Vers l'amélioration du processus décisionnel par l'intégration des données sociales

Soumia Benkrid*, Redouane Boucenna*, Dihia Boulegane*, Younes Sennadj*, Chahnez Zakaria*, Lynda Said L'hadj*

* Ecole nationale Supérieure d'Informatique (ESI), Alger, Algérie
(s_benkrid, br_boucenna,bd_boulegane, by_sennadj, c_zakaria, l_saidlhadj)@esi.dz

Résumé. Dans la vie réelle du Business Intelligence (BI), les faits sont très importants, mais l'opinion joue également un rôle crucial car elle peut influencer le processus de prise de décision. Aujourd'hui, de nombreuses sources d'information non structurées telles que les réseaux sociaux qui sont plus ou moins librement disponibles sur le web, leur volume est en constante croissance ce qui constitue une mine d'or gratuite en ce qui concerne la collecte de l'opinion publique. En effet, les entreprises ont toujours prêté une grande attention à l'opinion de leur clients et ne négligent en aucun cas l'importance de l'opinion publique issue des réseaux sociaux. Dans ce papier, nous proposons une architecture décisionnelle robuste qui permet de collecter, stocker et exploiter les données textuelles exprimées par les internautes sur les réseaux sociaux. Finalement, nous présentons une validation de nos propositions dans un contexte réel, il s'agit du domaine de la téléphonie mobile chez l'opérateur Oreedoo Algérie

1 Introduction

Ces dernières années, les données se développent plus rapidement que par le passé. Les organisations font face à un grand volume de données qui évolue rapidement et qui est de plus en plus varié (structuré, semi structuré et non structuré). Ces données proviennent du foisonnement de technologies et leur démocratisation dans de nombreux domaines, tels que les télécommunications, l'industrie, l'internet des Objets (Internet of things), les réseaux sociaux, la santé, . . .etc. Plus particulièrement, les réseaux sociaux sont devenus une source de grands volumes de données diverses. Effectivement, des millions de personnes utilisent les réseaux sociaux comme Facebook et Twitter pour s'exprimer. Leurs communications comprennent souvent des réflexions sur les bonnes et les mauvaises expériences ainsi que des avis sur des produits et des entreprises qu'ils aiment ou n'aiment pas. Cela fournit une précieuse source aux entreprises pour apprendre à connaître ses clients comme jamais auparavant. De ce fait, les entreprises doivent connaître les besoins réels des clients, leurs attentes ainsi que les éventuelles améliorations qui peuvent être apportées aux produits ou aux services. Toutefois, les chefs d'entreprise utilisent une combinaison de l'intuition, l'expérience et un certain niveau d'analyse pour prendre des décisions stratégiques et tactiques. La majorité des entreprises font

fréquemment appel à des programmes efficaces de collecte et d'écoute de la voix du client comme : (1) les sondages directs effectués par les boites de communication, (2) les groupes de discussion, (3) les questionnaires par Short Message Services (SMS), (4) les enquêtes téléphoniques, (5) le recueil par internet ou voie postale. Souvent, ces enquêtes traditionnelles coûteuses fournissent des renseignements dépassés.

Néanmoins, dans un secteur multinational concurrentiel où l'évolution est très rapide, la société devra toujours prendre les décisions satisfaisantes pour ses clients et avec un temps de réponse minimale. De ce fait, les réseaux sociaux constituent un excellent moyen que les entreprises ne doivent pas les sous-estimer. Le contenu des médias sociaux représente une trace numérique qui peut aussi entraver la réputation de l'entreprise pour une très longue période. Souvent, les entreprises détiennent leurs propres comptes de réseaux à partir desquelles elles peuvent compléter les Key Performance Indicator (KPI) issues des systèmes traditionnels comme le taux de fidélisation, taux de réclamations, le churn rate, …. Etc. Malheureusement, ces systèmes décisionnels n'intègrent fréquemment que des données internes. Ainsi, il est primordial d'intégrer des données externes comme celles issues des réseaux sociaux dans le processus métier pour améliorer le management de l'expérience client au sein de l'entreprise.

Suite à cette tendance, l'intégration des User Generated Content (UGC) a reçu, ces dernières années, un intérêt des industriels et des académiques. Les industriels se sont focalisés plus sur le stockage des données récoltées en intégrant l'écosystème Hadoop dans leurs plateformes (Oracle, IBM et Microsoft) pour traiter des données plus rapidement et soutenir de multiples formats. De plus, ils ont, d'une part, intégré les réseaux sociaux comme une source de données dans les éditeurs d'analyse et visualisation de données (QlickView, Tableau, …) et d'autre part, ils ont mis en place des outils dédiés pour analyser l'UGC en offrant des analyses portant sur l'engagement des internautes (germin8, MoodRaker, Trackur), la présence de la marque sur le web ou encore les performances des compagnes publicitaires sur les réseaux sociaux. Du coté académique, le domaine d'intégration de l'UGC est relativement nouveau. En analysant la littérature, nous constatons d'une part, l'absence d'un consensus ou d'une méthodologie pour la mise en place d'une plateforme BI qui booste le métier de l'entreprise. La majorité des travaux se focalise sur des cas réels et consacre une grande attention à la modélisation des données collectées. D'autre part, les outils disponibles sur le marché ne prennent pas en compte la langue la plus utilisée dans les réseaux sociaux comme les dialectes. En effet, ils proposent des solutions qui ne s'adaptent pas à la politique de mesure de la satisfaction client déjà adoptée au niveau de certaines entreprises multinationales. De plus, peu de travaux se sont intéressés à l'intégration de l'analyse d'opinion dans le business intelligence dans l'entreprise d'une manière particulière et le SI d'une manière générale vu que de nombreuses entreprises utilisent des réponses automatisées pour répondre au nombre élevé de commentaires des médias sociaux. Une telle automatisation peut réduire la tâche de plusieurs heures à quelques minutes. Cette automatisation n'élimine pas entièrement le travail mais elle accélére l'efficacité, ainsi, il est important de comprendre le potentiel de l'UGC. Par conséquent, le developpement d'un modèle d'analyse de sentiments est primordial vu la particularité du langage utilisé où uniquement un faible pourcentage est écrit correctement par contre, en contrepartie, le reste est mal écrit (faute d'orthographe) et il appartient à un dialecte

