique. L'e−réputation peut être positive ou négative sur une personne ou une entreprise. Pour cela les entreprises doivent veiller à ne pas avoir une mauvaise e-réputation et appliquer une stratégie adéquate à toute situation.

Les consommateurs connectés ont changé de comportements aussi depuis l'apparition d'internet. Ils sont ultra connectés, soif d'informations et méfiant aussi. Avant toute prise de décision, il consulte les informations disponibles, les avis des autres consommateurs. Il ne fait plus confiance aux informations émises par les commerciales.

Les médias sociaux :

Définition 2.1. Fred Cavazza, blogueur reconnu dans le domaine des médias sociaux, a publié une définition des médias sociaux très précise :

"Les médias sociaux désignent l'ensemble des services permettant de développer des conversations et des interactions sociales sur internet ou en situation de mobilité".

Ce qui est intéressant avec cette définition des médias sociaux, c'est qu'elle présente bien leurs 3 intérêts majeurs :

— Les médias sociaux permettent d'instaurer un dialogue avec sa communauté, et donc une relation concrète ;

— Les médias sociaux permettent de développer des interactions sociales (" like ", " retweet ", " partage "...) révélant un engagement de la part d'une communauté et entraînant de la virilité ;
— Les médias sociaux permettent de communiquer auprès de sa communauté à tout moment, même en situation de mobilité. A l'heure des smartphones et tablettes, votre cible est perpétuellement connectée et donc toujours susceptible de recevoir des informations.

Fred Cavazza, toujours lui, publie chaque année un panorama très complet des médias sociaux existants :

FIGURE 2.1 – Social média landscape 2021

Distinction entre médias sociaux et réseaux sociaux numériques :

Les technologies des médias sociaux prennent différentes formes telles que des blogs, des réseaux sociaux professionnels, des réseaux sociaux d'entreprise, des projets collaboratifs, des forums, des micro blogs, du partage de photos, de la revue de produits/services, du bookmarking social, du jeu social, des réseaux sociaux, du partage de vidéos et des mondes virtuels.

Les réseaux sociaux numériques (RSN) ne sont qu'une autre sous−partie des médias sociaux. Parmi les médias sociaux, il faut distinguer les outils de publication et de discussion des réseaux sociaux numériques que l'on peut diviser en deux types :

— les RSN de contact pour lesquels les fonctionnalités de mise en relation sont principales.
— les RSN de contenu pour lesquels les fonctionnalités de réseau sont secondaires et sont basées sur une activité particulière.

En 2007, Boyd et Ellison préfèrent parler de " sites de réseaux sociaux " et les définissent comme une plate-forme de communication basée sur le Web qui permet aux individus de :

— disposer de profils associés à une identification unique qui sont créés par une com-

binaison de contenus fournis par l'utilisateur, de contenus fournis par des " amis " et des données du système-exposer publiquement des relations susceptibles d'être visualisées et consultées par d'autres ;

— accéder à des flux de contenus incluant des contenus générés par l'utilisateur (notamment des combinaisons de textes, photos, vidéos, mises à jour de lieux et/ou liens) fournis par leurs contacts sur le site.

Un réseau social permet donc aux utilisateurs d'articuler et de rendre visible leur réseau, que ce soit pour établir de nouvelles connexions ou maintenir des liens existant hors ligne (latents).

En 2009, Thelwall catégorise les réseaux sociaux numériques selon leurs trois objectifs : la socialisation, le réseautage et la navigation (sociale).

L'utilisation des réseaux sociaux dans l'entreprise :

La transformation digitale a une portée considérable tant pour les entreprises et le monde du travail que pour la société dans son ensemble. Cette révolution numérique est si importante que des spécialistes la comparent à la naissance de l'imprimerie il y a plus de cinq siècles. Dans cette nouvelle ère, les canaux numériques se multiplient et leurs usages montent en puissance. Les réseaux sociaux sont devenus des outils incontournables de communication. Les organisations doivent aujourd'hui tirer avantages des opportunités digitales pour développer leur notoriété, leur chiffre d'affaires, adapter leur culture d'entreprise et fidéliser leurs collaborateurs. Augmentation de la visibilité de la marque, partage des actualités, fidélisassions des clients et des employés, recrutement. Facebook, Twitter, LinkedIn... présentent de nombreux atouts pour l'entreprise.

A l'ère du digital, les entreprises ne peuvent plus faire l'impasse sur la stratégie marketing des réseaux sociaux. Diffusion de contenus, promotion des nouveaux produits / services, le social média engage, fédère et rend viral les actions de communication de l'entreprise. Le web, devenu social, permet de créer une nouvelle relation client, plus valorisante et plus intime. Les collaborateurs d'entreprises deviennent des influenceurs, et permettent de promouvoir la marque, de la faire briller. Les barrières entreprises et consommateurs tombent, laissant place au partage des expériences. Avides du web, des nouvelles formes de consommation, grâce à internet, les clients comparent, partagent, recommandent un produit / service et cela modifie les stratégies marketing, de communication et commerciale.

La nouvelle donne induite par les réseaux sociaux s'inscrit pleinement dans la transformation digitale des entreprises et leur utilisation modifie les métiers avec des impacts à plusieurs niveaux :

— **Impact stratégique** : au niveau de la communication, de l'image de marque et de la notoriété ;

— **Impact commercial** : en favorisant la désintermédiation, en générant de nouvelles d'interactions des clients avec l'entreprise ;

— **Impact managérial** : en imposant de nouvelles organisations pour s'adapter à la transversalité du web, en générant de nouvelles attentes de relations entre les sala-

riés (réseaux sociaux d'entreprise), en impactant les liens hiérarchiques (les digital natives plus agiles et sachant que les anciens) ;
— **Impact RH** : en apportant de nouvelles techniques de recrutement et des attentes renouvelées des candidats.

Généralités sur la réputation numérique :

Définitions de la réputation numérique :

Définition 2.2. La réputation numérique, parfois appelée web-réputation, cyber-réputation, e-réputation, sur le Web, sur Internet ou en ligne, est la réputation, l'opinion commune (informations, avis, échanges, commentaires, rumeurs) sur le Web d'une entité (marque), personne morale (entreprise) ou physique (particulier), réelle (représentée par un nom ou un pseudonyme) ou imaginaire. Elle correspond à l'identité de cette marque ou de cette personne associée à la perception que les internautes s'en font.

Cette notoriété numérique, qui peut constituer un facteur de différenciation et présenter un avantage concurrentiel dans le cas des marques, se façonne par la mise en place d'éléments positifs et la surveillance des éléments négatifs. L'e-réputation peut aussi désigner sa gestion, via une stratégie globale et grâce à des outils spécifiques (activité à l'origine de nouveaux métiers) pour la pérennité de l'identité numérique.

Définition 2.3. L'E−réputation renvoie à : " L'ensemble des informations qu'il est possible de trouver sur une personne sur Internet, que ce soit via les moteurs de recherche, sites, blogs, réseaux sociaux, forums, messageries instantanées ou par simples courriers électroniques. Cette réputation numérique se construit de façon individuelle et volontaire, via l'information qu'on décide de publier en ligne, mais également de façon indirecte, par ce qui peut être publié par autrui sur nous."[8].

Définition 2.4. L'E-réputation est un terme récent qui synthétise tout ce qui touche à la réputation additionnée du " e " d'Internet. La réputation est vieille comme le monde. Le " e " d'Internet représente la modernité, un moyen d'échange entre personnes sur les réseaux sociaux. De tout temps, la réputation a été l'expression de l'opinion du public envers une personne, un groupe, ou une organisation[21].

En résumé, la réputation en ligne d'un établissement fait référence à toutes les informations qu'un client peut trouver sur Internet. Toutes les publications sur les réseaux sociaux, les sites de partenaires de réservation, les sites d'avis, etc sont concernées. Par contre l'E-réputation ne dépend pas que d'un contenu online : elle est également assujettie aux actions " offline ", qui veut dire la vie réelle, et ce concept s'applique aussi bien à l'individu qu'à une organisation.

Les concepts voisins de l'E-réputation :

Lorsque l'on parle d'E-réputation il y a trois (03) termes susceptibles de se confondre. Il est important de bien faire la différence entre l'E-réputation, l'image de marque et l'identité numérique.

1. **L'identité numérique de l'entreprise :**
   L'identité numérique est un lien technologique entre une entité réelle (personne, organisme ou entreprise) et des entités virtuelles (sa ou ses représentation(s) numériques).

   L'identité numérique peut être définit aussi comme " l'ensemble des contributions et des traces qu'une entreprise (ou une personne) laisse en ligne, volontairement ou non "[26].

   Elle est considérée comme une sorte de carte d'identité virtuelle et qui se forme à partir de deux grands types de données : des données formelles provenant d'organismes officiels : sites institutionnels (administrations, entreprise, etc.), des données informelles à l'origine de contributions volontaires : publication sur les blogs, les réseaux sociaux, les forums, etc.

   L'identité numérique est finalement assez comparable à l'identité traditionnelle dans son aspect multidimensionnel. Simplement, en raison de sa nature digitale, elle est Caractérisée par deux groupes d'informations distincts mais complémentaires :

   . *Les informations ou données incontestables et uniques* : coordonnées physiques, adresse IP, certificats numériques, comptes bancaires, numéro de téléphone, etc. Elles sont généralement attribuées par une autorité tierce (état civil, fournisseur d'accès Internet, opérateur de télécommunications, banque, etc.) ;

   . *Les informations réputées plus ambiguës et multiples* : pseudonymes, avatars, commentaires, blogs, photos, CV, etc., qui sont générées par l'individu lui-même ou par les individus composant son réseau.

   L'identité précède l'image. Il convient donc maintenant d'expliquer ce concept.

2. **Image de marque :**
   Le concept d'image de marque à fait l'objet de nombreuses recherches relativement récentes, de nombreuses définitions ont vu le jour restant tout de même cohérentes les unes avec les autres :

   Selon **LENDREVIE** et **LEVY** : " Une image de marque est un ensemble de représentations mentales, subjectives, stables, sélectives et simplificatrices à l'égard d'une marque."[22].

   Et **Jean−Jacques Lambin** apporte une définition précise de l'image de marque. Pour lui, il s'agit de " l'ensemble des représentations mentales, cognitives et affectives, qu'une personne ou un groupe de personnes se font d'une marque. " Il dégage trois (03) niveaux d'image de marque :

   . L'image perçue : c'est-à-dire la manière dont le segment cible (le public visé, sur lequel on projette l'image) voit et perçoit la marque ;

   . L'image vraie ou réalité de la marque avec ses forces et ses faiblesses, telle qu'elle est connue et ressentie par l'entreprise ;

   . L'image voulue : c'est la manière dont l'entreprise souhaite être perçue par le segment cible et qui résulte d'une décision de positionnement.

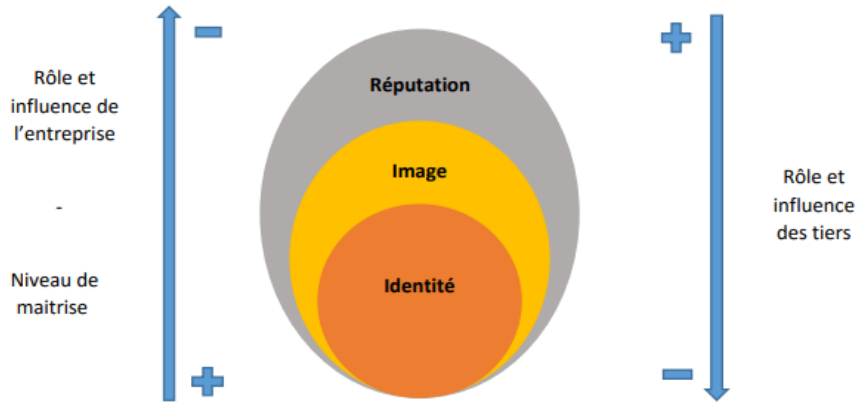


FIGURE 2.2 – Identité, Image, Réputation

L'image se situe quelque part entre l'identité et la réputation il s'agit de la façon dont est perçue une entreprise ou une marque : l'image se bâtit à l'intérieur de l'entreprise mais se dessine à l'extérieur. La réputation d'une organisation peut donc être associée à son " image perçue " et l'E-réputation à l'image que les internautes se font d'une organisation.

L'importance de l'E-réputation :

On ne doit surtout pas négliger l'importance de l'E−réputation. On va voir un bref aperçu sur l'importance de l'e-réputation pour une entreprise tout comme pour une personne :

1. **L'importance de l'E-réputation pour une entreprise ou une marque :**
   Internet a conquis tous les secteurs et il n'existe pas une seule société qui n'ait pas de présence digitale (site web, comptes sur les réseaux sociaux, avis, etc.). Les achats en ligne étant devenus une pratique extrêmement courante, les consommateurs ont de plus en plus tendance à vérifier l'opinion que les autres ont d'un produit avant de l'acheter. Selon une étude réalisée par l'IFOP pour Réputation VIP, sur internet, 85% des consommateurs réalisent des achats et 80% se renseignent avant d'acheter. Une mauvaise réputation sur le web peut désormais être synonyme de pertes colossales. En effet, toujours selon l'étude IFOP, 66% des consommateurs venus chercher un avis avant un achat diffèrent l'achat en cas de commentaires défavorables quand, dans 30% des cas, ils vont même jusqu'à renoncer à l'achat. Ainsi, 96% des internautes sont influencés par l'e−réputation d'une marque lors d'un achat.
   Un grand nombre d'entreprises sont aujourd'hui dotés d'un community manager, dont la fonction principale est de soigner l'image de la marque auprès des internautes. Un tweet maladroit ou une publication à charge sur Facebook, Twitter ou autre, peut faire perdre des milliers de clients en quelques heures et faire plonger le titre d'une entreprise en bourse.

Les entreprises ont désormais tout intérêt à soigner leur E-réputation si elles désirent à la fois garder leurs clients et en conquérir de nouveaux.

2. **L'importance de l'E-réputation pour une personne :**
   Que ce soit les " people ", les hommes politiques, les dirigeants, ou encore une personne en recherche d'emploi... tout le monde a intérêt à soigner sa réputation numérique. Certaines personnes ont vu leur destin basculer en postant de simples vidéos sur YouTube. Des hommes politiques ont dû faire des excuses publiques après des propos polémiques sur les réseaux sociaux. D'autres ont gagné des élections en peaufinant leur image auprès des internautes. Sur les réseaux sociaux, les comptes des hautes personnalités sont scrutés par les journalistes et les internautes. Nul ne peut désormais se permettre d'écorner son e-réputation sans en pâtir de façon directe.
   Bien valorisée, l'E-réputation peut avoir un impact important sur la vie réelle. Mal gérée, elle peut être source de cataclysmes.

2.1 Etablir une base

2.1.1 Comment établir un politique d'entreprise en matière de médias sociaux

Les réseaux sociaux font désormais partie du quotidien, et ce, au travail comme à la maison. Ces nouvelles technologies transforment le monde du travail. Les médias sociaux tels que LinkedIn, Facebook, Twitter, Pinterest, YouTube et autres sont de puissants leviers pour les entreprises du commerce de détail. En effet, ceux-ci contribuent à améliorer l'image de marque et la réputation en ligne de celles-ci, à recruter et à fidéliser des employés en plus d'être davantage à l'écoute de la clientèle. En contrepartie, bien que ces médias sociaux soient très présents, il n'est guère étonnant de constater que les employés peuvent avoir de la difficulté à tracer la ligne entre la vie personnelle et la vie professionnelle. Il est donc important, en tant qu'entreprise, de se doter d'une politique d'utilisation des réseaux sociaux afin de préserver le bien-être de tous.

Champ d'application

La présente politique d'utilisation des réseaux sociaux s'applique à tous les employés de l'entreprise. Elle s'applique aussi à toutes les publications et les commentaires qui impliquent directement ou indirectement l'organisation et/ou un collègue, et ce, quel que soit le médium employé (Facebook, Twitter, LinkedIn, Pinterest, YouTube, Instagram, blogue, etc.). Également, cette politique concerne l'utilisation des appareils mobiles et des cellulaires durant les quarts de travail.

Facebook est un réseau social qui permet aux usagers de publier du contenu et d'échanger des messages.

Twitter est un réseau social de microblogage qui permet d'envoyer des messages de 140 caractères tout en y rattachant des liens d'articles, des vidéos, des photos, etc.
LinkedIn est un réseau social professionnel permettant aux usagers de mettre leur CV en ligne. Cette plateforme offre également la possibilité aux entreprises de faire du recrutement 2.0.
Pinterest permet aux utilisateurs de démontrer leurs intérêts ainsi que de partager leurs trouvailles (photographies) avec d'autres usagers.
YouTube est un site web qui offre la possibilité aux utilisateurs de publier, de visionner et de partager des vidéos.
Instagram est un réseau social qui permet de partager des photographies et des vidéos entre utilisateurs en plus de pouvoir commenter les publications de ceux-ci.

Les différentes stratégies de l'E-réputation :

Cela peut se faire sur plusieurs volets :

1. **Déployer une stratégie " web social " :**
   Le Web social désigne avant tout la priorité à l'échange avec les internautes sur les espaces maitrisés par la marque mais aussi sur tous les autres espaces du web où la marque sera citée. Pour cela il est important que les entreprises mettent en place des Community-Managers qui seront capables d'échanger avec les internautes, d'animer les communautés, d'anticiper et de promouvoir la marque.
   La stratégie Web Social doit pouvoir s'adapter aux objectifs de communication de l'entreprise. Il ne suffit pas d'être présent sur les médias sociaux, il faut définir des objectifs précis comme la fidélisation client.
   Une notion importante est celle de la création de valeur. Il faut pouvoir se placer en tant qu'expert pour être vecteur de création de valeur et attirer les consommateurs. Il est également primordial que la marque connaisse les usages des membres d'une communauté.
2. **Trouver des mots clés pertinents :**
   L'étape du choix des mots clés est indispensable avant toute création de contenu. Il faut trouver les meilleurs mots clés, liés à votre secteur afin d'être trouvable par les internautes. Ces mots clés seront ensuite utilisés sur votre site, mais aussi dans les différents contenus que vous produirez sur la toile.
   Sachez que votre manière de chercher une information, n'est peut-être pas celle employée par la majorité des internautes. Vos mots clés ne sont peut-être pas ceux recherchés par les internautes. Il faut donc identifier les tendances, pour cela il est donc très utile d'avoir recours aux deux outils Google pour identifier les mots clés les plus pertinents :
   - *Google Tendances des recherches* : Cet outil est extrêmement efficace puisqu'il permet de comparer le taux de recherche entre plusieurs mots clés. Google Tendances de recherches montre également l'évolution des recherches dans le temps.

Les résultats peuvent être affinés à l'aide de critères temporels, géographiques et linguistiques ;

. *Google Adwords* : Google Adwords est un générateur de mots clés. A partir de la saisie d'un mot clé ou de l'adresse d'un site web, et de critères géographiques et linguistiques, Google va générer une liste de mots clés associés à la recherche. Le résultat de cette génération automatique de mots clés présentera en détail la concurrence sur un mot clé, ainsi que le volume de recherches mensuelles. On obtiendrait ainsi l'état concurrentiel des mots clés, c'est-à-dire si celui-ci est largement utilisé ou non. A savoir que plus un mot clé présente une concurrence élevée plus il sera difficile de se positionner sur ce mot clé en question.

3. **Optimiser le référencement** :

Dans une stratégie d'E-réputation, il est nécessaire d'optimiser sa présence sur son propre nom de marque afin de contrôler les contenus arrivant en premières positions lorsqu'un internaute recherche sa marque.

Il est donc du devoir des web-marqueteurs d'une entreprise de connaitre les techniques de référencement qui permettront de positionner les contenus qu'ils souhaitent dans les premières pages de résultats.

Il existe dix (10) techniques pour occuper la première page de résultat :

* *Optimiser le référencement " on-page " du site de l'entreprise*, cela se traduit par l'utilisation de balises html : Ces balises indiquent à Google les éléments importants du site de l'entreprise et permettent donc d'influencer son référencement ;
* *Associer un blog marketing à son site web.* Etant donné que les blogs sont mis à jour régulièrement, ceux-ci sont favorisés par Google et gagnent très vite les premières pages de résultats ;
* *Ouvrir des sous domaines ou des sites complémentaires* pour les espaces auxquels l'entreprise accorde de l'importance ;
* *Rédiger un article sur Wikipédia.* Ce wiki est généralement bien positionné par les moteurs de recherche. Attention toutefois à bien respecter les règles d'admissibilité sous peine de voir son article refusé ;
* *Utiliser de manière adaptée les réseaux sociaux* tels que Facebook, Twitter, Viadeo, LinkedIn. Ces réseaux se positionnent généralement dans les premiers résultats des moteurs de recherches ;
* *Prendre part à des échanges au sein de forums* concernant l'entreprise (au nom de l'entreprise bien sûr) ;
* *Adopter une stratégie de Netlinking*, c'est-à-dire obtenir des articles de sites tiers assurant la promotion de l'entreprise et pointant un lien vers le site de l'entreprise et ses différents contenus ;
* *Rédiger des communiqués de presse* afin de les publier sur des sites spécialisés ;
* *Inscrire le site de l'entreprise dans des annuaires d'entreprises* ;

* *Soumettre les meilleurs articles de l'entreprise aux DiggLike les plus fréquentés.* Les DiggLike sont des sites communautaires ou les internautes peuvent voter pour les articles soumis.

4. **Investir les réseaux sociaux et plateformes de contenus :**

* *Réserver son propre identifiant* : Il est primordial pour une marque d'identifier les noms de comptes qu'elles souhaitent détenir (nom de la marque, nom des produits, nom des personnalités) et de les réserver avant qu'ils ne soient ouverts par des internautes étrangers à la marque ;
* *Réseaux sociaux* : Ce regroupement d'internautes sur les réseaux sociaux nationaux et internationaux, présente pour les marques un moyen d'asseoir leur popularité et l'opportunité d'être visibles auprès d'un grand nombre d'internautes. La présence sur les réseaux sociaux est aujourd'hui indispensable dans une stratégie d'acquisition de trafic et de référencement (Social Media Optimisation) ;
* *Blogs de marque* : Pour une société, la création d'un blog officiel peut être pertinente, que ce soit pour améliorer la communication avec les internautes mais également pour contrôler son positionnement dans les moteurs de recherche. Un blog possède la qualité de bien se positionner sur les moteurs de recherche car celui-ci est alimenté plus régulièrement qu'un site institutionnel, lui permettant d'obtenir un Page Rank important plus rapidement. Dans l'esprit du Web 2.0, le blog de marque peut même permettre une communication moins officielle, plus libre et interactive qu'un site Internet traditionnel ;
* *Les forums* : Les forums sont des lieux d'échanges entre internautes partageant des centres d'intérêts communs. De nombreuses études ont démontré que les internautes se fient davantage aux avis laissés sur les forums par les internautes que par les dires de la marque. En aucun cas une société ne doit se faire passer pour un internaute en publiant un commentaire sur un forum. Les messages publiés par les marques sur les forums sont facilement repérables et ces pratiques sont loin d'être du goût des internautes. Cependant pour les marques, il reste possible d'intervenir sur les forums dans le but d'aider la communauté et de fournir des informations pertinentes répondant aux questions des membres du forum.

  Il est indispensable de bien la gérer et mettre en place différentes stratégies pour optimiser son influence sur son image.

2.1.2 La socialisation des entreprises : des relations publiques aux relations humaines

Relations humaines et relations publiques sont des termes très fréquemment rencontrés dans le monde de l'entreprise. Les deux sont utilisés par une organisation pour maximiser le retour sur investissement.

RH signifie ressources humaines et concerne les travailleurs ou les employés d'une organisation, bien qu'il soit désormais fait référence au potentiel humain de toute une nation. Les relations publiques manquent dans les relations publiques et il s'agit d'utiliser efficacement les politiques et les stratégies pour créer une bonne image de la société auprès de la population. Il existe des différences entre les deux termes.

Comme son nom l'indique, les ressources humaines traitent l'être humain comme une ressource, au même titre que la matière première. Les politiques et stratégies de plans de gestion visent à accroître l'efficacité de cette ressource de manière à générer plus de profits pour l'organisation. Ceci est également connu sous le nom de gestion humaine ou humaine qui essaie d'augmenter la productivité des employés en répondant à leurs besoins et en élaborant des plans pour veiller à leur bien-être. Des employés heureux et satisfaits sont un atout pour toute entreprise et les résultats sont visibles pour tous en termes d'augmentation de la productivité, entraînant en définitive une production accrue...

Entretenir de bonnes relations avec les personnes extérieures à l'organisation, en particulier la presse et les médias, est aujourd'hui une fonction importante pour toute entreprise. La relation publique est un vaste sujet qui englobe la projection des travaux réalisés par l'organisation dans le domaine de la protection sociale afin de créer une image favorable de la société dans l'esprit des gens. Les relations publiques constituent en effet un moyen de maintenir un dialogue ouvert avec le monde extérieur au moyen de communiqués de presse, de campagnes médiatiques et de publicités afin de rester à la vue du public. L'image aujourd'hui est très importante pour toute entreprise et aucun moyen n'est épargné pour atteindre cet objectif.

2.2 Veille digitale

2.2.1 Outils de surveillance des médias sociaux

Il existe des outils spécialisés et gratuits pour la plupart ; grâce à eux, la veille se fait de façon automatique. Ces logiciels permettent de scanner, de façon très rapide et complète, la totalité des informations disponibles sur les médias sociaux et le web en général.

Se priver d'une telle mine d'informations reviendrait à avancer à l'aveugle, alors que de nombreuses informations peuvent aider l'entreprise à s'améliorer, à corriger ses erreurs et à mieux comprendre le consommateur et l'environnement global. Il ne s'agit pas ici de présenter tous les outils existants mais de proposer les plus pertinents pour un suivi efficace de l'E-réputation d'une entité.

. *Les recherches dans les moteurs de recherche* (Google, Bing, etc.) qui se font à l'aide de mots-clés ou d'association de mots-clés. Pour que la recherche soit efficace, il ne faut pas hésiter à décliner et multiplier les mots-clés et les associations de mots-clés en prenant en compte les fautes d'orthographe ;

- *Alerti* est un outil avec un bon rapport performances/prix, qui permet de réaliser des recherches par mots-clés, sur une sélection de réseaux sociaux. Des bilans automatiques sont générés à des fréquences réglable ;
- *Kurrently* permet de suivre en temps réel ce qui se dit sur une marque sur les principaux réseaux sociaux : Facebook, Google + et Twitter. Cela donne une très bonne visibilité du bruit généré autour d'une marque à un instant donné ;
- *Howsociable* donne une note entre 0 et 10 qui indiquent le niveau d'activité autour d'une marque sur une semaine donnée. La version gratuite ne donne accès qu'à un certain nombre de réseaux sociaux ;
- *Socialmention* va, quant à lui, un peu plus loin en proposant un mix entre une analyse en temps réel et une analyse qualitative des résultats du bruit ;
- *Mention* est une plateforme qui, selon les mots-clés que vous avez saisis et décidé d'analyser (nom de votre entreprise, produit, opération), repère les conversations sociales sur l'ensemble des contenus publics diffusés sur les médias sociaux et restitue ces contenus pour permettre un accès quotidien et facile à ces discussions ;
- *Netvibes* est un tableau de bord complet qui permet de gérer plusieurs marques en même temps, de réaliser une veille permanente très efficace et de traquer toutes les conversations sociales des consommateurs et prospects sur les médias sociaux ;
- *Synthesio, Radian.*6*, Social Bro* sont des outils premium et coûteux mais très efficaces et complets, etc.

Conclusion

Les marques ont une multitude de moyen d'être présente sur Internet. Cependant elles doivent avant tout établir de vraies stratégies « Social Media » et réfléchir à l'image qu'elles souhaitent dégager et entretenir auprès de leur communauté. A l'heure où les internautes accordent une grande importance aux avis de leurs pairs pour leur décision, les entreprises se doivent de veiller à ce qui se dit à leur sujet afin de pouvoir réagir. L'E-réputation se fabrique et s'entretient sur le long terme, il est donc primordial qu'un poste de l'entreprise soit dédié à l'E-réputation.

Non seulement l'E-réputation exerce une influence significative sur la satisfaction du consommateur mais aussi sur sa fidélité à l'organisation. Face à cette prise de conscience autour de ce concept, il est certain que le marché de l'e-réputation se développera de manière importante dans les années à venir.

3
L'analyse des sentiments

3.1 Introduction

L'analyse des sentiments et l'Opinion Mining est un domaine très populaire pour analyser et trouver des informations à partir de données textuelles provenant de diverses sources telles que Facebook, Twitter et Amazon, etc. Elle joue un rôle essentiel en permettant aux entreprises de travailler activement sur l'amélioration de la stratégie commerciale et d'obtenir un aperçu approfondi des commentaires des acheteurs sur leur produit. Elle implique l'étude computationnelle du comportement d'un individu en termes d'intérêt d'achat, d'extraction de données et d'exploitation de ses opinions sur une entité commerciale d'une entreprise. Cette entité peut être visualisée comme un événement, un individu, un article de blog ou une expérience de produit.

L'analyse des sentiments fait appel à l'étude de l'analyse des textes, du traitement du langage naturel et de la linguistique informatique pour identifier, extraire et étudier scientifiquement les informations subjectives des données textuelles. Le sentiment ou l'opinion est l'attitude des clients provenant des avis, des réponses aux enquêtes, des médias sociaux en ligne, etc. La signification générale de l'analyse des sentiments est de déterminer l'insolence d'un orateur, d'un écrivain ou d'un autre sujet par rapport à un thème particulier ou à une polarité contextuelle d'un événement spécifique, d'une discussion, d'un forum, d'une interaction ou de tout autre document, etc.

La polarité d'une opinion exprime la positivité, la négativité ou une information de cette dernière. On dit d'une opinion positive qu'elle possède une polarité positive, et inversement, on dit d'une opinion négative qu'elle possède une polarité négative ou neutre possède une information.

La tâche essentielle de l'analyse des sentiments est de déterminer cette polarité d'un

texte donné au niveau de la caractéristique, de la phrase et du document. En raison de l'augmentation de l'utilisation d'Internet, chaque utilisateur est intéressé à mettre son opinion sur le Web par le biais de différents médias et les résultats des données d'opinions générés par cet opinion sur la toile. L'analyse des sentiments aide à analyser ces données d'opinions et d'en extraire des informations importantes qui aideront les autres utilisateurs à prendre une décision. Les données des médias sociaux peuvent être de différents types comme critiques de produits, critiques de films, critiques de compagnies aériennes, critiques d'hôtels, l'interaction avec les employés, les revues de santé, les nouvelles et les articles, etc.

Dans ce deuxième chapitre nous allons présenter quelques définitions du domaine d'études qu'est l'analyse des sentiments, ses caractéristiques, ses difficultés, les problèmes liés à ce domaine et le Machine Learning et ses types.

3.2 Analyses des sentiments :

L'analyse des sentiments est souvent désignée sous le nom d'extraction d'opinion, car l'opinion recueillie auprès du client sera extraite pour révéler la note du produit ensuite elle sera exploitée pour révéler l'évaluation du produit. Elle fait partie du Machine Learning. Étant donné que les données en ligne augmentent considérablement de jour en jour, elle est considérée comme très importante dans la situation actuelle, car de nombreux textes contenant l'opinion des utilisateurs sont disponibles sur le Web. L'analyse des sentiments est considérée comme l'étude des pensées et des sentiments des utilisateurs à l'égard d'un produit. Les deux termes **SA** (**S**entiment **A**nalysis) et **OM** (**O**pinion **M**ining) sont interchangeables.

L'importance de l'analyse des sentiments ou de l'extraction d'opinions augmente chaque jour, car les données s'accroissent de jour en jour. Les machines doivent être fiables et efficaces pour interpréter et comprendre les émotions et les sentiments humains.[25]

L'analyse des sentiments est un domaine multidisciplinaire, qui englobe la psychologie, la sociologie, le traitement du langage naturel et le Machine Learning. Récemment, la croissance exponentielle des quantités de données et de la puissance de calcul a permis de mettre en place des formes d'analyse plus avancées. Le Machine Learning est donc devenu un outil dominant pour l'analyse des sentiments. Il existe une abondance de littérature scientifique sur l'analyse des sentiments et plusieurs études secondaires ont été menées sur le sujet.[5]

3.2.1 Caractéristiques :

L'analyse des sentiments est un domaine de recherche vaste et complexe. Dans ce qui suit, les principales caractéristiques qui constituent le processus d'analyse des sentiments sont décrites et discutées en détail.

A) Catégorisation des sentiments : *Phrases objectives versus phrases subjectives*

Le premier objectif de l'analyse des sentiments consiste généralement à distinguer les phrases subjectives des phrases objectives. Si une phrase donnée est classée comme objective, aucune autre tâche fondamentale n'est requise, alors que si celle-ci est classée comme subjective, sa polarité (positive, négative ou neutre) doit être estimée. La classification de la subjectivité [12] est la tâche qui distingue les phrases qui expriment des informations objectives des phrases qui expriment des opinions et des avis subjectives.

Un exemple de phrase objective est L'iPhone est un smartphone, tandis qu'un exemple de phrase subjective est L'iPhone est génial. La classification de polarité est la tâche qui distingue les phrases qui expriment des polarités positives, négatives ou neutres. Notant qu'une phrase subjective peut ne pas exprimer un sentiment positif ou négatif (par exemple, Je suppose qu'il est arrivé), pour cette raison, il doit être classé comme neutre.

B) Niveaux d'analyse :

Comme mentionné précédemment, le but de l'analyse des sentiments est de définir des outils automatiques capables d'extraire des informations subjectives à partir des textes en langage naturel. Le premier choix lorsqu'on applique l'analyse des sentiments est de définir ce que signifie le texte (c'est-à-dire l'objet analysé) dans le cas d'étude considéré.

En général, l'analyse des sentiments dans la réputation numérique peut être étudiée principalement à trois niveau :

* *Niveau de texte* : L'objectif est de détecter la polarité d'un texte d'opinion. Par exemple, dans le cadre d'une revue de produit, le système détermine si le texte exprime une opinion globale positive, négative ou neutre au sujet du produit. L'hypothèse est que l'ensemble du texte n'exprime qu'une seule opinion sur une seule entité (un seul produit).
* *Niveau phrase* : Le but est de déterminer la polarité de chaque phrase contenue dans un texte. L'hypothèse est que chaque phrase, dans un texte donné, désigne une seule opinion sur une seule entité.
* *Niveau d'entité et d'aspect* : Effectue une analyse plus fine que le niveau du texte et de la phrase. Elle repose sur l'idée qu'une opinion est constituée d'un sentiment et d'une cible (d'opinion). Par exemple, la phrase L'iPhone est très bien, mais ils ont encore besoin de travailler sur la durée de vie de la batterie et la sécurité évalue trois aspects : iPhone (neutre), la durée de vie de la batterie (négatif) et la sécurité (négatif).

C) Opinion régulière versus opinion comparative :

Une opinion peut prendre différentes nuances et peut être assignée à l'un des groupes suivants :

* *Opinion régulière* : Une opinion régulière est souvent désignée dans la littérature comme une opinion standard, elle a deux sous-types principaux :

— Opinion directe : Une opinion directe fait référence à une opinion exprimée directement sur une entité (par exemple, La luminosité de l'écran de l'iPhone est impressionnante).
— Opinion indirecte : Une opinion indirecte est une opinion qui est exprimée indirectement sur une entité sur la base de ses effets sur d'autres entités. Par exemple, la phrase Après être passé à l'iPhone, j'ai perdu toutes mes données décrit un effet indésirable du passage à l'iPhone sur les données, ce qui donne indirectement un sentiment négatif à l'iPhone.

* *Opinion comparative* : Une opinion comparative exprime une relation de similitude ou de différence entre deux ou plusieurs entités et/ou une préférence du détenteur d'opinion basée sur certains aspects communs des entités. Par exemple, les phrases iOS est plus performant qu'Android et iOS est le système d'exploitation le plus performant expriment deux opinions comparatives[3]. Une opinion comparative est habituellement exprimée en utilisant la forme comparative ou superlative d'un adjectif ou d'un adverbe.

D) Opinions explicites versus opinions implicites :

Parmi les différentes nuances qu'une opinion peut prendre, nous distinguons les opinions explicites et implicites :

* *Opinion explicite* : Une opinion explicite est une déclaration subjective qui donne une opinion régulière ou comparative (par exemple, La luminosité de l'écran de l'iPhone est impressionnante).
* *Opinion implicite* : Une opinion implicite est un énoncé objectif qui implique une opinion régulière ou comparative qui exprime habituellement un fait désirable ou indésirable (par exemple, " Samedi soir, j'irai au cinéma pour regarder 'I am legend'. J'ai hâte de le regarder ! " Et " 'Saving Private Ryan' est plus violent que 'I am legend' "). Le premier exemple suggère qu'il y a de bonnes attentes à propos du film, bien qu'il ne soit pas expliqué en mots, alors que la compréhension de l'opinion cachée dans le second exemple est difficile même pour les humains. Pour certaines personnes, la violence dans les films de guerre pourrait être une bonne caractéristique qui rend le film plus réaliste, alors qu'elle pourrait être une caractéristique négative pour d'autres.

  Il est clair que les opinions explicites sont plus faciles à détecter et à classer que les opinions implicites. Une grande partie de la recherche actuelle s'est concentrée sur des opinions explicites. Relativement moins de travail a été fait sur les opinions implicites.

3.2.2 Disciplines en relation avec l'analyse des sentiments :

Plusieurs disciplines ont une relation directe ou moins directe avec l'analyse des sentiments, l'opinion mining, l'intelligence artificielle, le traitement automatique du langage naturel, le texte mining et même le data mining offrent des outils et algorithmes indispensables pour le traitement et la classification des sentiments.

