Université Paris Nanterre

modyco Modèles, Dynamiques, Corpus UMR 7114

Thèse présentée pour obtenir le grade de docteur
Université Paris Nanterre

❧

Modyco
École doctorale Connaissance, langage, modélisation - ED 139

Discipline : Traitement Automatique des Langues

Amélioration du système de recueils d'information de l'entreprise Semantic Group Company grâce à la constitution de ressources sémantiques

---

PAR : **YAHAYA ALASSAN MAHAMAN SANOUSSI**

Sous la direction de JEAN-LUC MINEL, professeur des Universités

MEMBRES DU JURY:

Rapporteur : Iris Eshkol-Taravella Maître de Conférences (HDR), Université d'Orléans

Rapporteur : Guillaume Cleuziou ,Maître de Conférences (HDR) ,Université d'Orléans

Examinateur : Delphine Battistelli, Professeur des Universités, Université Paris Nanterre

Co-encadrant de thèse : Thierry Charnois - professeur des universités, LIPN Univ Paris 13

Directeur de thèse : Jean-Luc Minel - professeur des universités, Modyco Université Paris Nanterre

Invité : Philippe Van den Bulke - Président, Succeed Together

Mes publications

Mes travaux de recherches ont donnés lieu à deux publications :

1. *Construction de ressources sémantiques pour améliorer le clustering de messages*, Ingénierie de Connaissances (IC) 2016 (`https://hal.archives-ouvertes.fr/IC_2016/halshs-01359862`).
2. *Enhancing Short Text Clustering by Using categorized data*, CICLing 2017 (`http://www.cicling.org/2017/accepted-abstracts.html?`).

Résumé

Prendre en compte l'aspect sémantique des données textuelles lors de la tâche de classification s'est imposé comme un réel défi ces dix dernières années. Cette difficulté vient s'ajouter au fait que la plupart des données disponibles sur les réseaux sociaux sont des textes courts, ce qui a notamment pour conséquence de rendre les méthodes basées sur la représentation "bag of words" peu efficientes. L'approche proposée dans ce projet de recherche est différente des approches proposées dans les travaux antérieurs sur l'enrichissement des messages courts et ce pour trois raisons. Tout d'abord, nous n'utilisons pas des bases de connaissances externes comme Wikipedia [45] parce que généralement les messages courts qui sont traités par l'entreprise proveniennent des domaines spécifiques. Deuxièment, les données à traiter ne sont pas utilisées pour la constitution de ressources à cause du fonctionnement de l'outil. Troisièment, à notre connaissance il n'existe pas des travaux d'une part qui exploitent des données structurées comme celles de l'entreprise pour constituer des ressources sémantiques, et d'autre part qui mesurent l'impact de l'enrichissement sur un système interactif de regroupement de flux de textes. Dans cette thèse, nous proposons la création de ressources permettant d'enrichir les messages courts afin d'améliorer la performance de l'outil du regroupement sémantique de l'entreprise Succeed Together. Ce dernier implémente des méthodes de classification supervisée et non supervisée. Pour constituer ces ressources, nous utilisons des techniques de fouille de données séquentielles.

Abstract

Taking into account the semantic aspect of the textual data during the classification task has become a real challenge in the last ten years. This difficulty is in addition to the fact that most of the data available on social networks are short texts, which in particular results in making methods based on the "bag of words" representation inefficient. The approach proposed in this research project is different from the approaches proposed in previous work on the enrichment of short messages for three reasons. First, we do not use external knowledge like Wikipedia [45] because typically short messages that are processed by the company come from specific domains. Secondly, the data to be processed are not used for the creation of resources because of the operation of the tool. Thirdly, to our knowledge there is no work on the one hand, which uses structured data such as the company's data to constitute semantic resources, and on the other hand, which measure the impact of enrichment on a system Interactive grouping of text flows. In this thesis, we propose the creation of resources enabling to enrich the short messages in order to improve the performance of the tool of the semantic grouping of the company Succeed

Together. The tool implements supervised and unsupervised classification methods. To build these resources, we use sequential data mining techniques.

REMERCIEMENTS

Cette thèse ayant été réalisée dans le cadre d'une convention CIFRE, je tiens à remercier monsieur Philippe Van den Bulke, responsable de la société Succeed Together, pour m'avoir confié la conduite des travaux qui font l'objet de cette thèse. Les trois années qui ont été nécessaires à l'élaboration de cette thèse n'auraient pas été les mêmes sans la présence de mes collègues de travail. Je tiens à les remercier pour leur accueil et leurs conseils notamment lors des réunions de travail. Mention spéciale à tous les membres de l'équipe R&D : vous m'avez accueilli les bras ouverts et je garderai une place pour vous tous dans ma petite tête, je ne pouvais pas rêver mieux comme équipe pour une première expérience professionnelle. Mes discussions avec Yuqian, Albert, Patrick, Cécile me manqueront.

Je tiens par ailleurs à remercier les membres du jury : Iris Eshkol-Taravella et Guillaume Cleuzio pour avoir accepté de rapporter mon travail ainsi que Delphine Battistelli et Philippe Van den Bulke pour leur participation au jury de soutenance.

Ce travail a été élaboré sous la direction de Jean-Luc Minel, Thierry Charnois et Delphine Batistteli, à qui je tiens à exprimer toute ma gratitude d'avoir accepté de diriger mes travaux de thèse. Merci, pour m'avoir fait confiance et surtout cru en moi. Nos discussions, parfois animées, ont toujours été constructives et vraies. Vos connaissances, votre participation active, vos critiques et vos conseils, votre patience et votre disponibilité durant tout le temps passé ensemble ont énormément inspiré mon évolution professionnelle et humaine.

Mes remerciements les plus sincères vont également aux membres des laboratoires Modyco au sein duquel j'ai été intégré pendant ces trois ans. Je les remercie tous pour leur accueil et leurs conseils notamment lors des séminaires doctorales. J'ai vraiment appris plein de choses sur vous et me suis régalé des histoires et expériences qu'ils partageaient pendant la pause repas. J'oublie surtout pas de mentionner *la Dream team*, c'est comme ça que je l'appelle. C'est une équipe qui m'a accompagné lors de ma rédaction de thèse pour affiner mon rapport : Adèle Désoyer avec ses « virgules », Elise Guy-Guyenet avec sa phrase magique « il faut qu'on regarde ça ensemble » et Amal GUHA avec ses exigences pour la bibliographie. Je les remercie pour leur disponibilité, leurs remarques ont été pertinentes et ont contribué à rendre ce document lisible.

Je pense aussi à mes amis qui ont été présents pendant cette aventure, proche ou à distance, ponctuellement ou en continu, dans les moments difficiles ou pour le plaisir, mais tous toujours

là. Ahmed, Alirou, Adamou, Ismael, Marah, Souleymane, Moctar, je vous remercie pour votre soutien, votre écoute, votre folie, nos retrouvailles et bien d'autres moments passés avec vous.

Ce travail n'aurait pas atteint son ampleur sans la contribution et le soutient des personnes qui m'ont mis au monde et qui m'ont donné un magnifique modèle de labeur et de persévérance. Mes parents, j'espère que vous trouverez dans ce travail toute ma reconnaissance et tout mon amour.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

Le besoin particulier de l'entreprise Succeed Together

Cette thèse a été réalisée au sein de l'entreprise Succeed Together® dans le cadre d'un contrat CIFRE. Succeed Together® conçoit des plateformes collaboratives mettant les échanges de l'entreprise au service de la performance. Cette technologie est deployée dans le cadre de :

— Séminaires, forums, workshops

— Enquêtes et révélations d'opinion des collaborateurs ou des clients

— Déploiement des projets

— Formations à de nouveaux comportements

— Mise en oeuvre de processus d'innovation

Succeed Together® est l'éditeur de la performance collective, aujourd'hui leader sur son marché, et offre une prestation de service globale : de l'ingénierie pédagogique à l'animation en passant par la conception des supports, le training des intervenants, la logistique de mise en œuvre, jusqu'au suivi post-réunion. Afin de pouvoir répondre à la diversité des demandes de ses clients et de proposer des réponses toujours performantes quels que soient les objectifs du séminaire et son contexte logistique, Succeed Together® se doit d'améliorer les performances de son logiciel, Meeting Software®.

Lors d'une réunion ou d'une conférence, les participants sont regroupés par tables de 6 à 8 collaborateurs. Chaque table dispose d'une tablette reliée à un serveur Meeting Software® qui est piloté par un expert humain (pilote). La salle est animée par un facilitateur. Les figures 1 et 2 illustrent une réunion à laquelle Succeed Together met à disposition son outil.

FIGURE 1 – Participants autour des tablettes

FIGURE 2 – Un animateur présentant la sortie de l'outil

Au début de l'intervention, une question est envoyée à chaque tablette, précédemment définie conjointement par l'entreprise cliente et le chef de projet de Succeed Together suivant le sujet du séminaire (cf. Figure 3).

FIGURE 3 – Question posée sur une tablette

Les participants ont un temps de réponse entre 5 à 10 minutes maximum. Ils ont comme contrainte de n'envoyer qu'une seule idée par champ de réponse, mais ils ne sont pas limités en nombre de réponses. Ci-dessous un extrait de la production brute liée à la question affichée sur la Figure 3.

Fixer trois objectifs concrets lies à la formation	Les responsabiliser en leur demandant de démultiplier
Savoir déléguer et gérer la délégation	Accompagnement individuel
Autonomie	Suivre les points de progrès du collaborateur

TABLE 1 – Exemple de réponses à la question « Citez les actions que vous, managers, allez mettre en oeuvre pour faire progresser le professionnalisme de vos collaborateurs »

L'outil Meeting Software® collecte les réponses instantanément afin de les classer dans des groupes. Le pilote humain veille, en temps réel, aux processus de traitement des réponses en corrigeant les erreurs du système. Ces erreurs correspondent aux mauvais classements de l'outil. On peut distinguer deux sortes de traitement : la classification non supervisée et la classification supervisée. Le pilote n'attend pas la fin du temps imparti [1], pour l'envoi de réponses, pour actionner le traitement. Il commence par effectuer une classification non supervisée sur les premières réponses envoyées, généralement, au bout d'une minute. Les réponses restantes sont classées, au fur et à mesure qu'elles sont envoyées, en faisant une classification supervisée jusqu'à la fin du temps imparti. La Figure 4 shématise le fonctionnement de l'outil.

1. lors des séminaires un temps est donné aux participants pour envoyés leurs réponses liées à un sujet (ou question) donné.

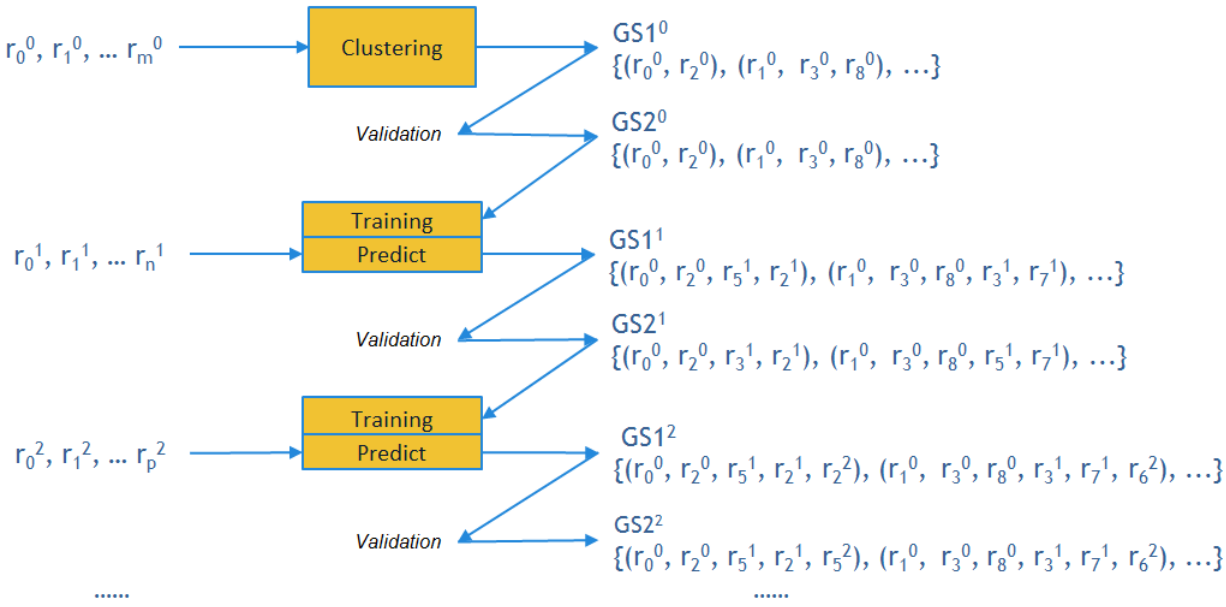

FIGURE 4 – Fonctionnement de Meeting Software®

Où :

— r_j^i correspond à la $j^{ième}$ réponse reçue par l'outil à l'étape i.

— $GS1^i$ est le regroupement des réponses effectué par l'outil à l'étape i.

— $GS2^i$ est le regroupement des réponses effectué par l'outil à l'étape i puis valider (correction manuelle) par le pilote humain.

À la fin de ces traitements, les réponses sont classées en plusieurs groupes de sens commun. Une synthèse est ensuite générée par le pilote. La figure 5 montre la synthèse produite à partir des réponses liées à la question précédente.

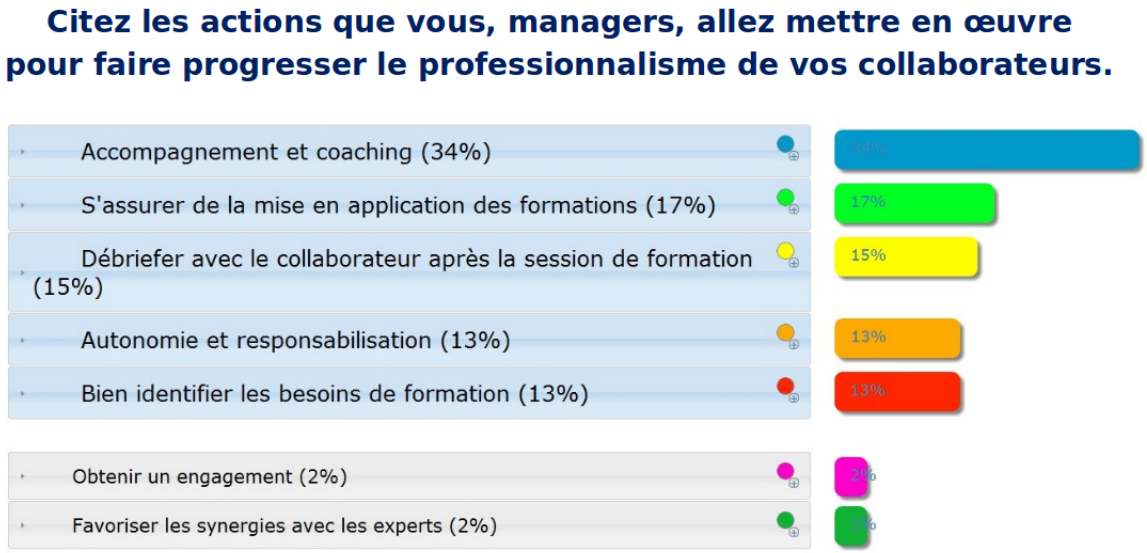

FIGURE 5 – Synthèse des réponses

Plusieurs contraintes liées au dispositif socio-technique se dégagent :

— La session interactive : Les participants interagissent avec le système via des tablèttes en envoyant leurs réponses liées à un sujet donné. La taille de chaque réponse est limitée à 255 caractères.