Dans ce papier, nous présentons une architecture évolutive, robuste, performante et agile à moindre coût afin d'absorber, analyser et réagir d'une façon efficace et proactive envers ces données. Cette architecture représente un système complet capable d'interroger les réseaux so-

ciaux pour principalement recueillir les données en temps réel ou différé. Elle permet d'assurer leur bonne circulation, et enrichir les données collectées par l'interprétation des sentiments des consommateurs avec le plus haut degré de précision. A cet effet, nous nous sommes appuyés sur des dictionnaires conçus à partir du corpus afin de proposer une démarche basée sur la phonétique des mots. Cette démarche permet de minimiser les fautes d'orthographe et le bruit.

Le papier est structuré comme suit. Une discussion de la littérature est présentée dans la section 2. Dans la section 3, nous décrivons les architectures décisionnelles possibles puis dans la section 4, nous présentons notre méthodologie et ses activités. Dans la section 5, nous discuterons notre étude de cas, tandis que dans la section 6 nous tirons les conclusions.

2 Etat de l'Art

L'intégration de l'UGC n'a pas été profondément étudiée dans le domaine des bases de données. En nous penchant sur la littérature seuls quelques articles se sont penchés sur l'image complète ce domaine à ce jour.

Dans le contexte décisionnel, (Dinter et Lorenz, 2012) ont énoncé un agenda de recherche pour la sociale BI, tandis que (Rosemann et al., 2012) ont cherché à faire progresser la conception conceptuelle du BI avec des données identifiées à partir de réseaux sociaux entre autres par une discussion de Social Customer Relationship Management SCRM et la sociale BI. Par la suite, des architectures complètes pour SBI ont été proposées ; (Gallinucci et al., 2013; Francia et al., 2014, 2016) ont proposé une architecture de référence pour supporter le processus SBI où les données récupérées des réseaux sociaux sont stockées après leur traitement (enrichissement) dans un data mart sous forme de cubes multidimensionnels interrogeables par le biais de techniques OLAP. Une nouvelle approche pour la modélisation des topics dans les systèmes ROLAP a été proposée nommée Meta-Star qui utilise la méta-modélisation couplée aux tables de navigation et aux tables de dimensions traditionnelles, ce qui permet d'avoir un modèle dynamique et une sémantique très flexible. Le principal atout de cette architecture est la capacité native de fournir des informations historiques, ce qui permet de surmonter les limites des approches traditionnelles à l'analyse de l'UGC textuelle, où seuls des rapports statiques sont fournis et les données historiques ne sont pas disponibles. En particulier, (Rehman et al., 2012; Cuzzocrea et al., 2015; Kraiem et al., 2014) ont proposé des solutions pour l'extraction et l'analyse de flux Twitter.

D'autres travaux se sont intéressés à l'enrichissement sémantique et compréhension du texte. L'enrichissement consiste à l'identification des parties pertinentes en utilisant du Traitement Automatique du Langage (TAL) ou des techniques d'analyse de texte pour interpréter chaque texte et, si possible, lui attribuer un sentiment (opinion) (Liu et Zhang, 2012). L'ensemble de ces travaux peut être réparti en trois grandes familles à savoir,(1) les techniques basées sur le lexique (Kennedy et Inkpen, 2006; Turney et Littman, 2002; Taboada et al., 2011; Dave et al., 2003), (2) les techniques basées sur l'apprentissage automatique (Dave et al., 2003; Pang et Lee, 2005; Pang et al., 2002) ainsi que (3)les techniques hybrides (Mudinas et al., 2012; Bahrainian et Dengel, 2013; Liu, 2012; Medhat et al., 2014). L'approche basée sur le lexique repose entièrement sur un lexique de mots d'opinion prédéfinis, l'approche basée sur l'apprentissage automatique est implémentée en construisant un classifieur (Naïve Bayes, SVM, Régression, Metric Labelling) tandis que l'approche hybride tire le meilleur de la combinaison

des deux précédentes méthodes pour atteindre une précision plus élevée lors de la classification.

Enfin, dans (García-Moya et al., 2013), un modèle de données multidimensionnel est proposé pour intégrer les données de sentiment extraites à partir de postes dans un entrepôt de données d'entreprise

3 Architecture alternative d'une plateforme décisionnelle

Dans cette section, nous discutons un certain nombre d'architectures pour intégrer l'UGC dans le métier de l'entreprise. Pour chaque architecture, nous soulignons les points forts et les défauts. Dans la figure 1, nous avons représenté trois architectures possibles.