Fouille de texte

La fouille de texte ou Text mining est l'analyse des données contenues dans un texte en langage naturel. L'application de techniques d'exploration de texte pour résoudre des problèmes métier est appelée analyse de texte.

Le Text Mining est un domaine passionnant qui englobe de nouvelles méthodes de recherche et des outils logiciels qui sont utilisés dans le milieu universitaire ainsi que par des entreprises et des organismes gouvernementaux. Aujourd'hui, les chercheurs utilisent les outils d'exploration de texte dans des projets ambitieux pour tenter de prédire tout, de la direction des marchés boursiers à l'occurrence de protestations politiques.

L'exploration de texte est également couramment utilisée dans la recherche marketing et de nombreuses autres applications commerciales, ainsi que dans le travail du gouvernement et de la défense[18].

Les processus de fouille de textes comprennent généralement la recherche d'informations (méthodes d'acquisition de textes) et des applications de méthodes statistiques avancées et de traitement du langage naturel (TLN) telles que le marquage des parties de la parole et l'analyse syntaxique. L'exploration de texte comprend aussi souvent la reconnaissance d'entités nommées (REN), qui est l'utilisation de techniques statistiques pour identifier les caractéristiques de textes nommés tels que les personnes, les organisations et les noms de lieux ; la désambiguïsation, qui est l'utilisation d'indices contextuels pour décider où les mots se réfèrent à l'une ou l'autre de leurs multiples significations et l'analyse des sentiments, qui implique de discerner le matériel subjectif et d'extraire des informations attitudinales telles que le sentiment, l'opinion, l'humeur et l'émotion[20].

Traitement automatique du langage naturel (TALN) :

Le Deep Learning et le Machine Learning continuent de proliférer dans diverses industries, et a révolutionné le sujet abordé dans ce titre : Le traitement du langage naturel (TLN). Le TLN est un sous-domaine de l'informatique qui se concentre sur la possibilité pour les ordinateurs de comprendre un langage d'une manière naturelle, comme le font les humains. En général, il s'agit de tâches telles que la compréhension du sentiment d'un texte, la reconnaissance vocale et la génération de réponses à des questions.

Le TLN est devenu un domaine en évolution rapide, dont les applications ont représenté une grande partie en intelligence artificielle (IA). Quelques exemples d'applications utilisant le Deep Learning sont les Chatbots qui traitent les demandes du service clientèle, la vérification automatique de l'orthographe sur les téléphones portables et les assistants d'intelligence artificielle tels que Cortana et Siri, sur les smartphones.

Pour ceux qui ont de l'expérience en Machine Learning et en Deep Learning, le traitement du langage naturel est l'un des domaines les plus passionnants pour les individus qui souhaitent appliquer leurs compétences. Cependant, afin de fournir un

contexte particulier, on se réfère au développement du traitement du langage naturel en tant que domaine[2].

3.3 Problèmes liés à l'analyse des sentiments :

De nos jours, l'analyse de sentiments est un domaine de recherche très populaire. De nombreux travaux sont réalisés, mais il n'existe pas encore de méthode suffisamment bonne pour classer les sentiments. Pour de nombreux auteurs, la moyenne des résultats est légèrement supérieure à 85%, mais cela ne suffit pas si nous avons besoin de résultats plus précis.

L'objectif principal de l'analyse des sentiments est d'analyser les avis et de tester les scores des sentiments. Cette analyse est divisée en trois niveaux[24] : Niveau document[16], niveau phrase[19], niveau mot/terme[7] ou niveau aspect[17]. Les processus séquentiels sont l'évaluation de l'analyse des sentiments et la détection de la polarité des sentiments.

Plusieurs enjeux doivent être pris en compte lors de la conduite du l'AS. Deux enjeux majeurs sont abordés. Premièrement, le point de vue (ou l'opinion) observé comme négatif dans une situation peut être considéré comme positif dans une autre situation. Deuxièmement, les gens n'expriment pas toujours leurs opinions de la même manière. La plupart des techniques de traitement de texte courantes utilisent le fait que des modifications mineures entre les deux fragments de texte ne sont pas susceptibles de changer le sens réel[13].

L'analyse des sentiments des données des médias sociaux a également été appliquée pour évaluer les produits, comme expliqué dans[1]. Chaque auteur propose ses propres méthodes pour évaluer les opinions. Malheureusement, la plupart des outils ou algorithmes d'analyse des sentiments sont encore au stade de la recherche. Jusqu'à présent, il n'existe aucun algorithme qui puisse fournir des résultats 100% précis pour l'analyse de sentiments. Il y a encore plusieurs débats entre différents chercheurs qui tentent de prouver que leur solution est plus parfaite que les autres.

L'extraction du sentiment ou d'opinion consiste à déterminer la polarité de ce dernier. Dans ce qui suit nous citerons quelques difficultés de cette procédure :

- Ambiguïté de certains mots positifs ou négatifs selon les contextes et qui ne peut pas toujours être levée.
- Difficulté due aux structures syntaxiques et sémantiques d'une phrase et l'expression de l'opinion. Par exemple : l'histoire du film est intéressante mais les acteurs étaient mauvais. Dans ce cas la polarité de la deuxième partie est opposée à la première.
- Difficulté due au contexte : La nécessite d'une bonne analyse syntaxique du texte ; analyse qui peut se révéler particulièrement difficile dans des cas de coordination entre plusieurs parties d'une phrase. Par exemple : ma tante a bien préparé le gâteau, son décor est beau mais je n'ai pas aimé le goût, l'opinion de la dernière partie de la phrase est la plus importante.
- Difficulté due à l'analyse de la phrase par paquets de mots. Les deux phrases sui-

vantes contiennent les mêmes paquets de mots sans pour autant exprimer les mêmes sentiments. La première phrase contient un sentiment positif alors que la deuxième est négative : Je l'ai apprécié pas seulement à cause de ..., Je ne l'ai pas apprécié seulement à cause de ... où se présente la gestion de négation.

— Difficulté due au langage qu'utilisent les internautes pour s'exprimer. Les ponctuations ne sont pas forcément utilisées pour marquer les fins de phrases, des mots spécifiques sont utilisés tel que : "Ha ha ha", "Bieenn", "Super".

— Difficulté de déterminer un lexique adapté à l'analyse de l'ensemble des textes d'opinions.

3.4 Comment réaliser une analyse des sentiments sur les médias sociaux :

Supposons que votre entreprise vienne tout
juste de sortir un nouveau produit et qu'il déchaîne les foules sur les médias sociaux.
Des milliers de publications Instagram, Facebook et Twitter en parlent,
et cet emballement n'est pas près de s'arrêter. S'agit-il toutefois d'un effet positif ou négatif ?
Vous avez besoin
de plus de contexte.C'est là que le sentiment sur les médias sociaux entre en jeu.
Le sentiment sur les médias sociaux représente la perception que les internautes
ont de votre marque et apporte du contexte à chaque commentaire, partage ou publication dans lesquels vous êtes mentionné.
Pour déterminer
votre position sur le spectre positif/négatif, vous devez analyser ces conversions.

Définition 3.1. Une analyse des sentiments sur les médias sociaux vous renseigne sur ce que pensent les internautes à propos de votre marque. Plutôt que de simplement comptabiliser les mentions et les commentaires, l'analyse des sentiments prend en compte les émotions et les opinions. Elle implique la collecte et l'analyse des informations contenues dans les publications que les individus partagent à propos de votre marque sur les médias sociaux.

Mesurer les sentiments sur les médias sociaux est une partie importante de tout plan de surveillance des médias sociaux.

3.5 L'importance de sentiment sur les médias sociaux :

L'analyse des sentiments sur les médias sociaux est parfois appelée
" fouille d'opinions " (de l'anglais " opinion mining "). En effet, il s'agit d'étudier en

profondeur les mots et le contexte des publications sur les médias sociaux pour comprendre les opinions qu'elles révèlent.

Découvrez

pourquoi votre marque doit assurer un suivi des sentiments sur les médias sociaux.

1. **Cernez votre public :** Les professionnels du marketing donnent le meilleur d'eux-mêmes lorsqu'ils parviennent à comprendre leur public. Cela signifie que vous devez comprendre l'opinion du public vis-à-vis de votre marque, de vos publications sur les médias sociaux et de vos campagnes de publicité, et pas uniquement vous fier au nombre de mentions.

   À l'époque de Mad Men, dans les années 1960, les publicitaires réunissaient des groupes cibles pour comprendre comment les gens réagiraient à une nouvelle campagne ou à un nouveau slogan publicitaire. Aujourd'hui, il suffit d'être à l'écoute de ce qu'ils affirment sur les médias sociaux.

   Une analyse continue des sentiments sur les médias sociaux peut vous alerter rapidement lorsque les désirs et préférences des clients changent.

   The Edelman Trust Barometer, dans son rapport spécial Brand Trust and the Coronavirus Pandemic, révèle que lorsque le COVID-19 a provoqué une crise mondiale en mars, 57% de la population souhaitait que les marques cessent de faire du marketing " au ton humoristique ou trop léger ".

   Les outils d'analyse des sentiments sur les médias sociaux peuvent vous aider à garantir que les évolutions de votre marque satisfont pleinement les attentes de votre public.

2. **Améliorez le service client :** La surveillance des opinions bénéficie énormément à l'assistance et au service client.

   D'abord, elle permet d'alerter vos équipes de service et d'assistance de tout nouveau problème dont elles doivent être informées. Votre entreprise peut ensuite préparer une réponse, une stratégie ou un texte approprié. Vous pourriez même en savoir plus sur des problèmes rencontrés avec un produit en particulier, ou son fonctionnement.

   Ensuite, la surveillance des mentions comportant une opinion négative sur les médias sociaux permet à votre équipe de contacter les personnes qui témoigneraient d'une expérience désagréable à l'égard de votre marque. Bien souvent, une simple réponse ou un suivi peut grandement contribuer à la résolution d'une réclamation client.

3. **Ajustez le message de la marque et le développement de produits** Au fur et à mesure que vous surveillerez les sentiments sur les médias sociaux, vous commencerez à comprendre la façon dont vos messages peuvent influencer ce que vos abonnés pensent de vous.

   En suivant les tendances et en étudiant les pics d'opinion positive ou négative, vous en saurez plus sur les souhaits réels de votre public. Vous aurez ainsi une idée plus précise du type de message que vous devriez publier sur chaque réseau social.

   Vous pourriez même obtenir des informations qui peuvent influer sur la stratégie générale de votre marque et le développement de vos produits. Ces informations

pourraient également vous aider à comprendre comment les changements que vous avez opérés hors ligne résonnent dans la sphère des médias sociaux.

4. **Déterminez où vous vous situez dans votre niche** Les marques ne peuvent pas satisfaire pleinement tout le monde. L'analyse des sentiments sur les médias sociaux peut vous aider à savoir où vous vous situez dans votre marché de niche. Vous pourrez ensuite toucher les bonnes personnes avec les bons messages, au bon moment.

   Vous pourrez également identifier les domaines dans lesquels vous excellez particulièrement, et ceux dans lesquels vous devez progresser.

   Par exemple, grâce à l'analyse des sentiments sur les médias sociaux, des chercheurs ont constaté que l'aéroport d'Heathrow, à Londres, est réputé pour la qualité de son réseau Wi-Fi, ses toilettes, ses restaurants et ses salons agréables. En revanche, les internautes n'étaient pas satisfaits du parking, des temps d'attente, des procédures de contrôle des passeports à la douane, ainsi que du personnel de l'aéroport.

   Grâce à ces informations, Heathrow pourrait tâcher d'améliorer les aspects qui n'apportent pas satisfaction aux clients. Ou bien il pourrait choisir de se concentrer sur les domaines dans lesquels il excelle, se démarquant ainsi comme un aéroport confortable et bien équipé.

5. **Repérez les crises de la marque le plus tôt possible** Vous souhaitez que votre marque ne connaisse jamais de crise. Néanmoins, si cela se produit, l'analyse des sentiments sur les médias sociaux peut vous aider à repérer le problème au plus tôt. Vous pouvez alors mettre en place votre plan d'urgence afin de limiter, voire d'éviter, les opinions négatives.

   Nous sommes à une époque où il est particulièrement important que les marques soient à l'écoute des sentiments de leurs clients. Le rapport spécial Edelman a révélé qu'à la fin du mois de mars, 33

   L'analyse des sentiments sur les médias sociaux aurait aidé ces entreprises à corriger le tir à temps pour limiter ces pertes de clients.

3.6 Caractéristiques :

L'analyse des sentiments est un domaine de recherche vaste et complexe. Dans ce qui suit, les principales caractéristiques qui constituent le processus d'analyse des sentiments sont décrites et discutées en détail.

A Catégorisation des sentiments : *Phrases objectives versus phrases subjectives*

Le premier objectif de l'analyse des sentiments consiste généralement à distinguer les phrases subjectives des phrases objectives. Si une phrase donnée est classée comme objective, aucune autre tâche fondamentale n'est requise, alors que si celle-ci est classée comme subjective, sa polarité (positive, négative ou neutre) doit être estimée. La classification de la subjectivité [3] est la tâche qui distingue les phrases

qui expriment des informations objectives des phrases qui expriment des opinions et des avis subjectives.

Un exemple de phrase objective est L'iPhone est un smartphone, tandis qu'un exemple de phrase subjective est L'iPhone est génial. La classification de polarité est la tâche qui distingue les phrases qui expriment des polarités positives, négatives ou neutres. Notant qu'une phrase subjective peut ne pas exprimer un sentiment positif ou négatif (par exemple, Je suppose qu'il est arrivé), pour cette raison, il doit être classé comme neutre.

B Niveaux d'analyse :

Comme mentionné précédemment, le but de l'analyse des sentiments est de définir des outils automatiques capables d'extraire des informations subjectives à partir des textes en langage naturel. Le premier choix lorsqu'on applique l'analyse des sentiments est de définir ce que signifie le texte (c'est-à-dire l'objet analysé) dans le cas d'étude considéré.

En général, l'analyse des sentiments dans le E-Commerce peut être étudiée principalement à trois niveaux :

* *Niveau de texte :* L'objectif est de détecter la polarité d'un texte d'opinion. Par exemple, dans le cadre d'une revue de produit, le système détermine si le texte exprime une opinion globale positive, négative ou neutre au sujet du produit. L'hypothèse est que l'ensemble du texte n'exprime qu'une seule opinion sur une seule entité (un seul produit).
* *Niveau phrase :* Le but est de déterminer la polarité de chaque phrase contenue dans un texte. L'hypothèse est que chaque phrase, dans un texte donné, désigne une seule opinion sur une seule entité.
* *Niveau d'entité et d'aspect :* Effectue une analyse plus fine que le niveau du texte et de la phrase. Elle repose sur l'idée qu'une opinion est constituée d'un sentiment et d'une cible (d'opinion). Par exemple, la phrase L'iPhone est très bien, mais ils ont encore besoin de travailler sur la durée de vie de la batterie et la sécurité évalue trois aspects : iPhone (neutre), la durée de vie de la batterie (négatif) et la sécurité (négatif).

C Opinion régulière versus opinion comparative :

Une opinion peut prendre différentes nuances et peut être assignée à l'un des groupes suivants :

* *Opinion régulière :* Une opinion régulière est souvent désignée dans la littérature comme une opinion standard et elle a deux sous-types principaux :
  — Opinion directe : Une opinion directe fait référence à une opinion exprimée directement sur une entité (par exemple, La luminosité de l'écran de l'iPhone est impressionnante).
  — Opinion indirecte : Une opinion indirecte est une opinion qui est exprimée indirectement sur une entité sur la base de ses effets sur d'autres entités. Par

exemple, la phrase Après être passé à l'iPhone, j'ai perdu toutes mes données décrit un effet indésirable du passage à l'iPhone sur les données, ce qui donne indirectement un sentiment négatif à l'iPhone.

* *Opinion comparative :* Une opinion comparative exprime une relation de similitude ou de différence entre deux ou plusieurs entités et/ou une préférence du détenteur d'opinion basée sur certains aspects communs des entités. Par exemple, les phrases iOS est plus performant qu'Android et iOS est le système d'exploitation le plus performant expriment deux opinions comparatives. Une opinion comparative est habituellement exprimée en utilisant la forme comparative ou superlative d'un adjectif ou d'un adverbe.

D Opinions explicites versus opinions implicites

Parmi les différentes nuances qu'une opinion peut prendre, nous distinguons les opinions explicites et implicite :

* *Opinion explicite :* Une opinion explicite est une déclaration subjective qui donne une opinion régulière ou comparative (par exemple, La luminosité de l'écran de l'iPhone est impressionnante).
* *Opinion implicite :* Une opinion implicite est un énoncé objectif qui implique une opinion régulière ou comparative qui exprime habituellement un fait désirable ou indésirable (par exemple, " Samedi soir, j'irai au cinéma pour regarder 'I am legend'. J'ai hâte de le regarder ! " Et " 'Saving Private Ryan' est plus violent que 'I am legend' "). Le premier exemple suggère qu'il y a de bonnes attentes à propos du film, bien qu'il ne soit pas expliqué en mots, alors que la compréhension de l'opinion cachée dans le second exemple est difficile même pour les humains. Pour certaines personnes, la violence dans les films de guerre pourrait être une bonne caractéristique qui rend le film plus réaliste, alors qu'elle pourrait être une caractéristique négative pour d'autres.

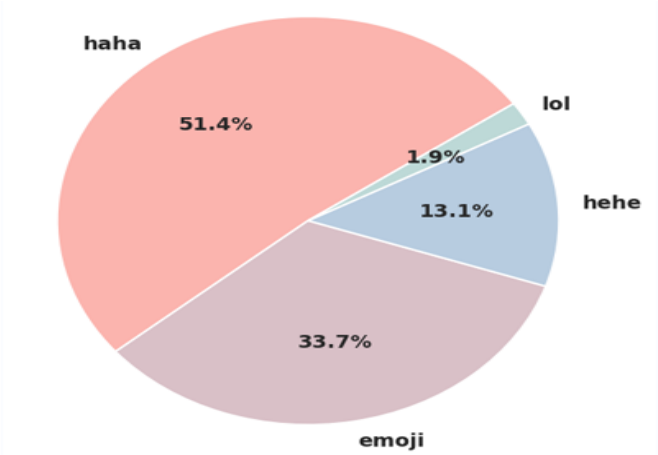
Il est clair que les opinions explicites sont plus faciles à détecter et à classer que les opinions implicites. Une grande partie de la recherche actuelle s'est concentrée sur des opinions explicites. Relativement moins de travail a été fait sur les opinions implicites.

3.7 Facebook Reactions :

Facebook propose de nouveaux émoticônes pour permettre à l'internaute de réagir aux posts des amis et au fil d'actualités via une palette de sentiments. Facebook étend le spectre de ses boutons d'émotions, l'utilisation du seul Like étant limitée et le "je n'aime pas", longtemps (voire toujours) désiré par les internautes, n'a pas été retenu, pour éviter les débordements ou le déferlement de haine, déjà trop courant sur les réseaux sociaux sans instituer un bouton. Et comme l'explique le Product Design Manager de Facebook, *Geoff Teehan* sur Medium : Tout ce qui est partagé n'a pas

vocation à être Liké. La mort du chien de votre ami ou un post sur un politicien peu recommandable ne vont pas susciter qu'une seule émotion chez les internautes, aussi, pourquoi se limiter au seul Like ? ".

Parce que les sentiments vieillissent sur Facebook comme sur le web. Facebook, avait mené, à l'été 2015, une étude auprès de ses utilisateurs actifs, montrant que les internautes voulant traduire leur amusement sur Facebook préféraient à 50% le "haha" puis à 34%, les emojis, et le "héhé", plus subtile, à 13%. Le fameux Lol (Laughing Out Loud), emblématique du web des années 2000 n'est plus utilisé qu'à 1, 9%, bref, complétement ringardisé.



Parce que vous allez être accrocs et que vous êtes surtout de la data Fort de ses 30 millions d'utilisateurs en France (1, 59 milliard dans le monde), Facebook va ainsi renforcer sa gigantesque base de données des émotions des internautes.Le pari de Facebook, au-delà de la collecte de vos data, est avant tout de vous rendre encore plus captif : plus de possibilités d'engagement, c'est potentiellement plus d'interactions envers le contenu et plus de temps encore passé sur les fils de vos amis.Surtout, comme ces "Reactions" vont impacter l'algorithme d'affichage de vos fils d'actualités de la même manière que le Like, vous verrez encore davantage ce que vous aimez voir ou aimez détester.Et pour les pages de marques, c'est aussi un engagement potentiellement plus qualitatif.

3.7.1 L'ultime outil pour tester ses campagnes sur le web social :

Un plus qualitatif

Plus de clics, moins de commentaires ? Si l'on peut penser que davantage d'émotions disponibles enrichiront qualitativement l'engagement, on peut aussi assister à un effet pervers conduisant à l'inverse à une diminution des commentaires, les internautes pouvant maintenant simplement cliquer pour exprimer des émotions plus nuancées.

Un plus quantitatif

Quoiqu'il en soit, avec cette palette de sentiment enrichie, c'est certainement plus d'engagement en volume que les marques verront à leur égard, là où auparavant, de nombreux internautes restaient muets, jugeant que le Like n'était pas approprié pour traduire leur ressenti.

Les fondamentaux des sentiments

L'analyse des sentiments sur le web social (positif, négatif, neutre) à partir du earned media (conversation des internautes et médias) est déjà couramment utilisée par les agences et les entreprises et permet de nombreuses analyses précieuses, à fortiori lorsque les données de tonalité sont croisées avec des critères socio-démographiques et des critères de produits ou services (ex : la perception du SAV par la tranche $25-35$ ans des villes de plus de 100000 habitants est majoritairement positive sur Facebook...).L'analyse des métriques d'engagement de vos propres comptes sociaux (owned media) consistait jusquici à analyser les données d'interactions limitées aux partages, vues, commentaires et aux seuls "Like" en termes démotions sociales. L'ajout des 5 réactions de Facebook développe la granularité des sentiments exprimées et, l'intérêt pour les marques danalyser ce type dengagement est réel même sil faut relativiser.

L'accueil des Facebook Reactions

À en juger par les réactions des marques et des agences et par les premiers "sondages", les internautes aiment ces nouvelles émotions Facebook. Logique dans la mesure où la frustration des limitations du Like sur le réseau le plus grand public du web était ancienne. Ainsi, un sondage YouGov sur les utilisateurs américains montre que 66% d'entre eux aiment les nouveaux boutons Reactions.

3.7.2 Que faire des Facebook Reactions ?

De la data...

Il y a peu de chances...Éventuellement, ces données figureront au sein de vos Social Media non objectivés.Vous aurez certainement à suivre davantage des objectifs liés à l'engagement global, au taux de réponse et pourquoi pas, à la tonalité moyenne des commentaires. Si la mesure des Likes n'avait de sens que comparée aux Partages du même contenu sur Facebook, la mesure d'un Wouah peut apporter un insight complémentaire si elle traduit un plébiscite lors d'une nouvelle campagne de pub par exemple.

Aussi, la collecte de certaines nouvelles émotions de Facebook peut permettre de sonder une certaine catégorie d'internautes sur un produit, une vidéo, un visuel. Mais cela ne suffit pas. Comme pour l'analyse des sentiments, l'analyse des Facebook Reactions n'est viable que croisée avec des critères socio, démo, géo-graphiques et produits. Vous obtiendrez ainsi des données connexes pour analyser certaines de vos communautés, connexes car insuffisantes en soi afin de mieux connaitre vos clients.

Les 5 nouvelles émotions sont en fait une corde à rajouter à l'arc des outils de sondages sociaux en ligne (à l'instar des sondages Twitter) non représentatifs puisque liés aux utilisateurs actifs de votre audience Facebook mais indicatifs sur une tendance exprimée vis à vis d'un nouveau produit/service.

L'analyse des Facebook Reactions
n'est viable que croisée avec des critères socio, démo, ou géo-graphiques et produits.

Une mesure intéressante d'abord pour la pub et le contenu.

Si ces Facebook Reactions peuvent mettre un terme à la simple course aux Likes et inciter à la production d'un contenu plus qualitatif grâce à un impact plus facilement mesurable, leur utilisation essentielle concerne surtout :- une mesure de la satisfaction d'une campagne- un support précieux pour les contents marketers, afin de les aider à répondre à cette question : est-ce que le contenu proposé est bien en adéquation avec mon audience pour une résonance optimale ?Exemples :

- tester quel type de contenu génère le plus de Wouah : jusqu'ici, la seule mesure était le Partage ou le Like. Hiérarchiser vos posts -notamment en fonction des sentiments exprimés-peut être intéressant.
- détecter des ambassadeurs de votre marque. Si le "Jadore" est utilisé à bon escient, vous pouvez, sur plusieurs mois, détecter de véritables ambassadeurs, authentiques amoureux et connaisseurs intimes de vos produits.

Comprendre les émotions qui favorisent le partage puis la viralité : j'adore, j'abhorre Les 2 moteurs les plus puissants du partage et donc à terme de la viralité sont la réaction psychologique (comment le contenu nous touche) et la motivation sociale (pourquoi on veut le partager). C'est ce que révèle l'étude de Unruly, sur 430 milliards de vues de vidéos et 100000 données de consommateurs. Plus l'intensité de lémotion suscitée par le contenu est grande, plus les internautes partagent.

Il sera donc intéressant, à moyen terme, de croiser les émotions Facebook avec les données de partages : partage-t-on uniquement ce que l'on adore, ou partage-t-on aussi ce que l'on déteste (Grrr) ? Certaines études ont déjà la réponse, il sera instructif de les corroborer avec des données du web social : ainsi, une étude menée par *Jonah Berger*, auteur de "Contagious : Why Things Catch On", a constaté que le contenu qui déclenche une réaction de colère chez les lecteurs est à 34% plus susceptible de se situer sur la page du New York Times des "articles les plus partagés", tandis que les messages qui rendent les internautes anxieux le sont à 21% (Source Hubspot).

Gardons à l'esprit que ces Reactions constituent une mesure parmi d'autres, et non une fin en soi : l'utilisation de ces boutons est la résultante de tout un processus qui fait le parcours et l'expérience du client vis-à-vis de votre marque (notoriété, traitement des questions, dynamisme des communautés). Ce parcours se traduit au final par un "Wouah", un "Jadore" ou un "Grrr" sur votre Page Facebook, parce que lexpérience client a été un succès ou un échec en amont.

Ce sont ces métriques du tunnel marketing Social Media (part de voix, acquisition, croissance des communautés) qu'il faudra analyser dabord puis compléter éventuellement par ce supplément d'âme sociale que sont les Facebook Reactions. Celle-ci nous l'espérons, donneront accès a une analyse de sentiments plus segmentée, améliorée et plus performante.

3.8 Machine Learning :

Le Machine Learning ou apprentissage automatique est un domaine scientifique, et plus particulièrement une sous-catégorie de l'intelligence artificielle. Elle consiste à laisser des algorithmes découvrir des " patterns ", à savoir des motifs récurrents, dans les ensembles de données. Ces données peuvent être des chiffres, des mots, des images, des statistiques

Tout ce qui peut être stocké numériquement peut servir de données pour le Machine Learning. En décelant les patterns dans ces données, les algorithmes apprennent et améliorent leurs performances dans l'exécution d'une tâche spécifique.

Pour résumer, les algorithmes de Machine Learning apprennent de manière autonome à effectuer une tâche ou à réaliser des prédictions à partir de données et améliorent leurs performances au fil du temps. Une fois entraîné, l'algorithme pourra retrouver les patterns dans de nouvelles données.

3.8.1 fonctionnement de la Machine Learning :

Le développement d'un modèle de Machine Learning repose sur quatre étapes principales. En règle générale, c'est un Data Scientist qui gère et supervise ce procédé.

La première étape consiste à sélectionner et à préparer un ensemble de données d'entraînement. Ces données seront utilisées pour nourrir le modèle de Machine Learning pour apprendre à résoudre le problème pour lequel il est conçu.

Les données peuvent être étiquetées, afin d'indiquer au modèle les caractéristiques qu'il devra identifier. Elles peuvent aussi être non étiquetées, et le modèle devra repérer et extraire les caractéristiques récurrentes de lui-même.

Dans les deux cas, les données doivent être soigneusement préparées, organisées et nettoyées. Dans le cas contraire, l'entraînement du modèle de Machine Learning risque d'être biaisé. Les résultats de ses futures prédictions seront directement impactés.

La deuxième étape consiste à sélectionner un algorithme à exécuter sur l'ensemble de données d'entraînement. Le type d'algorithme à utiliser dépend du type et du volume de données d'entraînement et du type de problème à résoudre.

La troisième étape est l'entraînement de l'algorithme. Il s'agit d'un processus itératif. Des variables sont exécutées à travers l'algorithme, et les résultats sont comparés avec ceux qu'il aurait du produire. Les " poids " et le biais peuvent ensuite être ajustés pour accroître la précision du résultat.

On exécute ensuite de nouveau les variables jusqu'à ce que l'algorithme produise le résultat correct la plupart du temps. L'algorithme, ainsi entraîné, est le modèle de Machine Learning.

La quatrième et dernière étape est l'utilisation et l'amélioration du modèle. On utilise le modèle sur de nouvelles données, dont la provenance dépend du problème à résoudre. Par exemple, un modèle de Machine Learning conçu pour détecter les spams sera utilisé sur des emails.

De son côté, le modèle de Machine Learning d'un aspirateur robot ingère des données résultant de l'interaction avec le monde réel comme le déplacement de meubles ou l'ajout de nouveaux objets dans la pièce. L'efficacité et la précision peuvent également s'accroître au fil du temps.

3.8.2 les principaux algorithmes de Machine Learning :

Il existe une large variété d'algorithmes de Machine Learning. Certains sont toutefois plus couramment utilisés que d'autres. Tout d'abord, différents algorithmes sont utilisés pour les données étiquetées.

Les algorithmes de régression, linéaire ou logistique, permettent de comprendre les relations entre les données. La régression linéaire est utilisée pour prédire la valeur d'une variable dépendante base sur la valeur d'une variable indépendante. Il s'agirait par exemple de prédire les ventes annuelles d'un commercial en fonction de son niveau d'études ou de son expérience.

La régression logistique est quant à elle utilisée quand les variables dépendantes sont binaires. Un autre type d'algorithme de régression appelé machine à vecteur de support est pertinent quand les variables dépendantes sont plus difficiles à classifier.

Un autre algorithme ML populaire est l'arbre de décision. Cet algorithme permet d'établir des recommandations basées sur un ensemble de règles de décisions en se basant sur des données classifiées. Par exemple, il est possible de recommander sur quelle équipe de football parier en se basant sur des données telles que l'âge des joueurs ou le pourcentage de victoire de l'équipe.

Pour les données non étiquetées, on utilise souvent les algorithmes de " clustering ". Cette méthode consiste à identifier les groupes présentant des enregistrements similaires et à étiqueter ces enregistrements en fonction du groupe auquel ils appartiennent.

Auparavant, les groupes et leurs caractéristiques sont inconnus. Parmi les algorithmes de clustering, on compte les K-moyennes, le TwoStep ou encore le Kohonen.

Les algorithmes d'association permettent quant à eux de découvrir des patterns et des relations dans les données, et à identifier les relations " si / alors " appelées " règles d'association ". Ces règles sont similaires à celles utilisées dans le domaine du Data Mining ou forage de données.

Enfin, les réseaux de neurones sont des algorithmes se présentant sous la forme d'un réseau à plusieurs couches. La première couche permet l'ingestion des données, une ou plusieurs couches cachées tirent des conclusions à partir des données ingérées, et la dernière couche assigne une probabilité à chaque conclusion.

Un réseau de neurones " profond " est composé de multiples couches cachées permettant chacune de raffiner les résultats de la précédente. On l'utilise dans le domaine du Deep Learning.

3.8.3 types de Machine Learning :

On distingue trois techniques de Machine Learning : l'apprentissage supervisé, l'apprentissage non-supervisé, et l'apprentissage par renforcement. Dans le cas de l'apprentissage supervisé, le plus courant, les données sont étiquetées afin d'indiquer à la machine quelles patterns elle doit rechercher.

Le système s'entraîne sur un ensemble de données étiquetées, avec les informations qu'il est censé déterminer. Les données peuvent même être déjà classifiées de la manière dont le système est supposé le faire.

Cette méthode nécessite moins de données d'entraînement que les autres, et facilite le processus d'entraînement puisque les résultats du modèle peuvent être comparés avec les données déjà étiquetées. Cependant, l'étiquetage des données peut se révéler onéreux. Un modèle peut aussi être biaisé à cause des données d'entraînement, ce qui impactera ses performances par la suite lors du traitement de nouvelles données.

Au contraire, dans le cas de l'apprentissage non supervisé, les données n'ont pas d'étiquettes. La machine se contente d'explorer les données à la recherche d'éventuelles patterns. Elle ingère de vastes quantités de données, et utilise des algorithmes pour en extraire des caractéristiques pertinentes requises pour étiqueter, trier et classifier les données en temps réel sans intervention humaine.

Plutôt que d'automatiser les décisions et les prédictions, cette approche permet d'identifier les patterns et les relations que les humains risquent de ne pas identifier dans les données. Cette technique n'est pas très populaire, car moins simple à appliquer. Elle est toutefois de plus en plus populaire dans le domaine du cyber sécurité.

L'apprentissage " semi-supervisé " se situe entre les deux et offre un compromis entre apprentissage supervisé et non-supervisé. Pendant l'entraînement, un ensemble de données étiqueté de moindre envergure est utilisé pour guider la classification et l'extraction de caractéristiques à partir d'un ensemble plus large de données non étiquetées.

Cette approche s'avère utile dans les situations où le nombre de données étiquetées est insuffisant pour l'entraînement d'un algorithme supervisé. Elle permet de contourner le problème.

Enfin, l'apprentissage par renforcement consiste à laisser un algorithme apprendre de ses erreurs pour atteindre un objectif. L'algorithme essayera de nombreuses approches différentes pour tenter d'atteindre son but.

En fonction de ses performances, il sera récompensé ou pénalisé pour l'inciter à poursuivre dans une voie ou à changer d'approche. Cette technique est notamment utilisée pour permettre à une IA de surpasser les humains dans les jeux.

3.8.4 Cas d'usage et applications :

Ces dernières années, on entend parler de nombreuses avancées dans le domaine de l'intelligence artificielle. De même, les applications de l'IA se multiplient. En réalité, la vaste majorité des progrès effectués dans ce domaine sont directement liés au Machine

Learning.

Il en va de même pour les moteurs de recherche web de Google et Baidu, pour les fils d'actualité de réseaux sociaux tels que Facebook et Twitter, ou pour les assistants vocaux comme Siri et Alexa. Ainsi, le Machine Learning peut être considéré comme une innovation phare de ce début de XXIème siècle. C'est la raison pour laquelle les plateformes citées ci-dessus et les autres géants du web collectent de vastes quantités de données personnelles sur leurs utilisateurs : le genre de films que vous préférez, les liens sur lesquels vous cliquez, les publications auxquelles vous réagissez toutes ces données peuvent être utilisées pour nourrir un algorithme de Machine Learning et lui permettre de prédire ce que vous voulez.

Le Machine Learning est également ce qui permet aux aspirateurs robots de faire le ménage seuls, à votre boite mail de détecter les spams, et aux systèmes d'analyse d'image médicale d'aider les médecins à repérer les tumeurs plus efficacement. Les voitures autonomes, elles aussi reposent sur l'apprentissage automatique.

Les assistants numériques, comme Apple Siri, Amazon Alexa ou Google Assistant, reposent sur la technologie de traitement naturel du langage (NLP). Il s'agit d'une application du Machine Learning permettant aux ordinateurs de traiter des données vocales ou textuelles afin de " comprendre " le langage humain. Cette technologie propulse aussi la voix de votre GPS ou encore les Chatbots et les logiciels de type " speech-to-text ".

À mesure que le Big Data continuera à se développer, avec toujours plus de données générées, et alors que l'informatique continuera à gagner en puissance, le Machine Learning offrira encore davantage de possibilités

3.8.5 Machine learning et analyse de données :

Le Machine Learning est massivement utilisé pour la Data Science et l'analyse de données. Il permet de développer, de tester et d'appliquer des algorithmes d'analyse prédictive sur différents types de données afin de prédire le futur.

En automatisant le développement de modèle analytique, le Machine Learning permet d'accélérer l'analyse de données et de la rendre plus précise. Il permet d'assigner aux machines des tâches au cur de l'analyse de données comme la classification, le clustering ou la détection d'anomalie.

Les algorithmes ingèrent les données et délivrent des inférences statistiques, et peuvent s'améliorer de manière autonome au fil du temps. Lorsqu'ils détectent un changement dans les données, ils sont capables de prendre des décisions sans intervention humaine.

Pour l'heure, un humain reste toutefois nécessaire pour passer en revue les résultats des analyses produites par les algorithmes de Machine Learning. Son rôle est de donner du sens à ces résultats, ou encore de s'assurer que les données traitées par l'algorithme ne soient ni biaisées ni altérées.

Conclusion

De nos jours, la recherche sur l'analyse des sentiments et l'extraction d'opinions est très importante. La plupart des industries créent différents types de données et ont besoin d'analyser ces données pour prendre des décisions qui sont bénéfiques pour l'industrie. Les médias sociaux génèrent également d'énormes quantités de données et il est nécessaire de les analyser et d'en tirer des enseignements de ces données en question.

Dans le chapitre suivant, nous allons faire eclarcé la notion d'exploration de données.

4
Exploration de données

Introduction

L'exploration de données, connue aussi sous l'expression de fouille de données, forage de données, prospection de données, data mining, ou encore extraction de connaissances à partir de données, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.

Elle se propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, et d'en extraire un maximum de connaissances.

L'utilisation industrielle ou opérationnelle de ce savoir dans le monde professionnel permet de résoudre des problèmes très divers, allant de la gestion de la relation client à la maintenance préventive, en passant par la détection de fraudes ou encore l'optimisation de sites web. C'est aussi le mode de travail du journalisme de données.