— Les actions du pilote : Le pilote est soumis à un stress permanent car il doit veiller au bon regroupement des réponses réalisées de l'outil pendant le temps imparti en déplaçant les réponses mal classées ou en construisant des nouveaux groupes.

— Les caractéristiques des réponses : Les réponses sont de textes très courts (messages courts) maquant de contextes. Elles contiennent souvent des fautes d'orthographes ou des abréviations.

A mon arrivé en Avril 2014, la chaîne de traitement Meeting Software® était composée de quatre modules :

— Module de pré-traitement : ce module contenait uniquement un processus lemmatisation.

— Module de vectorisation : permettant la répresentation vectorielle des messages courts. Ce module intégrait aussi le filtrage des mots vide de sens.

— Module de clustering : module composé des algorithmes de classification non supervisé. C'est le clustering de Ward qui est utilisé.

— Module de classification : module composé des algorithmes de classification supervisé. Le classifieur Random Forest [18] était utilisé pour la tâche de la classification supervisée.

La Figure 6 présente l'interface pilote de l'outil. On observe les questions posées ainsi que différentes fonctionnalités disponibles pour chacune des questions. Une fois les questions

publiées, les participants répondent. Le pilote construit les groupes en cliquant sur « Run clustering » après avoir récolté une partie des réponses. Afin d'accéder aux groupes de réponses d'une question, on clique sur « Gérer groupe ». La Figure 7 montre un exemple de groupement des réponses correspondantes à une question apparaissant sur la page du regroupement de l'outil.

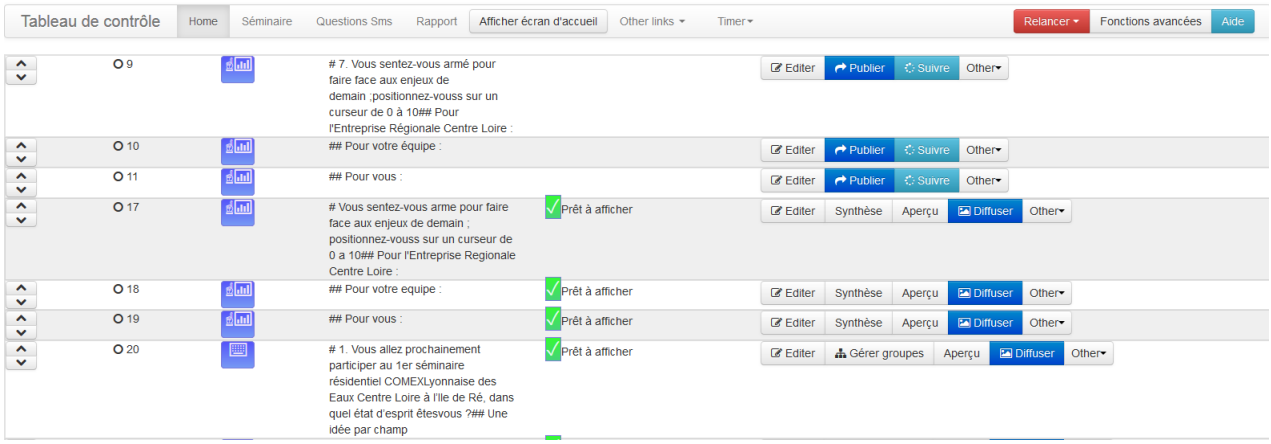

FIGURE 6 – Interface Meeting Software®

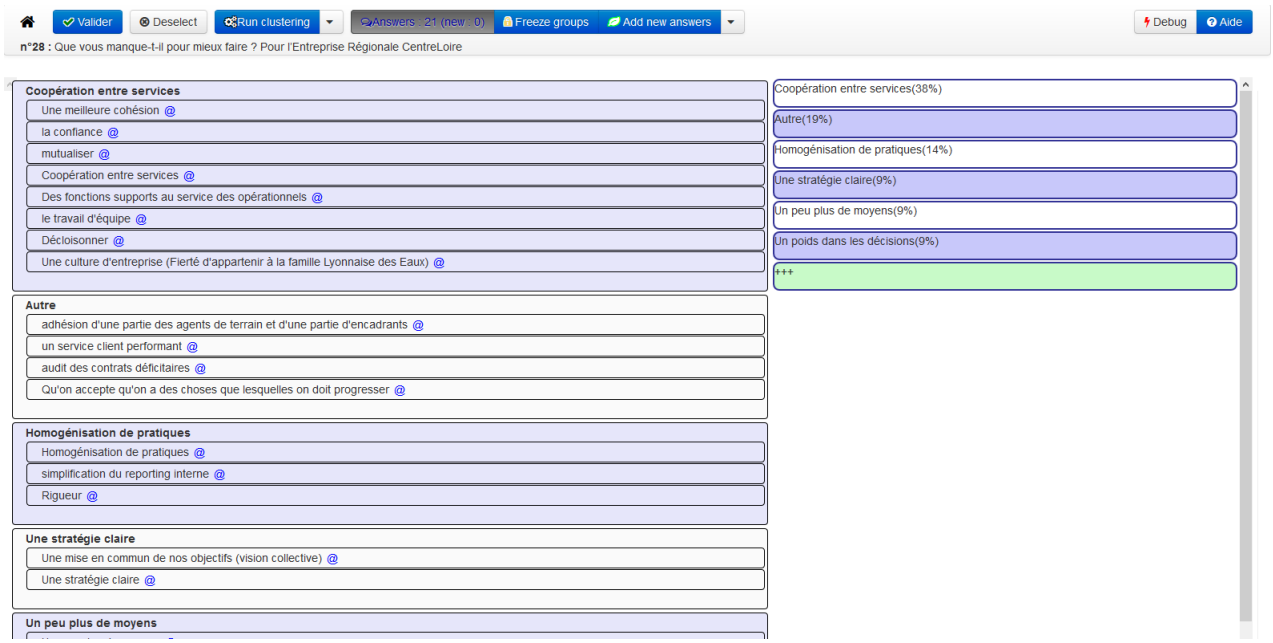

FIGURE 7 – Page pilote Meeting Software®

Les méthodes et les algorithmes de cette chaîne de traitement sont contenus dans deux dépôts :

— Bakfu : dépôt contenant les algorithmes de clustering et de classification supervisée ainsi que les méthodes de vectorisation.

— Bakfu_ms : dépôt contenant les méthodes de prétraitement des données.

Cette structure de l'outil n'est pas très lisible. Elle a le désavantage de ne pas être modulable : il est difficile de changer une méthode ou de connecter une nouvelle méthode à l'outil. De plus, les deux modules Bakfu et Bakfu_ms ne sont pas dans le même dépôt et ne sont pas connectés. Ce qui rend la détection d'éventuelles erreurs compliquée.

Le problème majeur de l'outil est qu'il est basé uniquement sur la puissance statistiques des algorithmes implémentés. Lorsque les messages courts ne partagent aucun ou peu de mot, l'outil a du mal à produire des regroupements cohérents. C'est le cas des exemples de messages courts suivants :

— **message 1** : *il faut responsabiliser les salariés*

— **message 2** : *les collaborateurs doivent être autonomes*

À ce jour, l'utilisation du logiciel exige toujours une intervention humaine (pilote) pour la production d'un regroupement cohérent. On peut distinguer trois sortes d'intervention humaine :

— Déplacement, d'un groupe à un autre, d'un ou plusieurs messages courts suite à une mauvaise classification de l'outil,

— Fusion d'un ou plusieurs groupes existants

— Création d'un ou plusieurs nouveaux groupes avec les réponses des groupes existants.

L'objectif est de minimiser le nombre de déplacements des messages courts liés à une mauvaise classification de l'outil.

Le regroupement sémantiques de messages courts

Les variations morphologiques, lexicales ou syntaxiques des termes dans la langue naturelle constituent des défis pour des applications informatiques où textes et connaissances jouent un rôle privilégié. C'est le cas de l'outil Meeting Software®, une application dédiée au regroupement et à la classification de messages courts. Les méthodes intégrées dans cet outil sont basées sur la représentation "sac de mots" ("bag of words"). Cette représentation ne prend pas en compte l'aspect sémantique des données. Par exemple les deux messages courts *être proche de ses clients* et *nous devons avoir une proximité avec nos collaborateurs* ne seront pas considérés comme appartenant à une même thématique car ils ne partagent aucun mot en commun. En effet, les scores de similarité de ces messages courts vont tendre vers 1, ce qui ne permet pas aux méthodes traditionnelles d'avoir une bonne précision.

Pour faire face à ces problèmes, la plupart des solutions proposées ont pour objectif d'enrichir la représentation de ces messages courts en injectant de la sémantique. Cette dernière peut provenir de la collection de messages courts elle-même [27] (par exemple, des entités nommées, des phrases) ou être extraite d'une base de connaissance importante externe comme Wikipedia

et WordNet . La première approche d'extraction de la sémantique requiert peu de traitements et techniques, tandis que la deuxième est exigeante en quantité de données appropriées aux messages courts traités. L'utilisation des ressources externes peut comporter des risques. Si les données ne sont pas appropriées, cela peut conduire à un ajout excessif d'information ou un ajout de "bruit".

L'approche proposée dans ce projet de recherche est différente des approches proposées dans les travaux antérieurs sur l'enrichissement des messages courts et ce pour trois raisons. Tout d'abord, nous n'utilisons pas des bases de connaissances externes comme Wikipedia parce que généralement les messages courts qui sont traités par l'entreprise proveniennent des domaines spécifiques. Or Wikipedia est une base large et très générale. Dans notre cas, c'est l'historique de données de l'entreprise qui est utilisé. Deuxièment, les données à traitées ne sont pas utilisées pour la constitution de ressources à cause du fonctionnement de l'outil. En effet, les données sont traitées par paquet dans l'outil. Cela ne permet pas pour une étape de traitement d'avoir une masse de données conséquente pour la constitution de ressource. Troisièmement, à notre connaissance il n'existe pas des travaux exploitant des données structurées comme celles dispose Succeed Together pour la création de ressources sémantique puis mesurer l'impact sur un système dynamique.

Le déroulement de la thèse et guide de lecture

Cette thèse a été réalisée dans le cadre d'un contrat CIFRE associant l'entreprise Succeed Together et l'université Paris Nanterre. L'encadrement a été assuré par Jean-Luc Minel [2] et Thierry Charnois [3] au niveau de l'université et le chef du pôle R&D au niveau de l'entreprise. La première année ma presence en entreprise a été de quatre jours par semaine. L'objectif étant de s'impregner des méthodes et outils utilisés en entreprise. Ma présence a évolué pendant les années suivantes, je passais deux jours en entreprise en deuxième année et un jour en troisième année pour me consacrer à la redaction de la thèse.

Ce document est composé de trois parties :

— état de l'art, qui contient quatre chapitres :
  — Chapitre 1 : Ce chapitre décrit les caractéristiques des messages courts et le mode de fonctionnement des algorithmes de classification (supervisée et non supervisée). Il définit aussi une taxonomie des algorithmes en présentant ceux qui sont fréquemment utilisés.
  — Chapitre 2 : Ce chapitre présente un état de l'art sur la fouille de motifs séquentiels, plus particulièrement les motifs fréquents et les motifs émergents. Les mo-

2. Modyco
3. LIPN

tifs séquentiels ont été utilisés ces dix dernières années dans les tâches d'extraction d'information [24] [23]. Ils sont utilisés dans nos travaux pour la constitution de ressources.

  - Chapitre 3 : Ce chapitre présente une taxonomie des ressources sémantiques utilisées dans la litterature. Il s'agit de situer les ressources sémantiques construites parmi les ressources présentent dans la littérature.
  - Chapitre 4 : Ce chapitre décrit l'architecture des logiciels de traitement sémantique de textes. On y présente la particularité de l'architecture de l'outil Meeting Software®.

- Construction des ressources sémantiques : cette partie détaille l'approche utilisée pour la construction des ressources et leur utilisation. Elle est composée de deux chapitres :
  - Chapitre 5 : Ce hapitre présente les données avec lesquelles, nous construisons nos ressources. On trouvera les différentes sources de données (les opérations) ainsi que des statistiques descriptives sur l'historique des données de l'entreprise.
  - Chapitre 6 : ce chapitre décrit l'approche que nous proposons pour la construction des ressources.
- Expérimentation et bilan : Cette partie détaille le banc de test mis en place pour évaluer l'impact de l'utilisation de ressources sémantiques construites sur la chaîne de traitement Meeting Software. Elle est composée de 2 chapitres :
  - Chapitre 7 : Ce chapitre présente les chaînes de traitement comparées ainsi que les données sur lesquelles l'évaluation est effectuée.
  - Chapitre 8 : Ce chapitre présente les expérimentations effectués. Pour chaque expérimentation, on présente les données sur lesquelles la ressource est construite, les données tests ainsi que les résultats obtenus.

PREMIÈRE PARTIE

état de l'art

CLASSIFICATION DES MESSAGES COURTS

1.1 Introduction

Grâce aux nouvelles technologies, les messages courts sont devenus omniprésents dans notre société. Ils prennent la forme de SMS sur téléphones mobiles, de micro-blogs comme sur Twitter, de commentaires sur les réseaux sociaux comme Facebook ou Google+, etc. Leur particularité consiste en une brièveté imposée à la fois par le médium et une volonté d'échanger l'information brute et instantanée. Ces messages représentent une richesse, en termes de quantité d'information, qui pourrait être utilisée pour analyser un climat politique, prédire des crises ou corriger les défauts d'un produit. En particulier, ils sont devenus un nouvel outil de communication directe entre un vendeur et ses acheteurs, entre les politiques et les électeurs, entre les dirigeants d'entreprise et leurs salariés. Par exemple, Morinaga et al. expliquent dans [54] comment ils vérifient les réputations de produits ciblés en analysant les critiques des clients. Le nombre de messages, la vitesse à laquelle ils sont produits et leur nature spontanée nécessitent de nouveaux moyens d'analyse pour en faire ressortir des tendances globales utiles. Les méthodes de classification font parties des méthodes les plus utilisées dans la littérature pour rendre cette masse de données exploitable.