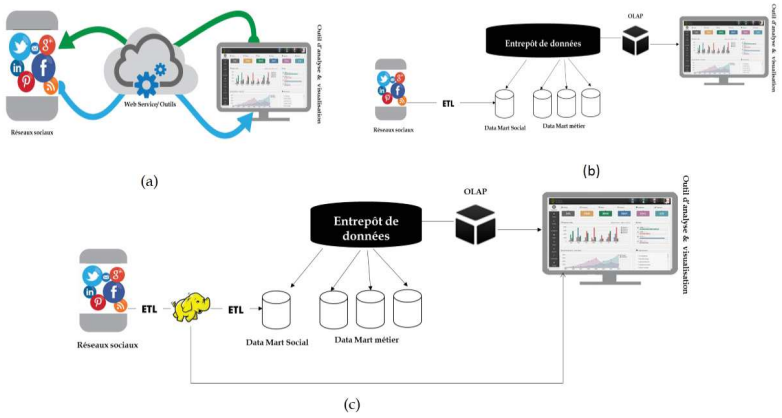

FIG. 1 – *Architecture alternative d'une plateforme décisionnelle*

Dans la figure 1 (a), l'idée de base de cette architecture consiste à utiliser un service web qui joue un rôle intermédiaire entre l'outil de recommandation et/ou visualisation et les API des différents médias sociaux. Son rôle consiste donc à gérer l'accès aux données sociales selon l'utilisateur authentifié et de lui fournir un mécanisme pour la consommation de données agrégées d'une manière simple et pratique. Le principal avantage provenant de cette architecture est qu'elle est facile à mettre en place, non couteuse en infrastructure matérielle et logicielle, ainsi que la possibilité d'évolution analytique selon les besoins métiers et l'évolution des services. Cependant, son principal inconvénient est le fait qu'elle fournit des analyses très restreintes surtout sur les dimensions Date et Temps. De plus il n'est ni possible de faire des analyses sur plusieurs dimensions au même temps. Ni de combiner entre les données des médias sociaux et celles de l'entreprise vu leur degré d'agrégation élevé.

Dans la figure 1 (b), la deuxième solution n'est rien d'autre qu'une implémentation des architectures classiques. Les médias sociaux sont considérés comme une source de données semi structurées. L'extraction de ces dernières peut être effectuée avec des composants spécialisés des ETL. Une fois les données extraites, elles sont transformées pour être stockées dans une base de données relationnelle. La dernière étape consiste à explorer ces données avec un outil

de data mining et un tableau de bord contenant les différents indicateurs clés de performance. L'avantage de cette architecture est en adéquation avec l'existant technique de l'entreprise et elle permet la combinaison entre les données des médias sociaux et les données de l'entreprise ce qui offre une analyse puissante et flexible. Cependant, l'inconvénient est la capacité de stockage et de traitement limité par rapport aux données sociales volumineux.

Dans la figure 1 (c), l'architecture consiste à utiliser un écosystème Hadoop, qui va assurer l'extraction, le traitement et le stockage des données provenant des médias sociaux et faciliter l'accès aux outils de restitution de données à savoir les outils de visualisation et ceux du Data-mining. Une telle architecture est adaptée pour un stockage d'une quantité massive de données et leur traitement rapide. Bien que cette architecture soit difficile à mettre en place et coûteuse en infrastructure matérielle et logicielle, elle reste très utile pour l'intégration de données vu qu'elle permet le développement d'un SBI, et de bons outils comme la recommandation et la réponse automatique personnalisée.

Dans la suite de cet article, nous élaborerons et mettrons en œuvre l'architecture (c) pour les raisons exposées ci-dessus, à savoir que nous allons détailler les composantes de notre plateforme.

4 Architecture proposée

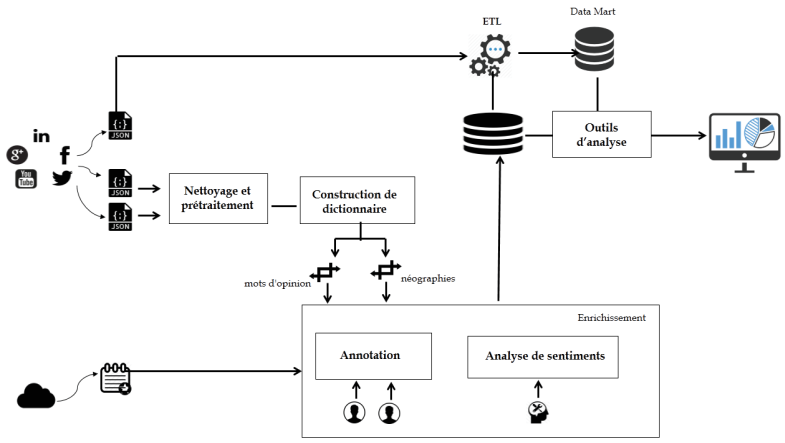

FIG. 2 – *Architecture proposée*

4.1 Collecte de données

Les informations collectées peuvent être : textuelles (Comment, Post et hashtags) ou agrégées (statistiques). La récupération des données textuelles se fait en *temps réel* et ce pour donner plus de perspectives et de possibilités pour notre système (un système de recommandation et un système d'actions instantanées). Comme les statistiques évoluent lentement, elles sont récupérées à des intervalles éloignés. Des tâches planifiées s'occupent de la récupération

de ces données à partir des réseaux sociaux sous format clé-valeur sauvegardées en mémoire au niveau du synchroniseur puis stockées sur disque.

Pour assurer la bonne collecte de données, cette partie est principalement décomposée en 3 parties :

— **Proxy d'accès Rest**. Pour pouvoir accéder aux informations publiées sur les réseaux sociaux, nous avons conçu un système autour d'un proxy pour faire abstraction sur les accès aux réseaux sociaux, en particulier, Facebook. L'objectif est d'adapter les nouvelles modifications imposées par les réseaux d'une façon transparente sans perturber l'architecture technique globale et ceci à tout moment.