L'exploration de données fait suite, dans l'escalade de l'exploitation des données de l'entreprise, à l'informatique décisionnelle. Celle-ci permet de constater un fait, tel que le chiffre d'affaires, et de l'expliquer comme le chiffre d'affaires décliné par produits, tandis que l'exploration de données permet de classer les faits et de les prévoir dans une certaine mesure ou encore de les éclairer en révélant par exemple les variables ou paramètres qui pourraient faire comprendre pourquoi le chiffre d'affaires de tel point de vente est supérieur à celui de tel autre.

4.1 Historique :

La génération de modèles à partir d'un grand nombre de données n'est pas un phénomène récent. Pour qu'il y ait création de modèle il faut qu'il y ait collecte de données. En Chine on prête à l'Empereur mythique *Yao*, la volonté de recenser les récoltes en 2238 av. J.-C. ; en Égypte le pharaon *Amasis* organise le recensement de sa population au V^e siècle av. J.-C. Ce n'est qu'au $XVII^e$ siècle qu'on commence à vouloir analyser les données pour en rechercher des caractéristiques communes. En 1662, *John Graunt* publie son livre " Natural and Political Observations Made upon the Bills of Mortality " dans lequel il analyse la mortalité à Londres et essaie de prévoir les apparitions de la peste bubonique. En 1763, *Thomas Bayes* montre qu'on peut déterminer, non seulement des probabilités à partir des observations issues d'une expérience, mais aussi les paramètres relatifs à ces probabilités. Présenté dans le cas particulier d'une loi binomiale, ce résultat est étendu indépendamment par Laplace, conduisant à une formulation générale du théorème de Bayes. *Legendre* publie en 1805 un essai sur la méthode des moindres carrés qui permet de comparer un ensemble de données à un modèle mathématique. Les calculs manuels coûteux ne permettent cependant pas d'utiliser ces méthodes hors d'un petit nombre de cas simples et éclairants.

De 1919 à 1925, *Ronald Fisher* met au point l'analyse de la variance comme outil pour son projet d'inférence statistique médicale. Les années 1950 voient l'apparition de calculateurs encore onéreux et des techniques de calcul par lots sur ces machines. Simultanément, des méthodes et des techniques voient le jour telles que la segmentation, classification (entre autres par la méthode des nuées dynamiques), une première version des futurs réseaux de neurones qui se nomme le Perceptron, et quelques algorithmes auto-évolutifs qui se nommeront plus tard génétiques. Dans les années 1960 arrivent les arbres de décision et la méthode des centres mobiles ; ces techniques permettent aux chercheurs d'exploiter et de découvrir des modèles de plus en plus précis. En France, *Jean-Paul Benzécri* développe l'analyse des correspondances en 1962. On reste cependant dans une optique de traitement par lots.

En 1969 paraît l'ouvrage de *Myron Tribus* "Rational descriptions, décisions and designs" qui généralise les méthodes bayésiennes dans le cadre du calcul automatique (professeur à Dartmouth, il utilise assez logiquement le langage BASIC, qui y a été créé quelques années plus tôt, et son interactivité). La traduction en français devient disponible en 1973 sous le nom Décisions rationnelles dans l'incertain. Une idée importante de l'ouvrage est la mention du théorème de Cox-Jaynes démontrant que toute acquisition d'un modèle soit se fait selon les règles de Bayes (à un homomorphisme près), soit conduit à des incohérences. Une autre est que parmi toutes les distributions de probabilité satisfaisant aux observations (leur nombre est infini), il faut choisir celle qui contient le moins d'arbitraire (donc le moins d'information ajoutée, et en conséquence celle d'entropie maximale. La probabilité s'y voit considérée comme simple traduction numérique d'un état de connaissance, sans connotation fréquentiste sous-jacente. Enfin, cet ouvrage popularise la notation des probabilités en décibels, qui rend la règle de Bayes additive et permet de quantifier de façon unique l'apport d'une observation en la rendant

désormais indépendante des diverses estimations a priori préalables.

L'arrivée progressive des micro-ordinateurs permet de généraliser facilement ces méthodes bayésiennes sans grever les coûts. Cela stimule la recherche et les analyses bayésiennes se généralisent, d'autant que Tribus a démontré leur convergence, au fur et à mesure des observations, vers les résultats des statistiques classique tout en permettant d'affiner les connaissances au fil de l'eau sans nécessiter les mêmes délais d'acquisition.

L'affranchissement du protocole statistique classique commence alors : il n'est plus nécessaire de se fixer une hypothèse et de la vérifier ou non a posteriori. Au contraire, les estimations bayésiennes vont construire elles-mêmes ces hypothèses au fur et à mesure que s'accumulent les observations.

L'expression " data mining " avait une connotation péjorative au début des années 1960, exprimant le mépris des statisticiens pour les démarches de recherche de corrélation sans hypothèses de départ. Elle tombe dans l'oubli, puis *Rakesh Agrawal* l'emploie à nouveau dans les années 1980 lorsqu'il entamait ses recherches sur des bases de données d'un volume de 1 Mo. Le concept d'exploration de données fait son apparition, d'après *Pal et Jain*, aux conférences de l'IJCAI en 1989. *Gregory Piatetsky-Shapiro* chercha un nom pour ce nouveau concept dans la fin des années 1980, aux GTE Laboratories. " Data mining " étant sous la protection d'un copyright, il employa l'expression " Knowledge discovery in data bases " (KDD).

Puis, dans les années 1990, viennent les techniques d'apprentissage automatique telles que les SVM en 1998, qui complètent les outils de l'analyste.

4.2 Applications industrielles :

4.2.1 Par objectifs :

De nos jours, les techniques d'exploration de données peuvent être utilisées dans des domaines complètement différents avec des objectifs bien spécifiques. Les sociétés de vente par correspondance analysent, avec cette technique, le comportement des consommateurs pour dégager des similarités de comportement, accorder des cartes de fidélité, ou établir des listes de produits à proposer en vente additionnelle (vente croisée).

Un publipostage (mailing) servant à la prospection de nouveaux clients possède un taux de réponses de 10% en moyenne. Les entreprises de marketing utilisent la fouille de données pour réduire le coût d'acquisition d'un nouveau client en classant les prospects selon des critères leur permettant d'augmenter les taux de réponses aux questionnaires envoyés.

Ces mêmes entreprises, mais d'autres aussi comme les banques, les opérateurs de téléphonie mobile ou les assureurs, cherchent grâce à l'exploration de données à minimiser l'attrition de leurs clients puisque le coût de conservation d'un client est moins important que celui de l'acquisition d'un nouveau.

Les services de polices de tous les pays cherchent à caractériser les crimes (répondre à la question : " Qu'est-ce qu'un crime " normal " ? ") et les comportements des

criminels (répondre à la question : " qu'est-ce qu'un comportement criminel " normal " ? ") afin de prévenir le crime, limiter les risques et les dangers pour la population.

Le scoring des clients dans les banques est maintenant très connu, il permet de repérer les " bons " clients, sans facteur de risque (Évaluation des risques-clients) à qui les organismes financiers, banques, assurances, etc., peuvent proposer une tarification adaptée et des produits attractifs, tout en limitant le risque de non-remboursement ou de non-paiement ou encore de sinistre dans le cas des assurances.

Les centres d'appel utilisent cette technique pour améliorer la qualité du service et permettre une réponse adaptée de l'opérateur pour la satisfaction du client.

Dans la recherche du génome humain, les techniques d'exploration de données ont été utilisées pour découvrir les gènes et leur fonction.

D'autres exemples dans d'autres domaines pourraient être trouvés, mais ce qu'on peut remarquer dès à présent, c'est que toutes ces utilisations permettent de caractériser un phénomène complexe (comportement humain, expression d'un gène), pour mieux le comprendre, afin de réduire les coûts de recherche ou d'exploitation liés à ce phénomène, ou bien afin d'améliorer la qualité des processus liés à ce phénomène.

4.2.2 Par secteurs d'activités :

L'industrie a pris conscience de l'importance du patrimoine constitué par ses données et cherche à l'exploiter en utilisant l'informatique décisionnelle et l'exploration des données. Les compagnies les plus avancées dans ce domaine se situent dans le secteur tertiaire. Selon le site kdnuggets.com la répartition aux États-Unis, en pourcentage du total des réponses au sondage, de l'utilisation de l'exploration des données par secteurs d'activités s'effectue en 2010 comme ceci :

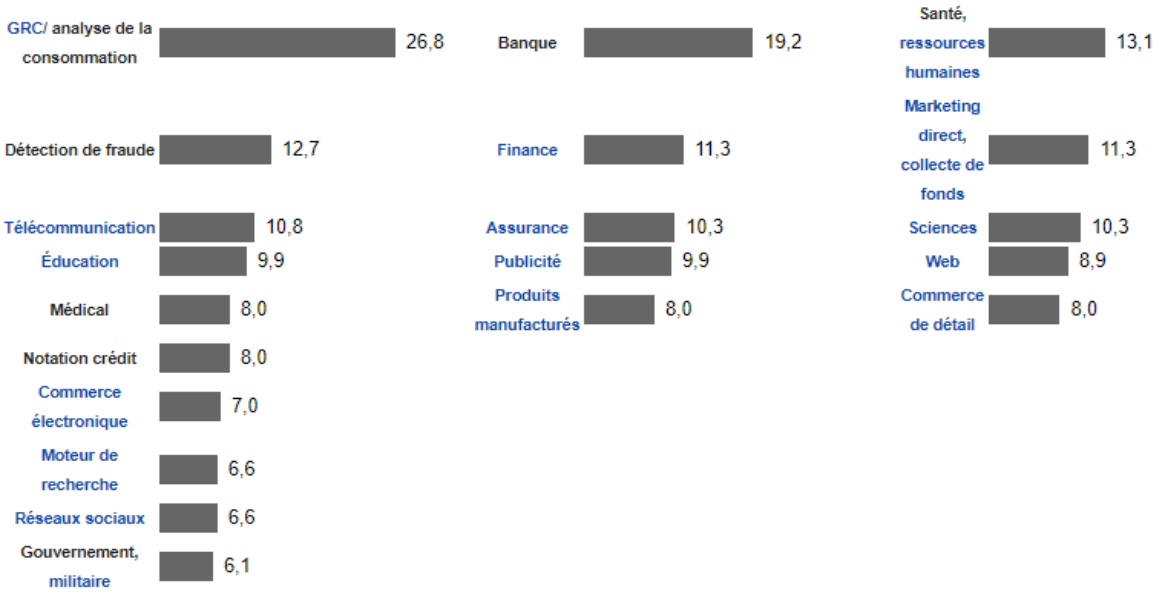


FIGURE 4.1 – Branches et domaines dans lesquels est utilisée l'exploration des données (%).

4.3 Méthode CRISP-DM :

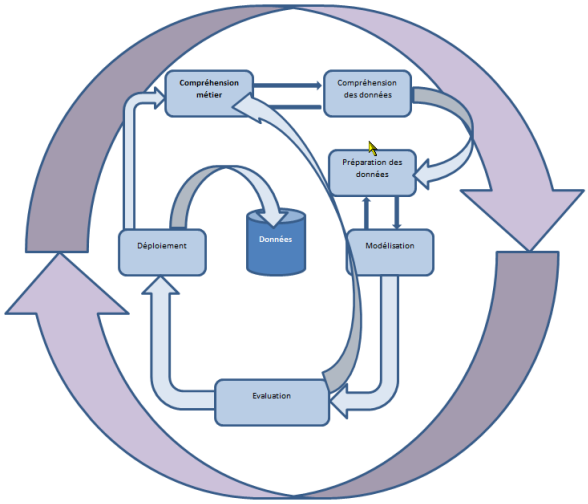


FIGURE 4.2 – Phases du processus CRISP-DM

La méthode CRISP-DM découpe le processus de fouille de données en six étapes permettant de structurer la technique et de l'ancrer dans un processus industriel. Plus qu'une théorie normalisée, c'est un processus d'extraction des connaissances métiers.

Il faut d'abord comprendre le métier qui pose la question à l'analyste, formaliser

le problème que l'organisation cherche à résoudre en ce qui concerne les données, comprendre les enjeux, connaître les critères de réussite du projet et mettre en place un plan initial pour réaliser cet objectif.

Ensuite, l'analyste a besoin de données appropriées. Dès que l'équipe de projet sait ce qu'il faut faire, elle doit se mettre en quête des données, des textes et tout le matériel qui lui permettra de répondre au problème. Il lui faut ensuite en évaluer la qualité, découvrir les premiers schémas apparents pour émettre des hypothèses sur les modèles cachés. Les données que l'équipe de projet a collectées sont hétérogènes. Elles doivent être préparées en fonction des algorithmes utilisés, en supprimant les valeurs aberrantes, ou valeurs extrêmes, en complétant les données non renseignées, par la moyenne ou par la méthode des K plus proches voisins, en supprimant les doublons, les variables invariantes et celles ayant trop de valeurs manquantes, ou bien par exemple en discrétisant les variables si l'algorithme à utiliser le nécessite, comme c'est par exemple le cas pour l'analyse des correspondances multiples ACM, l'analyse discriminante DISQUAL, ou bien la méthode de Condorcet.

Une fois les données prêtes, il faut les explorer. La modélisation regroupe des classes de tâches pouvant être utilisées seules ou en complément avec les autres dans un but descriptif ou prédictif. La segmentation est la tâche consistant à découvrir des groupes et des structures au sein des données qui sont d'une certaine façon similaires, sans utiliser des structures connues a priori dans les données. La classification est la tâche de généralisation des structures connues pour les appliquer à des données nouvelles. La régression tente de trouver une fonction modélisant les données continues, c'est-à-dire non discrètes, avec le plus petit taux d'erreur, afin d'en prédire les valeurs futures.

L'association recherche les relations entre des items. Par exemple un supermarché peut rassembler des données sur des habitudes d'achats de ses clients. En utilisant les règles d'association, le supermarché peut déterminer quels produits sont fréquemment achetés ensemble et ainsi utiliser cette connaissance à des fins de marketing. Dans la littérature, cette technique est souvent citée sous le nom d'" analyse du panier de la ménagère ".

Il s'agit d'évaluer ensuite les résultats obtenus en fonction des critères de succès du métier et d'évaluer le processus lui-même pour faire apparaître les manques et les étapes négligées. À la suite de ceci, il doit être décidé soit de déployer, soit d'itérer le processus en améliorant ce qui a été mal ou pas effectué.

Puis vient la phase de livraison et de bilan de fin de projet. Les plans de contrôle et de maintenance sont conçus et le rapport de fin de projet est rédigé. Afin de déployer un modèle prédictif, le langage PMML, basé sur le XML, est utilisé. Il permet de décrire toutes les caractéristiques du modèle et de le transmettre à d'autres applications compatibles PMML.

Résoudre un problème par un processus d'exploration de données impose généralement l'utilisation d'un grand nombre de méthodes et d'algorithmes différents plus ou moins faciles à comprendre et à employer. Il existe deux grandes familles d'algorithmes : les méthodes descriptives et les méthodes prédictives.

4.4 Méthodes descriptives :

Définition 4.1. Les méthodes descriptives permettent d'organiser, de simplifier et d'aider à comprendre l'information sous-jacente d'un ensemble important de données. Elles permettent de travailler sur un ensemble de données, organisées en instances de variables, dans lequel aucune des variables explicatives des individus n'a d'importance particulière par rapport aux autres. Elles sont utilisées par exemple pour dégager, d'un ensemble d'individus, des groupes homogènes en typologie, pour construire des normes de comportements et donc des déviations par rapport à ces normes telles que la détection de fraudes nouvelles ou inconnues à la carte bancaire ou à l'assurance maladie, pour réaliser de la compression d'informations ou de la compression d'image, etc.

Exemple 4.1. *Parmi les techniques disponibles, celles qui sont issues de la statistique peuvent être exploitées. Sont regroupées sous le vocable analyses factorielles, des méthodes statistiques qui permettent de dégager des variables cachées dans un ensemble de mesures ; ces variables cachées sont appelées " facteurs ". Dans les analyses factorielles, on part du principe que si les données sont dépendantes entre elles, c'est parce qu'elles sont liées à des facteurs qui leur sont communs. L'intérêt des facteurs réside dans le fait qu'un nombre réduit de facteurs explique presque aussi bien les données que l'ensemble des variables, ce qui est utile quand il y a un grand nombre de variables. Les techniques factorielles se décomposent principalement en analyse en composantes principales, analyse en composantes indépendantes, analyse factorielle des correspondances, analyse des correspondances multiples et positionnement multidimensionnel.*

Pour fixer les idées, l'analyse en composantes principales fait correspondre à m variables quantitatives décrivant p individus, n facteurs, les composantes principales, de telle manière que la perte d'information soit minimum. En effet, les composantes sont organisées dans l'ordre croissant des pertes d'information, la première en perdant le moins. Les composantes sont non corrélées linéairement entre elles et les individus sont projetés sur les axes définis par les facteurs en respectant la distance qui existe entre eux. Les similitudes et les différences sont expliquées par les facteurs.

L'analyse factorielle des correspondances et l'ACM font correspondre à m variables qualitatives décrivant les caractéristiques de p individus, n facteurs en utilisant le tableau de contingence, ou le tableau de Burt dans le cas de l'ACM, de telle manière que les facteurs soient constitués des variables numériques séparant le mieux les valeurs des variables qualitatives initiales, que deux individus soient proches s'ils possèdent à peu près les mêmes valeurs des variables qualitatives et que les valeurs de deux variables qualitatives soient proches si ce sont pratiquement les mêmes individus qui les possèdent.

On peut aussi utiliser des méthodes nées dans le giron de l'intelligence artificielle et plus particulièrement dans celui de l'apprentissage automatique. La classification non supervisée est une famille de méthodes qui permettent de regrouper des individus en classes, dont la caractéristique est que les individus d'une même classe se ressemblent, tandis que ceux de deux classes différentes sont dissemblables. Les classes de la classification ne sont pas connues au préalable, elles sont découvertes par le processus. D'une manière générale,

les méthodes de classification servent à rendre homogènes des données qui ne le sont pas à priori, et ainsi permettent de traiter chaque classe avec des algorithmes sensibles aux données aberrantes. Dans cette optique, les méthodes de classification forment une première étape du processus d'analyse.

Ces techniques empruntées à l'intelligence artificielle utilisent le partitionnement de l'ensemble des informations mais aussi le recouvrement. Le partitionnement est l'objectif des algorithmes utilisant par exemple des méthodes telles que celles des k-means (les " nuées dynamiques " en français), des k-medoids (k-médoïdes), k-modes et k-prototypes, qu'on peut utiliser pour rechercher les aberrations, les réseaux de Kohonen, qui peuvent aussi servir à la classification, l'algorithme EM ou l'AdaBoost. La classification hiérarchique est un cas particulier de partitionnement pour lequel les graphiques produits sont facilement compréhensibles. Les méthodes ascendantes partent des individus qu'on agrège en classes, tandis que les méthodes descendantes partent du tout et par divisions successives arrivent aux individus qui composent les classes. Ci-contre le graphique d'une classification ascendante a été tracé pour montrer comment les classes les plus proches sont reliées entre elles pour former des classes de niveau supérieur.

Le recouvrement à logique floue est une forme de recouvrement de l'ensemble des individus représentés par les lignes d'une matrice où certains d'entre eux possèdent une probabilité non nulle d'appartenir à deux classes différentes. L'algorithme le plus connu de ce type est le FCM (Fuzzy c-means).

Il faut aussi mentionner l'Iconographie des corrélations associée à l'utilisation des Interactions logiques, méthode géométrique qui se prête bien à l'analyse des réseaux complexes de relations multiples.

En bio-informatique, des techniques de classification double sont employées pour regrouper simultanément dans des classes différentes les individus et les variables qui les caractérisent.

Pour rendre compte de l'utilité de ces méthodes de recouvrement, il faut se rappeler que la classification est un problème dont la grande complexité a été définie par Eric Bell. Le nombre de partitions d'un ensemble de n objets est égal à : $B_n = \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^n}{k!} > e^n$. *Il vaut donc mieux avoir des méthodes efficaces et rapides pour trouver une partition qui répond au problème posé plutôt que de parcourir l'ensemble des solutions possibles.*

Enfin, quand l'analyse se porte non pas sur les individus, les items ou les objets, mais sur les relations qui existent entre eux, la recherche de règles d'associations est l'outil adapté. Cette technique est, à l'origine, utilisée pour faire l'analyse du panier d'achats ou l'analyse de séquences. Elle permet, dans ce cas, de savoir quels sont les produits achetés simultanément, dans un supermarché par exemple, par un très grand nombre de clients ; elle est également appliquée pour résoudre des problèmes d'analyse de parcours de navigation de sites web. La recherche de règles d'association peut être utilisée de manière supervisée ; les algorithmes APriori, GRI, Carma, méthode ARD ou encore PageRank se servent de cette technique.

4.5 Méthodes prédictives :

Définition 4.2. La raison d'être des méthodes prédictives est d'expliquer ou de prévoir un ou plusieurs phénomènes observables et effectivement mesurés. Concrètement, elles vont s'intéresser à une ou plusieurs variables définies comme étant les cibles de l'analyse. Par exemple, l'évaluation de la probabilité pour qu'un individu achète un produit plutôt qu'un autre, la probabilité pour qu'il réponde à une opération de marketing direct, celles qu'il contracte une maladie particulière, en guérisse, les chances qu'un individu ayant visité une page d'un site web y revienne, sont typiquement des objectifs que peuvent atteindre les méthodes prédictives.

En exploration de données prédictive, il y a deux types d'opérations : la discrimination ou classement, et la régression ou prédiction, tout dépend du type de variable à expliquer. La discrimination s'intéresse aux variables qualitatives, tandis que la régression s'intéresse aux variables continues.

Les méthodes de classement et de prédiction permettent de séparer des individus en plusieurs classes. Si la classe est connue au préalable et que l'opération de classement consiste à analyser les caractéristiques des individus pour les placer dans une classe, la méthode est dite " supervisée ". Dans le cas contraire, on parle de méthodes " non-supervisées ", ce vocabulaire étant issu de l'apprentissage automatique. La différence entre les méthodes descriptives de classification que l'on a vues précédemment, et les méthodes prédictives de classement provient du fait que leur objectif est divergent : les premières " réduisent, résument, synthétisent les données " pour donner une vision plus claire de l'amas de données, alors que les secondes expliquent une ou plusieurs variables cibles en vue de la prédiction des valeurs de ces cibles pour les nouveaux arrivants.

Exemple 4.2. *On peut référencer quelques exemples de méthodes prédictives, et les présenter selon le domaine d'où elles proviennent.*

Parmi les méthodes issues de l'intelligence artificielle, l'analyste pourra utiliser les arbres de décision, parfois pour la prédiction, parfois pour discrétiser les données quantitatives, le raisonnement par cas, les réseaux de neurones, les neurones à base radiale pour la classification et l'approximation de fonctions, ou peut-être les algorithmes génétiques, certains en appui des réseaux bayésiens, d'autres comme Timeweaver en recherche d'évènements rares.

Si l'analyste est plus enclin à utiliser les méthodes issues de la statistique et des probabilités, il se tournera vers les techniques de régressions linéaires ou non linéaires au sens large pour trouver une fonction d'approximation, l'analyse discriminante de Fisher, la régression logistique, et la régression logistique PLS pour prédire une variable catégorielle, ou bien le modèle linéaire généralisé (GLM), le modèle additif généralisé (GAM) ou modèle log-linéaire, et les modèles de régression multiple postulés et non postulés afin de prédire une variable multidimensionnelle.

Quant à l'inférence bayésienne et plus particulièrement les réseaux bayésiens, ils pourront être utile à l'analyste si celui-ci cherche les causes d'un phénomène ou bien cherche la probabilité de la réalisation d'un évènement.

S'il souhaite compléter les données manquantes, la méthode des k plus proches voisins (K-nn) reste à sa disposition.

La liste des algorithmes évolue chaque jour, car ils n'ont pas tous le même objet, ne s'appliquent pas aux mêmes données en entrée et aucun n'est optimal dans tous les cas. En outre, ils s'avèrent complémentaires les uns aux autres en pratique et en les combinant intelligemment en construisant des modèles de modèles ou méta modèles, il est possible d'obtenir des gains en performance et en qualité très significatifs. L'ICDM-IEEE a fait en 2006 un classement des 10 algorithmes ayant le plus d'influence dans le monde de l'exploration de données : ce classement est une aide efficace au choix et à la compréhension de ces algorithmes.

Avec les moyens modernes de l'informatique l'une ou l'autre de ces deux solutions peut s'envisager dans chaque projet, mais d'autres techniques sont apparues qui ont prouvé leur efficacité pour améliorer la qualité des modèles et leur performance.

4.6 L'analyse exploratoire des données :

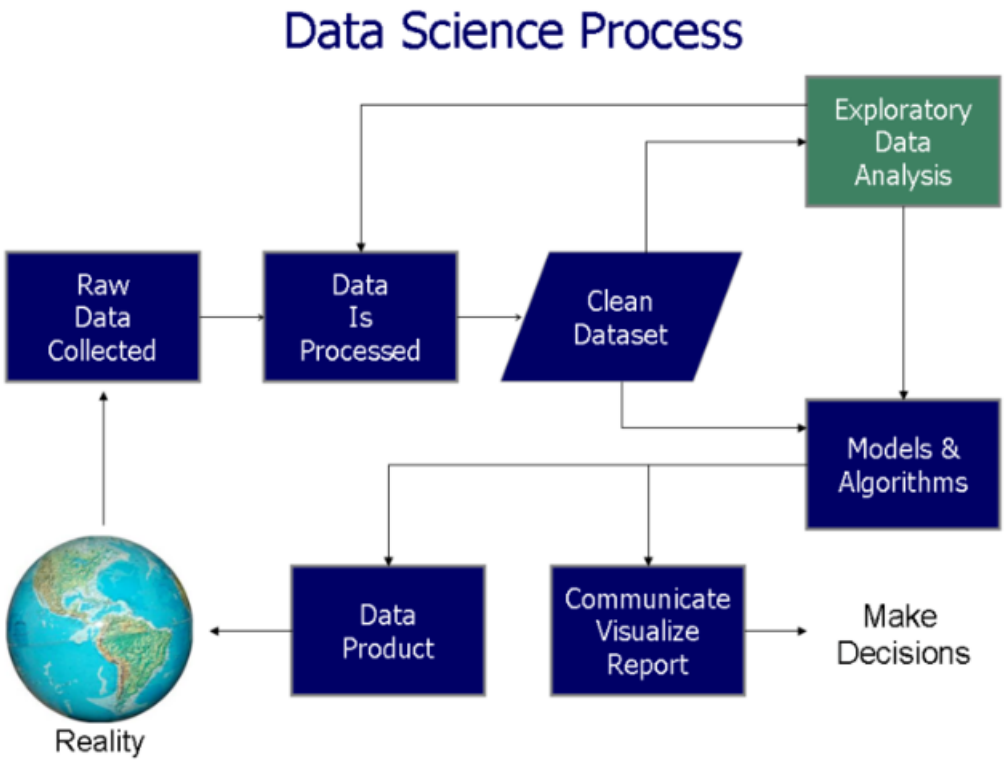


FIGURE 4.3 – Organigramme du processus de science des données

L'analyse exploratoire des données (AED) est utilisée par les spécialistes des données pour analyser et étudier les ensembles de données puis résumer leurs principales

caractéristiques, souvent en employant des méthodes de visualisation des données. Elle permet de déterminer la meilleure façon de manipuler des sources de données pour obtenir les réponses dont vous avez besoin. Ainsi, les spécialistes des données peuvent découvrir plus facilement des modèles (patterns), identifier des anomalies, tester une hypothèse ou vérifier des suppositions.

L'AED est principalement utilisée pour identifier ce que les données peuvent révéler au-delà de la tâche formelle de modélisation ou de test d'hypothèse, et permet de mieux comprendre les variables d'un ensemble de données et les relations entre elles. Elle peut en outre permettre de déterminer si les techniques statistiques que vous envisagez pour l'analyse des données sont appropriées. Développées à l'origine par le mathématicien américain

emphJohn Tukey dans les années 1970, les techniques d'AED restent aujourd'hui une méthode largement utilisée dans le processus de reconnaissance de données.

4.6.1 Objectif de l'EDA

L'objectif principal de l'AED est d'aider à examiner les données avant de formuler des hypothèses.Elle peut permettre d'identifier les erreurs évidentes, mais aussi de mieux comprendre les modèles (patterns) au sein des données, de détecter les valeurs aberrantes ou les événements anormaux, de trouver des relations intéressantes entre les variables.

Les spécialistes des données peuvent utiliser l'analyse exploratoire pour s'assurer que les résultats qu'ils produisent sont valides et applicables à tous les résultats et objectifs métier souhaités. L'AED aide également les parties prenantes en confirmant qu'elles posent les bonnes questions. L'AED peut aider à répondre à des questions sur les écarts-types, les variables catégorielles et les intervalles de confiance. Une fois l'AED terminée et les conclusions tirées, ses fonctions peuvent être utilisées dans des analyses de données ou des modélisations plus sophistiquées, y compris l'apprentissage automatique.

4.6.2 Outils d'analyse des données exploratoires

Les fonctions et techniques statistiques spécifiques que vous pouvez réaliser avec les outils d'AED incluent :

* Les techniques de regroupement et de réduction de dimension qui permettent de créer des représentations graphiques de données hautement dimensionnelles contenant de nombreuses variables.
* La visualisation univariée de chaque zone dans l'ensemble de données brutes, avec des statistiques sommaires.
* Les visualisations bivariées et les statistiques sommaires qui permettent d'évaluer la relation entre chaque variable de l'ensemble de données et la variable cible que vous étudiez.
* Les visualisations multivariées, pour cartographier et comprendre les interactions entre les différents champs des données.

* Le regroupement en k-moyennes, méthode de regroupement de l'apprentissage non supervisé où les points de données sont assignés à K groupes, c'est-à-dire le nombre de groupes, sur la base de la distance au centroïde de chaque groupe. Les points de données les plus proches d'un centroïde particulier sont regroupés dans une même catégorie. Le regroupement en k-moyennes est couramment utilisé dans la segmentation des marchés, la reconnaissance des formes et la compression d'image.
* Les modèles prédictifs, tels que la régression linéaire, qui utilisent les statistiques et les données pour prévoir les résultats.

4.6.3 Outils d'analyse exploratoire des données

Les outils
de science des données les plus couramment utilisés pour créer une AED incluent :

Python : Langage de programmation interprété, orienté objet, avec une sémantique dynamique. Ses structures de données intégrées de haut niveau, combinées au typage dynamique et à la liaison dynamique, le rendent très attrayant pour le développement rapide d'applications, ainsi que pour une utilisation en tant que langage de script ou langage de liaison pour connecter des composants existants. Python et AED peuvent être utilisés ensemble pour identifier des valeurs manquantes dans un ensemble de données, ce qui est important pour pouvoir décider de la manière de traiter les valeurs manquantes pour l'apprentissage automatique.

R : Langage de programmation open source et environnement logiciel libre pour le calcul statistique et les graphiques, soutenu par la R Foundation for Statistical Computing. Le langage R est largement utilisé par les statisticiens en science des données pour développer des observations statistiques et des analyses de données.

4.6.4 Partitionnement de données

Définition 4.3. La classification non supervisée, appelée aussi regroupement en français (clustering en anglais), est un processus qui permet de trouver des groupes d'objets (appelé, clustters) en fonction des variables ou des attributs qui les décrivent. Ainsi, dans un problème de clustering, l'ensemble D des données est composé de m objets (ou observations) sans étiquettes (ou classes prédéfinies), chacun décrit par plusieurs variables. On note $X = (x_1, x_2, ..., x_n)$ l'ensemble des n variables décrivant l'ensemble des m objets.

Les données peuvent donc être représentée par une matrice de taille $(m * n)$ avec m lignes d'objets et n colonnes de variables.

Le clustering a pour objectif de regrouper dans un même cluster les objets jugés similaires selon une certaine métrique de similarité (homogénéité intra-classe) et séparer les objets dissimilaire dans des clusters distincts (hétérogénéité inter-classe). le résultat final du clustering est appelé *scéma de clustering*[15].

Les principales étapes du clustering

Plusieurs phases sont à considérer lors du développement d'un outil d'extraction et d'analyse de l'information à base de clustering. En effet, le processus de clustering s'effectue en trois étapes principales : (1) La préparation des données, (2) Le choix de l'algorithme de clustering et (3) validation et interprétation des résultats[9].

La préparation des données La préparation des données est une étape indispensable en amont du processus de clustering. Elle consiste à filtrer, formatter et présenter ces données afin de ne retenir que les paramètres de description les plus discriminants. En effet, une mauvaise préparation produit des résultats difficilement exploitables. Ainsi, un même algorithme de clustering peut produire un résultat satisfaisant et un résultat aberrant sur les même données représentées différemment.

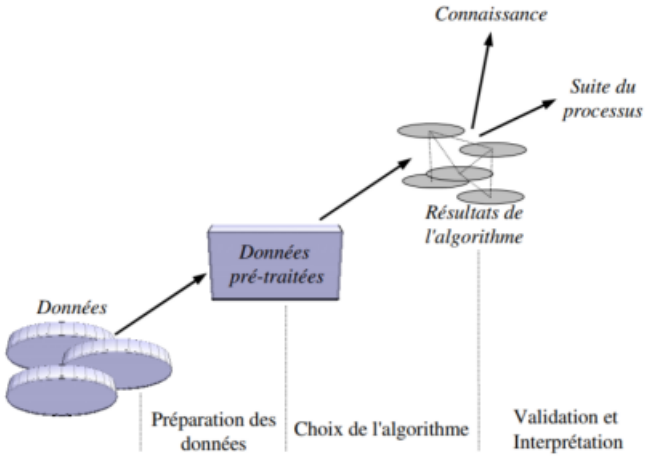


FIGURE 4.4 – Les principales étapes du prceessus de clustering[9]

Le choix de l'algorithme L'objectif des méthodes de clustering est de produire une structure permettant d'organiser les données. celle-ci peut être sous forme de dendrogramme ou de partition. Ainsi, les études ont montré qu'aucune méthode n'est intrinsèquement meilleure que d'autres sur l'ensemble des problèmes[10]. En effet, le choix de l'algorithme approprié dépend de l'application ou du contexte dans lequel les clusters sont créés et de la nature des données étudiées. Par exemple, si le problème est de réduire la taille d'un jeu de données, le meilleur schéma sera celui qui minimise la perte d'informations[6].

Validation et interprétation des résultats Cette étape dépend des deux étapes en amont. En effet, plus les données sont bien préparées et représentées et plus il sera aisé d'interpréter les résultats de la classification. Ainsi, la validation des résultats de clustering vise à déterminer si les clusters générés sont exploitable en utilisant un ensemble de critères permettant de déterminer la qualité des clusters[14].

Les méthodes de clustering de base

On distingue classiquement les grandes familles de méthodes en clustering suivantes :

* Les méthodes hiérarchique.
* Les méthodes par partitionnement.
* Les méthodes à base de densité.
* Les méthodes basées sur un modèle.

Dans notre travaille, nous avons travaillé par les méthodes par partitionnement.

Les méthodes par partitionnement

Les algorithmees de partitionnement construisent
directement, en sortie, une partition de l'espace des objets en k clusters.

* **Le principe de fonctionnement** Selon la définition d'une partition, le principe générale de ces algorithmes est que chaque cluster doit contenir au moins un objet, et que chaque objet doit appartenir à un cluster unique (cas des partitions strictes). Pour ce faire, étant donné le nombre de cluster k requis, ces algorithmes génèrent une partition initiale, puis recherchent à l'améliorer en rétribuant les objets d'un cluster à un autre. En pratique il est impossible de générer toutes les objets d'un partitions de clustering pour des raisons évidentes de complexité. On cherche alors une "bonne" partition correspondant à un optimum local en optimisant une fonction objective qui traduit que les objets doivent être "similaire" au sein d'un même cluster, et "dissimilaire" d'un cluster à un autre. Cette optimum est obtenu de façon itérative, en améliorant la partition initiale choisie plus au moins aléatoirement, par ré-allocation des objets autour de contres mobiles. Les clusters sont présentés par leur "centroïdes", qui correspond à la moyenne de l'ensemble des objets contenus dans le cluster. selon la manière dont les clusters sont construits, on présente dans ce qui suit, les algorithmes les plus cités dans littérature qui appliquennt le principe de partitionnement, à savoire : k-means[23] et k-medoitds(CALARA)[21].
* **Llgorithme des k-means** L'algorithme, ou l'algorithme des k-moyennes a été introduit initialement par J.MacQuenn[23]. il est sans aucun doute la méthode de partitionnement la plus connue et la plus utilisée dans divers domaines d'application scientifiques et industrielles. Ce succès est dû au fait que cet algorithme présentee un rapport coût :efficacité avantageux. Dans sa version classique, l'algorithme consiste à sélecyionner aléatoirement k objjets qui représentent les centroïde initiaux. Un objet est assigné au cluster pour lequel la distance entre les objets est le centroïde est minimale. Les centreoïdes sont alors recalculés et l'on passe à l'intération suivante :

$$\sum_{r}^{k} \sum_{x_i \in C_r} (x_i - g_r)^2 \tag{4.1}$$

Où :

. C_r est le cluster numéro r.

. x_i est un objet dans un cluster C_r.

. g_r est le centre de cluster C_r.

La figure suivante illuster l'algorithme des k-means sur un ensemble de quatre points $a = (-1, 1)$,$b = (0, 1)$,$c = (3, 0)$ et $d = (3, -1)$ qui doivent être classés dans 2 clusters.

L'étape 1 : On dispose de 4 points à classer en 2 classes.

L'étape 2 : à l'initialisation, deux de ces points sont choisis aléatoirement comme centre de classe.