L'objectif de ce chapitre est de faire un état de l'art non "exhaustif" des méthodes de classification (supervisée et non supervisée) et leurs critères d'évaluation. Il s'agit d'expliquer les principes sur lesquels les méthodes de classification sont construites ainsi que de mettre l'accent sur celles les plus couramment utilisées. Le lecteur désirant plus de détails pourra se référer à [35] [15] pour la classification non supervisée et à [5] pour la classification supervisée.

Notations :

— Soit $X = \{m_1, m_2, ..., m_N\}$ l'ensemble de N messages courts ;

— Soit $G = \{g_1, g_2, ..., g_K\}$ un ensemble de K groupes, résultat d'un algorithme de classification ;

— Soit $M = \{M_1, M_2, ..., M_T\}$ le regroupement manuelle appelé référence ;

— Soit $|g_i|$ le nombre d'objets contenus dans le cluster g_i ;

— Soit $d(m_1, m_2)$ la distance entre les objets m_1 et m_2 ;

— Soit $D(g_1, g_2)$ la distance entre les groupes g_1 et g_2.

Dans notre cas les objets sont les réponses liées à une question posée lors d'un séminaire ou une réunion professionnelle.

1.2 Caractéristiques des messages courts

L'objectif des messages courts est l'expression d'une idée ou opinion en utilisant un nombre limité des caractères. C'est le cas des messages mobiles qui ne dépassent pas 160 caractères, les messages tweeter avec moins de 140 caractères.

Dans [43], les auteurs auteurs décrivent les caractéristiques de ce nouveau type de texte :

— *Sparseness* : Les messages courts ont une densité faible en mots, les messages courts contiennent peu de mots (manque de contexte).

— Caractère immediat : Ce sont des messages échangés en temps réel en très grande quantité.

— Non standard : Le contenu d'un message court est concis, avec potentiellement des fautes d'orthographe. De plus, les messages sont souvent accompagnés de termes non standards et de bruit.

La plus part des méthodes traditionnelles (par exemple, KNN, le classifieur bayésien, SVM) sont utilisées généralement avec la représentation verctorielle standard des messages courts (bag of words). Cela les empêche d'avoir une meilleure précision vu la densité faible en mots des messages courts.

1.3 Classification non supervisée

L'objectif de la classification non supervisée, appélée aussi *clustering*, est de trouver une partrition (un ensemble des groupes) au sein d'un ensemble d'objets. Ces derniers sont caractérisés par des attributs qui décrivent leurs propriétés. Les objets doivent être similaires au sein d'un même groupe et dissimilaires lorsqu'ils appartiennent à des groupes différents. On trouve dans la littérature de très nombreuses méthodes de *clustering* permettant de créer ces groupes de manière automatique, chacune utilisant une stratégie et un objectif différents.

La tâche de classification non supervisée fait appel d'une part à la notion de similarité entre les objets pour la construction des groupes et d'autre part à la notion de probabilité. Trois éléments permettent de caractériser les différentes méthodes de *clustering* :

— La manière dont les objets sont regroupés : le *clustering* procède séquentiellement en groupant les objets les plus semblables en premier lieu (clustering hierarchique) ou en groupant simultanement tous les objets en k groupes.

— Le critère de similarité entre les objets

— Le critère de similarité entre les groupes

D'après [35], les méthodes de *clustering* peuvent être séparées en quatres groupes :

— Les méthodes basées sur une distance

— Les méthodes basées sur une grille

— Les méthodes probabilistes

— Les méthodes hiérarchiques

Pour présenter les méthodes de classification non supervisée, nous avons retenu les deux principales catégories : les méthodes de partitionnement et les méthodes de classification hiérarchique. La classification hiérarchique peut être ascendante ou descendante. Quant au partitionnement, c'est une classification non hiérarchique en un nombre fixe de groupes.

1.3.1 Les méthodes de partitionnement

Les méthodes de partitionnement [15] ne construisent pas une hiérarchie entre les groupes d'objets, ils construisent directement une partition de l'ensemble d'objets en k groupes. Chaque groupe devant contenir au moins un objet et un objet devant généralement appartenir à un groupe unique. Il faut noter que dans certain cas, on peut avoir un objet appartenant à plusieurs groupes. Ces méthodes effectuent une partition initiale en k groupes et ensuite cherchent à l'améliorer en reattribuant les objets d'un groupe à un autre. Les méthodes les plus répandues de partitionnement sont celles qui visent à minimiser la somme des carrés des erreurs. Parmi ces méthodes, nous retenons la méthode de k-means [47] [4].

Méthodes de k-means

Dans cette catégorie de méthodes de partitionnement, les groupes sont représentés par leurs "centroides". Un centroide d'un groupe g_i représente la moyenne de l'ensemble des objets contenus dans le groupe. L'algorithme consiste à chosir aléatoirement k objets initiaux qui représentent les centroides initiaux. Un objet x est assigné au groupe lorsque la distance entre cet individu et le centroide du groupe est minimale. La quantité à minimiser est :

$$\sum_{i=1}^{k} \sum_{x \in g_i} d(x, c_{g_i}) \tag{1.1}$$

Avec :

— g_i, le groupe i

— c_{g_i}, le centroide du groupe g_i

Les méthodes de partitionnement permettent de traiter rapidement de grands ensembles d'individus. Ces méthodes produisent directement une partition en un nombre de classes fixé au départ. Les groupes sont facilement interprétables et représentés naturellement par les centroides. Toutefois, ces techniques de partitionnement présentent un certain nombre des problèmes :

— Au niveau du nombre de classes qui doit être fixé au départ. Si le nombre de classes n'est pas connu ou si ce nombre ne correspond pas à la configuration véritable de l'ensemble d'objet, il faut presque toujours tester différentes valeurs, ce qui augmente le temps de calcul. C'est la raison pour laquelle, lorsque le nombre des individus n'est pas trop élevé, on fait appel aux méthodes hiérarchiques.

— Bon nombre d'algorithmes de cette catégorie sont sensibles aux objets aberrants lors de l'étape d'initialisation (c'est le cas des méthodes *k-means*). Un autre type d'algorithme a été dévelopé pour remedier à ce problème, *k-médoides* (*Partition Around Medoids*) [15].

1.3.2 Les méthodes hierarchiques

La classification hiérarchique, consiste à effectuer une suite de regroupements en classes de moins en moins fines en agrégeant à chaque étape les objets ou les groupes d'objets les plus proches. Le nombre d'objets n'est pas fixé a priori mais, sera fixé a posteriori. Elle fournit ainsi un ensemble de partitions de l'ensemble d'objets [14]. Il existe deux types de méthodes :

— les méthodes ascendantes

— les méthodes descendantes

Les méthodes ascendantes (CAH)

Ces méthodes sont les plus anciennes et les plus utilisées dans la classification automatique. Supposons que nous avons N objets à classer. Les algorithmes agglomératifs suivant cette approche, définissent d'abord une partition initiale en N groupes unitaires. Ils vont alors regrouper les groupes les plus proches par rapport aux caractéristiques des objets. Il existe à ce niveau $N - 1$ groupes, un groupe étant formé de deux groupes précédement fusionnés, les autres ne contenant qu'un seul objet. Le processus se poursuit en déterminant quelles sont les deux groupes le plus proches, et en les regroupant jusqu'à ce que tous les objets soient dans le même groupe. À chaque étape de fusion des classes, un recalcul des dissimilarités entre les nouvelles classes est nécessaire.

Les méthodes de cette catégorie diffèrent selon les méthodes de calcul de similarité entre les groupes appelées aussi critère d'agrégation, les plus connues sont :

— Le critère du saut minimum compare deux groupes g_i et g_j en considerant la distance minimale entre les objets de deux groupes.

$$D(g_i, g_j) = \min_{x \in g_i, y \in g_j} d(x, y) \tag{1.2}$$

— Le critère du saut maximum compare deux groupes g_i et g_j en considerant la distance maximale entre les objets de deux groupes.

$$D(g_i, g_j) = \max_{x \in g_i, y \in g_j} d(x, y) \tag{1.3}$$

— Le critère de la moyenne compare deux groupes g_i et g_j en calculant la distance moyenne entre tous les objets de g_i et tous les objets de g_j.

$$D(g_i, g_j) = \frac{1}{n_i \cdot n_j} \cdot \sum_{x \in g_i} \sum_{y \in g_j} d(x, y) \tag{1.4}$$

Avec :

— n_i, le nombre d'objets dans g_i

— n_j, le nombre d'objets dans g_j

— Le critère de Ward consiste à choisir à chaque étape du regroupement des groupes de sorte à minimiser l'inertie intra-groupe.

$$D(g_i, g_j) = \frac{n_i \cdot n_j}{n_i + n_j} \cdot d^2(c_{g_i}, c_{g_j}) \tag{1.5}$$

Avec :

— c_{g_i}, le centre de gravité du groupe g_i

— c_{g_j}, le centre de gravité du groupe g_j

— Le critère des centres de gravité compare deux groupes g_i et g_j en calculant la distance entre les centres de gravité de deux groupes.

$$D(g_i, g_j) = d(c_{g_i}, c_{g_j}) \tag{1.6}$$

Les méthodes descendantes (CHD)

À l'inverse des méthodes descendantes, celles-ci partent de l'ensemble d'objet constituant un seul groupe et construisent, de manière itérative une partition des objets. A chaque itération, deux processus sont réalisés :

— Scinder un groupe en deux

— Choisir du mode d'affectation des objets dans les sous-groupes

Les methodes de cette catégorie construisent la hiérachie des groupes en $N-1$ étapes si on considère N objets. Dans la première étape, les objets sont séparés en deux groupes en utilisant des méthodes de dissimilarité. À chaque étape suivante du processus, le groupe ayant le plus grand diamètre se divise de la même façon. La dissimilarité moyenne entre un objet x appartenant à un groupe g_i, contenant n objets, et tous les autres objets y du groupe est définie par :

$$d_x = \frac{1}{n-1} \cdot \sum_{\substack{x \in g_i \\ x \neq y}} d(x, y) \tag{1.7}$$

1.4 Classification supervisée

L'objectif de la classification supervisée est d'apprendre à l'aide d'un modèle d'apprentissage des règles apprises sur X (l'ensemble des messages courts dont les classes sont connues et étiquetées) et prédire les classes de Y (l'ensemble de nouveaux messages courts). Cela revient à déterminer une fonction qui, à partir des descripteurs des messages courts dont les classes sont connues, permet d'associer une classe à tout nouveau message court parmi les classes disponibles. Il existe plusieurs champs d'application de la classification supervisée :

— Reconnaissance de formes : reconnaissance de chiffres manuscrits (codes postaux), reconnaissance des visages.

— Catégorisation de textes : classification d'e-mails, classification de pages web.

— Diagnostic médical : évaluation des risques de cancer, detection d'arythmie cardiaque.

On peut distinguer deux grandes familles de méthodes de classification supervisée [16] : Les méthodes discriminatives et les méthodes génératives.

1.4.1 Les méthodes discriminatives

Elles permettent de définir une séparation entre les classes en créant des frontières de décision entre elles. La classification d'une nouvelle instance se fait selon la position de cette instance par rapport aux frontières de décision. Plus de détails sur le fonctionnement de ces méthodes sont donnés dans [70] [49]. Les méthodes discriminatives les plus utilisées sont présentées ci-dessous.

Support Vector Machine (SVM)

Les machines à vecteurs support on été introduites par Cortes et Vapnik en 1995 [26]. Ils projettent les données d'apprentissage dans un espace de plus grande dimension que leur espace d'origine. Dans ce nouvel espace, ils cherchent l'hyperplan qui permet une séparation linéaire optimale des données d'apprentissage, en utilisant les vecteurs de support et les marges définies

par ces vecteurs. Les SVM sont initialement définis dans le cas d'une classification binaire. On peut distinguer deux types de jeux de données :

— **Données linéairement séparables** : Soit l'ensemble d'apprentissage X tel que $X_1 \in Y_1$ et $X_2 \in Y_2$. Suppososns que les données d'apprentissage sont dans un espace à deux dimensions (A et B), $X_i(x_1, x_2)$ avec x_1 et x_2 les valeurs de X_i pour les attributs A et B. Pour tout élément dans l'ensemble d'apprentissage, le dégré d'appartenance à l'ensemble des groupes Y est binaire $\{1, 2\}$.