— **Collecteur de données**. Une couche de collection et d'extraction de données doit être mise en place pour recueillir les données et les dispatchers dans les différentes destinations à priori la couche enrichissement de données. En effet, les statistiques sur les publications sont récupérées selon un rythme qui suit la logique de fréquences d'activités sur les réseaux sociaux qui suit un modèle *Hot-to-cold*, où la plupart des interactions se passent durant les premiers moments avant qu'ils se stabilisent suivant une loi logarithmique. Rappelons que les données textuelles sont récupérées en temps réel, une écoute en continu pour absorber ce qui arrive.

— **Synchroniseur**. Il est nécessaire de pouvoir stocker les données collectées et s'assurer que chaque donnée a été stockée et envoyée vers leur cible convenablement. Cependant, dans un contexte Big Data, les rythmes de stockage et celui d'extraction peuvent différer. En effet, les données arrivent souvent avec une vélocité assez importante. Un système de messagerie s'impose pour garantir que les données soient mises dans une file d'attente et que chaque donnée a été stockée et/ou traitée. Pour cela, le synchroniseur se présente comme un système de messagerie distribué fonctionnant comme producteur-consommateur ayant un ensemble de composants de base. Ainsi, le synchroniseur permet de relancer les messages en cas de panne ou dans l'échec d'arrivée de message à sa destination.

4.2 Nettoyage et Pré-traitement

Une fois les commentaires récupérés, ils doivent passer par une phase de nettoyage pour ne garder que les commentaires jugés utiles. En effet, le flux collecté peut parfois contenir des messages dupliqués (même identifiant) en raison de l'API ou encore résultants de l'activité massive de certains internautes qu'on pourra qualifier de spam (identifiant différent). De plus, le système doit unifier les textes par la normalisation de la casse, désaccentuation des mots, suppression des lettres uniques isolées et également l'élimination maximale du bruit. En effet, il ne faut garder que les informations les plus représentatives et importantes. Pour cela il faut aussi : (1) suppression des mots vides, (2) substitution des liens hypertextes, (3) traitement des émoticônes, (4) substitution des hashtags (5) suppression des lettres répétées.

4.3 Construction du dictionnaire de néographies

Néographies désigne des graphies qui s'écartent délibérément de la norme orthographique. En analysant notre corpus, nous avons constatés l'existence de plusieurs écritures erronées pour une seule écriture correcte d'un mot donné. Vu l'absence de formalisme ou d'orthographe

pour le dialecte, nous avons constaté le besoin de mise en place d'un dictionnaire de néographies. Pour regrouper dans une même famille les mots du corpus selon leur prononciation *ou* s'écrivent pratiquement de la même manière. Ce processus a été mis en place afin de réduire le bruit généré par la diversité des écritures pour un même mot ; il passe par deux étapes :

— **Regroupement phonétique**. Nous regroupons tous les mots qui partagent le même code phonétique. Nous nous sommes basés sur un algorithme de Soundex (Hall et Dowling, 1980) qui permet d'indexer les mots par prononciation avec un même code. Nous aurons donc en sortie des familles composées des mots qui sont identiques en prononciation malgré des différences au niveau de leurs écritures respectives.

— **Regroupement des chaines de caractères par similarité**. Le fait de regrouper des mots en fonction de leur Soundex ne suffit pas pour les faire correspondre et les considérer comme étant un seul et même mot. En effet, nous devons également prendre en compte l'écriture des mots, ce qui implique une nouvelle étape de regroupement en fonction de la similarité des chaines de caractères. Pour ce faire, nous utilisons la mesure de distance de Jaro_Winkler (donnée par la formule ci-dessous) qui permet de juger de la ressemblance de deux chaines de caractères, cette mesure est particulièrement adaptée quand il s'agit de chaines courtes. Plus la valeur est élevée, plus les chaines sont similaires, le résultat est normalisé pour avoir une mesure entre 0 et 1, 0 indiquant l'absence totale de similarité.

$$D_jw(M_1, M_2) = D_j(M_1, M_2) + min(lpc(M_1, M_2)) * q * (1 - D_j(M_1, M_2)) \quad (1)$$

$$D_j(M_1, M_2) = (1/3) * (m/|M_1| + m/|M_2| + (m - t)/|m|) \quad (2)$$

Où $lpc(M_1, M_2)$ est la longueur du préfixe commun, $q = 0.1$, $D_j(M_1, M_2)$ est la distance de *Jaro*, m est le nombre de caractères correspondants, t est le nombre de transpositions.

Par la suite, ce dictionnaire sera utilisé pour substituer certains mots par leur forme la plus courante dans le dictionnaire. Ce type de substitution est fondé sur le fait que le langage utilisé (dialecte) ne possède pas de standard d'écriture pour distinguer une orthographe juste d'une orthographe fausse. En ce qui concerne les fautes d'orthographes, nous avons pris la décision de favoriser la forme la plus courante du mot au lieu de la forme correcte.

Cette étape du prétraitement a permis de réduire considérablement le nombre de mots que comporte le vocabulaire de tout le corpus, cette amélioration est estimée à presque 50% de la taille du vocabulaire.