L'étape 3 : Deux claasses sont créées en regroupant les autres points en fonction du centre de classe le plus proche.

L'étape 4 : on définit les nouveaux centres de classe comme étant le barycentre des classes nouvellement créées.

L'étape 5 : on regroupe à nouveau les points.

L'étape 6 : on définit les nouveaux centres de classes.

A l'étape suivante rien ne change, l'algorithme s'arrête.

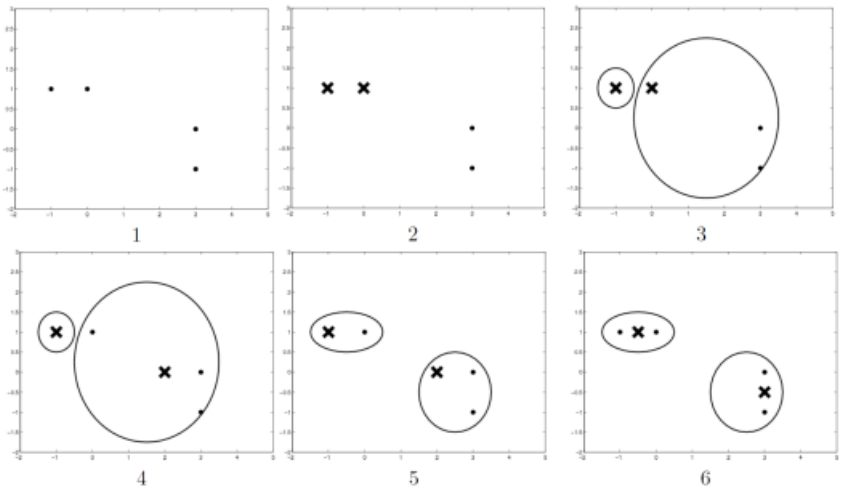


FIGURE 4.5 – Illustration de l'algorithme k-means

Conclusion

La technique de clustering présente l'une des techniques de data mining les plus utilisée et appliquée dans divers domaines. Dans ce cadre, nous avons étudier dans ce chapitre les déffirentes approches relatives à cette technique.

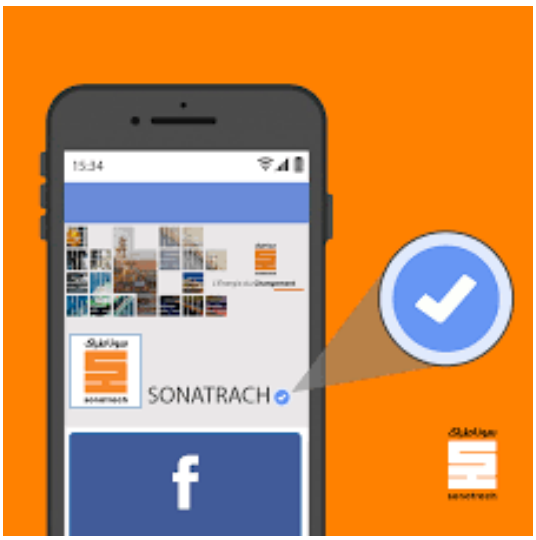
Dans le chapitre suivant, nous appliquons ces déffirents techniques et algorithmes sur notres données, afin de trouver une solution à notre problématique.

5

Implémentation, résultats et discussions

Introduction

Pour trouver une gestion de l'image de marque de l'entreprise SONATRACH au niveau des réseaux sociaux, on a choisi d'analyser sa page facebook :



La page officiel de l'entreprise a été créé le 26 septembre 2018, et elle est actif depuis le 24 février 2019 d'une manière juridique ; aujourd'hui la dernière contient plus de 225612 abonnées.
Et pour améliorer cette page facebook on doit d'abord répondre a ces questions :

? Quelle sont les critères les plus important pour dire qu'une publication a une bonne réaction ?

? Est-ce que y a-t-il certains types de publications qui est associé avec un nombre élevé de j'aime ?

? Sur qu'elle critère la réaction d'un internaute est considérée positive (ou négative) ?

? Qu'est-ce qu'il rend les internautes triste ou en colère envers une publication ?

? Le public réagit-il plus envers des publications écrites dans une certaine langue

hypothèses

* le type de la publication est le critère le plus important pour qu'une publication avoir une bonne réaction.
* y a certains types de publication qui a toujours un nombre de j'aime élever.
* la langue de la publication rend les internautes en colère.
* le nombre de réaction est élevé quand la publication est en arabe.
* le type de publication influence sur les réactions des internautes.

Dans ce chapitre, nous expliquons comment nous avons mené notre étude afin d'établir une stratégie de communication efficace pour une meilleure réputation de notre entreprise.

5.1 Protocole choisi

Le protocole que nous avons choisi est décrit comme suit :

Collecte de données facebook ne souhaite pas que vous collectiez ses données de manière abusive,facebook est un site privé,vous ne pouvez pas collecter de données à l'aide de moyens automatisés surtout les données privées qui ne sont pas publiques,alors comment faire ?

la réponse est simple, vous pouvez configurer un compte de développeur Facebook et créer un identifiant d'application Facebook pour collecter des données facebook sur une certaine page. tout ce qu'il faut, c'est un identifiant et un jeton d'accès.à l'aide de l'utile Graph API Explorer

à l'aide d'un simple code python, vous pouvez collecter le type de données que vous ciblerez au format json aprés le stocker dans un fichier csv

collection des commentaire d'une page facebook

```
pip install facebook_sdk
import facebook
import json
import pandas as pd
graph = facebook.GraphAPI('EAAFlZA43bTS4BANecijcQSVYVMepCtj7kRh8wZB
```

```
E6sDut56pwkI8WXucZBkgUpSnjB4a1OLIlDdB2twjtYIQ7Fnh28rgYksHuMwcrrkHwo

SpS3N456hqloXZAJasrOgyvaos0RwReLEEbZCaOnqbiWXNmoZBVGElaLazxZCauuQKM

oRK18iHDJ653EobEHja5wD8DF7gZAthGwZDZD')
posts=graph
.get_object(id='110749158308901_110809258302891',fields='comments')
print(json.dumps(posts,indent=4))
filecsv=pd.read_json(json.dumps(posts,indent=4))
filecsv.to_csv('FBcomments.csv',index=None)
```

output


```
{
    "comments": {
        "data": [
            {
                "created_time": "2022-05-21T14:54:31+0000",
                "from": {
                    "name": "ROM2 Boumerdes",
                    "id": "110749158308901"
                },
                "message": "good",
                "id": "110809258302891_389493669772766"
            }
        ]
    },
    "id": "110749158308901_110809258302891"
}
```


l'application nécessite des privilèges d'administrateur auxquels nous n'avions pas accès, nous avons donc décidé de collecter les données manuellement

Nous avons récolté les données sous format d'un fichier csv, contenant les champs suivants :

- *post_ date* : Les date de poster par publication.
- *post_ id* : Est le code ID de chaque publication.
- *post_ type* : signifie le type de la publication(condoléances, communiqué de presse, ...).

. *langauge* : La langue utilisé pour annoncer la publication (arabe/français).

. *num_ likes* : Nombre de j'aime par publication.

. *num_ loves* : Nombre de réaction "j'adore" par publication.

. *num_ solidarity* : Nombre de réaction "solidaire" par publication.

. *num_ hahas* : Nombre de réaction "rire" par publication.

. *num_ wows* : Nombre de réaction "étonné" par publication.

. *num_ sads* : Nombre de réaction "triste" par publication.

. *num_ angrys* : Nombre de réaction "énervé" par publication.

. *num_ comments* : Nombre de commentaire par publication.

. *num_ shares* : Nombre de partage par publication.

. *has_ photo* : La publication contient une photo ou pas (true = contient une photo / false= ne contient pas une photo).

. *has_ video* : La publication contient une vidéo ou pas (true = contient une vidéo / false= ne contient pas une vidéo).

. *score* : contient une moyenne obtenue d'après l'avis du responsable du service communication

Fonction scoring Dans le but de lier entre les réactions pour les utiliser en fonction du type de la publication, nous avons pensé à une fonction de score qui pondère toutes les réactions des abonnés selon le type de publication. Exemple : lors d'un message de félicitation, une réaction de type "*love*" a plus de poids qu'une réaction de type "*j'aime*". Par contre dans une publication de type "Condoléances", la réaction "*triste*" a plus de poids que toutes les autres, un "*love*" correspondrait à une réaction négative. Cette méthode de scoring s'inspire directement de l'avis du responsable du service communication de notre entreprise, ce qui correspond à la récupération des préférences du décideur.

modélisation :

Les variables : Soient les vecteur X et λ définies :

* $X = (x_1, x_2, ..., x_7)$ tel que :
  x_1 signifie le nombre de réaction "j'aime" d'une publication.
  x_2 signifie le nombre de réaction "j'adore" d'une publication.
  x_3 signifie le nombre de réaction "solidaire" d'une publication.
  x_4 signifie le nombre de réaction "rire" d'une publication.
  x_5 signifie le nombre de réaction "étonné" d'une publication.
  x_6 signifie le nombre de réaction "triste" d'une publication.
  x_7 signifie le nombre de réaction "énervé" d'une publication.
* $\lambda = (\lambda_1, \lambda_2, ..., \lambda_7)$ tel que : λ_i est la pondération associé à la variable x_i, $i = \overline{1..7}$.

La fonction score : soit la fonction score suivante $f(x)$ définie sur R comme suit : $f(x) = \frac{1}{\sum_{i=1}^{7} \lambda_i} * \sum_{i=1}^{7} (\lambda_i * x_i)$

Choix de pondération : après la discussion avec le responsable du service communication de notre entreprise et d'après l'explication de ces préférences on a conclu le vecteur de pondération suivant :$\lambda = (5, 7, 6, -2, 4, -3, -1)$

Préparation de données Dans le but d'appliquer les méthodes décrites dans les chapitres précédents, nous avons importé notre classeur sous format CSV, nous avons effectué des nettoyages de données à champs manquants, et arranger notre fichier de données.

Étude de corrélation Afin de valider les hypothèses de ce travail, nous avons étudié la corrélation entre les différents champs de notre étude.

Clustering Cette technique une étape très importante de notre analyse. Elle nous renvoie une visibilité globale de la performance des publications de SONATRACH.

Résultats, Discussions et recommandations Pour finir, d'après les résultats trouvées dans les dernières étapes, nous proposons les stratégies à mener pour SONATRACH afin d'améliorer son image de marque.

5.2 Les logiciels utilisés

5.2.1 Google Sheets : tableur en ligne

Google Sheets est un tableur en ligne qui permet de créer et de mettre en forme des feuilles de calcul, et de les modifier en collaboration avec d'autres personnes.

Sécurité des données :

La sécurité, c'est d'abord comprendre comment les développeurs collectent et partagent vos données. Les pratiques concernant leur confidentialité et leur protection peuvent varier selon votre utilisation, votre région et votre âge. Le développeur a fourni ces informations et peut les modifier ultérieurement.

Utiliser Google Sheets :

Étape 1 : créer une feuille de calcul
Pour créer une feuille de calcul :

1. On ouvre l'écran d'accueil de Sheets en accédant à l'adresse *sheets.google.com* .
2. On clique sur Nouveau +. La nouvelle feuille de calcul s'ouvre.

Étape 2 : modifier et mettre en forme une feuille de calcul

On pout ajouter, modifier
et mettre en forme du texte, des nombres et des formules dans une feuille de calcul.

Étape 3 : partager des fichiers et les modifier à plusieurs

On peut partager des fichiers et des dossiers avec d'autres personnes,
et déterminer si celles-ci peuvent les consulter, les modifier ou les commenter.

5.2.2 Python

Python est le langage de programmation open source le plus employé par les informaticiens. Ce langage s'est propulsé en tête de la gestion d'infrastructure, d'analyse de données ou dans le domaine du développement de logiciels. En effet, parmi ses qualités, Python permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font. Il a libéré les développeurs des contraintes de formes qui occupaient leur temps avec les langages plus anciens. Ainsi, développer du code avec Python est plus rapide qu'avec d'autres langages.

L'environnement de travail

ANACONDA

Anaconda est une distribution des langages de programmation **Python** et **R** pour le calcul scientifique (science des données, applications d'apprentissage automatique, traitement de données à grande échelle, analyse prédictive, etc.), qui vise à simplifier la gestion et le déploiement des packages. La distribution comprend des packages de science des données adaptés à Windows, Linux et macOS. Il est développé et maintenu par Anaconda, Inc., qui a été fondée par *Peter Wang* et *Travis Oliphant* en 2012. En tant que

produit Anaconda, Inc., il est également connu sous le nom d'Anaconda Distribution ou Anaconda Individual Edition, tandis que les autres produits de la société sont Anaconda Team Edition et Anaconda Enterprise Edition, qui ne sont pas gratuits.

La distribution Anaconda est livrée avec plus de 250 packages installés automatiquement, et plus de 7500 packages open source supplémentaires peuvent être installés à partir de PyPI ainsi que le package conda et le gestionnaire d'environnement virtuel. Il comprend également une interface graphique, Anaconda Navigator, comme alternative graphique à l'interface de ligne de commande (CLI).

Navigateur Anaconda

Anaconda Navigator est une interface utilisateur graphique (GUI) de bureau incluse dans la distribution Anaconda qui permet aux utilisateurs de lancer des applications et de gérer des packages, des environnements et des canaux conda sans utiliser de commandes de ligne de commande. Navigator peut rechercher des packages sur Anaconda Cloud ou dans un référentiel Anaconda local, les installer dans un environnement, exécuter les packages et les mettre à jour. Il est disponible pour Windows, macOS et Linux.

Les applications suivantes sont disponibles par défaut dans Navigator : JupyterLab, Cahier Jupyter, QtConsole , Espion, Colle, Orange, RStudio, Code Visual Studio.

Project Jupyter

Est un projet et une communauté dont le but est de "développer des logiciels, standards ouverts et services pour l'informatique interactive dans des dizaines de langages de programmation". Il a été dérivé d' IPython en 2014 par *Fernando Pérez* et *Brian Granger*. Le nom du projet Jupyter est une référence aux trois langages de programmation de base pris en charge par Jupyter, qui sont Julia, Python et R , ainsi qu'un hommage aux cahiers de Galileo enregistrant la découverte des lunes de Jupiter. Le

projet Jupyter a développé et pris en charge les produits informatiques interactifs Jupyter Notebook, JupyterHub et JupyterLab.

Bloc- notes Jupyter

Jupyter Notebook (anciennement IPython Notebooks) est un environnement de calcul interactif basé sur le Web permettant de créer des documents de bloc-notes. Un document Jupyter Notebook est un REPL basé sur un navigateur contenant une liste ordonnée de cellules d'entrée/sortie pouvant contenir du code, du texte (à l'aide de Markdown), des mathématiques, des tracés et des médias enrichis. Sous l'interface, un notebook est un document JSON, suivant un schéma visionné, se terminant généralement par l'extension ".ipynb".

Les notebooks Jupyter sont construits sur un certain nombre de bibliothèques open source populaires : IPython, ZéroMQ, Tornade, jQuery, Bootstrap (cadre frontal), Math Jax.

5.2.3 Les APIs utilisées

Il faut considérer les bibliothèques comme un ensemble d'outils prêts à l'emploi que quelqu'un d'autre a développés pour faciliter le système de codage. Ainsi, au lieu d'avoir la charge de créer une fonction qui effectue une certaine opération, on peut simplement aller dans une bibliothèque et utiliser une fonction déjà créée. Le côté unique de Python est que, comme il est si diffus et si répandu dans la communauté d'analyse de données, il existe des bibliothèques dédiées vraiment puissantes qu'on peut utiliser pour nos problèmes d'analyse de données. De plus, il y a beaucoup de documentation dans chaque bibliothèque. Ci-dessous, les principales bibliothèques dans le domaine de la science des données :

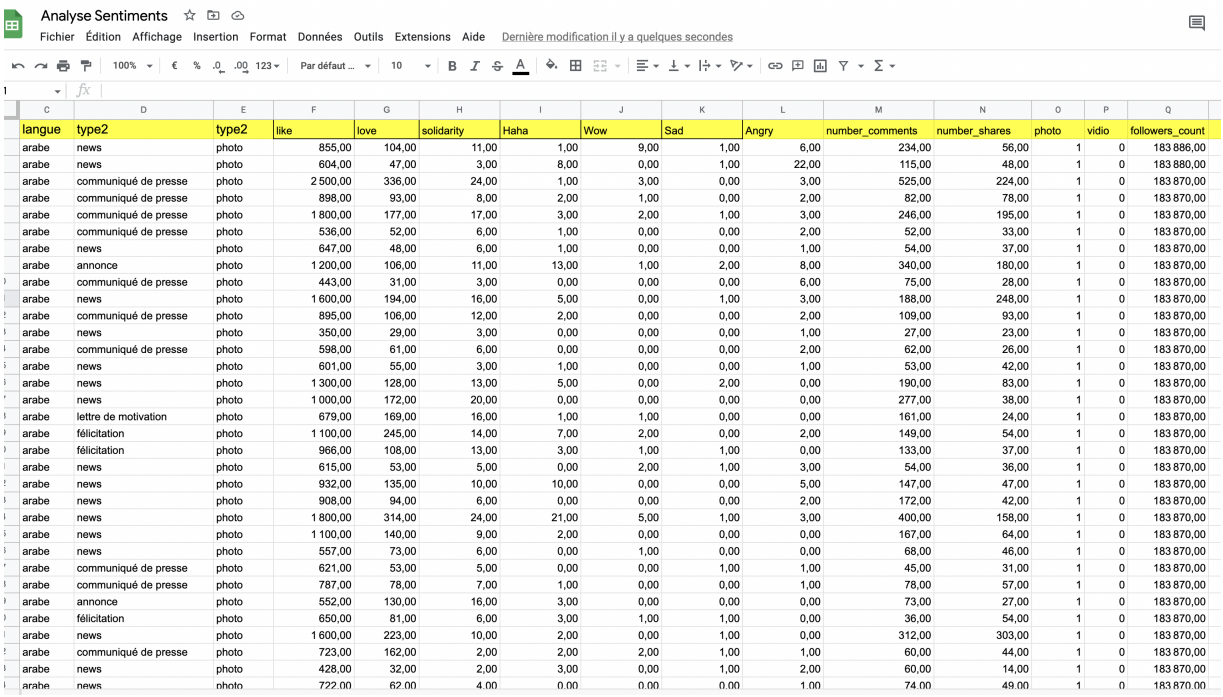
- **Numpy** : signifie "python numérique". Il offre des fonctions pré-compilées pour les routines numériques.
- **PANDAS** : C'est parfait pour l'analyse, la manipulation et la visualisation des données. Il permet aux structures de données de haut niveau et à certains outils de les manipuler.
- **MATPLOTLIB** Excellent pour la visualisation de données. Il peut exporter des graphiques et d'autres images vers des formats vectoriels.
- **SCIPY** Scipy est pour l'algèbre, les statistiques, l'algèbre linéaire.
- **Seaborn** est pour visualisation de données basée sur matplotlib . Il fournit une interface de haut niveau pour dessiner des graphiques statistiques attrayants et informatifs.
- **Scikit-learn** destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme Inria.
- **yellowbrick**Yellowbrick est une suite d'outils de visualisation et de diagnostic qui permettra une sélection plus rapide des modèles. C'est un package Python qui combine scikit-learn et matplotlib. Certains des outils de visualisation les plus populaires

incluent la sélection de modèles, la visualisation des caractéristiques, la classification et la visualisation de la régression.

. **NLTK** Natural Language Toolkit (NLTK) est une bibliothèque logicielle en Python permettant un traitement automatique des langues, développée par Steven Bird et Edward Loper du département d'informatique de l'université de Pennsylvanie. En plus de la bibliothèque, NLTK fournit des démonstrations graphiques, des données-échantillon, des tutoriels, ainsi que la documentation de l'interface de programmation (API).

5.3 Collecte de données

Dans cette partie, nous avons récolté les données de l'entreprise sur ses publication de 24 février 2019 jusq'à le 2 juin 2022, donc au total : 613 publications étudiées.
Comme l'entreprise n'a pas une base de données prête, et aussi l'absence de droits d'administration sur la page nous a empêchés d'utiliser différentes API pour l'extraction automatique des données.
L'objectif le plus important de la collecte de données est de s'assurer que les données sont riches en informations et fiables. Afin de prendre des décisions fondées sur ces données puissent pour la recherche.
Après la collecte des données, l'étape suivante est le pré traitement, c'est une étape importante dans l'analyse des sentiments, est un outil puissant pour traiter les données catégoriques et numériques en général. Cette étape est celle où les données sont préparées pour devenir des données prêtes à être analysées, en supprimant ou en modifiant les données qui sont incorrectes, incomplètes, non pertinentes, dupliquées ou mal formulé.
Nous avons obtenue le tableau de données suivant :


Analyse Sentiments

Fichier Édition Affichage Insertion Format Données Outils Extensions Aide Dernière modification il y a quelques secondes

langue	type2	type2	like	love	solidarity	Haha	Wow	Sad	Angry	number_comments	number_shares	photo	vidio	followers_count
arabe	news	photo	855,00	104,00	11,00	1,00	9,00	1,00	6,00	234,00	56,00	1	0	183 886,00
arabe	news	photo	604,00	47,00	3,00	8,00	0,00	1,00	22,00	115,00	48,00	1	0	183 880,00
arabe	communiqué de presse	photo	2 500,00	336,00	24,00	1,00	3,00	0,00	3,00	525,00	224,00	1	0	183 870,00
arabe	communiqué de presse	photo	898,00	93,00	8,00	2,00	1,00	0,00	2,00	82,00	78,00	1	0	183 870,00
arabe	communiqué de presse	photo	1 800,00	177,00	17,00	3,00	2,00	1,00	3,00	246,00	195,00	1	0	183 870,00
arabe	communiqué de presse	photo	536,00	52,00	6,00	1,00	0,00	0,00	2,00	52,00	33,00	1	0	183 870,00
arabe	news	photo	647,00	48,00	6,00	1,00	0,00	0,00	1,00	54,00	37,00	1	0	183 870,00
arabe	annonce	photo	1 200,00	106,00	11,00	13,00	1,00	2,00	8,00	340,00	180,00	1	0	183 870,00
arabe	communiqué de presse	photo	443,00	31,00	3,00	0,00	0,00	0,00	6,00	75,00	28,00	1	0	183 870,00
arabe	news	photo	1 600,00	194,00	16,00	5,00	0,00	1,00	3,00	188,00	248,00	1	0	183 870,00
arabe	communiqué de presse	photo	895,00	106,00	12,00	2,00	0,00	0,00	2,00	109,00	93,00	1	0	183 870,00
arabe	news	photo	350,00	29,00	3,00	0,00	0,00	0,00	1,00	27,00	23,00	1	0	183 870,00
arabe	communiqué de presse	photo	598,00	61,00	6,00	0,00	0,00	0,00	2,00	62,00	26,00	1	0	183 870,00
arabe	news	photo	601,00	55,00	3,00	1,00	0,00	0,00	1,00	53,00	42,00	1	0	183 870,00
arabe	news	photo	1 300,00	128,00	13,00	5,00	0,00	2,00	0,00	190,00	83,00	1	0	183 870,00
arabe	news	photo	1 000,00	172,00	20,00	0,00	0,00	0,00	0,00	277,00	38,00	1	0	183 870,00
arabe	lettre de motivation	photo	679,00	169,00	16,00	1,00	1,00	0,00	0,00	161,00	24,00	1	0	183 870,00
arabe	félicitation	photo	1 100,00	245,00	14,00	7,00	2,00	0,00	2,00	149,00	54,00	1	0	183 870,00
arabe	félicitation	photo	966,00	108,00	13,00	3,00	1,00	1,00	0,00	133,00	37,00	1	0	183 870,00
arabe	news	photo	615,00	53,00	5,00	0,00	2,00	1,00	3,00	54,00	36,00	1	0	183 870,00
arabe	news	photo	932,00	135,00	10,00	10,00	0,00	0,00	5,00	147,00	47,00	1	0	183 870,00
arabe	news	photo	908,00	94,00	6,00	0,00	0,00	0,00	2,00	172,00	42,00	1	0	183 870,00
arabe	news	photo	1 800,00	314,00	24,00	21,00	5,00	1,00	3,00	400,00	158,00	1	0	183 870,00
arabe	news	photo	1 100,00	140,00	9,00	2,00	0,00	0,00	0,00	167,00	64,00	1	0	183 870,00
arabe	news	photo	557,00	73,00	6,00	0,00	1,00	0,00	0,00	68,00	46,00	1	0	183 870,00
arabe	communiqué de presse	photo	621,00	53,00	5,00	0,00	0,00	1,00	1,00	45,00	31,00	1	0	183 870,00
arabe	communiqué de presse	photo	787,00	78,00	7,00	1,00	0,00	0,00	1,00	78,00	57,00	1	0	183 870,00
arabe	annonce	photo	552,00	130,00	16,00	3,00	0,00	0,00	0,00	73,00	27,00	1	0	183 870,00
arabe	félicitation	photo	650,00	81,00	6,00	3,00	1,00	1,00	0,00	36,00	54,00	1	0	183 870,00
arabe	news	photo	1 600,00	223,00	10,00	2,00	0,00	1,00	0,00	312,00	303,00	1	0	183 870,00
arabe	communiqué de presse	photo	723,00	162,00	2,00	2,00	2,00	1,00	1,00	60,00	44,00	1	0	183 870,00
arabe	news	photo	428,00	32,00	2,00	3,00	0,00	1,00	2,00	60,00	14,00	1	0	183 870,00
arabe	news	photo	722,00	62,00	4,00	0,00	0,00	0,00	1,00	74,00	49,00	1	0	183 870,00

FIGURE 5.1 – Le classeur contenant toutes nos données

Importation des bibliothèques

Scikit-learn (Sklearn) est la bibliothèque la plus utile et la plus robuste pour l'apprentissage automatique en Python. Il fournit une sélection d'outils efficaces pour l'apprentissage automatique et la modélisation statistique, notamment la classification, la régression, le regroupement et la réduction de la dimensionnalité via une interface de cohérence en Python.

Et pour sela

nous avons fait appel aux bibliothèques qu'on est besoin pour effectuer ce travail :

```
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns

# to visualize missing values
import missingno as msno
# to draw plot in notebook
%matplotlib inline


plt.style.use('ggplot')
```

```
mpl.rcParams['axes.unicode_minus'] = False

# for spliting train data andd test data
from mpl_toolkits.mplot3d import Axes3D
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, classification_report

from sklearn.ensemble import RandomForestClassifier

from sklearn.impute import SimpleImputer

from sklearn.metrics import mean_absolute_error
```

Chargement des données

Pour importer notre fichier sous Python avec ce script :

```
facebook_data = pd.read_csv("C:/sonadataset.csv")
pd.set_option('display.max_columns',None)
```

5.4 Préparation de données

Le terme "préparation des données" désigne les opérations de nettoyage et transformation qui doivent être appliqués aux données brutes avant leur traitement et analyse. Il s'agit d'une étape importante avant le traitement proprement dit, qui implique souvent de reformater et corriger les données et de combiner des datasets pour enrichir certaines données.

* Pour éliminer les cases vides (NAN) de notre tableau nous avons utilisé le scripte suivant :

```
fbsona.dropna(inplace=True)
```

* Nous avons utilisé ce scripte pour Convertir les variables catégorielles en variables factices.

```
genre_poke = pd.get_dummies(fbsona["post_type"])
language_poke = pd.get_dummies(fbsona['language'])

X = pd.concat([fbsona, genre_poke],axis = 1)
X = pd.concat([X, language_poke], axis=1)
```

```
X.head()
```

5.5 Visualisation de données

La visualisation des données (ou dataviz ou représentation graphique de données) est un ensemble de méthodes permettant de résumer de manière graphique de données statistiques qualitatives et surtout quantitatives afin de montrer les liens entre des ensembles de ces données. Cette visualisation fait partie de la science des données.

Connaître nos données

* Pour afficher notre tableau nous avons utilisé le scripte :

```
fbsona = pd.DataFrame(facebook_data)
fbsona.head()
```

Qui nous donne le résultat suivant

	post_date	post_id	post_type	language	num_likes	num_loves	num_solidarity	num_hahas	num_wows	num_sads	num_angrys	num_comments	num_s
0	06/02/2022	id0001	nouvelle	arabe	855.0	104.0	11.0	1.0	9.0	1.0	6.0	234.0	
1	06/01/2022	id0002	nouvelle	arabe	604.0	47.0	3.0	8.0	0.0	1.0	22.0	115.0	
2	06/01/2022	id0003	communiqué de presse	arabe	2500.0	336.0	24.0	1.0	3.0	0.0	3.0	525.0	
3	05/29/2022	id0004	communiqué de presse	arabe	898.0	93.0	8.0	2.0	1.0	0.0	2.0	82.0	
4	05/28/2022	id0005	communiqué de presse	arabe	1800.0	177.0	17.0	3.0	2.0	1.0	3.0	246.0	

FIGURE 5.2 – L'affichage du tableau par le python

* Pour savoir la dimension de notre tableau nous avons utilisé le scripte suivant :

```
fbsona.shape
```

Nous avons obtenu les résultats suivant : $(613, 16)$ Signifie que le tableau contient 613 ligne et 16 attributs.

* Et avec le scripte suivant nous avons extrait tous les informations sur chaque entrée du tableau :

```
fbsona.info
```

Résultat

```
<bound method DataFrame.info of       post_date       post_id             post_type language  num_likes  \
0     06/02/2022  1.408468e+15              nouvelle    arabe      855.0
1     06/01/2022           NaN              nouvelle    arabe      604.0
2     06/01/2022  1.407534e+15  communiqué de presse    arabe     2500.0
3     05/29/2022  1.405324e+15  communiqué de presse    arabe      898.0
4     05/28/2022           NaN  communiqué de presse    arabe     1800.0
..           ...           ...                   ...      ...        ...
608    2/26/2019           NaN                   NaN      NaN       98.0
609    2/25/2019           NaN                   NaN      NaN       71.0
610    2/24/2019           NaN                   NaN      NaN       70.0
611    2/24/2019           NaN                   NaN      NaN       83.0
612    2/24/2019           NaN                   NaN      NaN      109.0

     num_loves  num_solidarity  num_hahas  num_wows  num_sads  num_angrys  \
0        104.0            11.0        1.0       9.0       1.0         6.0
1         47.0             3.0        8.0       0.0       1.0        22.0
2        336.0            24.0        1.0       3.0       0.0         3.0
3         93.0             8.0        2.0       1.0       0.0         2.0
4        177.0            17.0        3.0       2.0       1.0         3.0
..         ...             ...        ...       ...       ...         ...
608       10.0             0.0        1.0       2.0       0.0         0.0
609        8.0             0.0        0.0       0.0       0.0         0.0
610        6.0             0.0        0.0       0.0       0.0         0.0
611        8.0             0.0        0.0       0.0       0.0         0.0
612        8.0             0.0        0.0       0.0       0.0         0.0
```

FIGURE 5.3 – Résultat de la description du tableau

Effectuer une analyse exploratoire avec des statistiques

* La fonction describe() est utilisée pour générer des statistiques descriptives qui résument la tendance centrale, la dispersion et la forme de la distribution d'un ensemble de données, à l'exclusion des valeurs NaN.

```
fbsona.describe()
```

Résultat

	post_id	num_likes	num_loves	num_solidarity	num_hahas	num_wows	num_sads	num_angrys	num_comments	num_shares	n
count	3.000000e+00	601.000000	601.000000	601.000000	601.000000	612.000000	612.000000	611.000000	612.000000	612.000000	
mean	1.407109e+15	539.304493	48.510815	4.299501	5.630616	0.921569	7.571895	1.396072	77.017974	60.261438	
std	1.614754e+12	401.179746	61.679010	10.039098	31.712878	2.243826	61.433234	4.305572	130.436161	106.270631	
min	1.405324e+15	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	1.406429e+15	257.000000	13.000000	0.000000	0.000000	0.000000	0.000000	0.000000	18.000000	19.000000	
50%	1.407534e+15	470.000000	31.000000	2.000000	2.000000	0.000000	0.000000	0.000000	43.000000	36.000000	
75%	1.408001e+15	722.000000	62.000000	6.000000	4.000000	1.000000	1.000000	1.000000	83.250000	66.000000	
max	1.408468e+15	3400.000000	659.000000	176.000000	642.000000	28.000000	1100.000000	67.000000	1600.000000	1700.000000	

FIGURE 5.4 – Résultat de la description du tableau

Interprétation

count : Le nombre de valeurs non vides.
moyenne : La valeur moyenne (moyenne).
std : L'écart type.
min : la valeur minimale.
25 % - Le centile 25%.
50 % - Le centile 50%.
75 % - Le centile 75%.
max - la valeur maximale.

* La langue la plus utilisée pour publier

```
 lang = fbsona['language'].value_counts()
plt.figure(figsize=(20,10))
colors = sns.color_palette('pastel')[0:5]

#create pie chart
plt.pie(x = lang.values, labels = lang.index, colors = colors, autopct='%.0f%%')
plt.title('La langue utilis e par la page Facebook SONATRACH')
plt.show()
```

Résultat

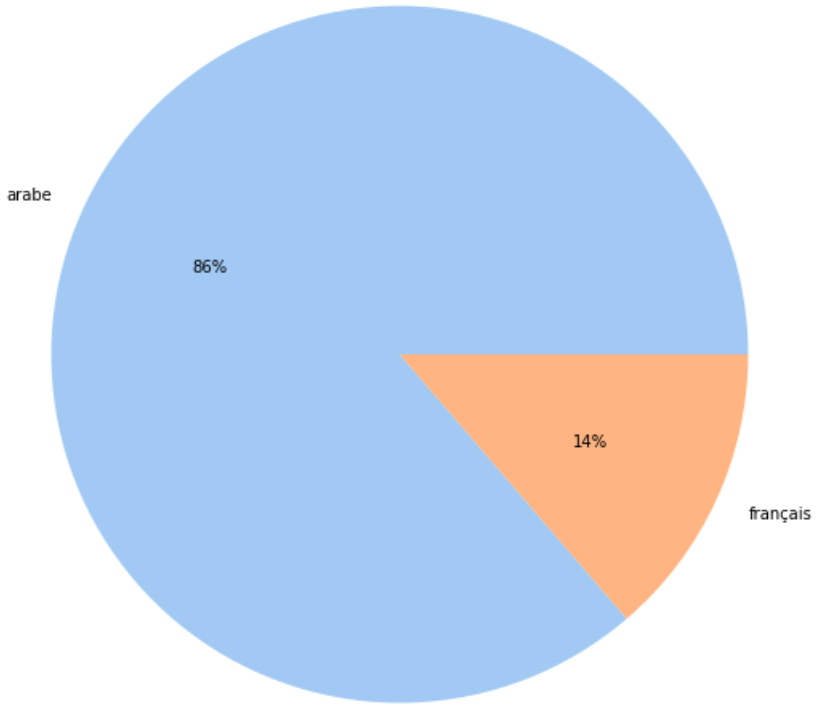


FIGURE 5.5 – Diagramme circulaire : les langues utilisée pour publier.

Interprétation : d'après le diagramme il est clair que la langue la plus utilisé pour publier sur la page facebook de la SONATRACH est l'arabe (86% des publications sont en arabe).

* La visualisation de nombres de publication qui contient une photo ou pas.

```
hasphoto = fbsona['has_photo'].value_counts()

plt.figure(figsize=(15,3))
sns.barplot(x=hasphoto.index,y=hasphoto.values).set_title('has_photo')
```

Résultat

FIGURE 5.6 – Le graphique à barres : nombre de publication qui contient une photo ou pas.

Interprétation : le diagramme signifie que la plus grande partie des publications de la page contient une photo.

* La visualisation de nombre de publication qui contient une vidéo ou pas.

```
hasphoto = fbsona['has_video'].value_counts()

plt.figure(figsize=(15,3))
sns.barplot(x=hasphoto.index,y=hasphoto.values).set_title('has_video')
```

Résultat

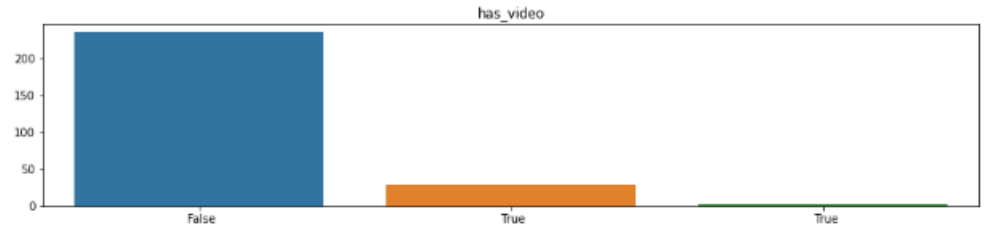


FIGURE 5.7 – Le graphique à barres : nombre de publication qui contient une vidé ou pas.

Interprétation : le diagramme signifie que la plus grande partie des publications de la page ne contient pas une vidéo.

* visualisation des types de publications.

```
 from wordcloud import WordCloud, STOPWORDS
plt.figure(figsize = (15,15))
stopwords = {'Length','dtype','Name','object','poste_type'}
wordcloud = WordCloud(
                      background_color = 'white',
                      colormap="copper_r",
                      stopwords=stopwords,
                      max_words = 100,
                      max_font_size = 120,
                      random_state = 42
                     ).generate(str(fbsona['post_type']))

#Plotting the word cloud
plt.imshow(wordcloud)
plt.title("SONATRACH type de publication", fontsize = 20)
plt.axis('off')
plt.show()
```

Résultat

FIGURE 5.8 – Wordcloud : les types de publications.

Interprétation : les communiqué de presse sont les publication qui existe plus dans la page, dans la deuxième classe on trouve "les nouvelle", et en troisième place motivation, après es roportage et félicitation.