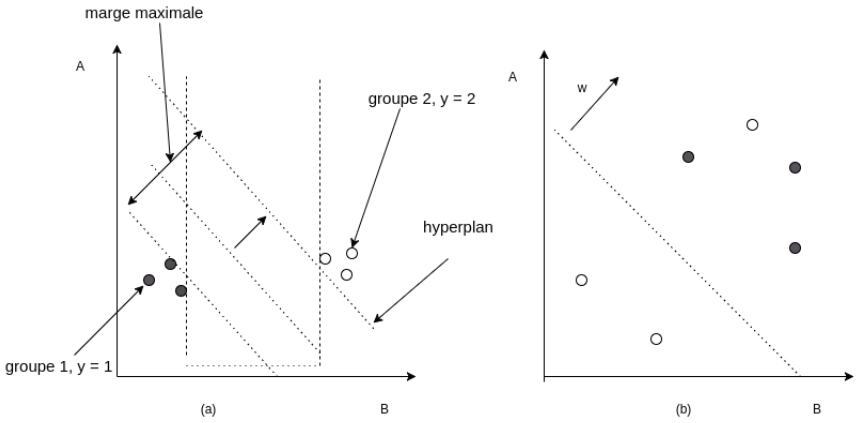

FIGURE 1.1 – Classification binaire par SVM

La Figure 1.1 montre l'existance d'une infinité des lignes qui permettent de séparer les deux groupes. Le but des SVM est de trouver le meilleur séparateur (taux d'erreur minimal). Il faut noter que dans le cas des données à trois dimensions, un plan est recherché et dans le cas de n dimensions, un hyperplan de $n - 1$ dimensions est recherché. Les machines à vecteurs support (SVM) cherchent l'hyperplan qui donne les résultats ayant les plus grandes marges. La fonction discriminante permettant de déterminer le groupe s'écrit de la manière suivante :

$$h(x) = w \cdot x + b \tag{1.8}$$

Avec w le vecteur du poids et b de x le biais. Si $h(x) \geq 0$, le groupe de x est 1 et si $h(x) \neq 0$, le groupe est 2. Le séparateur peut être alors définie par :

$$w \cdot x + b = 0 \tag{1.9}$$

Dans le cas de l'exemple précédent, l'équation (1.9) peut s'écrire sous la forme :

$$w_1 \cdot x_1 + w_2 \cdot x_2 + w_0 = 0 \tag{1.10}$$

— **Données non linéairement séparable** : Les données sont généralement linéairement inséparables, ce qui fait qu'il est difficile d'utiliser une droite pour avoir la meilleure clas-

sification. Pour résoudre ce problème, la méthode classique est de projeter les données dans un espace de dimension supérieur appelé espace de redescription. L'idée étant qu'en augmentant la dimensionnalité du problème on se retrouve dans le cas linéaire vu précédemment. Une transformation non linéaire Φ est appliquée aux données en entrées X_i tel que $X_i \in \mathbb{R}^n$ et $\Phi(X_i) \in \mathbb{R}^m$ avec $m \geqslant n$. Cette transformation permet de passer d'un produit scalaire dans l'espace d'origine $X_i \cdot X_j$ à un produit scalaire $\Phi(X_i) \cdot \Phi(X_j)$ dans le nouvel espace. Une fonction noyau notée K est utilisée pour faciliter le calcul de ce produit scalaire. Les fonctions noyaux les plus utilisées sont :

— noyau gaussien

$$K_\gamma(X_i, X_j) = \exp(-\gamma||X_i - X_j||^2) \tag{1.11}$$

— noyau polynomial

$$K_{\gamma,r,d}(X_i, X_j) = (\gamma \cdot X_i \cdot X_j + r)^d \tag{1.12}$$

Les SVM présentés traitent la classification de manière binaire. Dans notre situation, il s'agit d'une classification multi-groupes. Plusieurs méthodes ont été developpées pour étendre le schéma ci-dessus aux cas multi-groupes. Ces méthodes sont applicables à tous les classifieurs binaires [78]. Les deux méthodes les plus connues sont : *un contre un* et *un contre tous*. Supposons que nos données d'apprentissage sont reparties dans n groupes.

— *un contre un* : cette méthode consiste à construire $n \cdot \frac{(n-1)}{2}$ classifieurs en confrontant chacun des n groupes. En phase de tests, les données à classer sont analysées par chaque classifieur et un vote majoritaire permer de déterminer leur groupe respectif.

— *un contre tous* : cette méthode consiste à construire n classifieurs binaires en attribuant le label 1 aux données d'un des groupes et le label 2 aux données de tous les autres groupes. En phase de test, le classifieur donnant la valeur de confiance (la marge par exemple) la plus élévée remporte le vote.

Les arbres de décision

Les arbres de décision sont des classifieurs basées sur une approche logique. Les premiers algorithmes de classification par arbres de décision sont anciens. Les deux travaux les plus marquants sont la création de CART, par Breiman [19] et la création de C4.5 [62]. Un arbre de décision est un modèle représentant sous la forme d'un arbre dont chacun des noeuds internes représente un test sur un attribut, permettant de segmenter les données d'apprentissage en deux ou plusieurs sous-ensembles, sélon l'algorithme utilisé. Chaque noeud final, se présentant à la fin d'une branche, que l'on appelle feuille, représente une classification ou un résultat d'un test. Le noeud situé au debut de l'arbre est appelé racine. La Tigure 1.2 présente un arbre de décision contuit sur les données d'apprentissage de la Table 1.1.

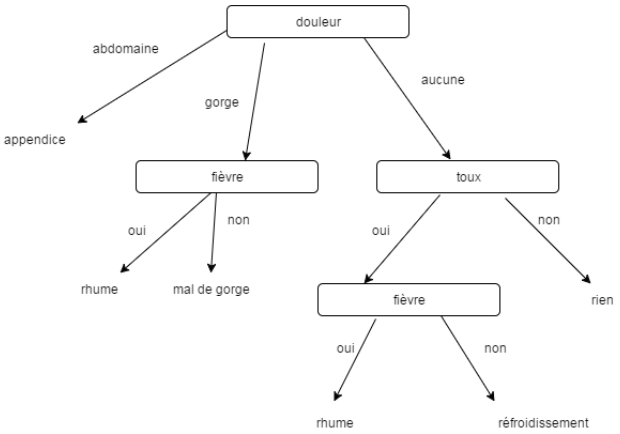

FIGURE 1.2 – Arbre de décision basé sur la table

Fièvre	Douleur	Toux	Maladie
oui	abdomaine	non	appendice
non	abdomaine	oui	appendice
oui	gorge	non	rhume
oui	gorge	oui	rhume
oui	non	non	aucune
oui	non	oui	rhume
non	non	oui	refroidissement
non	non	non	aucune

TABLE 1.1 – Exemple de données d'apprentissage pour la construction d'un arbre de décision

La construction d'un arbre de décision n'est pas faite au hasard. Le but est de déterminer les meilleurs attributs à placer au niveau de chaque noeud pour que l'arbre soit le plus petit possible et qu'il soit capable d'effectuer une meilleur prédiction. La recherche du meilleur attribut est assurée par le gain d'information par l'algorithme ID3 [7] et par l'indice de Gini par l'algorithme CART.

La rapidité et surtout la facilité d'intrépretation constituent les avantages principaux des arbres de décision. De plus, ils ne font aucune hypothèse sur les données, ce qui leur permet de traiter des ensembles d'apprentissage avec des données manquantes. Cependant, ils deviennent peu performants et complexes lorsque le nombre d'attributs et des groupes augmente.

Les réseaux de neurones

Le fonctionnement d'un réseau de neurones est inspiré de celui du cerveau humain. Il reçoit des impulsions, qui sont traitées, et en sortie d'autres impulsions sont émises. Le réseau de neurones repose alors sur la notion de neurones formel. Un réseau de neurones s'exprime sous forme d'un graphe composé de trois éléments [31] :

— Une architecture : Elle concerne le nombre et la disposition des neurones, le nombre de couches d'entrées de sorties et intermédiaires ainsi que les caractéristiques (pondération et direction) des arcs du réseau. Le nombre de neurones des différentes couches dépend du contexte d'application. Par ailleurs, la détermination du nombre de neurones à y associer demeure dans la plupart du temps arbitraire.

— Une fonction de transfert : Elle traduit le niveau d'activation d'un neurone en un état donné. Le niveau d'activation d'un neurone est obtenu en cumulant l'état de l'ensemble des entrées qui agissent sur lui. Par la suite, la fonction de transfert transforme le niveau d'activation en une valeur binaire ou continue, identifiant ainsi l'état du neurone. La fonction de transfert peut être linéaire, à seuil, stochastique et le plus souvent sigmoide

— Une règle d'apprentissage : C'est le processus d'ajustement des poids associés aux arcs lorsque le réseau est en situation d'apprentissage. La réduction de l'erreur entre la valeur de sortie du réseau et la valeur initiale dans l'ensemble d'apprentissage permet de déterminer les paramètres (poids) du réseau.

La Figure 1.3 représente un réseau de neurones multi-couches.

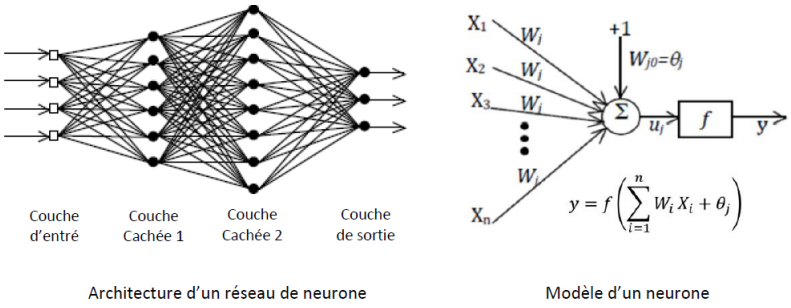

FIGURE 1.3 – réseaux de neurones multi-couches

Il existe différents types de réseaux, selon le nombre de couches, la fonction de transfert ou l'architecture elle-même du réseau : Perceptron, Adaline et le réseau de rétropropagation [80]). Les réseaux de neurones ont l'avantage d'être souple, ils peuvent traiter les problèmes de classification supervisée et non supervisée. Ils sont beaucoup plus puissant que la plupart des classifieurs. C'est le cas du modèle *Recurrent Neural Network* (RNN), qui à la différence des classifiers traditionnels ne nécessite pas forcement l'intégration de connaissances externes pour fournir des bons résultats sur la classification de textes. En effet, le modèle modèle prends en compte les informations contextuelles des mots composant les textes. Dans [50] les expérimentations montrent que le modèle RNN fournit des meilleurs résultats par rapport aux classifiers traditionnels sur plusieurs jeux de données.

1.4.2 Les méthodes génératives

Les méthodes génératives visent à trouver une structure pour la distribution jointe de X et Y noté $P(X/Y, \Theta)$. Le paramètre Θ est estimé en utilisant l'algorithme EM [21] ou CEM [22]. L'inconvénient des modèles génératifs est que, dans le cas où les hypothèses distributionnelles ne sont plus valides, leur utilisation tendra à détériorer leurs performances par rapport au cas où seuls les exemples étiquetés sont utilisés pour apprendre un modèle (Cohen et al. 2004). Nous verrons dans cette section, un peu en détails, quelques une des méthodes les plus simples (et plus souvent performantes).

La méthode Bayesienne

La méthode de classification Naïve Bayes est une approche probabiliste de la classification basée sur le théorème de Bayes. Elle est souvent utilisée dans la classification des documents pour sa simplicité et donne aussi des résultats comparables à ceux des méthodes plus sophistiquées. Elle stipule que les mots qui sont dans un document sont indépendants. Ce qui n'est pas évident car dans le langage naturel, les mots peuvent avoir plusieurs types de dépendance. La probabilité d'appartenance d'un message court $M = \cup_{i=1}^{T} m_j$, avec m_j un mot composant le message M, à un groupe g_i est donnée par la formule suivante :

$$P(g_i|M) = \frac{P(M|g_i)P(g_i)}{P(M)} \tag{1.13}$$

Où :

— $P(M|g_i)$ est la probabilité selon laquelle, les mots du message M proviennent du groupe g_i

— $P(g_i)$, la probabilité *a priori* associant le message M au groupe g_i

— $P(M)$, la probabilité propre au message M

Pour l'ensemble de groupes de messages courts G, le groupe associé au message M est celui qui fournit le max des probabilités $P(g_i|M)$. En utilisant l'hypothèse de l'indépendance des mots composant le message M, on cherche à calculer la valeur :

$$C(M) = \underset{g_i}{\operatorname{argmax}}\, P(g_i) \prod_{t=1}^{T} P(m_t|g_i) \tag{1.14}$$

Analyse discriminante linéaire (LDA)

L'analyse discriminante linéaire (ou *Linear Discriminant Analysis* en anglais) est le fruit des travaux de Fisher [81]. Elle est une méthode simple de discrimination basée sur une modélisation probabiliste des données. Elle fait partie de la famille de classifieurs binaires. L'objectif est de classer les messages courts dans deux groupes. On suppose pour cela que les messages

suivent une distribution normale. Elle présente l'avantage de pouvoir traiter des corpus de données de très grande taille. Le qualificatif linéaire fait référence à la combinaison linéaire des évènements, hyperplans, qui va être utilisée afin de séparer les groupes et de déterminer la classe d'un nouveau message court.

La construction de ces hyperplans de séparation peut être effectuée en utilisant plusieurs techniques, comme la méthode des moindres carrés ou la méthode du maximum de vraisemblance. Les hyperplans sont construits de manière à minimiser la dispersion des points d'une même catégorie autour du centre de gravité de celle-ci. L'utilisation d'une distance est alors nécessaire pour mesurer cette dispersion.

Intuitivement, nous pouvons qualifier la discrimination linéaire comme une fonction d'agrégation pondérée. Cette technique est considérée comme une méthode de classification très compacte. Le défi dans cette méthode consiste à déterminer les poids de la somme pondérée.

1.5 Critères d'évaluation

L'évaluation des méthodes de classification (supervisée ou non supervisée) exige des mesures indépendantes et fiables. Il s'avère qu'on trouve toujours un gap entre la théorie et la pratique. La masse de données d'une part et les détails subtils de la représentation des données et des algorithmes de classifications d'autre part rendent impossible un jugement intuitif. Il n'existe pas une mesure absolue pour l'évaluation des méthodes de classification, mais une variété des méthodes selon les caractéristiques des données ou des algorithmes. Dans cette section, nous présenterons les mesures d'évaluation les plus fréquemment utilisées. Ces méthodes peuvent se segmenter en deux familles : les méthodes d'évaluation interne et les méthodes d'évaluation externe.

1.5.1 Les méthodes d'évaluation interne

Ces méthodes visent à atteindre une similarité intra-groupes élevée (les messages courts au sein d'un groupe sont similaires) et une faible similarité inter-groupes (les messages courts de différents groupes sont dissemblables).

1.5.1.1 Somme de carrés des erreurs (SSE)

$$SSE(G) = \sum_{i=1}^{k} \sum_{m \in g_i} d(m - u_i) \tag{1.15}$$

avec u_i le centroide du groupe g_i et d une mesure des distance entre les messages courts. Plus la valeur est petite plus les groupes des messages courts sont compacts.

1.5.1.2 Coefficient Silhouette (CS)

Le coefficient silhouette [66] permet de vérifier si un message court est bien classé. Il est defini pour chacun des messages courts m_i :

$$CS(m_i) = \frac{b(m_i) - a(m_i)}{\max b(m_i), a(m_i)} \tag{1.16}$$

où $a(m_i)$ représente la distance moyenne séparant le message m_i des autres messages courts du groupe auquel il appartient et $b(m_i)$ la distance moyenne le séparant des autres messages courts appartenant au groupe le plus proche. Le message court m_i est bien classé si $CS(m_i)$ est proche de 1, c'est-à-dire la distance qui le sépare du groupe le plus proche est très supérieure à la distance le séparant de son groupe. Par contre, le document est mal classé si $CS(m_i)$ est proche de -1. Losque $CS(m_i)$ est proche de 0, le message court pourrait être classé dans le groupe le plus proche. Le coefficient silhouette CS pour le résultat du clustering des messages courts m_i est la moyenne des $CS(m_i)$.

1.5.1.3 Indice de Dunn (Du)

L'indice de **Dunn** [29] est définie par la formule suivante :

$$Du = \min_{1 \leq i \leq n} \left\{ \min_{1 \leq j \leq n, j \neq i} \left\{ \frac{d(g_i, g_j)}{\max_{1 \leq k \leq n} (d(g_k))} \right\} \right\} \tag{1.17}$$

Cet indice mesure la distance qui sépare deux groupes de message courts dans le résultats obtenu tout en tenant compte de la distribution des messages courts appartenent aux groupes. Plus cette distance est grande, meilleur est le résultats obetnu par l'algorithme de classification.