4.4 Construction du dictionnaire de mots d'opinion

Une des composantes essentielles de tout système d'analyse de sentiments est le dictionnaire de mots d'opinion utilisés pour la détection et la catégorisation des opinions. Nous avons décidé d'étendre des dictionnaires existants que nous enrichissons grâce à notre contribution pour une meilleure couverture du vocabulaire utilisé. Nous avons donc adopté une approche semi-automatique pour construire notre propre dictionnaire de mots d'opinion. Ces derniers seront divisés en deux familles à savoir les mots positifs et les mots négatifs. Pour ce faire, nous avons procédé en deux étapes complètement indépendantes qui sont :

— **Sélection de l'ensemble de mots graines**. Nous avons collecté manuellement, un ensemble initial de mots que nous avons jugés positifs ou négatifs parmi les mots les plus fréquents du corpus que nous avons parcouru. Notre sélection a pu dégager 305 mots positifs et 275 autres négatifs
— **Extension de l'ensemble initial grâce à SO-PMI**. A partir de cet ensemble initial, nous avons appliqué la formule de calcul mise en place par (Turney et Littman, 2002) et qui se base donc sur les cooccurrences des mots pour émettre l'hypothèse suivante : *Deux mots qui ont une très grande valeur de l'information mutuelle (PMI) ont tendance à avoir la même orientation sémantique (même sentiment)*. Nous avons donc calculé la valeur de la PMI entre chaque mot du vocabulaire et l'ensemble de mots positifs afin de déterminer la force de leur association. Nous calculons la même valeur avec l'ensemble de mots négatifs, l'écart entre ces deux valeurs détermine l'orientation du mot en question. En effet, plus l'écart tend vers des valeurs grandes et positives, plus le mot est jugé positif, inversement pour les mots d'opinion négatifs. Nous repassons manuellement chaque mot déterminé par la méthode de PMI avant de l'ajouter à l'un des ensembles afin d'être sûr de la véracité de l'orientation du mot. Grâce à cette technique, nous avons pu ajouter près de 200 mots positifs et 418 mots négatifs.
— **Traduction en langage dialectal à partir de dictionnaires**. Dans le but d'enrichir un peu plus notre dictionnaire et de porter des ressources déjà disponibles, nous avons utilisé un système de traduction automatique qui permet de traduire des mots de langue comme le français ou l'anglais vers des mots du langage dialectal. Les ressources de ce système sont principalement puisées de contributions d'internautes et de particuliers qui enrichissent le dictionnaire en ligne avec leurs connaissances. Le dictionnaire en question est GLOSBE [1] qui offre une API aux développeurs afin de leur permettre de porter des ressources disponibles issues du crowdsourcing. Les résultats obtenus comportent 148 positifs et 192 autres mots négatifs.

4.5 Traitement et Enrichissement des données

Cette couche (1) assure la communication des données entre la couche de récolte et celle d'intégration de données et (2) enrichit l'ensemble des données récupérées à partir de la couche de récolte avec des informations ultérieures. Tout cela peut être réalisé via un ensemble de jobs ayant principalement comme but d'insérer les données dans la base de données mais aussi d'enrichir le contenu extrait à partir des réseaux sociaux.

4.5.1 Annotation de corpus

La phase de l'annotation du corpus nous permet de construire le corpus labellisé. En effet, nous avons utilisé deux techniques d'annotation, manuelle et automatique.

— **Annotation manuelle.** Nous avons conçu une application collaborative afin d'accélérer et de faciliter le processus de l'annotation. L'application permet aux annotateurs de visualiser un nombre de documents de manière aléatoire ou en utilisant un filtre de mots clés et de choisir pour chaque document une classe en suivant un protocole d'annotation. Le protocole est utilisé pour choisir la classe que nous devons affecter à un

1. https ://glosbe.com/

document donné d. Chaque document est annoté par une seule personne et passe par trois nœuds de décision. Le premier nœud concerne le domaine, si le document traite un sujet d'analyse, il passe au deuxième nœud sinon il est considéré comme neutre. Le deuxième nœud porte sur la subjectivité du document, dans le cas où il ne porte aucune opinion (questions, suggestions, réponses à une question) il est considéré comme neutre. Dans le cas inverse, le document comporte une opinion qui peut être unique ; un seul sentiment envers le sujet, ou multiple où l'auteur exprime des sentiments envers plusieurs aspects d'un sujet. Si l'opinion est multiple nous prenons le sentiment le plus fort ou le sentiment dominant.

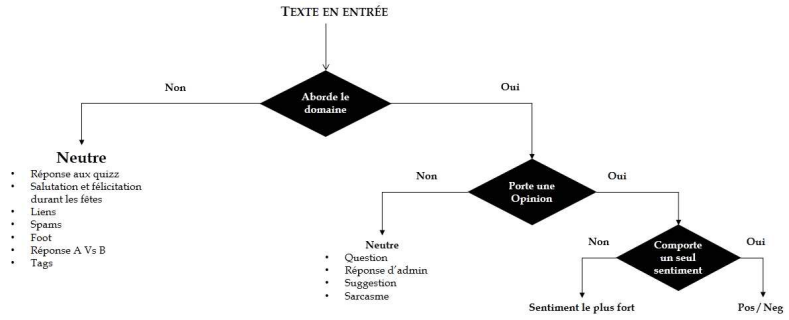

FIG. 3 – *Protocole d'annotation*

— **Annotation automatique.** Pour enrichir d'avantage le corpus par des documents annotés, une annotation automatique est proposée. Cette annotation exploite des ressources web comme les sites de comparaison et de vente en ligne qu'ils permettent aux utilisateurs de partager des critiques textuelles accompagnées d'une note représentée en étoiles. La note est généralement entre zéro et cinq. (Pang et al., 2002) proposent d'utiliser ces données comme données annotées, en transformant la note attribuée par l'auteur en polarité selon la Table 1. Ils préfèrent exclure les documents ayant une note de trois étoiles, car ils jugent cette note insuffisante pour assigner une classe au document. Le choix de la source est porté sur Google Play, il faut choisir des applications ayant une relation avec le domaine de l'entreprise

Nombre d'étoiles	Zero, une ou deux	Trois	Quatre ou cinq
Classe	Négative	Non pris	Positive

TAB. 1 – *Conversion de l'échelle d'étoiles en polarité*

4.5.2 Analyse de sentiments

A l'issue du prétraitement du texte des commentaires, ces derniers doivent passer par le processus d'apprentissage qui commence par la représentation en vecteur de caractéristiques.