Trouver les cases manquante

À l'aide de la fonction missingno
nous pouvons voire clairement les cases qui manque sans notre tableau des datas.

```
msno.matrix(fbsona)
```

Résultat Le résultat :

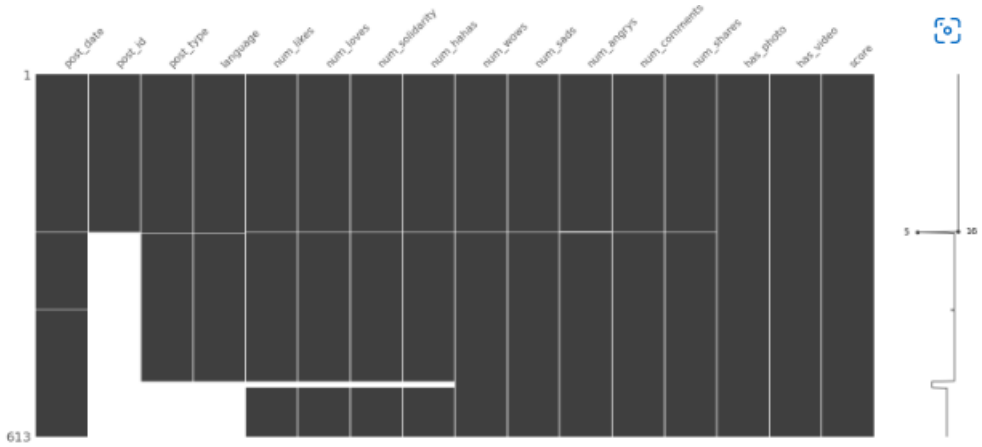


FIGURE 5.9 – schéma représente les cases vide du tableau

Interprétation : les cases en blanc dans le schéma représentent les cases vide dans le tableau des datas .

Regroupement des données

Le type de la publication et les différents ensembles de variables :

. Le type de la publication et la moyenne de nombre de réaction "j'aime" :

```
fbsona.groupby('post_type').num_likes.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

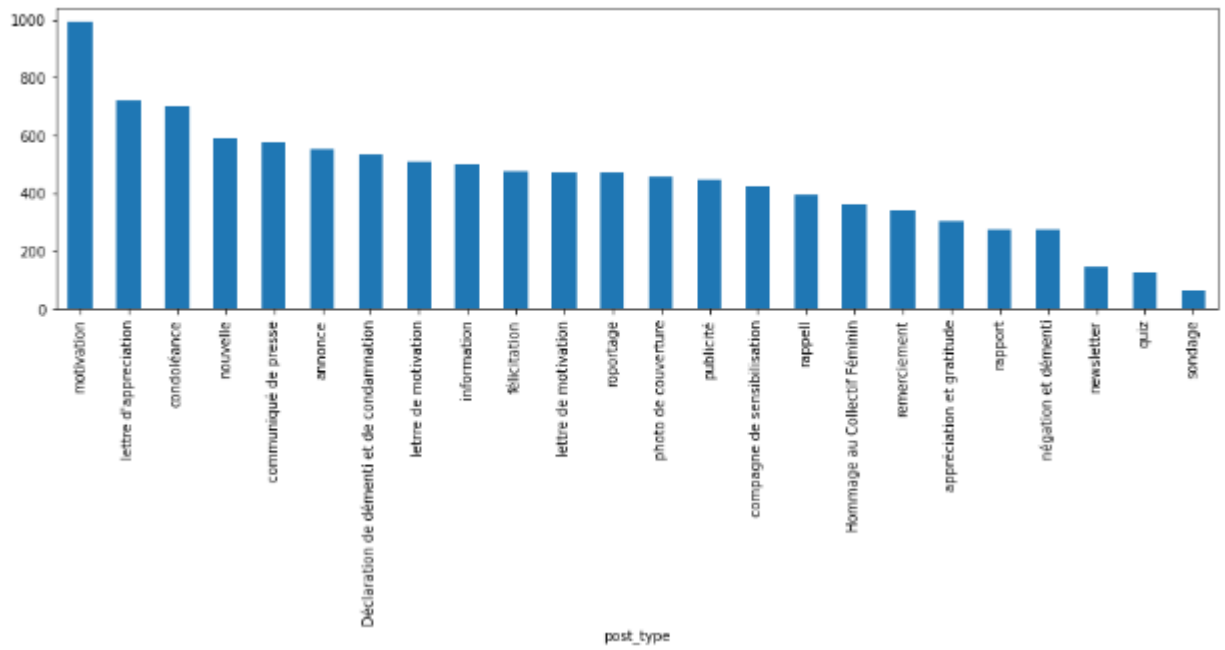


FIGURE 5.10 – Diagramme en bâtons :type de publication/nombre de j'aime

Interprétation : les publications de type "motivation" ont la plus grande moyenne ensuite on a "lettre de d'appréciation" et "condoléance", puis "nouvelle", "communiqué de presse","annonce" presque ont eu la même moyenne de nombre de réaction "j'aimee.

. Le type de la publication et la moyenne de nombre de réaction "j'adore" :

```
fbsona.groupby('post_type').num_loves.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

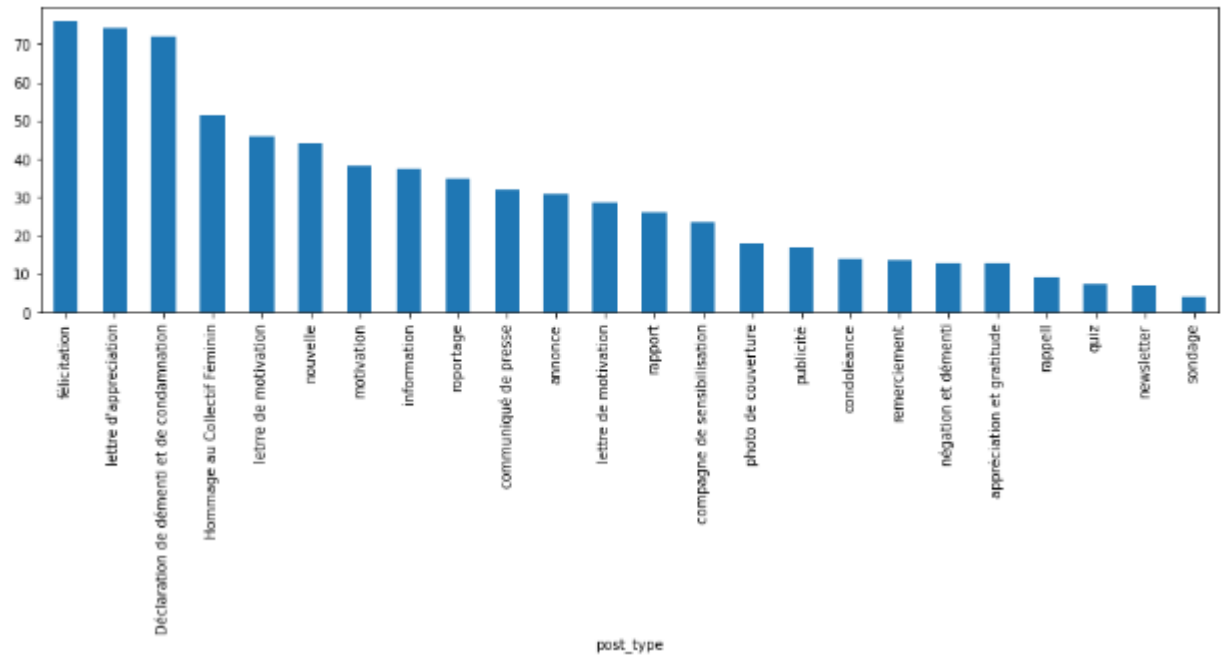


FIGURE 5.11 – Diagramme en bâtons :type de publication/nombre de j'adore

Interprétation En moyenne les types de publication qui adapte le plus grand nombres de réactions "j'adore" en premier sont "félicitation", "déclaration de démenti et de condamnation".

. Le type de la publication et la moyenne de nombre de réaction "solidaire" :

```
fbsona.groupby('post_type').num_solidarity.median()
.sort_values(ascending=False).head(613).plot.bar(figsize=(15,4))
```

Résultat

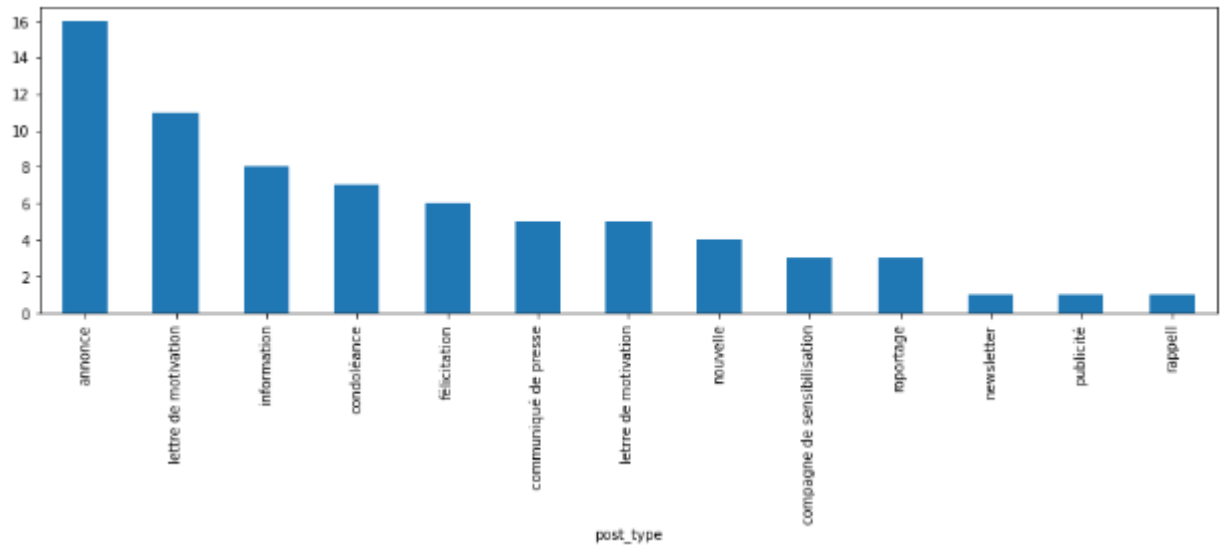


FIGURE 5.12 – Diagramme en bâtons :type de publication/nombre de solidaire

Interprétation En moyenne les types de publication qui adapte le plus grand nombres de réactions "solidaire" en premier sont "condoléance", "déclaration de démenti et de condamnation" et "motivation".

. Le type de la publication et la moyenne de nombre de réaction "rire" :

```
fbsona.groupby('post_type').num_hahas.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

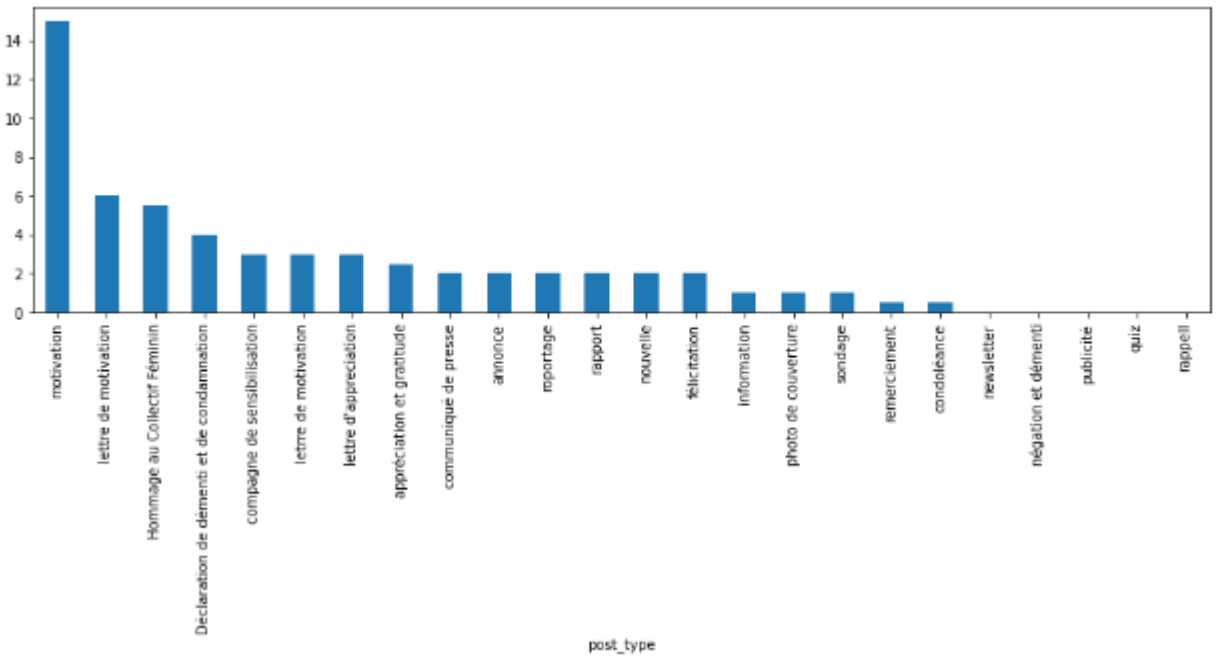


FIGURE 5.13 – Diagramme en bâtons :type de publication/nombre de rire

Interprétation En moyenne les types de publication qui adapte le plus grand nombres de réactions "rire" en premier est "motivation".

. Le type de la publication et la moyenne de nombre de réaction "étonné" :

```
fbsona.groupby('post_type').num_wows.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

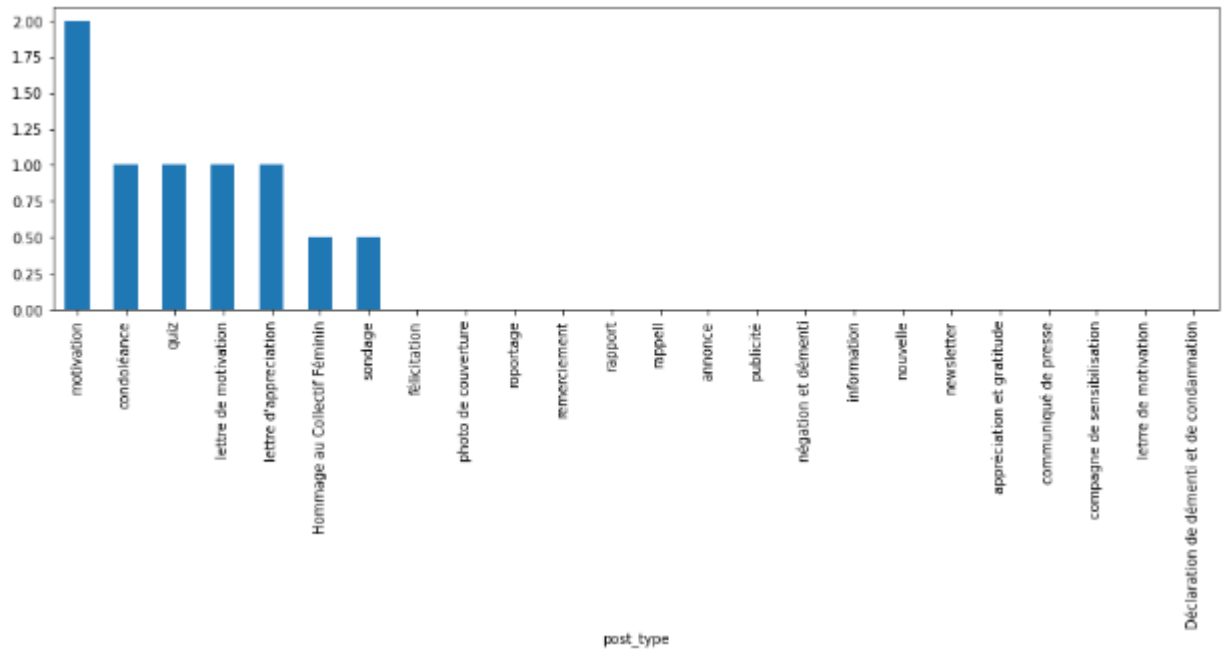


FIGURE 5.14 – Diagramme en bâtons :type de publication/nombre de étoné

Interprétation En moyenne les types de publication qui adapte le plus grand nombres de réactions "étonné" en premier sont "motivation".

. Le type de la publication et la moyenne de nombre de réaction "triste" :

```
fbsona.groupby('post_type').num_sads.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

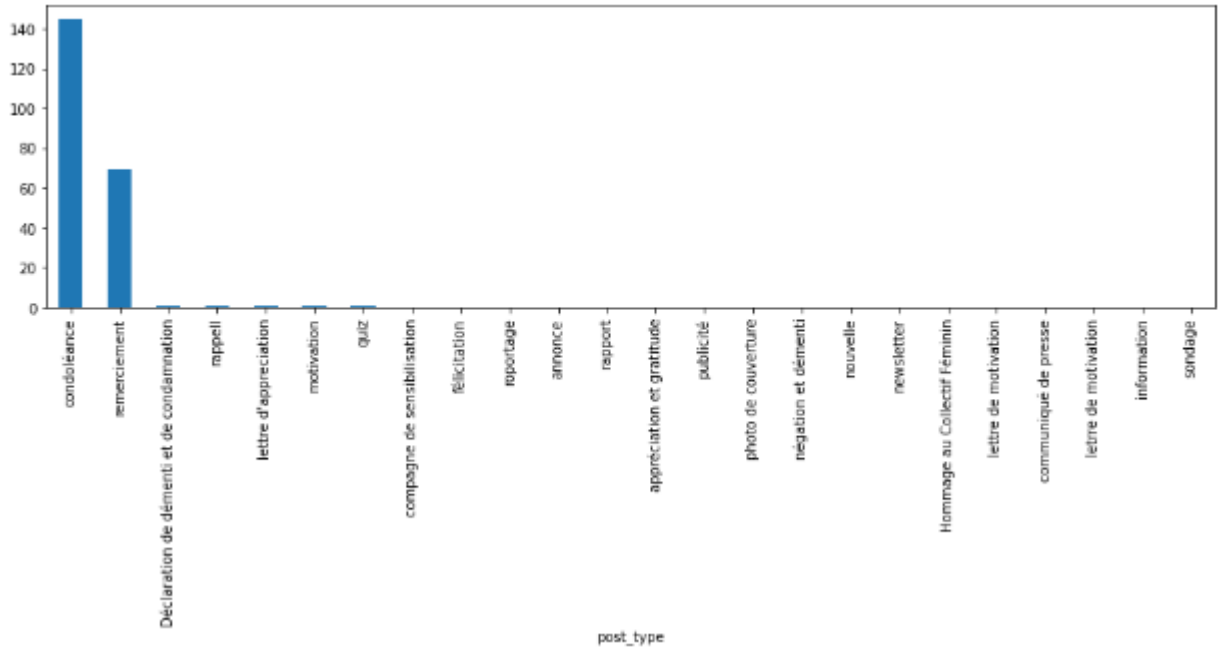


FIGURE 5.15 – Diagramme en bâtons :type de publication/nombre de triste

Interprétation En moyenne les types de publication qui adapte le plus grand nombres de réactions "triste" en premier sont "condoléance" et "remerciement".

. Le type de la publication et la moyenne de nombre de réaction "énervé" :

```
fbsona.groupby('post_type').num_angrys.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

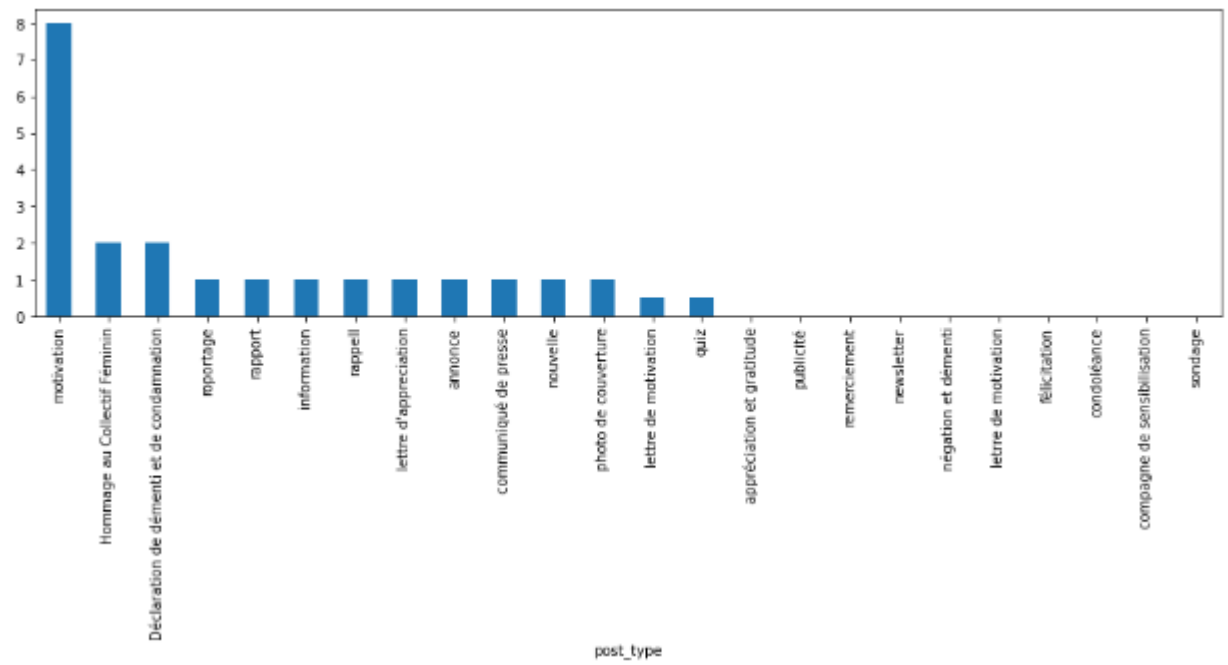


FIGURE 5.16 – Diagramme en bâtons :type de publication/nombre de énervé

Interprétation En moyenne les types de publication qui adapte le plus grand nombres de réactions "énervé" en premier est celle de type "motivation".

. Le type de la publication et la moyenne de nombre de commentaire :

```
fbsona.groupby('post_type').num_comments.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

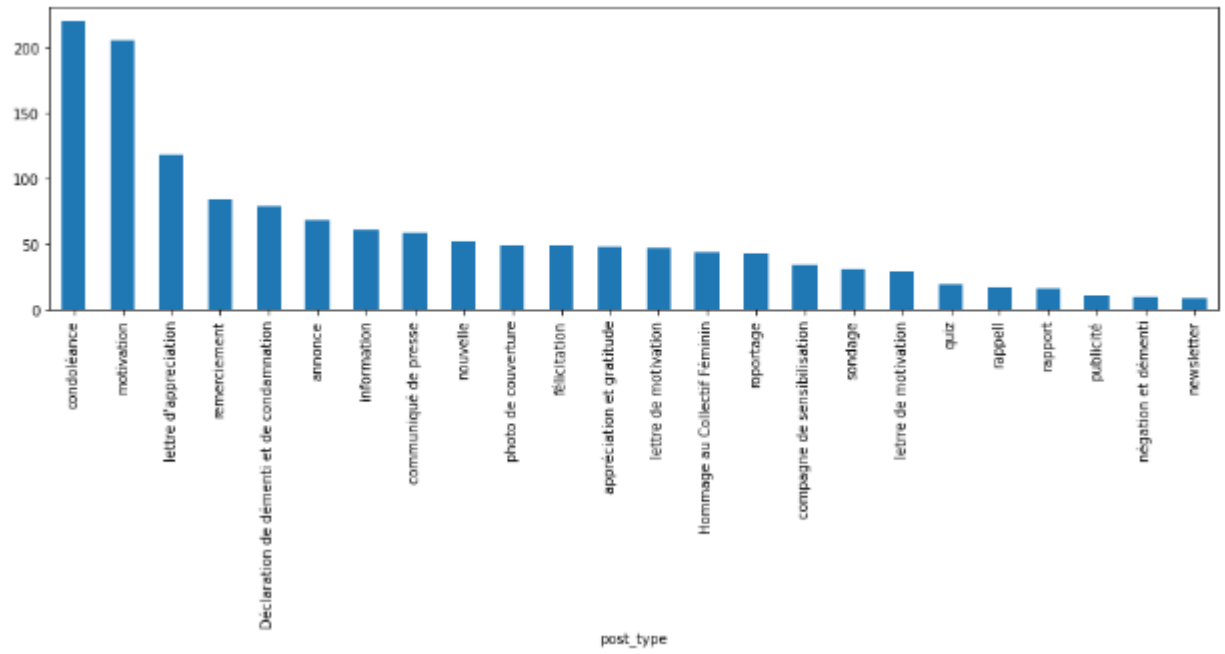


FIGURE 5.17 – Diagramme en bâtons :type de publication/nombre de commentaires

Interprétation En moyenne les types de publication qui adapte le plus grand nombres de commentaires en premier sont "condoléance", "motivation".

. Le type de la publication et la moyenne de nombre de partage :

```
fbsona.groupby('post_type').num_shares.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

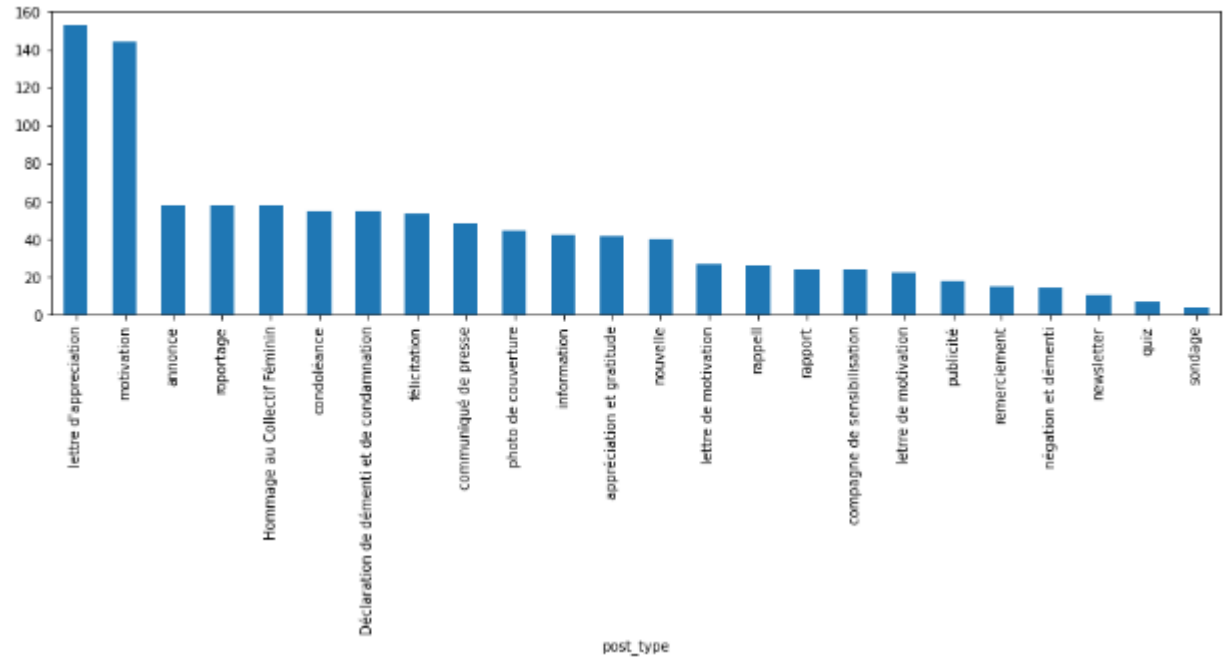


FIGURE 5.18 – Diagramme en bâtons :type de publication/nombre de partages

Interprétation En moyenne les types de publication qui adapte le plus grand nombres de partage en premier sont "lettre d'appréciation", "motivation".

. Le type de la publication et la moyenne de score :

```
fbsona.groupby('post_type').score.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

Interprétation

En moyenne les types de publication qui ont eu le meilleur score sont celle de type "motivation", "lettre d'appréciation" et "condoléance"

D'après cette analyse on conclue que les types de publication qui ont eu beaucoup de réaction sont "motivation", "lettre d'appréciation", "condoléance"

La langue et les différents ensembles de variables :

. La langue de la publication et la moyenne de nombre de réaction "j'aime" :

```
fbsona.groupby('language').num_likes.median()
.sort_values
.head(613).plot.bar(figsize=(15,4))
```

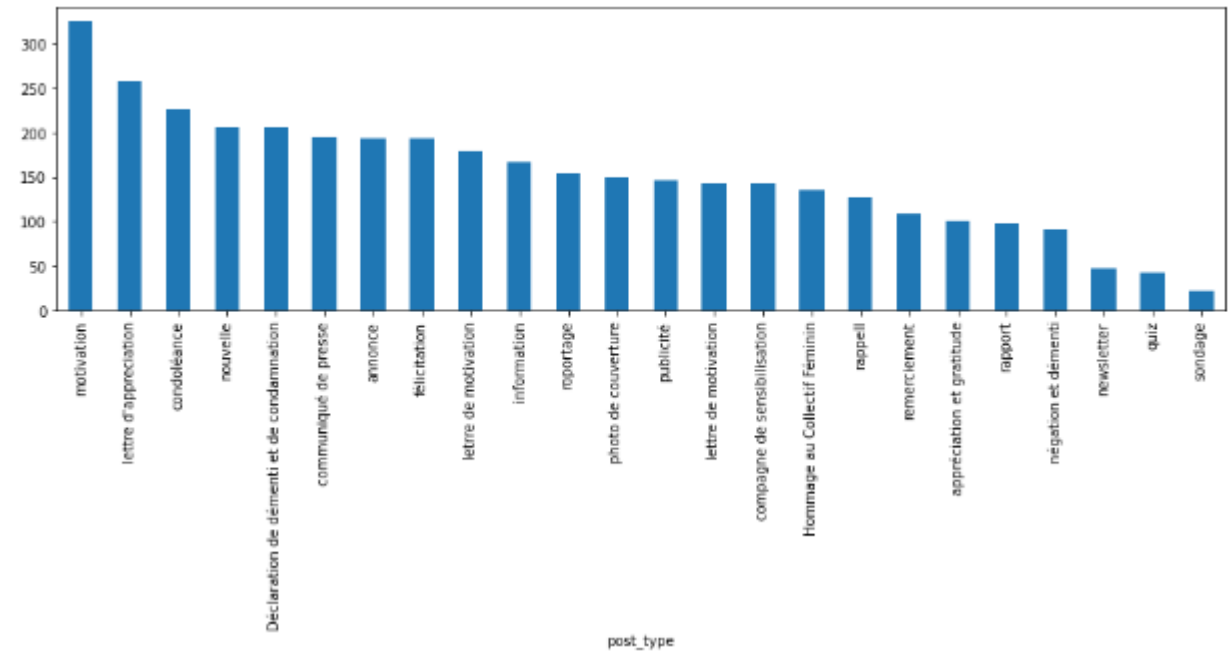


FIGURE 5.19 – Diagramme en bâtons :la langue/nombre de partage

Résultat

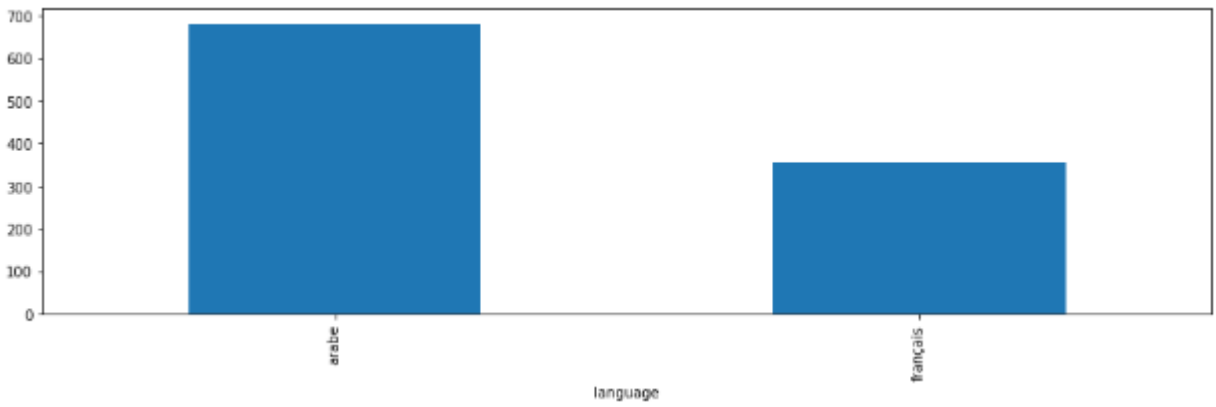


FIGURE 5.20 – Diagramme en bâtons :la langue/nombre de j'aime

Interprétation Il est clair que les publications en arabe ont plus de réaction "j'aime" que celle en français.

. La langue de la publication et la moyenne de nombre de réaction "j'adore" :

```
fbsona.groupby('language').num_loves.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

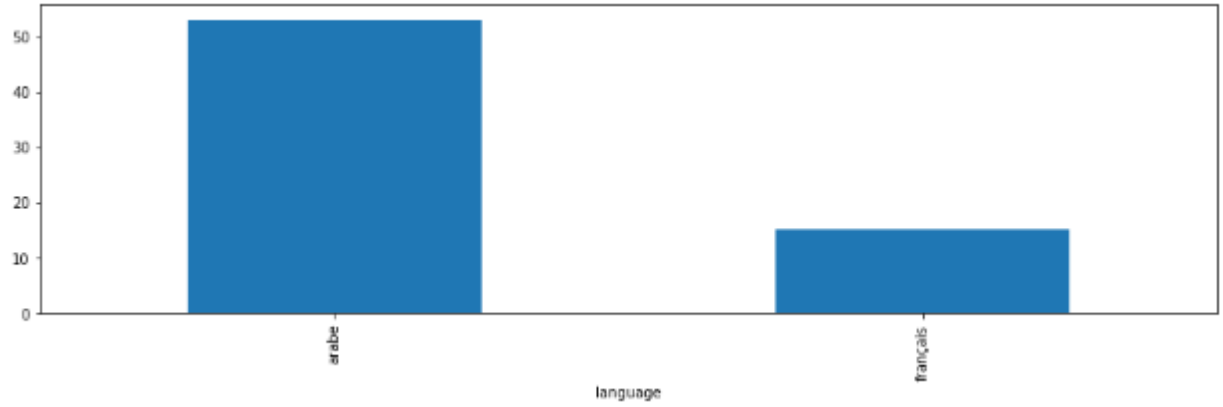


FIGURE 5.21 – Diagramme en bâtons :la langue/nombre de j'adore

Interprétation Il est clair que les publications en arabe ont beaucoup plus de réaction "j'adore" que celle en français.

. La langue de la publication et la moyenne de nombre de réaction "solidaire" :

```
fbsona.groupby('language').num_solidarity.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

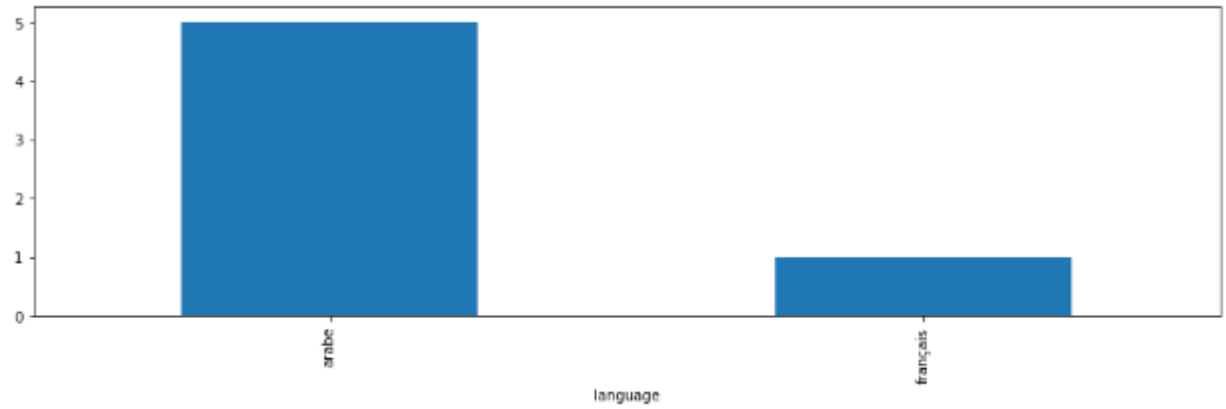


FIGURE 5.22 – Diagramme en bâtons :la langue/nombre de solidaire

Interprétation Il est clair que les publications en arabe ont beaucoup plus de réaction "solidaire" que celle en français.

. La langue de la publication et la moyenne de nombre de réaction "rire" :

```
fbsona.groupby('language').num_hahas.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

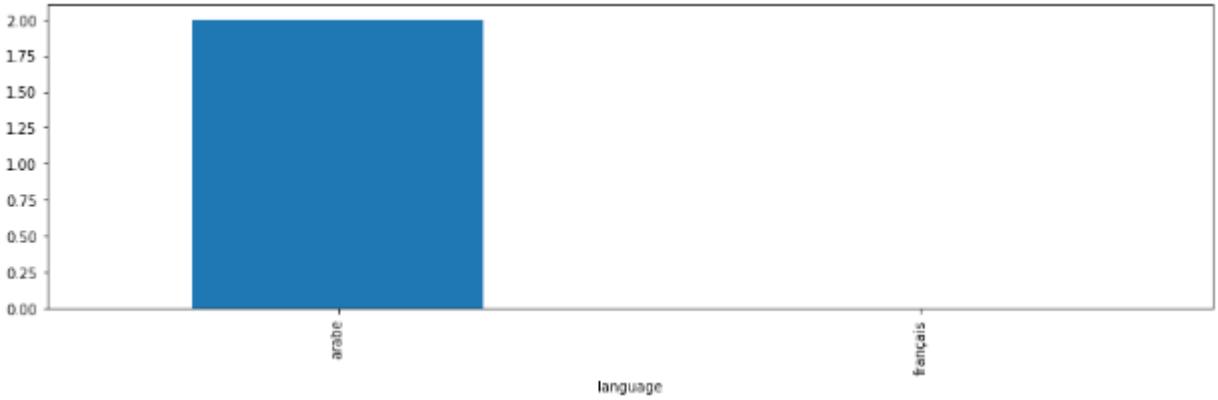


FIGURE 5.23 – Diagramme en bâtons :la langue/nombre de rire

Interprétation Il est clair que les publications en arabe ont plus de réaction "rire" que celle en français (presque toutes les réactions " rire" sont nulle pour les publications en français).