1.5.1.4 Indice de Davies et Bouldin (DB)

L'indice de Davies et Bouldin [28] traite individuellement les groupes de messages courts. Pour chaque groupe, l'indice mesure à quel point il est similaire au groupe le plus proche.

$$DB = \frac{1}{n} \sum_{i=1}^{n} \max_{j \neq i} \left\{ \frac{\tilde{d}(g_i) + \tilde{d}(g_j)}{D(g_i, g_j)} \right\} \tag{1.18}$$

où $\tilde{d}(g_i)$ représente la moyenne des distances entre les messages courts appartenant au groupe g_i et son centre. $D(g_i, g_j)$ représente la distance entre les centres des groupes g_i et g_j. Le meilleur regroupement est celui qui minimise la similarité entre les groupes, c'est-à-dire DB.

1.5.2 Les méthodes d'évaluation externe

Ces méthodes permettent de savoir jusqu'à quel point le regroupement issu d'un algorithme de classification correspond à la référence.

1.5.2.1 La pureté

Il s'agit d'une mesure simple et transparente. Chaque groupe prédit correspond au groupe de référence avec lequel il partage le maximum de messages courts en commun. La qualité du regroupement est mauvaise lorsque le score de pureté s'approche de 0 et meilleure plus le score s'approche de 1. La pureté est la somme des intersections maximales entre groupes prédits et groupes de référence, divisée par le nombre total des réponses traitées :

$$Purity(G, M) = \frac{1}{n} \sum_{i=1}^{n} \underset{j}{\operatorname{argmax}} \mid g_i \cap m_j \mid \tag{1.19}$$

Cependant, la pureté atteint 1 quand chaque objet constitue un groupe. Ainsi, la pureté ne peut pas être utilisée pour évaluer la qualité de la classification par rapport au nombre de groupes.

1.5.2.2 F-mesure

D'après Manning et al. [52], il est possible d'évaluer la qualité d'un regroupement en analysant toutes les paires *(i, j)* des objets regroupés et en répondant au problème de classification suivant : est-ce que les objets *i* et *j* font partie du même groupe ? Quatre ratios peuvent alors être définies :

— VP : nombre de vrais positifs (objets correctement regroupés et faisant partie d'un même groupe)

— VN : nombre de vrais négatifs (objets ne faisant pas partie d'un même groupe et annotés comme ne faisant pas partie d'un même groupe)

— FP : nombre de faux positifs (objets regroupés mais ne faisant pas partie d'un même groupe dans la vérité terrain)

— FN : nombre de faux négatifs (objets non regroupés mais faisant partie d'un même groupe dans la vérité terrain)

La valeur de F-mesure F_{PR} est obtenue par la formule suivante :

$$F_{PR} = 2 \cdot \frac{Precision \cdot Rappel}{Precision + Rappel} \tag{1.20}$$

Où $Precision = \frac{VP}{VP+FP}$ et $Rappel = \frac{VP}{VP+FN}$. La F-Mesure a quelques limites. Par exemple, dans le cas où touts les objets sont incorrectement regroupés dans un même groupe, le rappel est égal à 1 (aucun faux négatif, FN=0), pour une précision non nulle. Une F-Mesure élevée peut donc être atteinte par un système se contentant de regrouper touts les objets ensemble.

1.5.2.3 Adjusted Rand Index

Le *Rand Index* (RI) [84] est une métrique générique d'évaluation de regroupement proposée par Rand en197 [64]. Comme la F-Mesure, elle est basée sur le calcul des valeurs de VP, VN, FP et FN :

$$RI = \frac{VP + VN}{VP + VN + FP + FN} \tag{1.21}$$

C'est une mesure de performance qui correspond au à la proportion de paires de messages courts pour lesquels la le regroupement référence et le regroupement issu de l'algorithme sont d'accord. L'avantage du RI est que, contrairement à la F-Mesure, il prend en compte le nombre de VN, et donc un système regroupant toutes les messages courts en un unique groupe obtient un RI faible.

Dans [46] les auteurs proposent de réaliser un ajustement du Rand Index : l'*Adjusted Rand Index* (ARI). Cet ajustement utilise un calcul de μ qui représente l'espérance mathématique du Rand Index que peut atteindre un regroupement aléatoire des messages courts comparé au regroupement référence.

$$ARI = \frac{RI - \mu}{1 - \mu} \tag{1.22}$$

1.5.2.4 Information Mutuelle Normalisée (IMN)

L'information mutuelle est une autre mesure d'évaluation basée sur la théorie de l'information qui permet de mesurer la corrélation entre un regroupement de référence et regroupement issu d'un algorithme. Sa version normalisée [74] se présente comme suit :

$$IMN(G, M) = \frac{I(G, M)}{\frac{[H(G)+H(M)]}{2}} \tag{1.23}$$

où I est l'information mutuelle, telle que

$$I(G, M) = \sum_i \sum_j \frac{\mid g_i \cap m_j \mid}{n} \log \frac{\mid g_i \cap m_j \mid \cdot n}{\mid g_i \mid + \mid m_j \mid} \tag{1.24}$$

et H est l'entropie, telle que

$$H(G) = -\sum_j \frac{\mid g_j \mid}{n} \log \frac{\mid g_i \mid}{n} \tag{1.25}$$

Un regroupement issu d'un algorithme de classification est cohérent lorsque le coefficient IMN s'approche de 1 et mauvais lorsque le coefficient s'approche de 0. L'avantage de cette mesure est qu'elle n'est pas, contrairement à la pureté, dépendante du nombre de groupes comme la pureté.

1.5.2.5 Notre proposition pour l'évaluation du système développé : Nombre de déplacements

Les critères numériques présents dans la littérature sont un bon miroir de la qualité d'un processus de classification supervisée ou non. Cependant, comme nous l'avons évoqué, ils répondent de façon parcellaire à notre problématique. L'activité du pilote est caractérisée par le nombre de déplacements de réponses mal classées à chaque étape du processus. En conséquence, nous définissons un nouveau critère rendant compte de l'activité du pilote. Ce critère est basé sur le décompte des déplacements à chaque étape. Un regroupement de messages courts est meilleur losque le pilote effectue peu de déplacements. Il permet au pôle R&D et aux pilotes d'avoir une même approche qualitative du produit, ce qui facilite leur communication. Par ailleurs, les pilotes perçoivent concrètement les améliorations du système dans des termes qu'ils comprennent, contrairement à des chiffres peu parlants pour un public non initié.

Calcul du nombre de déplacement

Soit $G = \{g_1, g_2, ..., g_n\}$ le regroupement de messages courts tel que $g_i = \{m_{i1}, m_{i2}, ..., m_{in_i}\}$ où m_{ij} est un message court appartenant au groupe g_i. Il faut distinguer deux sortes de regroupements :

— G_{ref} : le regroupement des messages courts de référence.

— G_{pred} : le regroupement des messages courts fournit automatiquement par un algorithme de *machine learning*.

G_{ref} et G_{pred} peuvent ne pas avoir le même nombre de groupes, mais le même nombre de messages courts.

La mesure *Nombre de déplacement* répond à la question suivante : quel est le nombre de déplacements minimum à effectuer à partir de G_{pred} pour obtenir G_{ref} ? On distingue deux types de déplacements :

— Déplacements lié à une mauvaise classification : les déplacements de messages courts d'un groupe à un autre groupe existant.

— Déplacements liés au changement de décision humaine : ce sont les déplacements des groupes et les déplacements des réponses dans l'objectif de créer un ou plusieurs groupes.

Pour cette mesure, nous ne comptons que les déplacements liées à une mauvaise classification.

Pour calculer le nombre minimal de déplacements à effectuer à partir de G_{pred} pour obtenir G_{ref}, on procède de la manière suivante :

1. Trouver les correspondances des groupes à l'aide d'une table d'intersection. Cette table permet de trouver les correspondances entre les groupes de l'ensemble G_{ref} et l'ensemble G_{pred} en se basant sur le nombre de messages courts maximum en commun entre les

groupes. Les lignes de la table correspondent aux groupes de G_{pred} et les colonnes aux groupes de G_{ref}. La recherche de la correspondance d'un groupe g_{pred_i} dans l'ensemble G_{ref} est modélisée par :

$$\max_j \mid g_{pred_i} \cap g_{ref_j} \mid \tag{1.26}$$

2. Rechercher des combinaisons des correspondances nécessitant le moins de déplacements de messages courts.
3. Choisir parmi les combinaisons trouvées à l'étape précédente, celles dont le nombre de déplacements dédiés à la création de nouveaux groupes est minimal.

Exemple

Soient :

$G_{ref} = \{\{m_0, m_4, m_7\}, \{m_3, m_5\}, \{m_1, m_2, m_{10}\}, \{m_6, m_8, m_9\}\}$
$G_{pred} = \{\{m_0, m_6, m_7, m_{10}\}, \{m_1, m_2, m_3, m_4, m_5\}, \{m_8, m_9\}\}$

Table d'intersection

On obtient les correspondances suivantes (Table 1.2) :

— g_{pred_1} correspond au groupe g_{ref_1}
— g_{pred_2} correspond aux groupes g_{ref_2} et g_{ref_3}
— g_{pred_3} correspond au groupe g_{ref_4}

	g_{ref_1}	g_{ref_2}	g_{ref_3}	g_{ref_4}
g_{pred_1}	2	0	1	1
g_{pred_2}	1	2	2	0
g_{pred_3}	0	0	0	2

TABLE 1.2 – Correspondance entre les groupes prédits et les groupes références

Combinaison de correspondances

On trouve deux combinaisons possibles (Table 1.2) :

— **comb1** : g_{ref_1}, g_{ref_2}, g_{ref_4}
— **comb2** : g_{ref_1}, g_{ref_3}, g_{ref_4}

La combinaison choisie est *comb1* car elle nécessite moins de déplacements de messages courts, pour la création des nouveaux groupes que *comb2*. Le nombre total de déplacements est 5 et le nombre de déplacements lié à la mauvaise classification est 2 (cf. Table 1.3).

Nombre de déplacements	**comb1**	**comb2**
nb de déplacements pour création de groupes	2	3
nb total de déplacements	5	5

TABLE 1.3 – Nombre de déplacements pour chaque combinaison

1.6 Conclusion

Le mode de fonctionnement de l'outil Meeting Software a influencé le choix des algorithmes de classification (non supervisée et supervisée). L'outil est utilisé en temps réel et traite des petites quantité des données par flux. Le choix des algorithmes a été basé sur les critères suivants :

— Meilleur performance

— Coût en temps de clacul

— Lisibilité des résultats

Ce sont l'algorithme Ward et le classifieur ExtraTrees (Forêt aléatoire) qui sont implémenté dans l'outil car ils correspondent à nos critères de choix. Le premier est utilisé pour la tâche de la classification non supervisée et le deuxième pour la classification supervisée. Les forêts aléatoires font partie de la famille des méthodes discriminantes dont les performances sont meilleures que celles des méthodes de la famille générative. De plus les résultats sont facilement interprétables par l'humain. L'algorithme de Ward, un algorithme hièrarchique, est moins coûteux en temps de calcul sur des données de faible quantité. De plus il est très performant sur la détection des données aberrantes.

Meeting Software nécessite l'intervention d'un pilote humain qui veille à son bon classement. L'intervention du pilote correspond aux déplacements des messages courts mal classés dans les bons groupes. L'objectif de l'entreprise est de limiter cette intervention humaine. L'utilisation de critères traditionnels (*Rand Index*, pureté, ...), bien qu'il s'agisse d'approches scientifiques nécessaires pour nous comparer à la littérature, ne nous permet pas d'apprécier correctement l'avancée de nos travaux d'un point de vue utilisateur. C'est pourquoi nous nous sommes tournés vers une évaluation plus proche de l'expérience terrain, le decompte du nombre de déplacements.

CHAPITRE 2

LA FOUILLE DES MOTIFS SÉQUENTIELS

2.1 Introduction

La masse des données échangées sur internet augmente de manière exponentielle d'année en année. La firme IDC [1] mandatée par EMC [2] (spécialiste des logiciels et systèmes de stockage) a réalisé une étude qui montre la prolifération de ces données. D'après l'étude, cette explosion du volume mondiale des données est dû à l'augmentation du nombre d'entreprises et d'individus se connectant sur Internet et surtout aux objets connectés. Les principaux enseignements de cette étude sont les suivants :

— Les objets connectés contribueront à doubler le volume des données mondiale tous les 2 ans. Ce volume sera en 2020 de 44.000 milliards de gigaoctets, soit 10 % plus qu'en 2013.

— En 2013, seules 22 % des données numériques étaient exploitables et 5 % seulement d'entre elles ont été analysées. En 2020, l'expansion de l'Internet des objets devrait porter à 35 % la proportion des données exploitables.

Ces données constituent un enjeu majeur pour les industriels. D'après l'étude *étude big data analytics* [3] réalisée en 2016 en France, dans les 12 prochains mois :

— 26 % des entreprises prévoient d'acheter une analyse prédictive de leurs données.

— 26 % des entreprises prévoient d'avoir une visualisation de données

— 25 % des entreprises prévoient d'effectuer une analyse multi-dimentionnelle de données.

Pour satisfaire les besoins des industriels, la nécessité de mettre en place des techniques permettant de résumer automatiquement les données s'est imposé. Ces techniques ont pour objectif d'extraire l'essentiel de l'information contenue dans une masse de données importante. En 1989, Gregory Piatetsky-Shapiro [59] a donné un nom à cette discipline : Extraction de Connaissances à partir de Données (ECD), ou en Anglais, Knowledge Discovery in Databases

1. http://www.idc.fr/
2. https://www.emc.com/leadership/digital-universe/index.htm
3. http ://www.idc.fr/infographies

(KDD). Cette discipline puise ses racines des statistiques, de l'intelligence artificielle et de l'apprentissage automatique. Généralement le processus d'extraction de connaissance est composée des étapes suivantes [**AlMarascu**] :

— La sélection des données ;

— Le prétraitement des données ;

— La transformation des données ;

— La fouille de données ;

— L'interprétation et l'évaluation des modèles.