Vient ensuite le choix de la fonction de pondération pour accorder plus d'importance aux mots les plus significatifs avant de passer à la phase de sélection qui nous permet de ne garder que les caractéristiques les plus discriminantes entre les différentes classes. Pour finir, nous devons entrainer un algorithme sur le corpus de documents que nous avons collecté, annoté et prétraité dans le but de concevoir un modèle de classification. Dans le but d'être sûrs d'arriver aux meilleures performances possibles, nous avons décidé d'effectuer une série de tests sur différentes combinaisons de caractéristiques, de pondérations et d'algorithmes. La phase d'évaluation nous permettra de choisir le modèle le plus précis dans la prédiction des classes des documents du corpus de test.

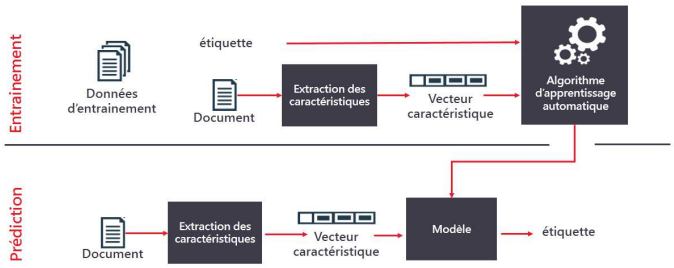

FIG. 4 – *Processus d'analyse de sentiments*

4.5.3 Détection du sujet

Notre approche pour détecter les sujets abordés par un texte est basée sur les mots clés. Le système maintient les mots clés dans un dictionnaire où il permet aux utilisateurs d'ajouter et de modifier les mots clés pour s'adapter aux besoins de l'entreprise. Afin de faciliter l'ajout des mots clés dans le système, nous proposons un module de recommandation des mots clés à partir du corpus. Le module permet d'extraire à partir d'un corpus tous les mots en relation avec un sujet donné *S* représenté par une instance de mot. On dit que le mot *w* est en relation avec le sujet *S* si ces deux apparaissent souvent ensemble, autrement dit, ils sont *corrélés*. Nous effectuons donc une étude des *cooccurrences* dans le texte afin d'extraire ces mots.

La corrélation entre deux mots (S, w) peut être mesurée en utilisant les deux fonctions *L'information Mutuelle (PMI)* et le test de *CHI-2*. Notre choix s'est porté sur la *PMI* pour des raisons de performance et temps d'exécution. Pour calculer l'information mutuelle entre le sujet *S* et un mot de corpus ***w*** nous utilisons la formule suivante :

$$PMI(w, S) = log_2 \frac{P(w, S)}{P(w) \times P(S)} \tag{3}$$

tel que : $P(w)$ *est la probabilité qu'un document* d *contienne* w et $P(w, S)$ *est la probabilité qu'un document contienne les deux mots ensemble.*

Dans le but d'accélérer le calcul de la PMI, nous utilisons un *index inversé*. Les documents du corpus sont indexés périodiquement, lorsque le système est interrogé il accédera à l'index pour calculer la PMI au lieux de refaire tous les calculs à chaque interrogation. La formule de

calcul devient :

$$PMI(w,S) = log_2 \frac{NbDoc(w \cap S)}{NbDoc(w) \times NbDoc(S)} \quad (4)$$

$NbDoc(w \cap S)$ *est le nombre de documents où les deux mots apparaissent ensemble.*

Enfin, nous normalisons la valeur par la formule suivante :

$$NPMI(w,S) = \frac{PMI(w,S)}{log_2(NbDoc(w \cap S))} \quad (5)$$

5 Etude de cas

Notre approche a été déployée dans un contexte algérien ; la langue utilisée est très souvent un mélange d'arabe dialectal et de français voire de l'arabe écrit en lettres latines plus connu sous le nom d'Arabizi. L'objectif de notre projet n'est pas de concevoir un système d'analyse de sentiments uniquement, mais de mettre en place une solution globale permettant l'extraction, la sauvegarde, l'analyse et la synthèse des sentiments relatifs aux données textuelles issues des réseaux sociaux. Ceci dit, l'évaluation ne concerne que le système d'analyse de sentiments qui fait appel à des algorithmes et techniques d'apprentissage automatique. La figure 5 illustre notre architecture technique [2]

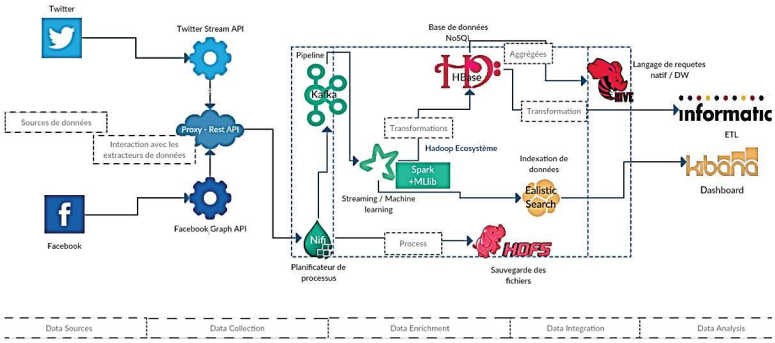

FIG. 5 – *Architecture technique de la solution*

La phase d'évaluation est une étape très importante dans la validation du modèle de prédiction pour une analyse de sentiments basée sur une approche d'apprentissage automatique. Pour ce faire, nous avons préparé un jeu de données composé de 9969 commentaires issus de la page facebook officielle d'Ooredoo Algérie. Certains commentaires ont été annotés manuellement tandis que les d'autres ont été annotés automatiquement. Nous avons découpé ce jeu de données en ensemble d'apprentissage et ensemble de test avec une part de 80% et 20% respectivement.