. La langue de la publication et la moyenne de nombre de réaction "triste" :

```
fbsona.groupby('language').num_sads.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

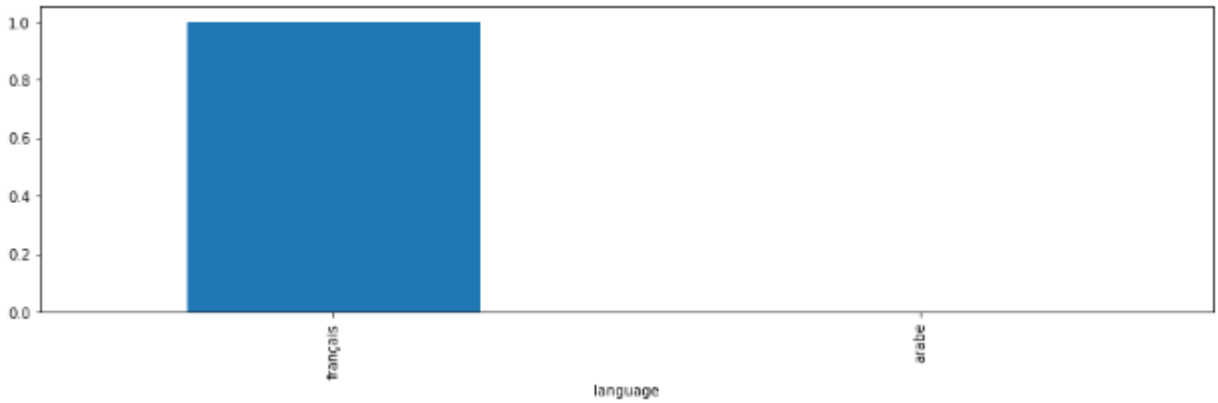


FIGURE 5.24 – Diagramme en bâtons :la langue/nombre de triste

Interprétation Il est clair que les publications en arabe ont plus de réaction "triste" que celle en français (presque tous les réaction "triste" sont nulle pour les publication en français).

. La langue de la publication et la moyenne de nombre de réaction "énervé" :

```
fbsona.groupby('language').num_angrys.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

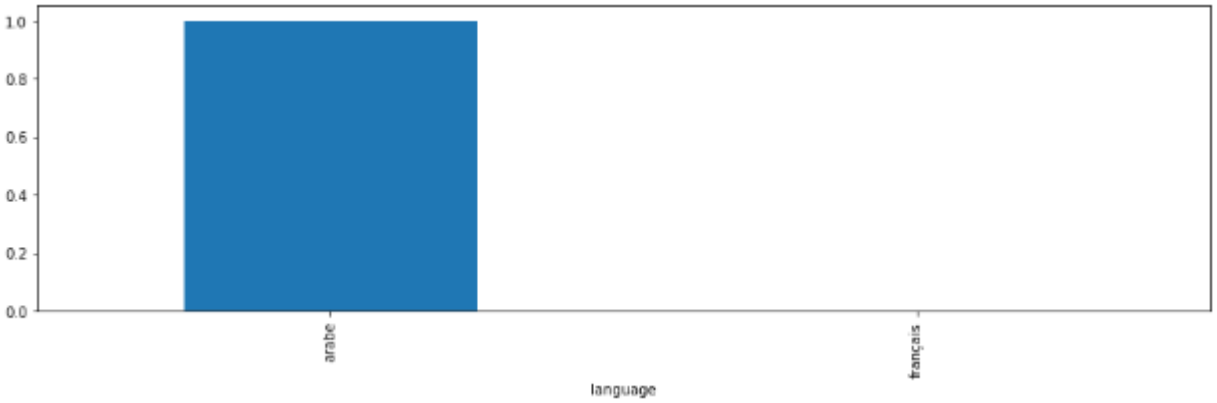


FIGURE 5.25 – Diagramme en bâtons :la langue/nombre de énervé

Interprétation Il est clair que les publications en arabe ont plus de réaction "énervé" que celle en français (presque tous les réaction "énervé" sont nulle pour les publication en français).

. La langue de la publication et la moyenne de nombre de commentaire :

```
fbsona.groupby('language').num_comments.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

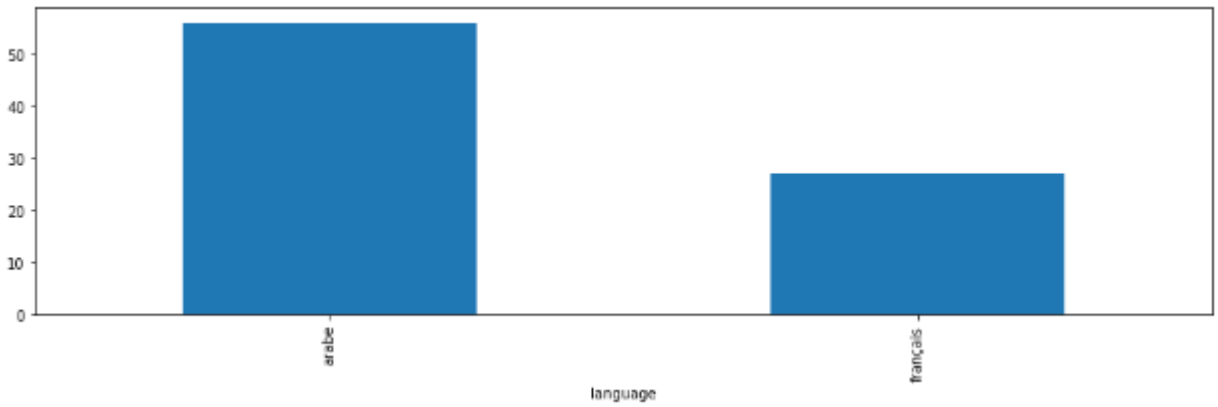


FIGURE 5.26 – Diagramme en bâtons :la langue/nombre de commentaires

Interprétation Il est clair que les publications en arabe ont plus de commentaire que celle en français.

. La langue de la publication et la moyenne de nombre de partages :

```
fbsona.groupby('language').num_shares.median()
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

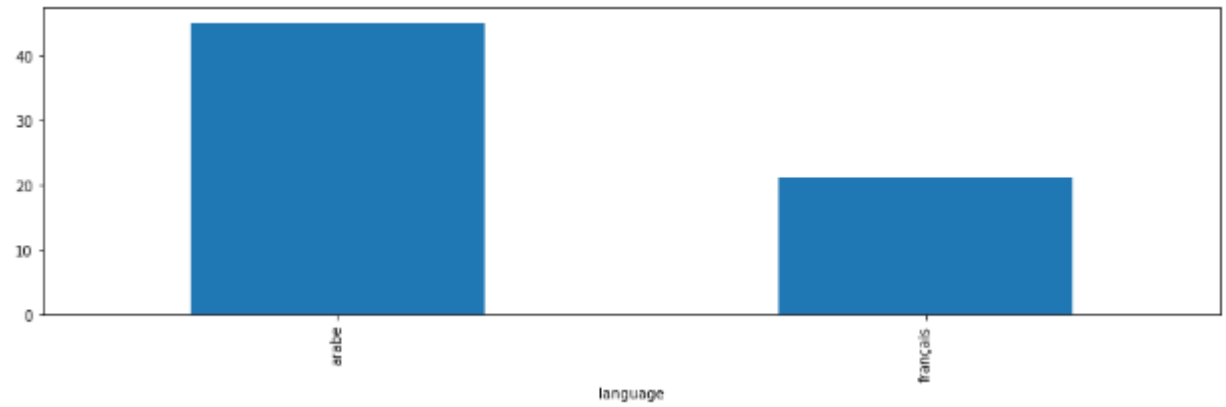


FIGURE 5.27 – Diagramme en bâtons :la langue/nombre de partage

Interprétation Il est clair que les publications en arabe sont les plus partagé que celle en français.

. La langue de la publication et la moyenne de score :

```
fbsona.groupby('language').score.median()
```

```
.sort_values(ascending=False)
.head(613).plot.bar(figsize=(15,4))
```

Résultat

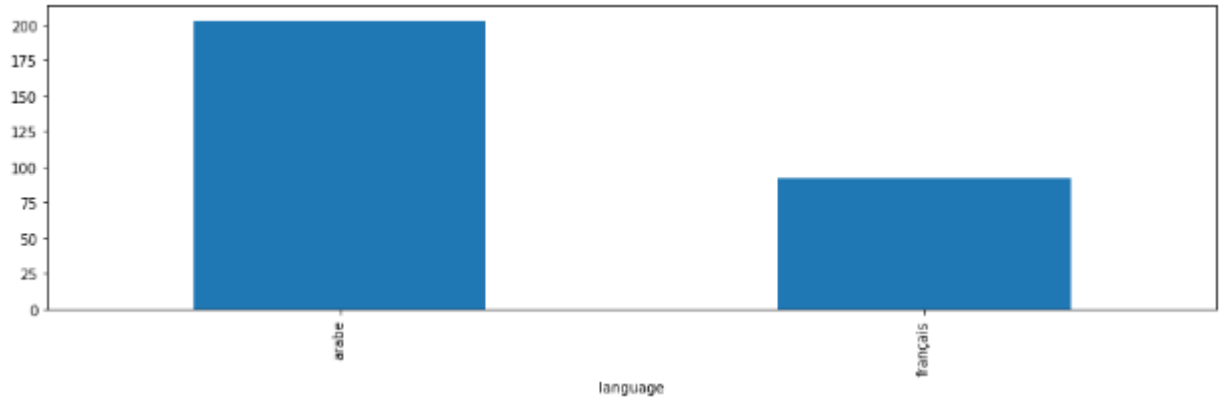


FIGURE 5.28 – Diagramme en bâtons :la langue/score

Interprétation Le meilleur score en moyenne est pour les publications en français.

D'après ces résultat en conclue que les publications en arabe sont les publication qui adaptent le plus grand nombre de réactions, commentaires et partage.

Corrélation :

. Relations potentielles entre variables du tableau

```
 mask = np.array(fbsona.corr())
mask[np.tril_indices_from(mask)] = False

fig, ax = plt.subplots()
fig.set_size_inches(20,10)
sns.heatmap(fbsona.corr(), mask = mask, vmax =
.8, square = True, annot = True)
```

Résultat

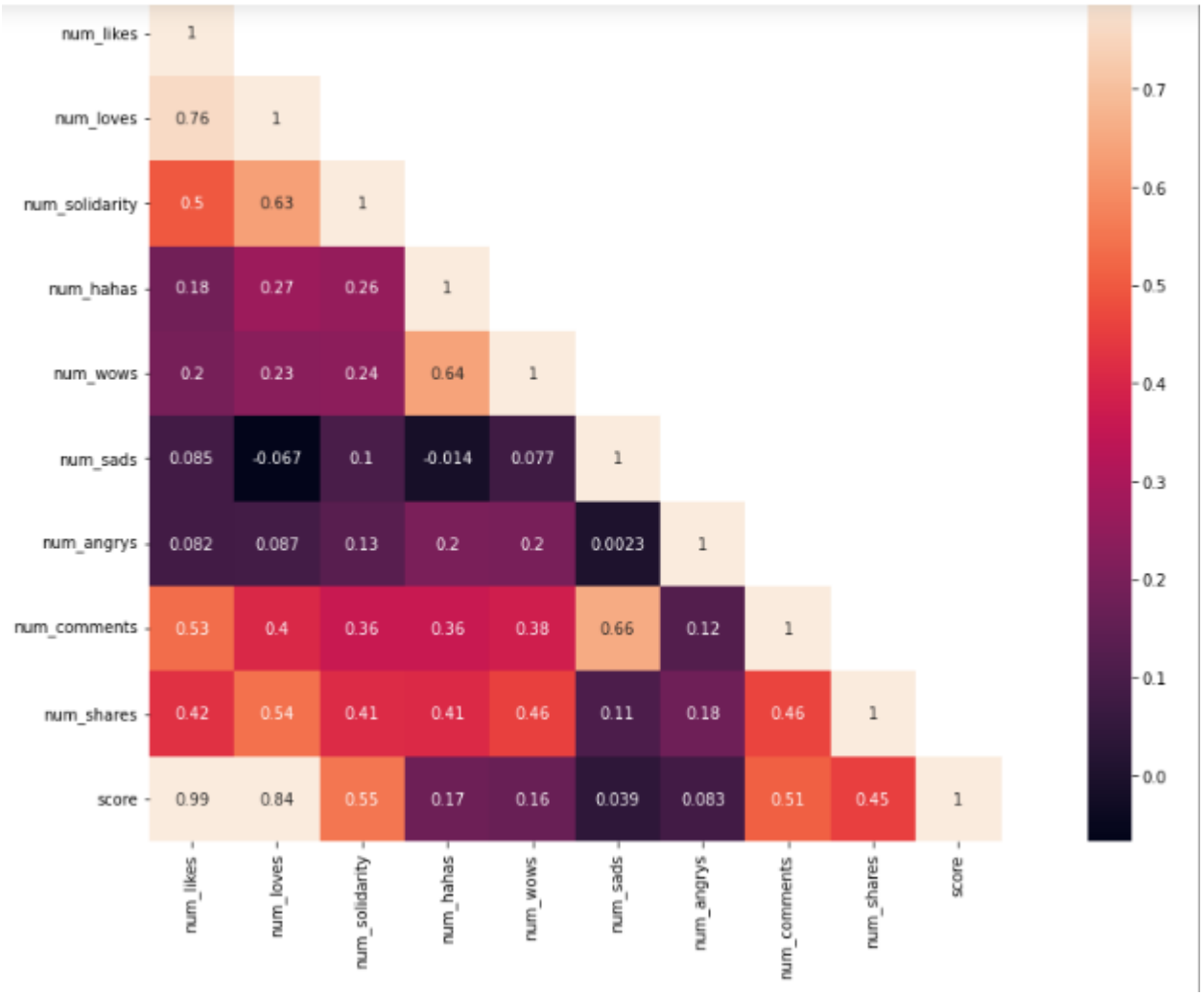


FIGURE 5.29 – Heatmap de corrélation entre les attributs de tableau

Interprétation D'après le heatmap on conclue :

— Le coefficient de corrélation entre le nombre de commentaires et le score est de 51% ce qui signifie une corrélation positive faible entre ces deux variables.

— Le coefficient de corrélation entre le nombre de partages et le score est de 45% ce qui signifie une corrélation positive faible entre ces deux variables.

— Le coefficient de corrélation entre le nombre de commentaires et réaction "j'aime" est de 53% ce qui signifie une corrélation positive faible entre ces deux variables.

— Le coefficient de corrélation entre le nombre de commentaires et réactions "triste" est de 66% ce qui signifie une corrélation positive faible entre ces deux variables.

— Le coefficient de corrélation entre le nombre de partages et réactions "j'adore" est de 54% ce qui signifie une corrélation positive faible entre ces deux va-

riables.
— Le coefficient de corrélation entre le nombre de partages et réactions "étonné" est de 46% ce qui signifie une corrélation positive faible entre ces deux variables.
— Le coefficient de corrélation entre le nombre de partages et réactions "solidaire" est de 41% ce qui signifie une corrélation positive faible entre ces deux variables.
— Le coefficient de corrélation entre le nombre de partages et réactions "rire" est de 41% ce qui signifie une corrélation positive faible entre ces deux variables.

D'après ces résultats on conclue que :

. les internautes expriment leurs sentiment sur les publications par des réactions, des commentaires et aussi par le partage.

. Les publication qui ont le nombre de réactions "j'aime" ou "triste" sont les publications qui ont le plus grand nombre de commentaires commentaire.

. Les publication qui ont le nombre de réaction "j'adore", "étonné", "solidaire", "rire" sont les publications qui ont le plus grand nombre de partages.

5.6 Clustering

5.6.1 Étapes de clustering à suivre

Regroupement de nos données procédez comme suit :
— Préparer les données.
— Créer une métrique de similarité.
— Exécutez l'algorithme de clustering.
— Interprétez les résultats et ajustez votre clustering.

Préparation des données Comme pour tout problème de ML, nous devons normaliser, mettre à l'échelle et transformer les données d'entité. Cependant, lors de la mise en cluster, nous devons également nous assurer que les données préparées nous permettons de calculer avec précision la similarité entre les exemples.

Normalisation

La normalisation est une technique souvent appliquée dans le cadre de la préparation des données pour l'apprentissage automatique. L'objectif de la normalisation est de modifier les valeurs des colonnes numériques dans l'ensemble de données pour utiliser une échelle commune, sans déformer les différences dans les plages de valeurs ni perdre d'informations.

L'objectif de la normalisation est d'éviter les données brutes et divers problèmes d'ensembles de données en créant de nouvelles valeurs et en maintenant une distribution générale ainsi qu'un ratio dans les données. En outre, il améliore également les performances et la précision des modèles d'apprentissage automatique à l'aide de diverses techniques et algorithmes.

Différents types de normalisation dans l'apprentissage automatique

Normalization Technique	Formula
Linear Scaling	x ′ = (x − x m i n) / (x m a x − x m i n)
Clipping	if x > max, then x' = max. if x < min, then x' = min
Log Scaling	x' = log(x)
Z-score	x' = (x - μ) / σ

FIGURE 5.30 – Types de normalisation des données

tel que

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Silhouette$_s$*core* : $Silhouettefaitrfrenceunemthoded'interprtationetdevalidationdelacohre$
Le score de silhouette pour un point de données i est donné par la relation suivante :

$$S_i = \frac{b_i - a_i}{\max(b_i, a_i)}$$

où,
b_i : est la distance inter-cluster définie comme la distance moyenne au cluster le plus proche du point de données i, sauf qu'il fait partie de C_k :

$$b_i = \min_{k \neq i} \frac{1}{\|C_k\|} \sum_{j \in C_k} d(i, j)$$

a_i : est la distance intra-cluster définie comme la distance moyenne à tous les autres points du cluster dont il fait partie :

$$a_i = \frac{1}{\|C_i\| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

Le score global de silhouette pour l'ensemble de données complet peut être calculé comme la moyenne du score de silhouette pour tous les points de données dans l'ensemble de données. Comme on peut le voir d'après la formule, le score de silhouette se situerait toujours entre -1 et 1. 1 représentant un meilleur regroupement.

Méthode du coude (Elbow method)

Dans l'analyse de clustering, la méthode du coude est une heuristique utilisée pour déterminer le nombre de clustering dans un ensemble de données. La méthode consiste à tracer la variation expliquée en fonction du nombre de clusters et à choisir le coude de la courbe comme nombre de clusters à utiliser. La même méthode peut être utilisée pour choisir le nombre de paramètres dans d'autres modèles basés sur les données, tels que le nombre de composants principaux pour décrire un ensemble de données.

Analyse en composantes principales (ACP) pour l'apprentissage automatique

L'analyse en composantes principales (ACP) est l'un des algorithmes d'apprentissage automatique non supervisé les plus couramment utilisés dans diverses applications : analyse exploratoire des données, réduction de la dimentionnalité, compression des informations, débruitage des données et bien plus encore.

Exploration de Text

dans cette partie de notre travaille on a essayé d'explorer et visualiser les publication de la page apartir de de text des publication (post message) ici on a utilisée un fichier CSV qui contient les colonnes suivantes :

la date : c'est la date de publication.

post : c'est le text de la publication.

afin d'analyser les réaction de la page facebook on veut savoir les sujet aboreé par la page.

Chargement des packadges (bibliothèque)

```
import numpy as np
import pandas as pd
from IPython.display import display
from tqdm import tqdm
from collections import Counter
import ast

import matplotlib.pyplot as plt
```

```
import matplotlib.mlab as mlab
import seaborn as sb

from sklearn.feature_extraction.text import CountVectorizer
from textblob import TextBlob
import scipy.stats as stats
import nltk



from sklearn.decomposition import TruncatedSVD
from sklearn.decomposition import LatentDirichletAllocation
from sklearn.manifold import TSNE

from bokeh.plotting import figure, output_file, show
from bokeh.models import Label
from bokeh.io import output_notebook

output_notebook()
```

Téléchargement de packadge (nltk)

```
nltk.download()
```

Téléchargement de tableau des données

```
datafile = "C:/Users/zineb/Downloads/Analyse Sentiments — sonatrachposts (5).csv"
raw_data = pd.read_csv(datafile, parse_dates=[0], infer_datetime_format=True)

reindexed_data = raw_data['post']
reindexed_data.index = raw_data['date']

raw_data.head()
```

Résultat

	date	post
0	2022-06-02	#sonatrach #visits_visits_and_inspected,\n# Mo...
1	2022-06-01	#sonatrach \n#Sponsorship of the Mediterranean...
2	2022-06-01	#press release \nSonatrach announces the start...
3	2022-05-29	#press release #Starting_Operation_Center_Supe...
4	2022-05-28	#press release \n#Signing_a_new_contract_in_th...

FIGURE 5.31 – tableau des données textuelle

Les mots les plus utiliser dans les publication (après le nettoyage.

```
    vectorized_headlines = count_vectorizer.fit_transform(text_data.values)
    vectorized_total = np.sum(vectorized_headlines, axis=0)
    word_indices = np.flip(np.argsort(vectorized_total)[0,:], 1)
    word_values = np.flip(np.sort(vectorized_total)[0,:],1)

    word_vectors = np.zeros((n_top_words, vectorized_headlines.shape[1]))
    for i in range(n_top_words):
        word_vectors[i,word_indices[0,i]] = 1

    words = [word[0].encode('ascii').decode('utf-8') for
             word in count_vectorizer.inverse_transform(word_vectors)]

    return (words, word_values[0,:n_top_words].tolist()[0])
```

```
count_vectorizer = CountVectorizer(stop_words='english')
words, word_values = get_top_n_words(n_top_words=15,
                                     count_vectorizer=count_vectorizer,
                                     text_data=reindexed_data)

fig, ax = plt.subplots(figsize=(16,8))
ax.bar(range(len(words)), word_values);
ax.set_xticks(range(len(words)));
ax.set_xticklabels(words, rotation='vertical');
ax.set_title('Top words in headlines dataset (excluding stop words)');
ax.set_xlabel('Word');
ax.set_ylabel('Number of occurences');
plt.show()
```

Résultat

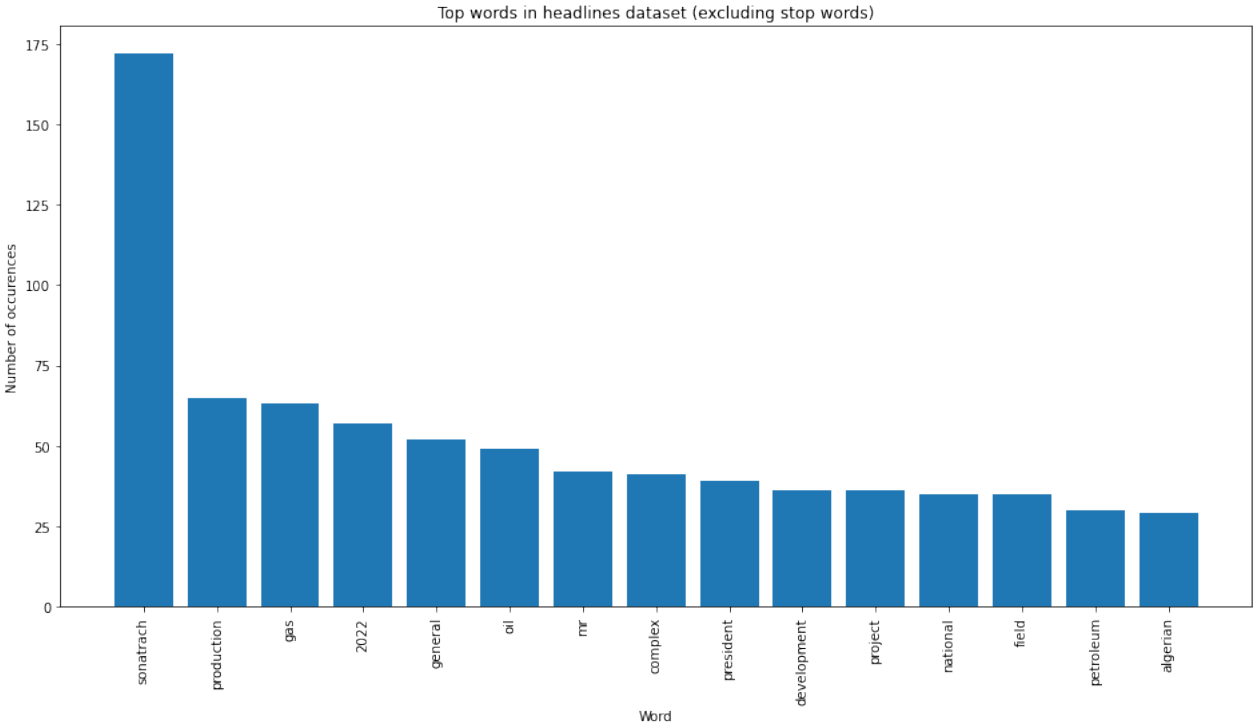


FIGURE 5.32 – Résultat des mots plus répéter dans les publications

Les mots les plus utiliser dans les publication (avant nettoyage).

```
import base64

# Plotly imports
import plotly.offline as py
py.init_notebook_mode(connected=True)
import plotly.graph_objs as go
import plotly.tools as tls


# Other imports
from collections import Counter
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer

from matplotlib import pyplot as plt
%matplotlib inline
```

```
all_words = sona['post'].str.split(expand=True).unstack().value_counts()
data = [go.Bar(
            x = all_words.index.values[2:50],
```

```
            y = all_words.values[2:50],
            marker= dict(colorscale='Jet',
                         color = all_words.values[2:100]
                        ),
            text='Word counts'
    )]

layout = go.Layout(
    title='Top 50 (Uncleaned) Word frequencies in the training dataset'
)

fig = go.Figure(data=data, layout=layout)

py.iplot(fig, filename='basic-bar')
```

résultats

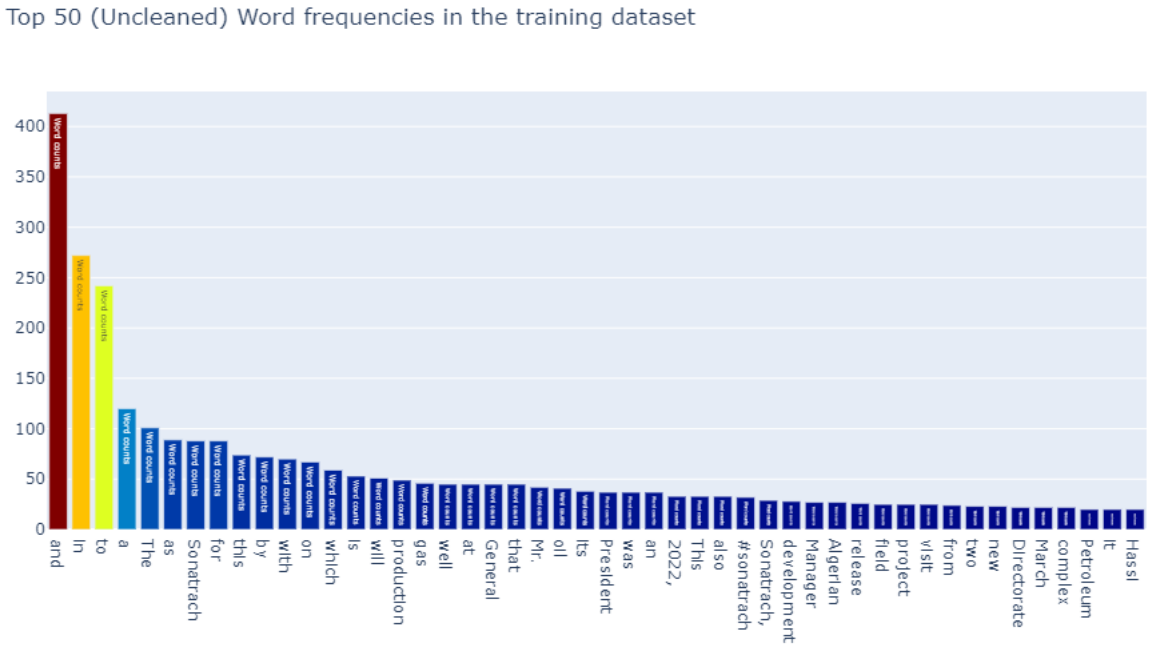


FIGURE 5.33 – Top 50 fréquences de mots (non nettoyées) dans l'ensemble de données d'entraînement

Le wordcloud des mots répéter dans les publications

```
from wordcloud import WordCloud, STOPWORDS , ImageColorGenerator
import pandas as pd
import matplotlib.pylab as plt
from PIL import Image
```

```
import numpy as np


stopwords = set(STOPWORDS)


#we will read the text
data_file = pd.read_csv("C:/Users/zineb/Downloads/Analyse Sentiments — sonatrachposts (5).csv")
#wordcloud
wordcloud = WordCloud(stopwords = stopwords , width=1600 , height=800,background_color="White",colorma
plt.figure(figsize=(20,10),facecolor='k')
plt.imshow(wordcloud,interpolation='bilinear')
plt.axis('off')
plt.tight_layout (pad=0)


plt.show()
```

Résultat

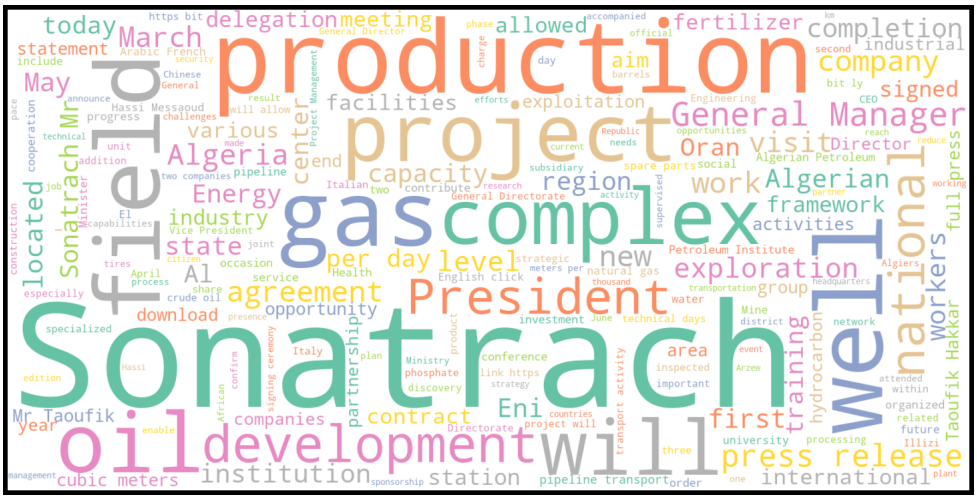


FIGURE 5.34 – Wordcloud des mots plus répéter dans les publications

Interprétation

d'après les trois résultats président on remarquent que le terme le plus fréquent dans les publication de la page est "Sonatrach" puis le mot "production" et "gaz" après "oil", "field","well" et "project" et les autres termes.

5.6.2 Chargement de L'ensemble de Données

```
import numpy as np  linear algebra
```

```
import pandas as pd data processing, CSV file I/O (e.g. pd.read_csv)
data = pd.read_csv("C:/Users/zineb/Downloads/Analyse Sentiments — sonadataset (32).csv")
df = data.copy()
pd.set_option('display.max_row',df.shape[0])
pd.set_option('display.max_column',df.shape[1])
df.head()
```

Résultat

language	num_likes	num_loves	num_solidarity	num_hahas	num_wows	num_sads	num_angrys	num_comments	num_shares	has_photo	has_video	scoring
arabe	855.0	104.0	11.0	1.0	9.0	1.0	6.0	234.0	56.0	True	False	318
arabe	604.0	47.0	3.0	8.0	0.0	1.0	22.0	115.0	48.0	True	False	206
arabe	2500.0	336.0	24.0	1.0	5.0	0.0	3.0	525.0	224.0	True	False	938
arabe	898.0	93.0	8.0	2.0	1.0	0.0	2.0	82.0	78.0	True	False	324
arabe	1800.0	177.0	17.0	3.0	2.0	1.0	3.0	246.0	195.0	True	False	646

FIGURE 5.35 – tableau des données

5.6.3 Exploration des données et préparation des données

Identification des valeurs manquantes dans l'ensemble de données : Afin de garder une meilleure vue d'ensemble, nous allons visualiser les valeurs manquantes par colonnes. L'objectif est de tracer les variables avec plus et moins de 10% de valeurs manquantes.

```
def missing_data(data, thresh=1, color='black', edgecolor='black', width=15, height=3):

plt.figure(figsize=(width,height))
percentage=(data.isnull().mean())*100
percentage.sort_values(ascending=False)[:10].plot.bar(color=color, edgecolor=edgecolor)
plt.axhline(y=thresh, color='r', linestyle='—')
plt.title('Missing values percentage per column', fontsize=20, weight='bold' )
plt.text(len(data.isnull()[:10].sum()/len(data))/15, thresh +8,
         f'Columns with more than {thresh}\% missing values', fontsize=12, color='darkblue',
         ha='left' ,va='top')
  plt.text(len(data.isnull()[:10].sum()/len(data))/15, thresh —3,
         f'Columns with less than {thresh}\% missing values', fontsize=12, color='green',
         ha='left' ,va='top')
plt.xlabel('Columns', size=15, weight='bold')
plt.ylabel('Missing values percentage')

return plt.show()
missing_data(df, 10, color=sns.color_palette('BrBG',10))
```

résultat

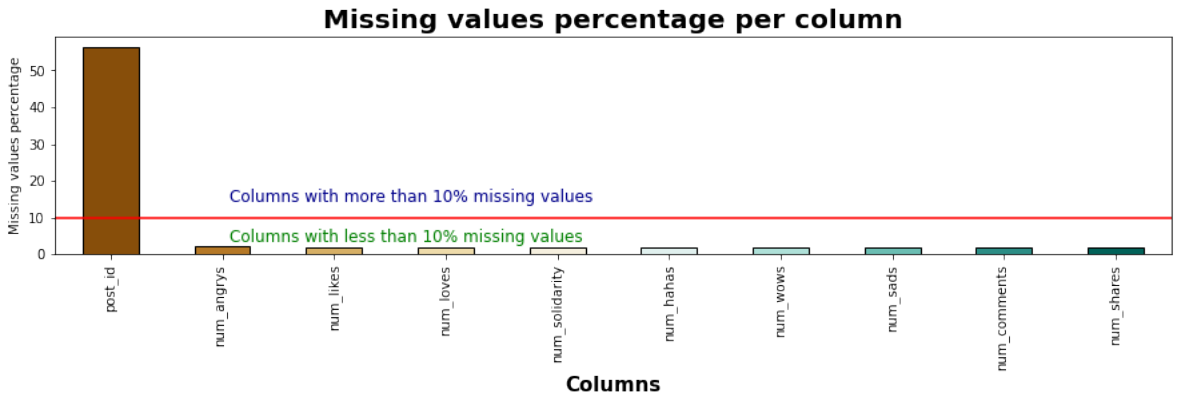


FIGURE 5.36 – Pourcentage de valeurs manquantes par colonne

Le seuil pour les valeurs manquantes est représenté par la ligne rouge, dans ce qui suit, toutes les colonnes avec plus de 10% de valeurs manquantes seront supprimées de la trame de données. Nous vérifierons la forme de l'ensemble de données à comparer avec le nouveau dataframe après avoir supprimé les colonnes.

```
df.shape
```

Résultat (613,16)

Déposez
les colonnes comme ci-dessus. Créez une nouvelle trame de données appelée "df1"

```
df1 = df.dropna(how='any',axis=1,thresh=df.shape[0]*0.9)
df1.shape
```

Résultat (613,15)

Au total,
nous avons supprimé 1 variables de la base de données d'origine qui est "post_id".
L'étape suivante consiste à remplacer
les valeurs nulles dans les variables par leurs valeurs moyennes respectives.

```
df2 = df1.fillna(df1.mean())
```

Vérification des doublons : Ensuite,
nous vérifierons s'il y a des données en double dans notre ensemble de données.
Nous allons créer un nouveau bloc de données appelé "duplicateRowsdf" pour vérifier et stocker les lignes dupliquées dans notre ensemble de données. Dans une prochaine étape, nous vérifierons la forme des valeurs dupliquées.

```
duplicateRowsdf = df2[df2.duplicated()]
print(duplicateRowsdf)
```

Résultat

	post_date	post_type	language	num_likes	num_loves	num_solidarity	\
526	06/10/2019	nouvelle	arabe	539.304493	48.510815	4.299501	

	num_hahas	num_wows	num_sads	num_angrys	num_comments	num_shares	\
526	5.630616	0.886855	7.710483	1.421667	78.427621	60.980033	

	has_photo	has_video	scoring
526	True	False	1

o

FIGURE 5.37 – duplicateRowsdf

Interprétation Le résultat nous donne que y a une seule ligne en double (526). Ensuite, nous appliquerons la fonction describe() pour obtenir les informations de résumé et de distribution sur nos variables catégorique.

```
df2.describe(include='object')
```

	post_date	post_type	language	has_photo	has_video
count	611	612	612	613	613
unique	526	23	2	6	4
top	12/12/2019	nouvelle	arabe	True	False
freq	4	231	539	531	482