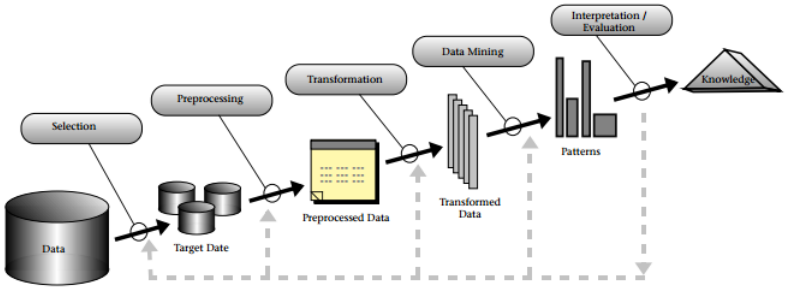

FIGURE 2.1 – Étapes de l'ECD

Dans notre cas, la fouille de données s'effectue sur des messages courts. Ce type de données fait parti des exemples de types de données, présentés par [85], sur lesquels s'applique la fouille de données. Ce sont des fichiers en format texte, contenant un enregistrement par ligne, avec des champs séparés par des tabulations. Chaque ligne correspond à une réponse lisée à une question posée.

Après cette présentation générale, nous allons nous focaliser dans les sections suivantes sur la fouille de données plus précisement sur les motifs séquentiels.

2.2 Les motifs fréquents

La fouille des motifs séquentiels est une technique de fouille de données dont l'objectif est d'extraire des connaissances sous forme de motifs (ou régularités) dans des bases de données dans lesquelles l'ordre temporel caractérise les données (Agrawal et Srikant, 1995). Les définitions suivantes proviennent du [63], elles ont été adaptées au cadre de cette thèse.

2.2.1 Définitions

Définition 1 (Environnement d'extraction)

Soit le triplet $\mathcal{K} = (\mathcal{B}, \mathcal{A}, \mathcal{R})$ un environnement d'extraction, où $\mathcal{B} = \{M_1, M_2, ..., M_n\}$ un ensemble de messages courts appelé base de messages courts, $\mathcal{A} = \{m_1, m_2, ..., m_n\}$ un ensemble de n attributs appelés des motifs et $\mathcal{R}$ une relation de support binaire entre $\mathcal{B}$ et $\mathcal{A}$ vérifiant $\mathcal{R} \subseteq \mathcal{B} \times \mathcal{A}$.

Définition 2 (Relation de support)

Soient M un message court et m un motif. m est supporté par M si m est inclus dans M.

Exemple 1

Soit une base $\mathcal{B}$ de messages courts composée de :

— M_1 : *former collaborateur méthode vente*

— M_2 : *former salarié*

— M_3 : *adaptation collaborateur système*

— M_4 : *former accompagner salarié*

et A une base composée des motifs suivants : *former*, *former salarié*, *système*. La Table 2.1 décrit la relation de support $\mathcal{R}$ entre $\mathcal{B}$ et $\mathcal{A}$.

	former	former salarié	système
M_1	X		
M_2	X	X	
M_3			X
M_4	X	X	

TABLE 2.1 – Représentation d'un environnement d'extraction $(\mathcal{B}, \mathcal{A}, \mathcal{R})$

D'après la Table 2.1 le message court M_1 supporte le motif *fomer* et le message court M_4 supporte les motifs *former* et *former salarié*. La relation de support permet de définir le support absolu d'un motif dans une base de messages courts.

Définition 3 (Support absolu d'un motif)

Le **support absolu** du motif $m \in \mathcal{A}$ dans $\mathcal{B}$, noté $Supp_{\mathcal{B}}(m)$, est défini comme le nombre de messages courts dans $\mathcal{B}$ qui supporte m :

$$Supp_{\mathcal{B}}(m) = \mid \{m \in \mathcal{A} / (M, m) \in \mathcal{R}\} \mid \tag{2.1}$$

Exemple 2 :

Soient $m_1 =$*former* et $m_2 =$*former salarié*. D'après l'exemple précédent, le support absolu du motif *former* est $Supp_{\mathcal{B}}(m_1) =\mid \{M_1, M_2, M_4\} \mid= 3$ et le support absolu du motif former salarié est $Supp_{\mathcal{B}}(m_2) =\mid \{M_2, M_4\} \mid= 2$

Le support absolu d'un motif, généralement appelé support, nous renseigne sur le nombre des messages courts le supportant. Lorsque l'on s'intéresse à la proportion de messages courts qui supportent un motif, c'est la fréquence du motif qui est utilisée.

Définition 4 (Fréquence d'un motif ou support relatif)

La **fréquence** d'un motif $m \in \mathcal{A}$ dans $\mathcal{B}$, notée $Freq_{\mathcal{B}}(m)$, est définie telle que :

$$Freq_{\mathcal{B}}(m) = \frac{Supp_{\mathcal{B}}(m)}{\mid \mathcal{B} \mid} \tag{2.2}$$

Exemple 3 :

Les fréquences des motifs de l'exemple précédent sont $Freq_{\mathcal{B}}(m_1) = \frac{Supp_{\mathcal{B}}(m_1)}{|\mathcal{B}|} = \frac{3}{4}$ et $Freq_{\mathcal{B}}(m_2) = \frac{Supp_{\mathcal{B}}(m_2)}{|\mathcal{B}|} = \frac{2}{4}$.

Définition 5 (Motif fréquent)

Soit σ un réel tel que $0 \leq \sigma \leq 1$, appelé **seuil minimum de fréquence**. Un motif $m \in \mathcal{A}$ dont la fréquence dans $\mathcal{B}$ est supérieure ou égale à σ est appelé un motif fréquent dans $\mathcal{B}$. En d'autres termes, m est fréquent dans $\mathcal{B}$ si :

$$Freq_{\mathcal{B}}(m) \geq \sigma \tag{2.3}$$

L'ensemble des motifs fréquents dans B pour un seuil minimum de fréquence σ, noté $MFreq(\mathcal{B}, \sigma)$, est défini comme :

$$MFreq(\mathcal{B}, \sigma) = \{m \in \mathcal{A} / Freq_{\mathcal{B}}(m) \geq \sigma\} \tag{2.4}$$

Exemple 4 :

Soit un seuil minimum de fréquence $\sigma = 0.5$, alors les motifs m_1 et m_2 de l'exemple 2 sont fréquents. En effet, $Freq_{\mathcal{B}}(m_1) = \frac{2}{4} = \sigma$ et $Freq_{\mathcal{B}}(m_2) = \frac{2}{4} = \sigma$. En revanche, le motif $m_3 = système$ n'est pas fréquent car $Freq_{\mathcal{B}}(m_3) = \frac{1}{4} < \sigma$

La litterature propose aussi des définitions liées au support du motif. Dans ce cas, un motif est fréquent si son support est supérieur ou égal à un seuil minimum défini comme un entier ρ,

$0 < \rho \leq \mid \mathcal{B} \mid$. Pour plus de détail sur les motifs fréquents ainsi que leurs propriétés, on pourra se référer à [38] [63].

2.2.2 Algorithmes d'extraction des motifs fréquents

Quatre grandes approches sont proposées pour l'extraction des motifs fréquents [10] :

— La première consiste à parcourir itérativement par niveau l'ensemble des motifs. À chaque itération ou niveau, un ensemble de motifs candidats est créé en joignant les motifs fréquents découverts durant l'itération précédente ; les supports de ces motifs sont calculés et les motifs non fréquents sont supprimés. L'algorithme de référence basé sur cette approche est l'algorithme Apriori [2], proposé concurremment à l'algorithme OCD [33].

— La seconde approche est basée sur l'extraction des motifs fréquents maximaux dont tous les sur-ensembles sont non fréquents et tous les sous-ensembles sont fréquents. Les algorithmes utilisant cette approche combinent un parcours par niveaux en largeur du bas vers le haut et un parcours en largeur du haut vers le bas de l'ensemble des motifs. Lorsque les motifs fréquents maximaux sont découverts, tous les motifs fréquents sont dérivés de ces derniers et un ultime balayage de la base de données est réalisé afin de calculer leur support. L'algorithme le plus efficace basé sur cette approche est l'algorithme *Max-Miner* [11].

— La troisième approche, représentée par l'algorithme Close [57], est basée sur le cadre théorique introduit dans [58] qui utilise la fermeture de la connexion de Galois [36] [30]. Ici, les motifs fermés fréquents (et leurs supports) sont extraits de la base de données en réalisant un parcours par niveaux. Un motif fermé est un ensemble maximal commun à un ensemble d'objets de la base de données. Tout motif non fermé est inclus dans le même ensemble d'objets et possède donc le même support que sa fermeture (le plus petit motif fermé qui le contient). Tous les motifs fréquents et leur support peuvent donc être déduits des motifs fermés fréquents avec leur support, sans accéder à la base de données. En conséquence, les motifs fréquents ne sont pas tous considérés durant la phase la plus coûteuse de l'algorithme, le comptage des supports. Par ailleurs l'espace de recherche est considérablement réduit, particulièrement dans le cas de données corrélées.

— La quatrième approche est basée sur des projections [53] [75]. Les algorithmes les plus cités sont *FreeSpan* [42] et *PrefixSpan* [41], le deuxième étant une amélioration du premier. Le principe de ces algorithmes consiste à proposer des projections récursives d'une base de données en fonction des items fréquents. En effet, la base est projetée en plusieurs bases plus petites et les séquences fréquentes grandissent avec le nombre de projections. Les temps de réponses sont alors améliorés car chaque base projetée est plus petite et facile à traiter.

Le lecteur intéressé par une description complète des algorithmes d'extraction des motifs fréquents peut se référer à la thèse de N. Pasquier [56]. Une typologie de structures de données

pour l'implantation de ces algorithmes, avec des évaluations expérimentales, se trouve dans la thèse de Y. Bastide [9].

L'un des défis majeurs auquel est confronté la fouille de données est la sélection de motifs potentiellement intéressant. Dans [12], les auteurs proposent un outil en ligne pour extraire des motifs séquentiels sous contraintes :

— La contrainte de support minimal (au moins 2 dans notre cas) : Le support minimal est le nombre minimal de phrases dans lequels ce motif occurre. Cette contrainte traduit une certaine régularité des motifs produits.

— La contrainte de gap (au maximun 1 dans notre cas) : Un motif séquentiel avec contrainte de $gap[M, N]$, noté $P[M, N]$ est un motif tel qu'au minimum M items et au maximum N items sont présents entre chaque item voisin du motif dans les séquences à partir desquelles il est extrait.

— La contrainte de longueur (au maximun 2 dans notre cas) : Il s'agit de ne concerver que les motifs de longueur maximale 2 mots.

2.3 Les motifs émergents

Initialement introduits dans [1], les motifs émergents permettent de caractériser une classe d'objets par rapport aux autres classes. En effet, ils représentent les caractéristiques fortement présentes dans une classe et rares dans les autres. Dans [61], un motif M d'un ensemble G_1 par rapport à un autre ensemble G_2 est émergent si $TauxCrioss(P) \geq \rho$ où

$$TauxCrioss(P) = \begin{cases} \infty & si\ supG_2(P) = 0 \\ \frac{sup_{G_1}(P)}{sup_{G_2}(P)} & sinon \end{cases}$$

$sup_{G_1}(P)$ et $sup\_G_2(P)$ désignent respectivement le support relatif du motif P par rapport à G_1 et celui par rapport à l'union des autres ensembles noté G_2.

2.4 Conclusion

L'approche de construction de ressources proposée dans nos travaux est basée sur la fouille de motifs séquentiels. Nous nous sommes inspirés des travaux réalisés dans [12] pour implémenter en Python notre méthode d'extraction sous contraintes (gap, longueur des motifs, support minimal) de motifs séquentiels (cf. chapitre 6). Les données utilisées imposent de prendre comme support pour l'extraction des motifs fréquents et le seuil pour la selection des motifs émergents des valeurs relativement pétites. En effet, nous disposons enormement des messages courts dont la taille en mots ne dépasse pas 5. En plus, nous ne disposons pas d'un grand nombre

de messages courts la quantité maximale des messages courts utilisée pour l'extraction est 18 000).

CHAPITRE 3

LES RESSOURCES SÉMANTIQUES

3.1 Introduction

Le traitement sémantique des documents, qu'il s'agisse d'indexation sémantique, d'alignement de textes, de désambiguïsation, de regroupements sémantiques, etc., requiert des connaissances. Ces connaissances existent actuellement sous forme de ressources de différents types, telles que les terminologies, les glossaires, les ontologies (générales ou de domaine), les dictionnaires multilingues ou encore les corpus de textes (simples ou parallèles). Chaque type de ressource dispose , de multiples formalismes, langages et formats de représentation. Elles peuvent être déduites des documents étudiés ou externes.

On assiste à un développement et à la mise à disposition d'un nombre croissant de bases de connaissances. Cela a permis à des système basé sur des connaissances externes de se développer. Dans le regroupement de messages courts ces ressources jouent un rôle très important. Elles permettent de mettre en évidence le sens commun des messages courts tout en les rendant facilement exploitables par les methodes tradtionnelles. Ce chapitre est consacré à un état de l'art, d'une part sur les différentes catégories des ressources sémantiques, et d'autre part sur l'utilisation faite de ces ressources dans le regroupement des messages courts en général.

3.2 Les différents types de ressources

Dans [37], les auteurs ont organisé les ressources en deux catégories principales : les ressources autonomes et les ressources d'enrichissement.

3.2.1 Les ressources autonomes

Les ressources autonomes désignent la catégorie des ressources dont l'existence est indépendante des autres ressources :

— Ressources ontologiques : une ontologie a pour but de représenter une conceptualisation d'un domaine [39]. Cette conceptualisation consiste essentiellement en une définition des concepts du domaine et des relations existant entre ces concepts. Les ontologies sont exprimées à l'aide de formalismes qui fournissent des constructeurs pour la définition des

entités ontologiques. Suivant le formalisme utilisé, les entités peuvent être des classes, propriétés, individus et axiomes (dans les logiques de description), des concepts et relations (dans les réseaux sémantiques), des classes, objets et associations (dans les modèles à objets), etc. Le choix du formalisme dépend de l'objectif pratique poursuivi lors de la construction de l'ontologie : échange de connaissances, référence commune, raisonnement automatique (inférences logiques), structuration de données, etc.

— Ressources terminologiques : elles représentent des termes rigoureusement définis pour un domaine spécifique [82]. Ces ressources sont le résultat d'une étude théorique des dénominations des objets ou des concepts utilisés par un domaine de l'activité humaine. Cette étude se focalise sur le fonctionnement dans la langue des unités terminologiques et sur les problèmes de traduction, de classement et de documentation. Beaucoup de travaux de recherche se sont focalisés sur l'étude des terminologies surtout dans le domaine biomédical. Parmi ces ressources, on trouve les thesaurus pour les systèmes d'indexation automatique, les référentiels terminologiques pour les systèmes de gestion de données techniques, les bases de données terminologiques pour l'aide à la traduction, etc. Les thesaurus sont généralement utilisés pour la recherche d'information. Chaque ressource de connaissances peut être associée à un ou plusieurs concepts représentés à l'aide d'un ensemble de termes. Dans les thesaurus les termes sont organisés suivant un nombre restreint de relations (hiérarchiques, d'équivalence et associatives).