2. Une démonstration de la partie collecte et stockage est disponible sur youtube (https ://www.youtube.com/watch ?v=Bd4cVvKQN1A)

Dans un but d'évaluation, nous adoptons une démarche de test qui nous permet de mesurer les métriques suivantes : précision, rappel, F-Mesure. Dans ce qui suit, nous allons donc présenter les différents paramètres qui interviennent dans les performances de notre système d'analyse de sentiments.

— **Composantes du vecteur de caractéristiques.** Nous avons décidé d'évaluer l'utilisation des caractéristiques suivantes : Unigrammes (U), Unigrammes ou bigrammes d'opinion (U ou BO), Unigrammes ou bigrammes les plus fréquents (U ou B-Top).

— **Pondération** Nous représenterons nos textes comme suit : Binaire (0-1) : présence / absence, TF : fréquence du mot dans le document, TF-IDF : fréquence-fréquence de document inversée.

— **Algorithme d'apprentissage automatique.** En ce qui concerne les algorithmes entrainés, nous avons opté pour les suivants : SVM : une classification de type 1 (C-SVM) Naïve Bayes : nous avons testé deux variantes qui sont le Multinomial (désignée par NBM) et le Bernoulli (désignée par NBB). Nous avons opté pour ces algorithmes en raison de leur vaste utilisation dans le problème d'analyse de sentiments ainsi que leurs performances.

Les tableaux 2 et 3 synthétisent les résultats obtenus. Nous constatons que le Naive Bayes Multinomial (NBM) est l'algorithme qui a surpassé dans tous les cas les autres algorithmes testés quand il est utilisé avec les caractéristiques de U ou B-Top. De même, la pondération de la TF-IDF est commune aux meilleures performances dans les trois modèles retenus.

	Pondération	**Précision**			**Rappel**			**F-Mesure**		
		0-1	TF	TF-IDF	0-1	TF	TF-IDF	0-1	TF	TF-IDF
NBM	U	0,88	0,89	0,89	0,88	0,89	0,89	0,88	0,89	0,89
	U ou BO	0,88	0,88	0,88	0,87	0,88	0,88	0,87	0,88	0,88
	U ou B-Top	0,88	0,88	**0,90**	0,88	0,88	**0,90**	0,88	0,88	**0,90**
SVM	U	0,85	0,83	0,84	0,84	0,83	0,84	0,84	0,83	0,84
	U ou BO	0,84	0,83	0,85	0,83	0,82	0,84	0,83	0,82	0,84
	U ou B-Top	0,86	0,87	0,88	0,85	0,86	0,88	0,85	0,86	0,88
NBB	U	0,78	-	-	0,73	-	-	0,72	-	-
	U ou BO	0,78	-	-	0,73	-	-	0,71	-	-
	U ou B-Top	0,78	-	-	0,71	-	-	0,69	-	-

TAB. 2 – *Résultats des évaluations de la classification sur les classes Positive-Négative*

Nous avons testé une approche à deux étages en cascade : elle détecte la subjectivité avant de déterminer la polarité du commentaire. Les résultats obtenus dégradent les performance de notre système et cela est dû au taux d'erreurs qui s'est amplifié entre les deux étages.

Pour améliorer les performances, nous avons essayé de donner plus d'importance aux mots d'opinion et les mots allongés en attribuant un poids à ces derniers en raison de leur importance dans l'expression des sentiments. Nous avons également effectué une sélection des caractéristiques en utilisant la méthode de χ^2 afin de réduire la dimension du modèle. Les internautes ont tendance à utiliser une suite de lettres répétées dans le but d'intensifier le sens du mot, cette façon de communiquer est très importante pour l'analyse de sentiments en raison de l'amplification des mots d'opinion et de l'aspect psychologique véhiculé. A cet effet, nous avons attribué un poids à un mot en fonction de la longueur de la série de lettres répétées qu'il comporte. Comme nous pouvons le voir, les performances globales du système se sont dégradées

	Pondération	Précision			Rappel			F-Mesure		
		0-1	TF	TF-IDF	0-1	TF	TF-IDF	0-1	TF	TF-IDF
NBM	U	0,80	0,81	0,81	0,80	0,81	0,81	0,80	0,81	0,81
	U ou BO	0,80	0,81	0,81	0,80	0,81	0,81	0,80	0,81	0,81
	U ou B-Top	0,81	0,81	0,82	0,81	0,81	0,82	0,81	0,81	0,82
SVM	U	0,77	0,77	0,78	0,77	0,77	0,78	0,77	0,77	0,78
	U ou BO	0,76	0,76	0,77	0,76	0,76	0,77	0,76	0,76	0,77
	U ou B-Top	0,78	0,77	0,81	0,77	0,77	0,81	0,77	0,77	0,81
NBB	U	0,80	-	-	0,78	-	-	0,78	-	-
	U ou BO	0,80	-	-	0,78	-	-	0,78	-	-
	U ou B-Top	0,81	-	-	0,76	-	-	0,74	-	-

TAB. 3 – *Résultats des évaluations de la classification sur les classes Subjective-Objective*

de près de 1%. Nous expliquons ce fait, par la nature du texte catégorisé, où nous avons remarqué une utilisation excessive de l'allongement de mots par les internautes sur nos réseaux sociaux

	Précision	**Rappel**	**F-mesure**
Négative	0,74	0,76	0,75
Positive	0,8	0,82	0,81
Neutre	0,73	0,69	0,71
Moyenne	0,76	0,76	0,76

TAB. 4 – *Résultat de l'amélioration*

Dans le but d'accorder plus d'importance aux mots d'opinions, nous avons attribué un poids p pour chaque mot d'opinion dans le vecteur caractéristique. La TF-IDF des mots appartenant au dictionnaire des mots d'opinions est multipliée par ce poids pendant la phase de pondération. Nous avons testé plusieurs valeurs de p avant d'opter pour la meilleure.