FIGURE 5.38 – Description des variables catégoriques

Nous supprimons la variable : "post_id" et "post_date" car il n'y a pas de valeurs utilent dans ces variables pour notre analyse. (la colone "post_id" est déja supprimer)

suppréssion de la colone "poste_date"

```
df2 = df2.drop('post_date', axis=1)
```

Préparer l'ensemble de données pour appliquer l'algorithme et les modèles de clustering La 1ère étape consiste à diviser le dataframe en variables indépendantes et en données cibles : Nous allons créer "X" incluant les variables indépendantes, et la variable "y" incluant uniquement la variable cible "scoring"

```
X = df2.drop('scoring',axis=1)
y = df2[['scoring']]
```

Pour les variables indépendantes du dataframe "X", nous utilisons la méthode get_dummies pour préparer les données catégorielles(codage des variables catégorielle en valeurs numérique).

```
X = pd.get_dummies(data=X,columns=["post_type","language", "has_photo","has_video"])
```

Affichage de résultats après le codage

```
X.info()
```

Résultat

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 613 entries, 0 to 612
Data columns (total 44 columns):
 #   Column                                                Non-Null Count  Dtype
---  ------                                                --------------  -----
 0   num_likes                                             613 non-null    float64
 1   num_loves                                             613 non-null    float64
 2   num_solidarity                                        613 non-null    float64
 3   num_hahas                                             613 non-null    float64
 4   num_wows                                              613 non-null    float64
 5   num_sads                                              613 non-null    float64
 6   num_angrys                                            613 non-null    float64
 7   num_comments                                          613 non-null    float64
 8   num_shares                                            613 non-null    float64
 9   post_type_Déclaration de démenti et de condamnation   613 non-null    uint8
 10  post_type_Hommage au Collectif Féminin                613 non-null    uint8
 11  post_type_annonce                                     613 non-null    uint8
 12  post_type_appréciation et gratitude                   613 non-null    uint8
 13  post_type_communiqué de presse                        613 non-null    uint8
 14  post_type_compagne de sensibilisation                 613 non-null    uint8
 15  post_type_condoléance                                 613 non-null    uint8
 16  post_type_félicitation                                613 non-null    uint8
 17  post_type_information                                 613 non-null    uint8
 18  post_type_letrre de motivation                        613 non-null    uint8
 19  post_type_lettre d'appreciation                       613 non-null    uint8
 20  post_type_lettre de motivation                        613 non-null    uint8
 21  post_type_motivation                                  613 non-null    uint8
 22  post_type_newsletter                                  613 non-null    uint8
 23  post_type_nouvelle                                    613 non-null    uint8
 24  post_type_négation et démenti                         613 non-null    uint8
 25  post_type_photo de couverture                         613 non-null    uint8
 26  post_type_publicité                                   613 non-null    uint8
 27  post_type_quiz                                        613 non-null    uint8
 28  post_type_rapport                                     613 non-null    uint8
 29  post_type_remerciement                                613 non-null    uint8
 30  post_type_roportage                                   613 non-null    uint8
 31  post_type_sondage                                     613 non-null    uint8
 32  language_arabe                                        613 non-null    uint8
 33  language_français                                     613 non-null    uint8
```

```
 34  has_photo_          False                    613 non-null    uint8
 35  has_photo_          Flase                    613 non-null    uint8
 36  has_photo_         False                     613 non-null    uint8
 37  has_photo_         True                      613 non-null    uint8
 38  has_photo_        False                      613 non-null    uint8
 39  has_photo_ False                             613 non-null    uint8
 40  has_video_         True                      613 non-null    uint8
 41  has_video_  True                             613 non-null    uint8
 42  has_video_ False                             613 non-null    uint8
 43  has_video_ True                              613 non-null    uint8
dtypes: float64(9), uint8(35)
memory usage: 64.2 KB
```

Pour préparer la frame de données "X" pour un traitement ultérieur, nous allons mettre les données à l'échelle : Nous utiliserons le MinMaxScaler importé de la bibliothèque sklearn. Nous allons créer un nouveau dataframe appelé "df_scaled".

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()

  df_scaled = scaler.fit_transform(X)
  df_scaled = pd.DataFrame(df_scaled)
  df_scaled.head()
```

Résultat après scalarésation

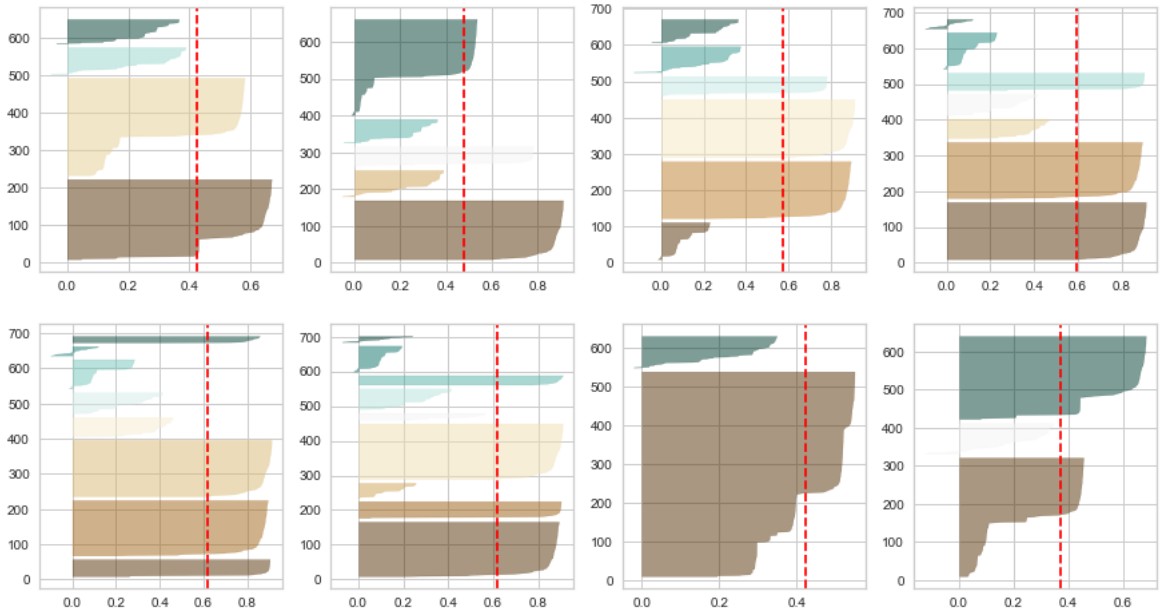


FIGURE 5.39 – Tableau des données après MinMaxScaler

La dimention de nouveau tableau après scalarésation

```
df_scaled.shape
```

Résultat on a 613 ligne et 44 colonne. (613,44)
*Modèles d'apprentissage automatique(Machine Learning)
A) Regroupement K-means Dans un premier temps, nous devrons choisir le nombre de clusters K. Nous utiliserons la méthode Elbow pour les déterminer.

Application de la méthode Elbow pour trouver le meilleur nombre de clusters

Dans ce qui suit, nous utiliserons le score d'inertie et de silhouette pour évaluer le meilleur modèle K-means a considérer pour l'interprétation : L'inertie mesure à quel point un ensemble de données a été regroupé par K-Means. Il est calculé en mesurant la distance entre chaque point de données et son centre de gravité, en élevant cette distance au carré et en additionnant ces carrés sur un cluster. Un bon modèle est un modèle à faible inertie ET à faible nombre de clusters (K). Cependant, il s'agit d'un compromis car à mesure que K augmente, l'inertie diminue.

À considérer pour l'interprétation du Silhouette Score : Le score de silhouette est une mesure de la cohésion et de la séparation des clusters. Le score de silhouette se situe dans la plage [-1,1] Le score de silhouette de 1 signifie que les grappes sont très denses et bien séparées et ont une cohésion et une séparation fortes. Le score o signifie que les clusters se chevauchent. Le score inférieur à 0 signifie que les données appartenant aux clusters peuvent être fausses/incorrectes.

Ensuite, nous analyserons le score Silhouette à l'aide de la méthode Silhouette Visualizer pour décider combien de clusters donnent le meilleur score. Nous installerons également sklearn. Utils et -Ubalanced-learn à partir de la bibliothèque sklearn.

téléchargement de package

```
pip install yellowbrick
```

Téléchargement des bibliothèques nécessaire pour le k-means

```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from yellowbrick.cluster import SilhouetteVisualizer
```

Visualisation de silhouette score

```
    fig, ax = plt.subplots(2,4, figsize=(15,8))
    for i in [2,3,4,5,6,7,8,9]:
       km = KMeans(n_clusters=i, init = 'k-means++', n_init=10, max_iter=100, random_state=42)
          q, mod = divmod(i, 4)

          visualizer = SilhouetteVisualizer(km, colors='BrBG', ax=ax[q-1][mod])
          visualizer.fit(df_scaled)
```

Résultat

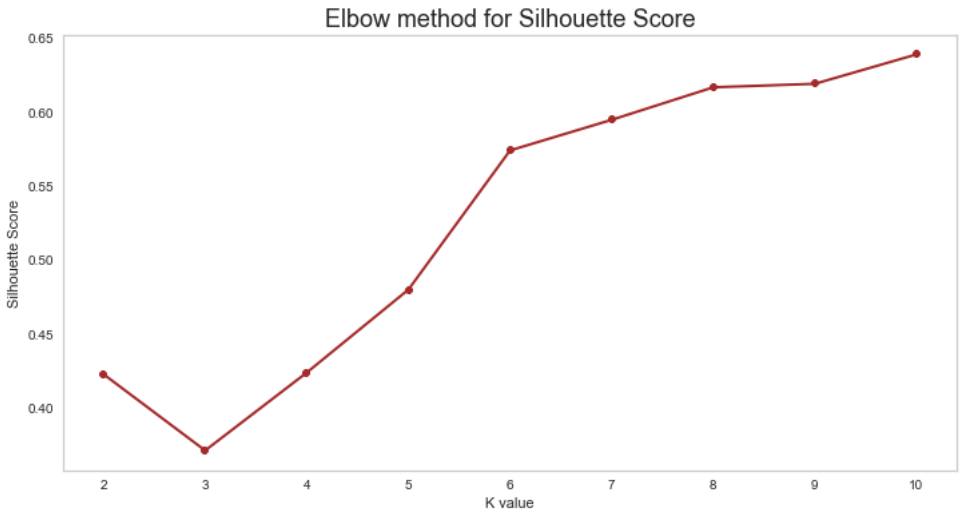


FIGURE 5.40 – Visualisation de silhouette score

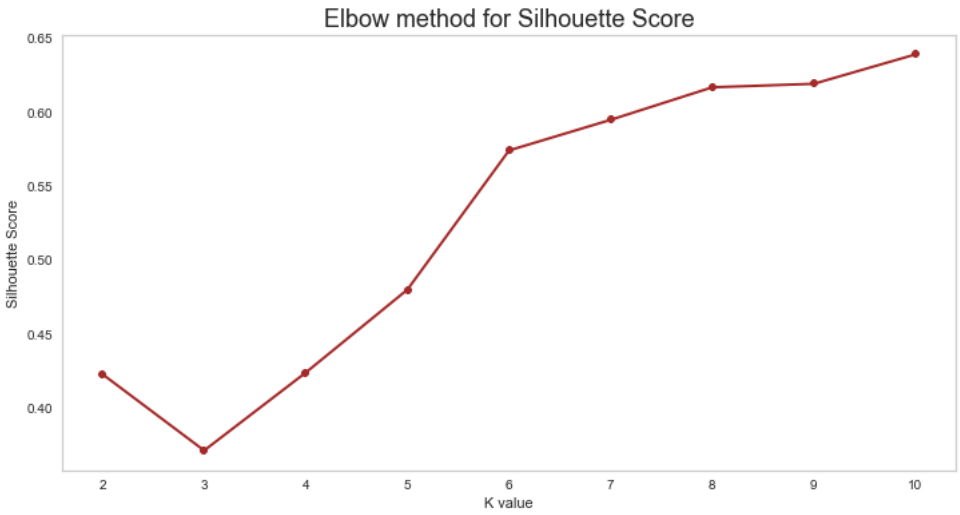
De plus, nous vérifierons le meilleur
score de silhouette à l'aide de la méthode Elbow sur le dataframe df_sclaed.

```
sil = []

for i in [2,3,4,5,6,7,8,9,10]:
   km = KMeans(n_clusters=i, init = 'k-means++', n_init=10, max_iter=100, random_state=42)
   km.fit(df_scaled)
    sil.append(silhouette_score(df_scaled, km.predict(df_scaled)))

plt.figure(figsize=(12,6))
plt.grid()
plt.plot([2,3,4,5,6,7,8,9,10], sil, linewidth=2, color='brown', marker = '8')
plt.xlabel('K value')
plt.ylabel('Silhouette Score')
plt.title('Elbow method for Silhouette Score', fontsize=18)
plt.show()
```

Résultat



hbc

FIGURE 5.41 – résultat de la méthode du coude (Elbow) pour silhouette score

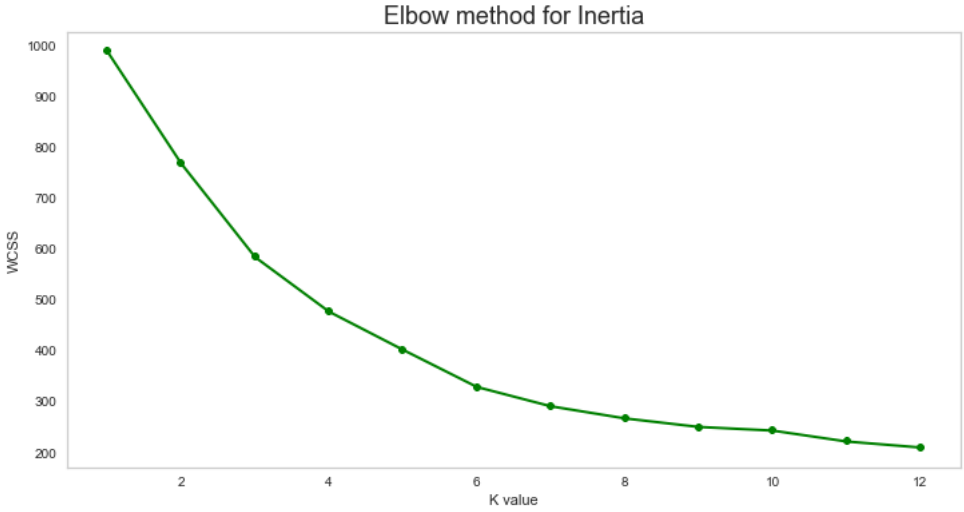
Sur la base de ce qui précède, nous pouvons clairement voir que le score de silhouette le plus élevé que nous pouvons obtenir lors de l'exécution du clustering K-means est 10 clusters.

Vérification du meilleur nombre de clusters avec Inertia :

```
wcss=[]

for k in range(1,13):
    kmeans = KMeans(n_clusters=k, init ='k-means++')
    kmeans.fit(df_scaled)
    wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,13), wcss, linewidth=2, color='green', marker = '8')
plt.xlabel('K value')
plt.ylabel('WCSS')
plt.title('Elbow method for Inertia', fontsize=18)
plt.show()
```

Résultat



hbc

FIGURE 5.42 – résultat de la méthode du coude pour l'inetie

Nous pouvons conclure que si nous prenons en considération à la fois le score de silhouette et l'inertie (WCSS), le meilleur nombre de clusters pour notre ensemble de données est de 10.

Appliquation de la méthode K-means sur 10 clusters

```
kmeans_1 = KMeans(n_clusters=6, random_state=42)
kmeans_1 = kmeans_1.fit(df_scaled)
inertia_1 = kmeans_1.inertia_
print('The clusters are:  ', kmeans_1.labels_)
print('The Inertia is:   ', kmeans_1.inertia_ )
```

Résultat

```
The clusters are:   [2 2 1 1 1 1 2 2 1 2 1 2 1 2 2 2 0 0 0 2 2 2 2 2 2 1 1 0 0 2 1 2 2 2 1 1 1
 1 2 2 2 0 1 2 2 1 3 2 2 2 2 2 2 2 2 2 1 2 2 1 1 2 2 2 1 1 1 1 2 2 1 3 0
 4 2 0 2 4 2 1 1 2 4 1 2 2 0 4 0 1 1 2 2 2 1 2 2 2 2 0 4 1 1 1 2 4 1 2 2 1
 1 1 1 2 2 0 1 2 2 2 5 2 4 2 1 1 2 2 1 0 1 1 4 0 1 1 0 0 0 1 0 0 0 2 0 4 1
 0 1 2 1 4 4 0 0 1 2 0 2 0 2 0 4 2 2 4 0 4 4 1 5 2 0 1 1 0 1 2 0 4 2 2 1 2
 2 0 0 1 2 1 0 1 1 0 4 0 2 1 0 4 0 0 2 1 5 0 1 0 4 2 2 1 2 1 0 1 1 2 1 2 2
 2 2 1 2 2 2 1 2 1 0 4 0 5 4 2 2 4 4 4 4 4 2 0 4 2 2 2 1 2 1 0 2 5 2 1 1 0
 5 4 1 0 4 0 2 0 0 5 2 1 1 5 5 5 4 5 0 4 4 4 5 5 1 1 0 5 1 0 0 4 4 4 0 5 1
 1 5 4 5 5 1 5 0 4 4 0 1 1 1 4 5 5 4 0 4 4 5 5 4 5 1 0 0 5 0 1 1 0 0 5 1 0
 4 5 1 1 1 0 2 5 4 4 1 5 4 1 1 1 1 4 1 0 1 1 1 0 1 5 0 1 1 1 1 1 1 0 1 1 4
 1 1 4 0 5 2 1 0 4 4 1 1 1 1 1 1 4 1 1 0 4 1 1 0 1 1 0 4 0 2 1 1 1 1 0 1 1
 1 1 4 1 4 4 1 1 0 0 0 1 4 4 4 0 4 5 0 1 1 0 4 1 0 0 1 0 1 0 0 0 3 2 5 2 5
 2 5 0 1 1 5 5 5 5 1 5 5 5 5 0 0 5 1 2 1 3 1 5 3 4 4 0 1 5 5 1 5 1 5 1 0 5
 1 5 5 2 1 2 5 0 2 2 5 5 5 0 5 0 5 2 5 5 5 1 0 4 2 4 0 2 4 0 5 5 4 1 0 2 5
 5 0 0 4 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3]
The Inertia is:    328.1600864418109
```

FIGURE 5.43 – Application de k-means sur 10 cluster

Dans ce qui suit, nous ferons des prédictions pour de nouvelles données, avec de nouveaux clusters créés. Nous allons d'abord calculer les comptages du cluster puis créer une frame de données appelée "countscldf_1". Dans une dernière étape, nous allons imprimer la nouvelle trame de données countscldf_1.

```
km_label_1 = kmeans_1.predict(df_scaled)
```

```
unique, counts1 = np.unique(km_label_1, return_counts=True)
counts1 = counts1.reshape(1,6)
```

```
countscldf_1 = pd.DataFrame(counts1, columns= ['Cluster 0', 'Cluster 1','Cluster 2', 'Cluster 3',
                                               'Cluster 4', 'Cluster 5'])
countscldf_1
```

Résultat

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
0	103	158	161	54	72	65

FIGURE 5.44 – countscldf_1

Silhouette score

```
silh_score_1 = silhouette_score(df_scaled, km_label_1)
print(f'Silhouette Score(n = 2): {silhouette_score(df_scaled, km_label_1)}')
```

Résultat Silhouette Score($n = 2$) : 0.6391089592687702.

Les résultats ne sont pas bons, l'inertie est très élevée et le score de silhouette pas loin de zéro. Ensuite, nous appliquerons l'algorithme PCA à nos données brutes pour voir si l'application du clustering à la nouvelle base de données améliorera les résultats.

Analyse en composantes principales (ACP) Les algorithmes de réduction de dimensionnalité capturent les informations saillantes dans les données d'origine tout en réduisant la taille de l'ensemble de données. Lorsque nous passons d'un nombre élevé de dimensions à un nombre inférieur, le bruit dans l'ensemble de données est minimisé car l'algorithme de réduction de la dimensionnalité doit capturer les aspects les plus importants des données d'origine et ne peut pas consacrer d'attention aux éléments peu fréquents.

L'objectif de PCA est d'extraire les informations les plus importantes de la table de données en compressant la taille des données et en ne conservant que les informations importantes.

Le premier composant principal doit avoir la plus grande variance possible (inertie) et donc ce composant expliquera la plus grande partie de l'inertie/variance des données, donc moins vous avez de composants principaux, plus l'inertie sera faible après PCA. Pour chaque nouvelle composante, l'inertie augmentera puisque les rotations sont toujours effectuées dans un sous-espace et les nouveaux axes expliqueront toujours moins d'inertie que les composantes d'origine. L'inertie ne doit pas être le critère pour choisir le nombre optimal de composantes principales puisque plus les composantes sont faibles, plus l'inertie sera faible.la règle d'or est qu'une variance expliquée de 60% devrait être le critère lors du choix du nombre de composantes principales

L'objectif est d'identifier le « meilleur » nombre de composants pour notre ensemble de données. Nous visons à extraire le nombre de composants qui expliqueront 60% de la variance de nos données :

Nous allons d'abord importer le PCA et Linear Discriminant Analysis (LDA) de la bibliothèque sklearn, puis nous adapterons le dataframe df_sclaed.

Importation des PCA

```
from sklearn.decomposition import PCA
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
```

Visualisation de la variance expliquée cumulée

```
pca = PCA().fit(df_scaled)
plt.plot(np.cumsum(pca.explained_variance_ratio_), color='teal')
plt.xlabel('number of components')
plt.ylabel('cumulative explained variance');
```

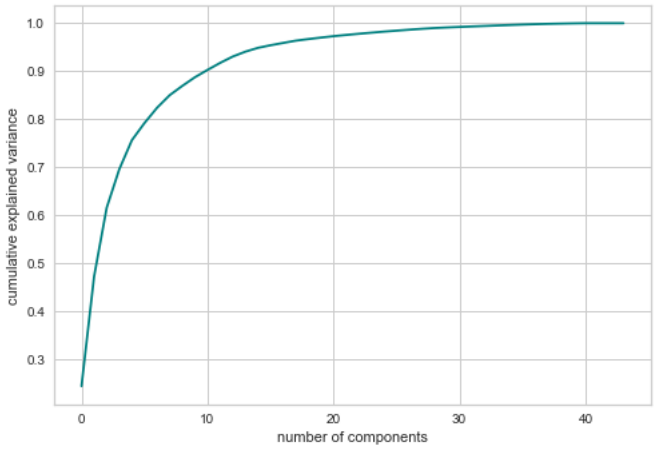


FIGURE 5.45 – variance expliquée cumulée

Calculons le nombre de composantes principale à garder :

```
pca = PCA(0.60).fit(df_scaled)
pca.n_components_
```

Résultat : 3

```
pca = PCA(0.30).fit(df_scaled)
pca.n_components_
```

Résultat : 2

Interprétation :

Nous avons découvert que pour conserver 60% de la variance expliquée de notre base de données, nous devrions effectuer une ACP avec 3 composantes. Au cas où nous déciderions d'implémenter l'ACP avec trois composants seulement (de cette façon, il serait facile de visualiser nos données, car nous n'aurions qu'un espace à 2 dimensions), il ne nous resterait que 30% de variance expliquée. Cela signifierait que nous perdons beaucoup de signal à partir de nos données d'origine et qu'il faut en tenir compte.

Appliquer K-means avec 10 clusters sur les données après PCA avec 3 composants

Pour comparer avec

nos précédentes implémentations de clustering, nous utiliserons le score d'inertie.

```
pca = PCA(n_components=3, random_state=42)
df_pca_1 = pca.fit(df_scaled).transform(df_scaled)

kmeans_2 = KMeans(n_clusters=10, random_state=42)
kmeans_2 = kmeans_2.fit(df_pca_1)
inertia_2 = kmeans_2.inertia_
print('The clusters are:  ', kmeans_2.labels_)
print('The Inertia is:   ', kmeans_2.inertia_ )
```

```
The clusters are:   [1 1 3 3 3 3 1 1 3 1 3 1 3 1 1 1 0 0 0 1 1 1 1 1 1 3 3 0 0 1 3 1 1 1 3 3 3
 3 1 1 1 0 3 1 1 3 5 1 1 1 1 1 1 1 1 1 3 1 1 3 3 1 1 1 3 3 3 3 1 1 3 5 0
 2 1 0 1 5 1 3 3 1 5 3 1 1 0 5 0 3 3 1 1 1 3 1 1 1 1 0 8 3 3 3 1 8 3 1 1 3
 3 3 3 1 1 0 3 1 1 1 4 1 5 1 3 3 1 1 3 0 3 3 8 0 3 3 0 0 0 3 0 0 0 1 0 5 3
 0 3 1 3 8 8 0 0 3 1 0 1 0 1 0 8 1 1 8 0 5 5 3 4 1 0 3 3 0 3 1 0 8 1 1 3 1
 1 0 0 3 1 3 0 3 3 0 8 0 1 3 0 8 0 0 1 3 4 0 3 0 8 1 1 3 1 3 0 3 3 1 3 1 1
 1 1 3 1 1 1 3 1 3 0 8 0 4 6 1 1 8 5 8 8 8 1 0 8 1 1 1 3 1 3 0 1 4 1 3 3 0
 4 8 3 0 8 0 1 0 0 0 1 3 3 9 9 9 8 4 0 8 2 6 9 9 3 3 0 4 3 0 0 8 6 8 0 9 3
 3 4 6 9 4 3 4 0 2 2 0 3 3 3 8 9 4 8 0 6 8 4 9 6 9 3 0 0 4 0 3 3 0 0 9 3 0
 8 4 3 3 3 0 1 9 2 2 3 9 2 3 3 3 3 2 3 0 3 3 3 0 3 4 0 3 3 3 3 3 3 0 3 3 8
 3 3 2 0 4 1 3 0 2 2 3 3 3 3 3 3 2 3 3 0 2 3 3 0 3 3 0 2 0 1 3 3 3 3 0 3 3
 3 3 2 3 8 8 3 3 0 0 0 3 2 2 2 0 8 4 0 3 3 0 2 3 0 0 3 0 3 0 0 0 5 1 6 1 4
 1 6 0 3 3 4 4 4 4 3 6 4 4 4 0 0 4 3 1 3 5 3 4 5 5 5 0 3 4 4 3 4 3 4 3 0 4
 3 4 6 1 3 1 4 0 1 1 4 4 4 0 4 0 4 1 4 4 4 3 0 6 1 5 0 1 5 0 4 6 8 3 0 1 4
 4 0 0 6 6 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7
 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 5 5 5]
The Inertia is:    14.934951999008549
```

FIGURE 5.46 – Clustering avec $k = 10$ et 2 composantes

L'inertie s'est légèrement améliorée après l'application de PCA avec 2 composants. Nous allons maintenant faire des prédictions pour de nouvelles données, avec de nouveaux clusters créés et imprimer le Silhouette Score pour la mise en œuvre.

```
km_label_2 = kmeans_2.predict(df_pca_1)

silh_score_2 = silhouette_score(df_pca_1, km_label_2)
print(f'Silhouette Score(n = 2): {silhouette_score(df_pca_1, km_label_2)}')
```

Silhouette Score($n = 2$) : 0.8973467208631747

Ensuite, nous allons calculer les comptages du cluster et créer un nouveau bloc de données appelé "countscldf_2"

```
unique, counts2 = np.unique(km_label_2, return_counts=True)
counts2 = counts2.reshape(1,6)

countscldf_2 = pd.DataFrame(counts2, columns= ['Cluster 0', 'Cluster 1','Cluster 2', 'Cluster 3',
                                               'Cluster 4', 'Cluster 5'])

countscldf_2
```

Résultat

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	cluster6	cluster7	cluster8	cluster9
0	104	161	19	158	46	20	14	46	32	13

FIGURE 5.47 – countscldf_2

Appliquer K-means avec 10 clusters sur les données après PCA avec
2 composantes (sachant que seulement 30% de la variance expliquée sera conservée)
Nous allons initialiser le modèle PCA pour deux composants et imprimer
le score d'inertie comme précédemment mis en œuvre pour deux composants.

```
pca = PCA(n_components=2, random_state=42)
df_pca = pca.fit(df_scaled).transform(df_scaled)

kmeans_3 = KMeans(n_clusters=6, random_state=42)
kmeans_3 = kmeans_2.fit(df_pca)
inertia_3 = kmeans_3.inertia_
print('The clusters are:  ', kmeans_3.labels_)
print('The Inertia is:   ', kmeans_3.inertia_ )
```

Résultat

```
The clusters are:   [2 2 3 3 3 3 2 2 3 2 3 2 3 2 2 2 1 1 1 2 2 2 2 2 3 3 1 1 2 3 2 2 2 3 3 3
 3 2 2 2 1 3 2 2 3 7 2 2 2 2 2 2 2 2 2 3 2 2 3 3 2 2 2 3 3 3 3 2 2 3 7 1
 8 2 1 2 6 2 3 3 2 6 3 2 2 1 6 1 3 3 2 2 2 3 2 2 2 2 1 5 3 3 3 2 5 3 2 2 3
 3 3 3 2 2 1 3 2 2 2 9 2 6 2 3 3 2 2 3 1 3 3 5 1 3 3 1 1 1 3 1 1 1 2 1 6 3
 1 3 2 3 5 5 1 1 3 2 1 2 1 2 1 5 2 2 5 1 6 6 3 9 2 1 3 3 1 3 2 1 5 2 2 3 2
 2 1 1 3 2 3 1 3 3 1 5 1 2 3 1 5 1 1 2 3 9 1 3 1 5 2 2 3 2 3 1 3 3 2 3 2 2
 2 2 3 2 2 2 3 2 3 1 5 1 9 5 2 2 5 6 5 5 5 2 1 5 2 2 2 3 2 3 1 2 2 2 3 3 1
 9 5 3 1 5 1 2 1 1 1 2 3 3 3 3 3 5 2 1 5 0 0 3 3 3 3 1 9 3 1 1 5 0 5 1 3 3
 3 9 5 3 9 3 9 1 0 0 1 3 3 3 5 3 9 5 1 0 5 9 3 5 3 3 1 1 9 1 3 3 1 1 3 3 1
 5 9 3 3 3 1 2 3 0 0 3 3 0 3 3 3 3 0 3 1 3 3 3 1 3 9 1 3 3 3 3 3 3 1 3 3 5
 3 3 0 1 9 2 3 1 0 0 3 3 3 3 3 3 0 3 3 1 0 3 3 1 3 3 1 0 1 2 3 3 3 3 1 3 3
 3 3 0 3 5 5 3 3 1 1 1 3 0 0 0 1 5 9 1 3 3 1 0 3 1 1 3 1 3 1 1 1 7 2 7 2 9
 2 8 1 3 3 2 2 9 9 3 8 2 9 9 1 1 9 3 2 3 7 3 9 7 8 8 1 3 9 2 3 9 3 9 3 1 2
 3 9 8 2 3 2 9 1 2 2 9 9 9 1 9 1 9 2 9 2 9 3 1 5 2 8 1 2 6 1 9 7 5 3 1 2 9
 9 1 1 5 5 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 7 7 7]
The Inertia is:    3.613656384929622
```

FIGURE 5.48 – résultat de k-means avec 60%

```
km_label_3 = kmeans_2.predict(df_pca)
silh_score_3 = silhouette_score(df_pca, km_label_3)
print(f'Silhouette Score(n = 2): {silhouette_score(df_pca, km_label_3)}')
```

Résultat Silhouette Score($n = 2$) : 0.880945219275541

Ensuite, nous allons calculer
les comptages du cluster et créer un nouveau bloc de données appelé "countscldf_3"

```
unique, counts3 = np.unique(km_label_3, return_counts=True)
counts3 = counts3.reshape(1,6)


countscldf_3 = pd.DataFrame(counts3, columns= ['Cluster 0', 'Cluster 1','Cluster 2', 'Cluster 3',
                                               'Cluster 4', 'Cluster 5'])

countscldf_3
```

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	cluster6	cluster7	cluster8	cluster9
0	21	104	169	171	46	38	9	10	7	38

FIGURE 5.49 – "countscldf_3"

Nous créons le dataframe
"df_pca_2" avec de nouveaux composants et des anciennes et nouvelles étiquettes

```
df_pca_2 = pd.DataFrame(df_pca, columns =['Component 1', 'Component 2'])
df_pca_2.head(2)
```

Résultat

	Component 1	Component 2
0	-0.253758	-0.667262
1	-0.250477	-0.676268

FIGURE 5.50 – df_pca_2

Résultat de cluster apr PCA_2

```
    columns = ['km_label_3']
km_label_3 = pd.DataFrame(data=km_label_3, columns=columns)
km_label_3.head()
print(km_label_3.value_counts())
```

Résultat
*Comparaison entre les deux composantes

```
    df_pca_2 = pd.concat([df_pca_2, km_label_3,y], axis=1)
df_pca_2.head()
```

Résultat
Visualisez les données d'origine par rapport aux résultats des données en cluster

```
km_label_3
3    171
2    169
1    104
4     46
5     38
9     38
0     21
7     10
6      9
8      7
dtype: int64
```

FIGURE 5.51 – Résultat de cluster apr PCA_2

	Component 1	Component 2	km_label_3	scoring
0	-0.253758	-0.667262	1	318
1	-0.250477	-0.676268	1	206
2	-0.422970	0.680539	0	938
3	-0.406283	0.651071	0	324
4	-0.414399	0.665796	0	646

FIGURE 5.52 – L'interface Graphique De Logiciel Tableau

```
plt.figure(figsize=(16, 6))
plt.subplot(1,2,1)

sns.scatterplot(x = 'Component 1', y = 'Component 2', data = df_pca_2, hue = 'scoring', palette='BrBG')
plt.title('Original Data', fontsize=18)
plt.legend(title='Clusters', bbox_to_anchor=(1.02, 1), loc='upper left', borderaxespad=0, fontsize='medium')

plt.subplot(1,2,2)

sns.scatterplot(x = 'Component 1', y = 'Component 2', data = df_pca_2, hue = 'km_label_3', palette='BrBG')
plt.title('K-means with PCA(2)', fontsize=18)
plt.legend(title='Clusters', bbox_to_anchor=(1.02, 1), loc='upper left', borderaxespad=0, fontsize='medium')

plt.show()
```

Résultat

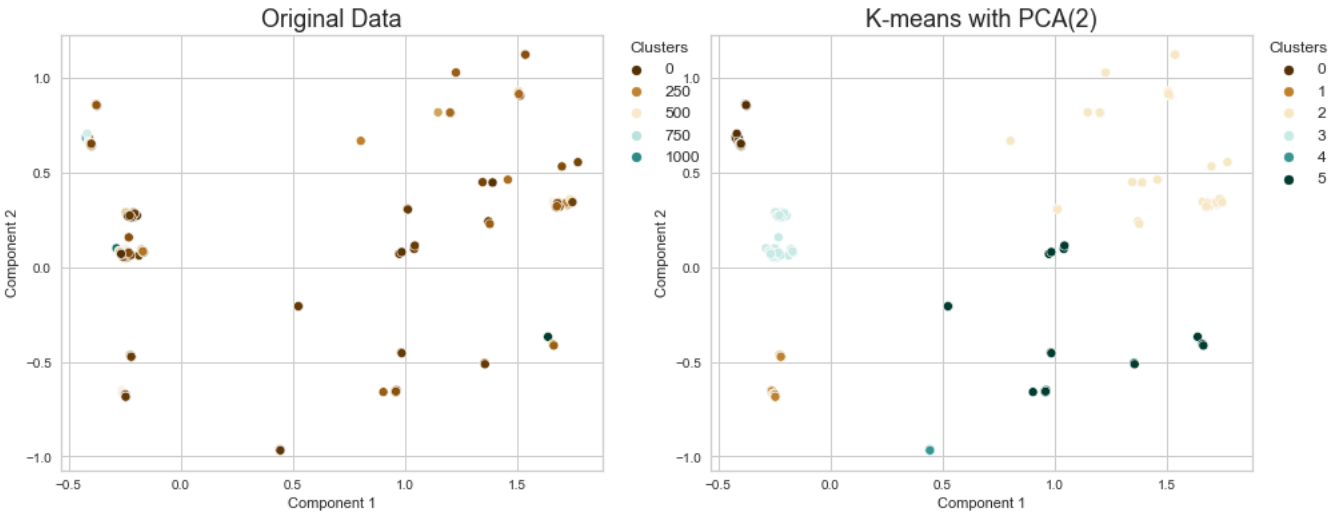


FIGURE 5.53 – Comparaison entre les données d'origine et les résultats de clusteur

Comparaison de modèles basée sur l'inertie et le score de silhouette :

```
Model_Comparison = pd.DataFrame({
'Model' : ['K-means_1', 'K-means_PCA(3)', 'K-means_PCA(2)'],
'Inertia_Score' : [inertia_1, inertia_2, inertia_3],
'Silhouette_Score' : [silh_score_1, silh_score_2, silh_score_3]})

Model_Comparison_df = Model_Comparison.sort_values(by='Inertia_Score', ascending=True)
Model_Comparison_df = Model_Comparison_df.set_index('Model')
Model_Comparison_df.reset_index()
```

	Model	Inertia_Score	Silhouette_Score
0	K-means_PCA(2)	14.395369	0.860617
1	K-means_PCA(3)	55.747189	0.844472
2	K-means_1	328.160086	0.574087

FIGURE 5.54 – Comparaison de modèles basée sur l'inertie et le score de silhouette

Conclusion

Sur K-Means : Après avoir exécuté la méthode Elbow sur WCSS et le score Silhouette, nous avons conclu que le meilleur nombre de clusters était de six. Après application de K-means sur des données brutes/inchangées, nos résultats étaient

médiocres : inertie très élevée plus Silhouette Score très proche de zéro. Ensuite nous avons appliqué l'ACP à 3 composantes après avoir vérifié qu'il restera 60% de variance expliquée. Les résultats se sont beaucoup améliorés. La dernière étape consistait à appliquer l'ACP avec 2 composantes seulement, sachant que les deux composantes ne détenaient que 30% de la variance expliquée. Les scores d'inertie et de silhouette sont considérablement améliorés cette fois.