— Ressources linguistiques : elles représentent les types de données et informations sur la langue. Ces ressources sont plus généralement utilisées pour le traitement automatique de la langue, l'apprentissage (pour entraîner les programmes de traduction automatique par des approches statistiques). Dans ce type de ressources on trouve les documents, les corpus, les hyperdocuments, etc.

3.2.2 Les ressources d'enrichissement

Les ressources d'enrichissement désignent les ressources résultant de l'application d'un processus (automatique ou humain) sur les ressources autonomes.

— Ressources d'indexation : Elles résultent d'un processus par lequel les ressources appartenant à une collection sont étiquetées pour représenter les caractéristiques des ressources et les rendre exploitables par des services de recherche d'information. Les index peuvent avoir plusieurs formes en fonction des ressources utilisées. Parmi ces ressources on trouve : (i) les index par mots-clés basés sur les ressources linguistiques et les index hypertextuels (tels que les cartes des sites) structurés pour la navigation dans les documentations techniques électroniques ou sur les sites web ; et (ii) les index ontologiques ou conceptuels (annotations sémantiques) qui enrichissent la ressource initiale en associant à son contenu des éléments conceptuels lui permettant d'être utilisable, accessible et reconnue par un ensemble d'acteurs ou d'agents. Une annotation sémantique est une

formalisation de l'interprétation du texte sous forme de métadonnées [48].

— Ressources d'alignement : Il s'agit des ressources ayant un degré d'expressivité variable et des formes simples ou complexes et résultant de l'application d'une procédure de mise en correspondance entre deux ressources de même type. Cette catégorie de ressource est utilisée dans les applications de gestion de connaissances. L'alignement sert à trouver des entités similaires dans des ressources différentes tout en préservant l'indépendance et l'intégrité de ces ressources. Parmi ces ressources on trouve :

  — Les alignements des termes et des ressources terminologiques ;

  — Les alignements des ressources linguistiques telles que les corpus de textes alignés dans différentes langues ;

  — Les alignements d'ontologies, qui servent à mettre en correspondance les concepts des deux ontologies. Ces correspondances peuvent être l'inclusion, l'équivalence, la disjonction etc. [32].

3.3 Utilisation des ressources

L'utilisation des ressources externes dans l'analyse des messages courts a été largement explorée cette dernière décenie. Les applications principales sont les suivantes [51] :

— L'extraction de concepts et l'indexation conceptuelle : il s'agit d'identifier les concepts correspondant aux termes et de les utiliser dans l'indexation.

— L'expansion de requêtes et de documents consiste à ajouter les concepts sémantiquement liés aux concepts des requêtes ou des documents.

— Le calcul de similarité sémantique entre concepts via les informations sur les concepts et leurs relations dans les ressources externes.

3.4 Conclusion

Les ressources construites selon l'approche proposée font partie des ressources autonomes, plus précisement des thesaurus. Pour une partition des messages courts, l'approche construit un ensemble des collections des motifs sémantiquement proches. Les collections sont comme des synsets, des composantes atomiques sur lesquelles repose WordNet [25]. Chaque collection est représentée par un concept qui est lié par un lien d'hyponymie avec des motifs qui le caractérisent. Les ressources sont utilisées pour enrichir les messages courts en leurs ajoutant les concepts sémantiquement liés.

CHAPITRE 4

ARCHITECTURE D'UNE CHAÎNE DE TRAITEMENT DE REGROUPEMENT SÉMANTIQUE : ILLUSTRATION AVEC MEETING SOFTWARE

4.1 Introduction

L'explosion des données due au developpement technologique a fait naître des outils de regroupement sémantique, permettant de rendre cette masse de données exploitable. On peut distinguer entre autres des outils open source comme Weka [72], et des outils privés Monkey Learning, IBM Watson etc... Les entreprises utilisent ces outils par exemple pour maîtriser les avis de leurs clients afin de repondre de manière adéquate à leurs exigences. D'autres entreprises s'en servent lors des enqêtes internes pour améliorer leur fonctionnement et la vie de leurs collaborateurs.

Dans la suite de ce chapitre nous utiliserons les notations suivantes :

— MC : messages courts

— MC labelisés : messages courts dont les groupes sont connus

— MC non labelisés : messages courts dont les groupes ne sont pas connus

4.2 Architecture générale

La chaîne de traitement de ces outils est composée, en général, des éléments suivants :

— Un module de pré-traitement, pour préparer les données pour la tâche visée ;

— Un module d'enrichissement des données, pour caractériser les données ;

— Un module de représentation ;

— Un module contenant l'algorithme de regroupement.

La Figure 4.1 décrit le mode de fonctionnement des outils effectuant le clustering et la Figure 4.2 le mode de fonctionnement des outils effectuant une classification supervisée. La

différence entre les deux modes de fonctionnement se trouve au niveau des données en entrée et de l'objectif visé. Le clustering consiste à prendre en entrée des messages courts dont les groupes ne sons pas connus pour construire une partition (classification non supervisée). Cette dernière est composée des groupes des messages courts, chaque groupe représentant un ensemble des messages courts sémantiquement proches. L'objectif de la classification supervisée est de d'apprendre un modèle sur des données dont les groupes sont connus pour classer ensuite des nouvelles données (classification supervisée).

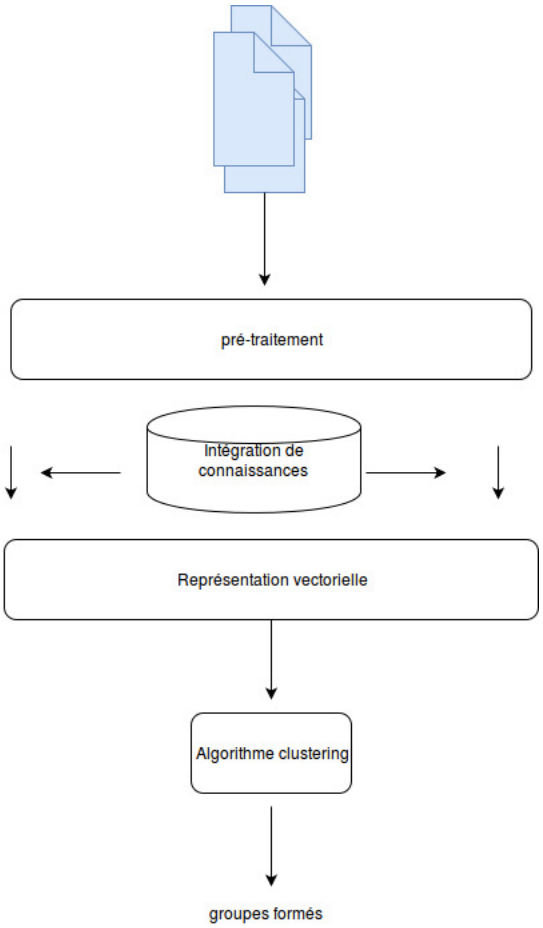

FIGURE 4.1 – Architecture classification non supervisée

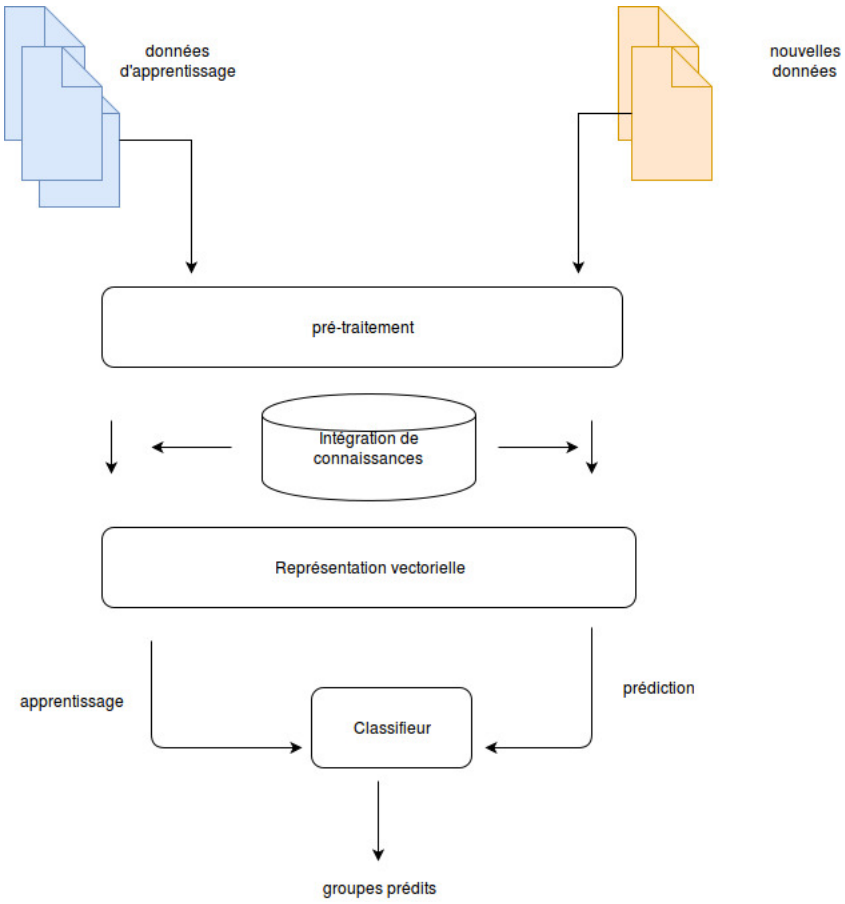

FIGURE 4.2 – Architecture classification supervisée

4.2.1 Module de pré-traitement

Le module de pré-traitement joue un rôle très important dans la performance des méthodes d'analyse des messages courts et leurs applications. Il constitue la première étape. Généralement, on distingue trois étapes clés à savoir, la tokenisation (segmentation), la suppression des mots outils et la normalisation (figure 4.1).

4.2.1.1 La tokenisation

Cette étape permet de segmenter un messages courts en mots le composant. Le module de tokenisation utilisé est basé sur les expressions regulières en Python. Ces dernières constituent un outil intéressant pour vérifier si le contenu d'une variable a la forme attendue. Les expressions regulières permettent de repérer des caractères ou chaîne de caractères dans un texte mais également de supprimer/modifier tous ceux que l'on ne souhaite pas conserver (les caractères spéciaux dans notre cas : les parenthèses, les guillemets, etc...). Les mots composés et les abréviations sont reconnus par notre module de tokenisation via une liste pré-établies spécifique à la langue utilisée. La tokenisation du message court « le client est satisfait » est donnée par la figure (cf. Figure 4.4).

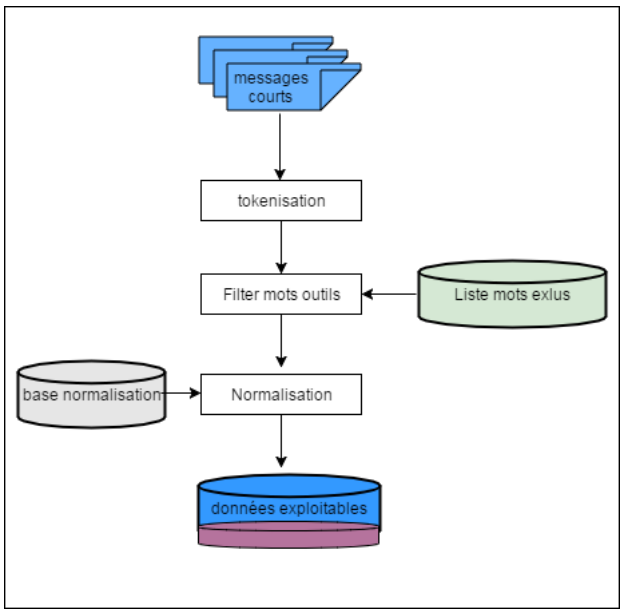

FIGURE 4.3 – Processus de prétraitement des messages courts

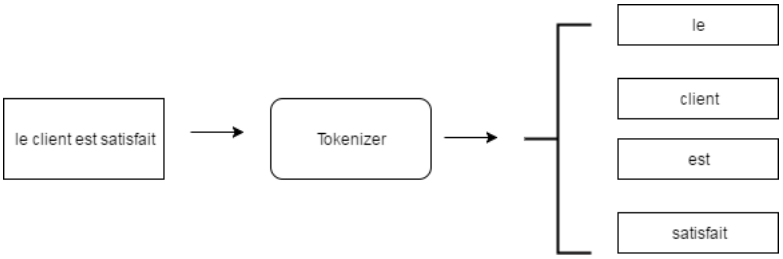

FIGURE 4.4 – La tokenisation

4.2.1.2 La suppression des mots outils

Il s'agit pour un message court d'enlever les mots outils (prépositions, articles, etc...). La méthode utilisée est celle basée sur une liste des mots outils préétablie. Cette dernière est construite en recupérant sur internet les éléments suivants :

— L'ensemble des flexions des verbres être et avoir

— Les déterminants

— Les pronoms (sauf les pronoms numéraux)

— Les prépositions

— Les conjonctions de coordination et de subordination

Chaque composant d'un message court figurant dans cette liste est automatiquement enlevé. La limite de cette méthode est que un mot outil ne figurant pas dans la liste préétablie sera considéré comme un mot plein (qui a un sens). Supposons que dans la liste préétablie figurent les mots « le » et « est ». Le message court « le client est satisfait »se transforme comme suit (figure 4.4) :

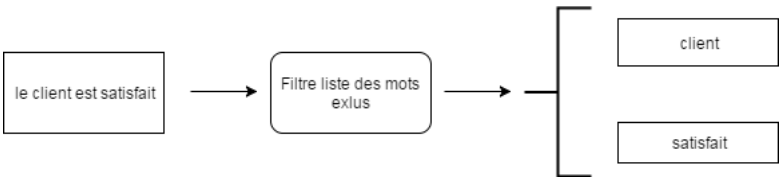

FIGURE 4.5 – Suppression des mots outils

4.2.1.3 La normalisation

Dans [6] et [76] les auteurs décrivent la lemmatisation et la racinisation comme étant des processusde normalisation morphologique. L'intérêt immédiat est de dimunier le nombre des mots à traiter en regroupant les variantes d'un mot (*ergonomie*, *ergonomique* remplacés par *ergonomie*). En effet, les messages courts peuvent avoir en commun différentes formes d'un mot désignant le même concept. L'objectif de ce processus est de les représenter par un seul mot qui porte un concept commun. Il existe plusieurs types de processus de normalisation :

— La lemmatisation : Elle consiste à trouver la racine des verbes fléchies et à ramener les mots pluriels et/ou féminins à la forme masculin singulier. Deux types de système de lemmatisation ont été identifiés :

  — *Treetagger* : disponible pour le français (en accès libre mais sans usage commercial)

  — Le module Python de web mining, *pattern* : basé sur le lefff (http ://www.clips.ua.ac.be/pattern) qui est libre d'utilisation.