Les résultats obtenus montrent une augmentation de 2% dans la précision des classes positive et négative. La classe neutre a connu une amélioration considérable de 8% et 1% en rappel et précision respectivement. Par conséquent, nous constatons une amélioration de 2% des performances globales du système.

6 Conclusion

Dans ce papier, nous avons présenté une architecture décisionnelle qui intègre l'UGC. La particularité de notre travail réside : (1) dans la construction d'un ETL robuste qui assure la bonne extraction et enregistrement des données collectées des réseaux sociaux, (2) la construction des dictionnaires pour améliorer l'enrichissement du texte collecté et le stocker dans un ODS qui sera utilisé comme une source pour un data mart dédiés aux réseaux sociaux. Le travail a été évalué dans un contexte algérien où les ressources TAL sont pauvres, les dictionnaires conçus ont données de bons résultats. Pour les travaux futurs, il serait intéressant de mettre en place des systèmes de recommandation et un système d'action (réponse automatique) pour améliorer la robustesse de l'architecture proposée.

Références

Bahrainian, S.-A. et A. Dengel (2013). Sentiment analysis using sentiment features. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, Volume 3, pp. 26–29. IEEE.

Cuzzocrea, A., C. D. Maio, G. Fenza, V. Loia, et M. Parente (2015). Towards OLAP analysis of multidimensional tweet streams. In *Proceedings of the ACM Eighteenth International Workshop on Data Warehousing and OLAP, DOLAP 2015, Melbourne, VIC, Australia, October 19-23, 2015*, pp. 69–73.

Dave, K., S. Lawrence, et D. M. Pennock (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pp. 519–528. ACM.

Dinter, B. et A. Lorenz (2012). Social business intelligence: a literature review and research agenda. In *Proceedings of the International Conference on Information Systems, ICIS 2012, Orlando, Florida, USA, December 16-19, 2012*.

Francia, M., E. Gallinucci, M. Golfarelli, et S. Rizzi (2016). Social business intelligence in action. In *Advanced Information Systems Engineering - 28th International Conference, CAiSE 2016, Ljubljana, Slovenia, June 13-17, 2016. Proceedings*, pp. 33–48.

Francia, M., M. Golfarelli, et S. Rizzi (2014). A methodology for social BI. In *18th International Database Engineering & Applications Symposium, IDEAS 2014, Porto, Portugal, July 7-9, 2014*, pp. 207–216.

Gallinucci, E., M. Golfarelli, et S. Rizzi (2013). Meta-stars: multidimensional modeling for social business intelligence. In *Proceedings of the sixteenth international workshop on Data warehousing and OLAP, DOLAP 2013, San Francisco, CA, USA, October 28, 2013*, pp. 11–18.

García-Moya, L., S. Kudama, M. J. Aramburu, et R. B. Llavori (2013). Storing and analysing voice of the market data in the corporate data warehouse. *Information Systems Frontiers 15*(3), 331–349.

Hall, P. A. et G. R. Dowling (1980). Approximate string matching. *ACM computing surveys (CSUR) 12*(4), 381–402.

Kennedy, A. et D. Inkpen (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence 22*(2), 110–125.

Kraiem, M. B., J. Feki, K. Khrouf, F. Ravat, et O. Teste (2014). OLAP of the tweets: From modeling toward exploitation. In *IEEE 8th International Conference on Research Challenges in Information Science, RCIS 2014, Marrakech, Morocco, May 28-30, 2014*, pp. 1–10.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies 5*(1), 1–167.

Liu, B. et L. Zhang (2012). A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pp. 415–463.

Medhat, W., A. Hassan, et H. Korashy (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal 5*(4), 1093–1113.

Mudinas, A., D. Zhang, et M. Levene (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pp. 5. ACM.

Pang, B. et L. Lee (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 115–124. Association for Computational Linguistics.

Pang, B., L. Lee, et S. Vaithyanathan (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86. Association for Computational Linguistics.

Rehman, N. U., S. Mansmann, A. Weiler, et M. H. Scholl (2012). Building a data warehouse for twitter stream exploration. In *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012, Istanbul, Turkey, 26-29 August 2012*, pp. 1341–1348.

Rosemann, M., M. Eggert, M. Voigt, et D. Beverungen (2012). Leveraging social network data for analytical CRM strategies - the introduction of social bi. In *20th European Conference on Information Systems, ECIS 2012, Barcelona, Spain, June 10-13, 2012*, pp. 95.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, et M. Stede (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics 37*(2), 267–307.

Turney, P. et M. L. Littman (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus.

Remerciements

Ce travail est soutenu par Ooredoo Algerie, nous remercions Messieurs Allouche Badredine, Baba-Ahmed Djillali et Benreskellah Oussama pour leur collaboration

Summary

In the BI world, the importance of facts is undeniable. Although, opinion plays a crucial role too due to its tendency to influence the decision making process. Nowadays, multiple unstructured information sources are free and open to use via the Web, such as social networks. In fact, the amount of these data is in a constant growth creating a free gold mine that can be used to collect public opinion. Companies are aware of the importance of their clients' opinions; they do not ignore in any case the importance of the public opinion that is available on social networks since the standing and growth of these platforms. In this paper we propose a robust decisional platform that collects, stores, and exploits the textual data generated by users through social networks.Finally, we present a validation of our proposals in a real-life, it's about the telecommunication area at Oreedoo Algeria.