5.7 Application sur le logiciel tableau

5.7.1 Présentation du logiciel Tableau

Tableau est la plate-forme d'analyse de bout en bout la plus puissante, la plus sécurisée et la plus flexible. Tableau a été fondé en 2003 à la suite d'un projet informatique à Stanford qui visait à améliorer le flux d'analyse et à rendre les données plus accessibles aux utilisateurs grâce à la visualisation. Les cofondateurs Chris Stolte, Pat Hanrahan et Christian Chabot ont développé et breveté la technologie fondamentale de Tableau, VizQL, qui exprime visuellement les données en traduisant les actions de glisser-déposer en requêtes de données via une interface intuitive.

*Rechercher des clusters dans les données

L'analyse de cluster partitionne les repères de la vue en clusters, où les repères de chaque cluster sont plus similaires les unsaux autres qu'ils ne le sont avec les repères des autres clusters.

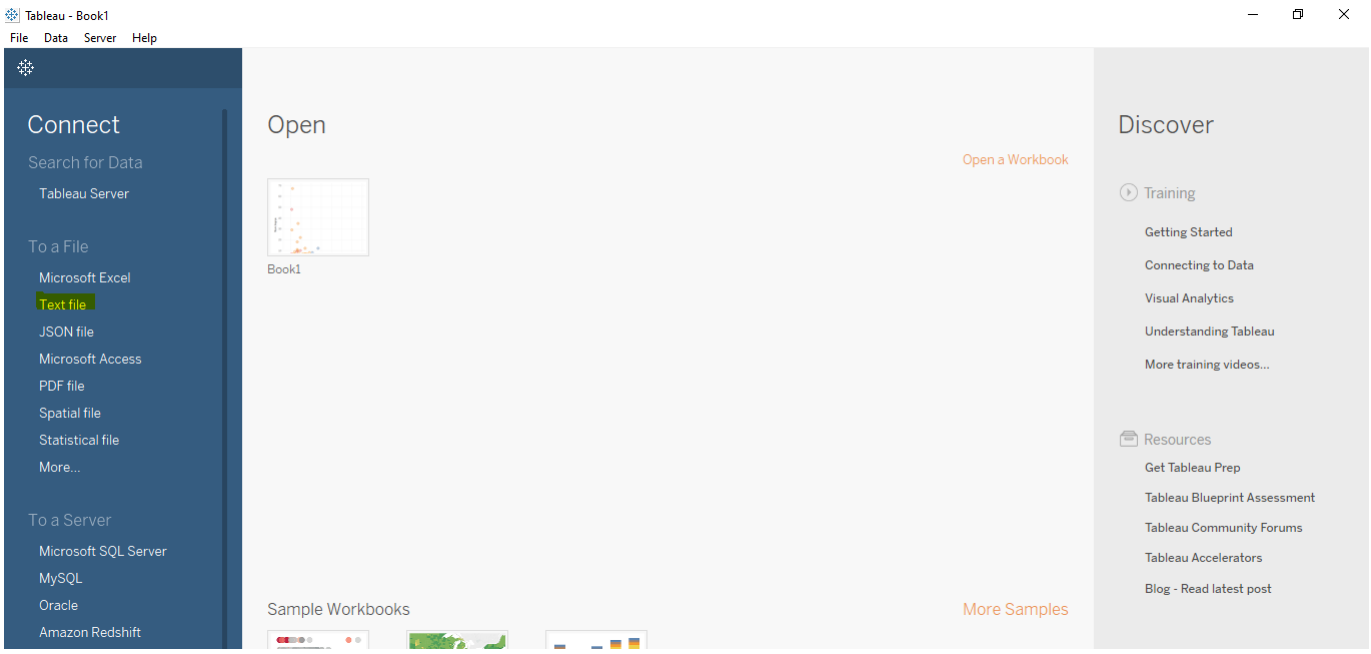


FIGURE 5.55 – L'interface Graphique De Logiciel Tableau

chargement des données dans Tableau choisir (Text file) cas d'un fichier csv

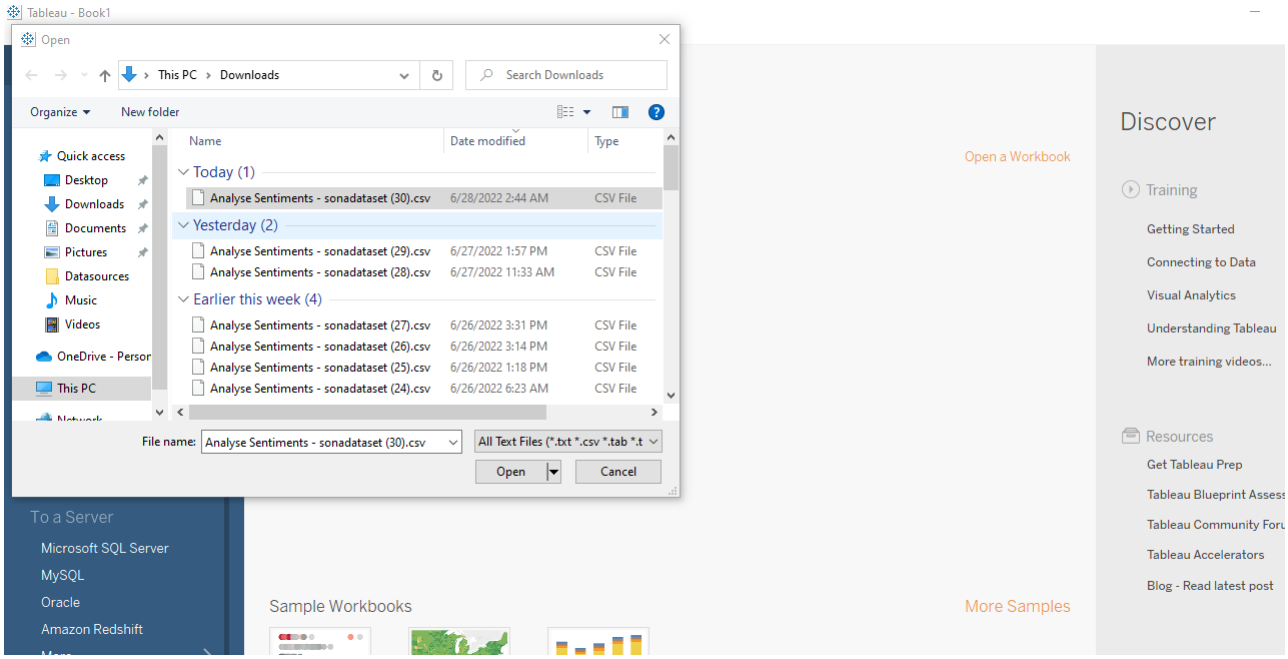


FIGURE 5.56 – chargement des données dans Tableau

aprés chargement des données on clique sur (Extract) aprés on choisie (sheet1)

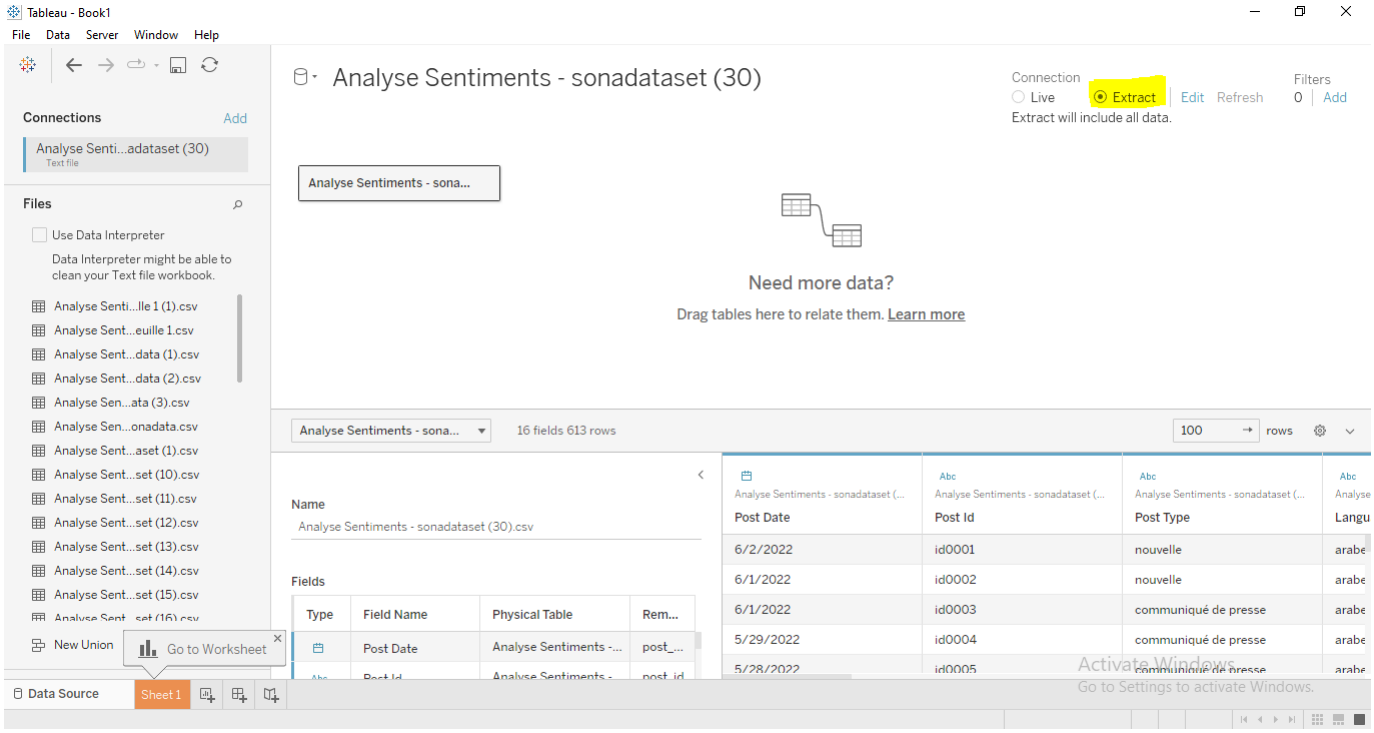


FIGURE 5.57 – Lancement de traitement de données

on sauvgarde le fichier

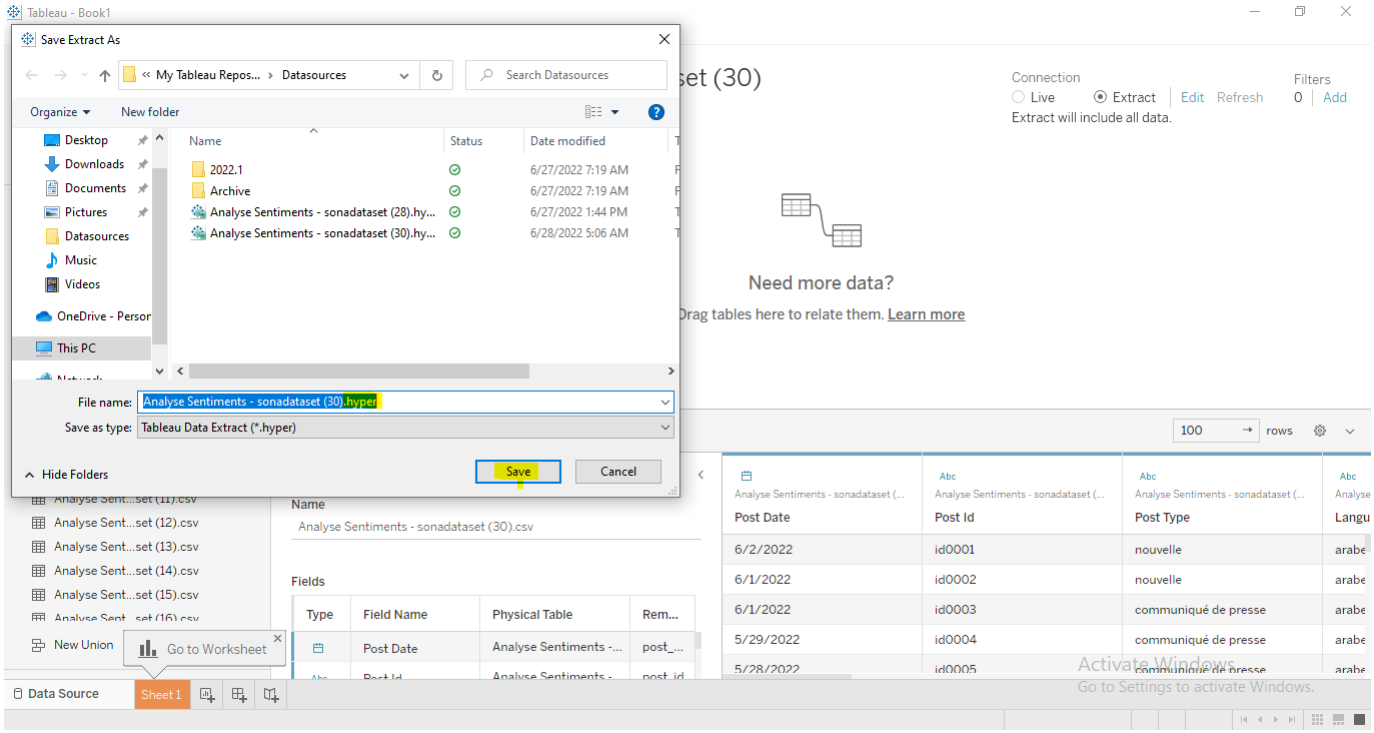


FIGURE 5.58 – L'enregistrement de nouveau fichier

Créer des clusters
Pour rechercher des clusters dans une vue dans Tableau, procédez comme suit.
Créez une vue.
Faites glisser
le cluster du volet Analytics vers la vue et déposez-le dans la zone cible de la vue :

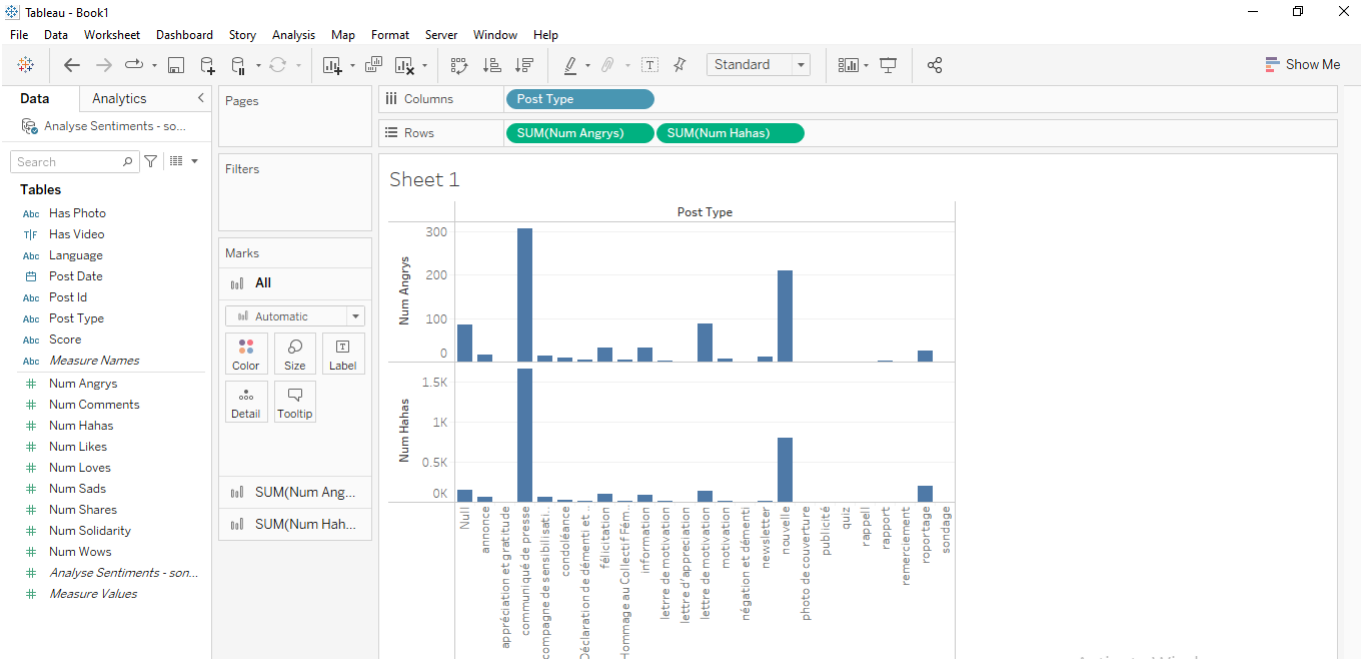


FIGURE 5.59 – début de clustering

Vous pouvez
également double-cliquer sur Cluster pour rechercher des clusters dans la vue.
Lorsque vous déposez ou double-cliquez sur Cluster :
Tableau crée un groupe Clusters sur Couleur et colore les repères de votre vue par cluster. S'il existe déjà un champ sur Couleur, Tableau déplace ce champ vers Détail et le remplace sur Couleur par les résultats de regroupement.
Tableau attribue chaque repère de la vue à l'un des clusters. Dans certains cas, les marques qui ne s'intègrent pas bien dans un cluster sont affectées à un cluster "Pas regroupé".
Tableau affiche
la boîte de dialogue Clusters, dans laquelle vous pouvez personnaliser le cluster.
Personnalisez les résultats du
cluster en effectuant l'une des actions suivantes dans la boîte de dialogue Clusters.
Faites glisser de nouveaux champs du volet Données vers la zone Variables de la boîte de dialogue Clusters. Vous pouvez également faire glisser des champs hors de la zone Variables pour les supprimer.
Lorsque vous ajoutez des variables, les mesures sont agrégées à l'aide de l'agrégation par défaut pour le champ ; les dimensions sont agrégées à l'aide de l'ATTR, qui est la méthode standard utilisée par Tableau pour agréger les dimensions.
Pour modifier l'agrégation d'une variable, cliquez dessus avec le bouton droit.
Cliquez pour ouvrir le volet Analytics :

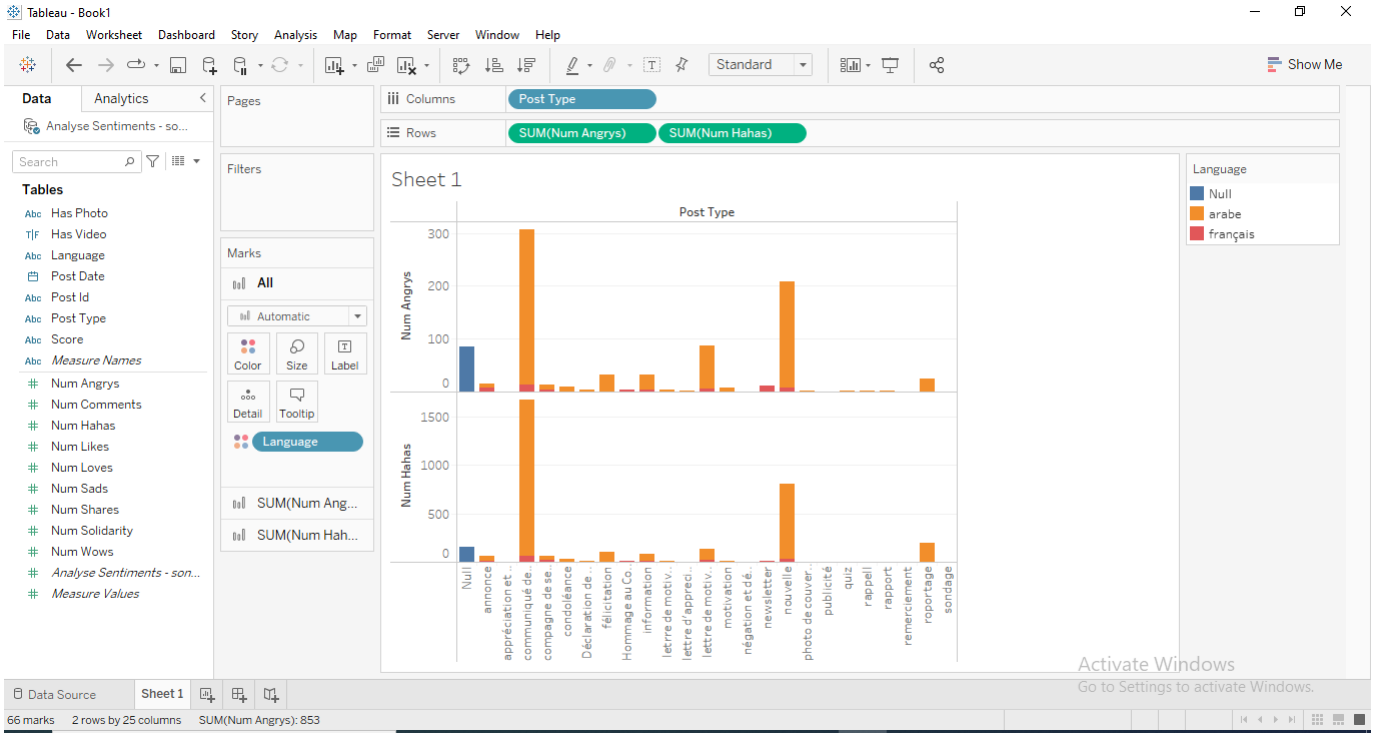


FIGURE 5.60 – affectation des attribues colones

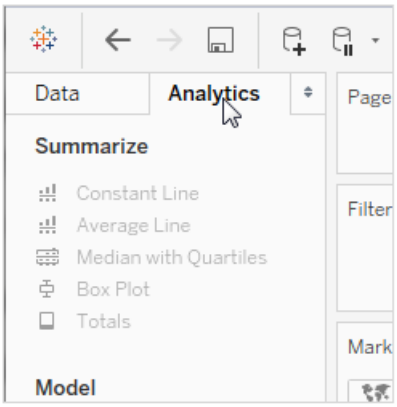


FIGURE 5.61 – Initialisation pour faire les analytiques

Faites glisser le cluster depuis le volet Analytics et déposez-le dans la vue :

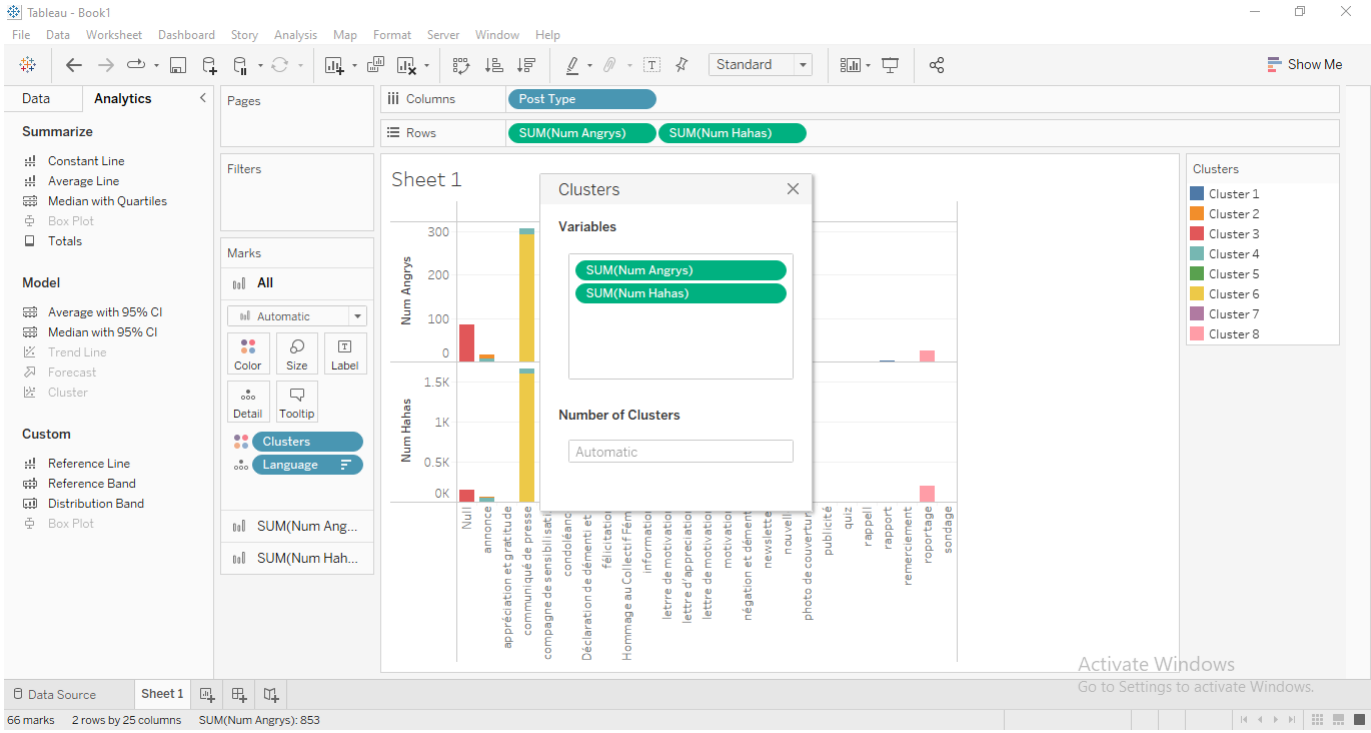


FIGURE 5.62 – Définir les attributs des axes

Resultats de clustering

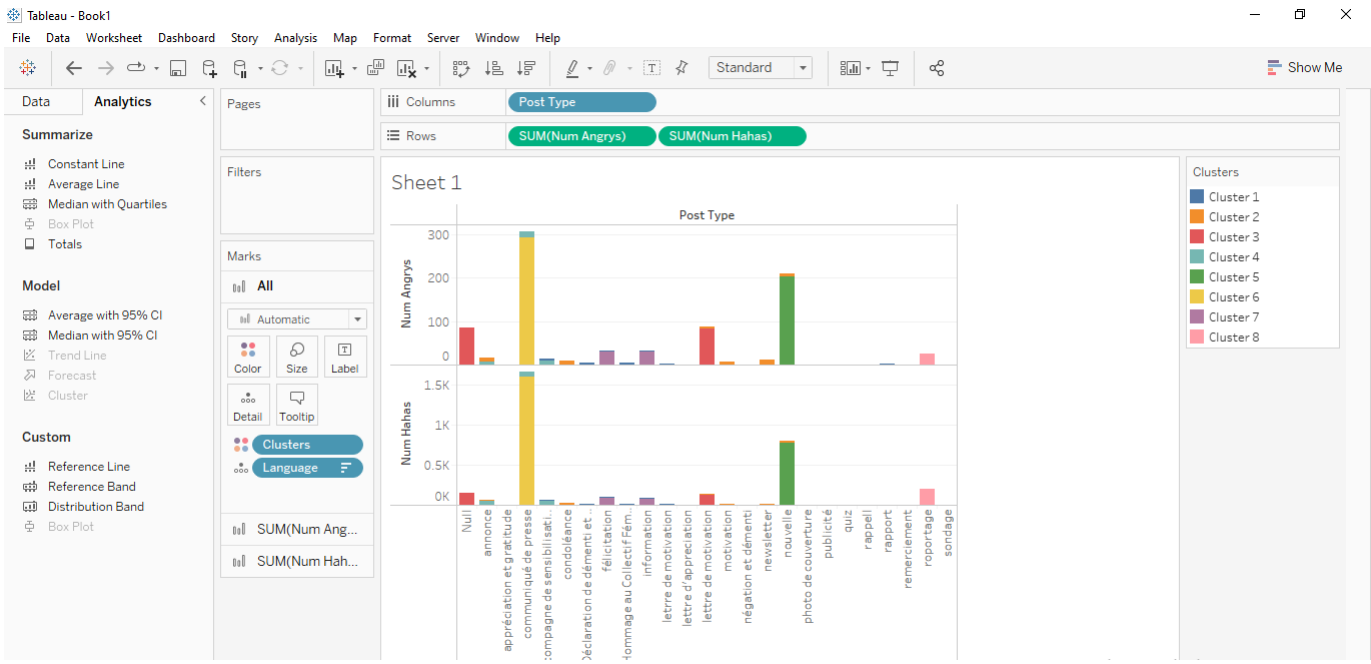


FIGURE 5.63 – Schéma De Clustering

Tableau utilise l'algorithme k-means pour le clustering. Pour un nombre donné de clusters k, l'algorithme partitionne les données en k clusters. Chaque cluster a un centre (centre de gravité) qui est la valeur moyenne de tous les points de ce cluster. K-means localise les centres grâce à une procédure itérative qui minimise les distances entre les points individuels d'un cluster et le centre du cluster. Dans Tableau, vous pouvez spécifier un nombre de clusters souhaité ou demander à Tableau de tester différentes valeurs de k et de suggérer un nombre optimal de clusters

K-means nécessite une spécification initiale des centres de cluster. Partant d'un cluster, la méthode choisit une variable dont la moyenne est utilisée comme seuil pour scinder les données en deux. Les barycentres de ces deux parties sont ensuite utilisés pour initialiser les k-means afin d'optimiser l'appartenance des deux clusters. Ensuite, l'un des deux groupes est choisi pour le fractionnement et une variable au sein de ce groupe est choisie dont la moyenne est utilisée comme seuil pour diviser ce groupe en deux. K-means est ensuite utilisé pour partitionner les données en trois clusters, initialisés avec les centroïdes des deux parties du cluster divisé et le centroïde du cluster restant. Ce processus est répété jusqu'à ce qu'un nombre défini de clusters soit atteint.

Tableau utilise l'algorithme de Lloyd avec des distances euclidiennes au carré pour calculer le regroupement des k-moyennes pour chaque k. Combiné avec la procédure de division pour déterminer les centres initiaux pour chaque $k > 1$, le regroupement résultant est déterministe, le résultat dépendant uniquement du nombre de groupes.

Validation et Interprétation des Résultats

5.8 Discussions et recommandations

Arrivé à l'issue de cette parie, nous pouvons remarquer :

- Les internautes qui "aiment" ou "sont tristes" commentent et partagent plus que les autres.
- Les abonnés qui "adorent", "étonnés", "solidaires" ou "rient" ont tendance à partager plutôt qu'à commenter les publications.
- La langue qui suscite le plus de réactions est l'arabe, mais cela peut avoir une autre interprétation dans le fait que le publique qui comprend l'arabe est plus important que celui qui comprend le français.
- Concernant les types de publications qui sont le plus partagées sont : 'lettres d'appréciation' et 'lettres de motivation'. Les types qui sont les plus commentés sont : 'condoléances' et 'motivation'.
- Concernant le score, nous avons remarqué ce classement, du meilleur au moins bons selon le types de publication :
  1. Motivation
  2. Lettre d'appréciation
  3. Condoléances

Nous recommandons donc à l'entreprise de favoriser
la langue arabe dans les publication et les types motivation, lettre d'appréciation.

Conclusion générale

Dans ce travail, nous avons considéré un problème bien particulier et qui touche toute entreprise qui cherche à se moderniser et avoir une place, une bonne place dans le monde industriel. Le problème considéré porte sur la réputation numérique ou la e-réputation. Nous avons étudier cette question avec la plus grande entreprise nationale : Sonatrach. En plus de sa réputation de leader, Sonatrach jouit d'une bonne réputation à l'international, aussi. Nous nous sommes focalisé dans notre étude sur la page Facebook de cette entreprise et avons étudié les différentes réaction, commentaires de leurs abonnées durant ces deux dernières années.

Nous avons étudié les sentiments des gens qui réagissent aux publication de Sonatrach sur plusieurs aspects pour répondre aux hypothèses de départ que nous pouvons résumer avec cette phrase : 'Qu'est ce qui fait qu'une publication de la page officielle de Sonatrach est bonne ou pas ?'. La réponse à cette question nous a conduit à :

- Récolter des données : plus de 600 publication ont été minucieusement étudiées (les réactions, les commentaires, le nombre de partages, le nombre de personnes touchées)
- Préparer un dataset : filtrer les données dans le but d'enlever les doublant et les cases vides.
- Visualiser nos données : pour avoir des aperçus globaux sur le comportement de notre dataset.
- Etude de correlation : a fin de répondre à notre hypothèse de départ, nous avons introduit un score pondéré avec les différents types des de réactions. Ensuite, nous nous sommes demandé quel paramètre a fait qu'une publication a un meilleur score qu'une autre.
- Clustering : dans cette partie, nous avons employé deux techniques l'une avec le logiciel Tableau et l'autre avec Python pour dégager le nombre optimal de clusters qui permettra de catégoriser les différents datasets de ce genre d'études. Nous avons trouvé 8 clusters avec la méthode Tableau et 10 avec Python.

Comme recommandation finales à notre entreprise, nous avons déduit qu'il faut se diriger vers des publications orientées plus vers la motivation et l'appréciation/condoléances en langue arabe. Cette recommandation n'est pas complète sans une étude sur la population qui représente les abonnées de la page de Sonatrach, en effet, la préférence de la langue arabe dans les publication peut être dû à l'incompréhension des abonnés. Ce qui peut être une bonne perspective de travail pour les études futures.

Bibliographie

[1]

[2] A. J, G. K, V. S. Natural language processing. *International Journal of Computer Sciences and Engineering, Vol.06, N°1, pp.161-167* (2018).

[3] A. R, P. R., AND V, T. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation, 47(1), p. 239-268.* (2013).

[4] A.G. *Le web social et la e-réputation.*. P 55. Lextenso éditions, France, 2013.

[5] A.L, C., AND B.T, E. Systematic reviews in sentiment analysis : a tertiary study. *Artificial Intelligence Review, Information Technology Group, Wageningen University Research, Wageningen, The Netherlands, p.2.* (2021).

[6] B, R. *Minimisation des désagrément dans les clusters agrégés.* PhD thesis, , 2015.

[7] B. P, G. E, J. A. M. T. Enriching word vectors with subword information. *arXiv preprint arXiv :1607.04606*.

[8] B.A, A.M, M. *Bien gérer sa réputation sur Internet : E-réputation personnelle mode d'emploi.*. P 175. Paris, Dunod, 2011.

[9] C, G. *Une méthode de classification non supeviisée pour l'apprentissage de règles et la recherche d'information.* PhD thesis, Université d'Orléans, 2004.

[10] C, L. *Contextualisation, visualisation et évaluation en apprentissage non supervisé.* PhD thesis, Université Charles de Gaulle-Lille 3, 2006.

[11] D.J-M, D.J, F. *E-Réputation des marques, des produits et des dirigeants.* Vuibert,Paris, 2013.

[12] F. B, H. S. Modelling irony in twitter. *in : Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Sweden, Gothenburg, pp. 56-64.* (2014).

[13] G. V, R. C. Sentiment analysis and opinion mining :a survey. *International Journal, 2(6), p.282–292.*.

[14] H, M, B. Y., AND V. Cluster validity methods : part i. acm sigmod record. *31(2) :40-45* (2002).

[15] H, J. A. Statitiscal theory in clustering. *Journal of classification, 2(1) :63-76* (1985).

[16] H. P, S. B, G. Sentiment mining of movie reviews using random forest with tuned hyperparameters. *Conference : International Conference on Information ScienceAt : Kerala.*.

[17] H. Z, F. S. Aspect-level sentiment analysis based on a generalized probabilistic topic and syntax model. *The Twenty-Eighth International Flairs Conference..*

[18] J .B, H. M, X. Z. Twitter mood predicts the stock market. *Journal of Computational Science* (2011).

[19] J. W, M. B, K. G. K. Towards universal paraphrastic sentence embeddings. *In Proceedings of International Conference on Learning Representations..*

[20] K, N. On the predictive power of web intelligence and social media the best way to predict the future is to tweet it. *Massachusetts Institute of Technology, USA* (2014).

[21] K.L, AND J, R. Clustering large applications (program clara). *finging groups inn data : an antroduuction ti cluster analysis, page 126-163* (2008).

[22] L.J, L. Mercator tout le marketing à l'ère numérique. *11ème édition, Dunod, Paris,P811* (2014).

[23] M, J. E. A. somme methodes for classification and analysis of mulicariate. *In proceeding of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 2881-297.Oakland,CA,USA* (1967).

[24] P. B, O.N, D. levels of sentiment analysis and its challenges : A literateure review. *Conference on big data analytics and computational, intelligence (ICBDAC)* (2017).

[25] R.S, K.R.S, T. S. E. Sentiment analysis on online product review. *International Research Journal of Engineering and Technology (IRJET)* (Vol.04, N° 04, p. 2381).

[26] T.R. Le rôle des réseaux sociaux dans l'amélioration de l'e-réputation de l'entreprise. Mémoire master, Ecole des Hautes Etudes Commerciales d'Alger, Alger, 2017. P 44.

[27] WIKIPÉDIA. Sonatrach — wikipédia, l'encyclopédie libre, 2022. [En ligne ; Page disponible le 7-mars-2022].

Résumé – Abstract

Résumé

Dans ce travail, nous avons réalisé une étude sur l'E-repuration de la plus grande entreprise algérienne : Sonatrach. Nous avons considéré sa page officielle et étudié ses publications depuis le début de l'activation de la page. A l'aide de techniques statistiques, nous avons essayé d'extraire les paramètres d'une publication réussie pour aider considérablement les administrateurs de cette page. Nous avons formulé des recommandations et des perspectives pour des travaux futurs.

Abstract

In this work, we have realized a study on the E-repuration of the biggest Algerian company : Sonatrach. We considered its official page and studied its publications during the last two years. By means of statistical techniques, we try to extract the parameters of a successful publication to help considerably the administrators of this page. We have come up with recommendations and perspectives for future work.

Présenté par : Chentoufi Roza, Djelouah Zineb *UMBB*