  Notre choix s'est porté sur le module *pattern*.

— La stemmatisation : C'est le processus d'élimination des suffixes des mots afin d'obtenir leur racine comune. Nous utilisons l'implémentation *NLTK* de l'algorithme de Porter [60] appelé **Snowball** [1].

— La lexematisation : Dans [76], l'auteur la présente comme une alternative à la lemmatisation et à la stemmatisation. Elle permet de regrouper les mots d'une même famille au moyen de leur graphie. Par exemple, elle permet de ramener à la même forme graphique les mots : *chant*, *chantaient*, *chanté*, *chanteront*, *chanteuse*, *chanteuses*, *chanteurs*, *chants*. Cette méthode de normalisation n'est pas utilisé dans l'outil Meeting Software.

La Figure ci-dessous montre comment le processus de lemmatisation permet de mettre en évidence des mots en communs entre deux messages courts :

Dans [71], les auteurs ont effectué des expériences dans le but d'évaluer l'impact des méthodes de pré-traitement sur certains algorithmes de *machine learning*. Les expériences consistaient à evaluer l'impact des combinaisons des méthodes de pré-traitements (correction orthographique, filtrage des mots vide de sens, stemmatisation) sur les *arbres de décision*, *SVM* et

1. http ://snowball.tartarus.org/

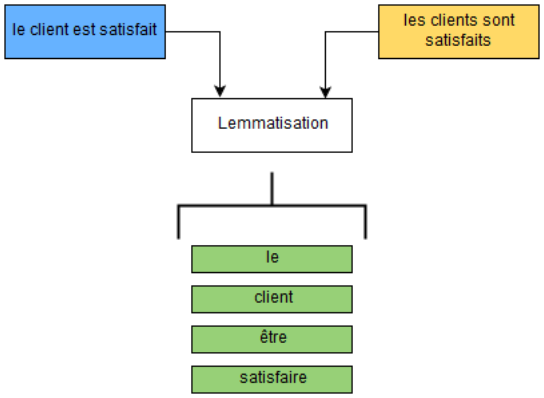

FIGURE 4.6 – Exemple de lemmatisation de deux messages courts

k-means. Les données utilisées ont presque les mêmes caractéristiques que celles de l'entreprise. On y trouve des fautes d'orthographes et d'autres sources de bruit. La seule différence se trouve au niveau de la taille des messages courts. La taille moyenne des données qu'ils ont utilisées est de 20 mots alors qu'elle est de 2 à 3 mots pour les données de l'entreprise. Il ressort de ces expériences que la stemmatisation a un impact positif sur la performance des trois algorithmes (voire expérimentations).

4.2.2 Module d'enrichissement des données

Ce module permet soit d'utiliser une base des connaissances externes soit les données traitées pour intégrer de la sémantique dans les messages courts. Dans notre cas, l'intégration de la sémantique consiste à remplacer un lexème . Cela permet de calculer la similarité sémantique entre deux messages courts qui ne partagent aucun mot en commun. C'est l'exemple des réponses de la table suivante :

Fixer trois objectifs concrets lies à la formation	Les responsabiliser en leur demandant de démultiplier
Savoir déléguer et gérer la délégation	Accompagnement individuel
Autonomie	Suivre les points de progrès du collaborateur

TABLE 4.1 – Extrait des réponses liée à la question posée sur la figure 2

Les réponses « Autonomie »et « Les responsabiliser en leur demandant de démultiplier »sont deux messages courts sémantiquement proches mais n'ayant pas des mots en commun. L'utilisation d'une base des connaissances dans laquelle les mots « responsabilité »et « autonomie »sont représentés par le même concept, permettra d'améliorer le score de similarité entre les deux messages courts (figure 4.6).

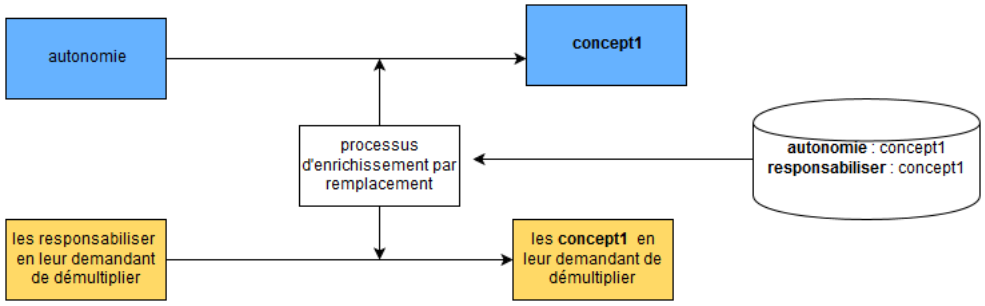

FIGURE 4.7 – Exemple d'enrichissement de deux messages courts

4.2.3 Module de représentation de messages courts

Quelle que soit la méthode de classification (supevisée et non supervisée) utilisée la première opération consiste à représenter les documents de façon à ce qu'ils puissent être traités automatiquement. En général, c'est la représentation vectorielle des documents qui est la plus utilisée car s'appuyant sur les notions de dissimilarité et de similarité entre deux messages courts. Les implémentations utilisées dans l'outil Meeting Software proviennent de *Sckit-learn* [2]. La représentation vectorielle est utilisée dans de nombreux autres domaines connexes de l'apprentissage automatique, par exemple la fouille de texte ou la recherche d'information [67].

La représentation vectorielle d'un object suppose un certain nombre fini d'attributs constituant les dimensions. Un message court peut être considéré d'une part comme étant une suite fini des caractères (ou combinaisons des caractères), et d'autre part comme une suite fini de mots (ou combinaisons de mots). La limite de la suite des caractères est qu'elle ne porte aucune information directe sur la sémantique. Cela n'est pas adapté aux tâches visées par Meeting Software, ce qui nous conduit à considérer un message court comme une suite finie de mots (ou combinaisons des mots).

On peut distinguer deux famille de représentation vectorielle des messages courts : la représentation standard et la représentation sémantique. Nous nous somme concentrés uniquement la représentation vectorielle standard, l'autre représentation faisant parti d'un d'autre projet de l'entreprise.

Soit $M = \{m_1, m_2, ..., m_N\}$, un message court tel que m_i $(i = 1...N)$ représentant les mots composant le message M. Soit $D = \{w_1, w_2, ..., w_k\}$ le vocabulaire constitué de l'ensemble des mots w_j des messages courts traités avec $k \gg N$. On distingue trois types de représentation standard en fonction des poids attribués aux mots correspondants : la booléenne, fréquentielle et pondérée.

2. `http://scikit-learn.org/stable/modules/feature_extraction.html`

4.2.3.1 La représentation booléenne

Le message M est codé de la façon suivante : $M = [x_1 x_2 ... x_N]$ où $x_i = 1$ si w_i apparaît dans M et 0 sinon. Cette représentation à l'avantage d'être simple. Cependant elle ne prend pas en compte la fréquence des mots et ignore la longueur des messages courts.

4.2.3.2 La représentation fréquentielle

Dans [68], les auteurs représentent un texte constitué de mots distincts par un vecteur dont la valeur de chaque composante représentative d'un mot vaut le nombre d'occurrence de ce mot dans ledit texte. Le message M est codé de la façon suivante : $M = [x_1 x_2 ... x_N]$ où x_i est la fréquence de w_i dans M. Elle a l'avantage de prendre en compte des fréquences des mots et de prendre en compte la longueur des messages courts. En revanche, les messages courts traités doivent avoir à peu près la même longueur, ce qui n'est pas toujours le cas.

4.2.3.3 La représentation pondérée

La méthode *tf-idf* (*term frequency inverse document frequency*) est très souvent utilisée pour attribuer des poids aux mots d'un corpus. C'est une référence dans le domaine, intégrée dans les premiers modèles développés [20]. Elle est basée sur le principe de donner plus d'importance aux mots qui apparaissent souvent dans un texte, mais peu dans le corpus (composé d'un ensemble de textes). Elle donne moins d'importance aux mots qui apparaissent dans beaucoup de documents différents ; en effet, un mot qui apparaît souvent dans plusieurs documents implique qu'il est peu représentatif des spécificités des textes, et que nous ne pourrons nous baser sur celui-ci pour différencier les documents. Ainsi, cette méthode émet l'hypothèse qu'un terme est important pour un certain document s'il apparaît dans ce document, et que parallèlement peu de documents du corpus le contiennent. Le message M est codé de la façon suivante : $M = [x_1 x_2 ... x_N]$ où $x_i = tf \cdot idf(w_i, M)$. Le poids tf-idf d'un terme dans un document est donnée par la formule suivante :

$$tf - idf(w_i, M) = tf(w_i, M) \cdot idf(w_i) \tag{4.1}$$

où $tf(w_i, M) = \frac{n_i}{\sum n_i}$ est la fréquence du terme w_i dans le document M et $idf(w_i) = \ln \frac{m}{m_i}$ sa fréquence inverce avec n_i la fréquence de w_i dans M, m le nombre total des documents dans la collection et m_i le nombre de documents dans la collection où w_i apparaît.

4.3 Cas de Meeting Software

La particularité de Meeting Software est la dynamique induite par l'activité du pilote. Les données (voir chapitre 5 section 5.1) ne sont pas collectées en même temps par l'outil. Comme indiqué dans l'introduction, le pilote lance une classification non supervisée avec les premières

données et puis itérativement une classification supervisée avec les paquets de données suivants. Mesurer la performance de Meeting Software revient alors à mesurer la performance de sa chaîne de traitement au niveau de chacune des étapes de son fonctionnement. Une chaîne de traitement Meeting Software est composée :

1. D'une liste de pré-traitement,
2. D'un algorithme de clustering,
3. D'un algorithme de classification,
4. D'un ensemble de mesures de qualité.

Architecture Meeting Software

L'architecture de l'outil Meeting Software est compsée de six packages (Figure 4.7) :

— **Preprocessing** : Package de pré-traitement des données. C'est à ce niveau que les méthodes d'enrichissement des données sont implémentées.

— **Vectorizer** : Package permettant de faire une représentation vectorielle des messages courts.

— **Classification** : package dans lequel sont implémentées les algorithmes de classification supervisée et non supervisée.

— **Analyse** : Package dans lequel sont implémentées les mesures de qualité des algorithmes de classification (supervisée et non supervisée).

— **Benchmark** : Package qui definit une chaîne de traitement. Cette dernière est composée d'une liste des pré-traitements, d'une méthode de vectorisation, d'un algorithme de clustering et d'un algorithme de classification supervisée.

— **Comparaison** : Package dans lequel on compare différentes chaînes de traitement.

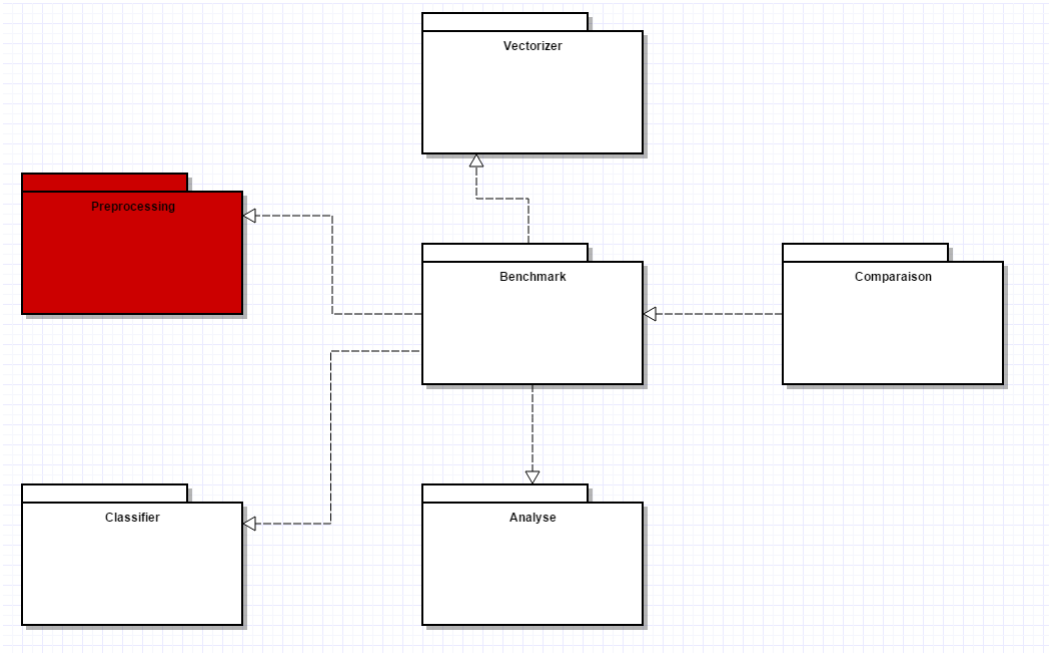

FIGURE 4.8 – Architecture Meeting Software

Cette architecture a les avantages suivants :

— **Modulaire** : On peut facilement supprimer ou modifier un module.

— **Connectivité** : On peut connecter des modules ou applications externes à l'outil.

— **Robustesse et fiabilité** : Les modules ont été développés en utilisant la méthode de programmation "pair-programming". Cette méthode de développement permet d'avoir des codes plus agiles et plus sûrs. Chaque module possède des tests unitaires, 300 tests contre moins de 10 tests pour l'ancienne architecture. Un test de régression a été mis en place pour éviter que la modification ou la création d'un module particulier crée d'éventuelles erreurs au niveau de la chaîne de traitement.

4.4 Conclusion

L'architecture de l'outil Meeting Software® correspond aux exigences de l'entreprise. Cette dernière s'est fixé l'objectif d'avoir un outil fiable, robuste et modulaire. Cela facilitera les échanges entre l'entreprise et les laboratoires avec lesquels elle collabore et d'industrialiser à terme l'outil. De plus, elle permet de facilement évaluer différentes chaînes de traitement. Nous avons vu ,par exemple, que le module du pré-traitement a un impact sur la chaîne du traitement de l'outil Meeting Software®. En effet, la combinaison filtrage de mots vides de sens et stemmatisation fournit les meilleurs résultats pour les deux classifieurs utilisés (*ExtraTrees* et *SVM linéaire*). On minimise de 2.13% à 5.58% de déplacements par rapport à la chaîne de traiment sans pré-traitements avec le classifieur *SVM* et entre 2.39% et 5.31% avec le classifieur *ExtraTrees*.

DEUXIÈME PARTIE

Construction des ressources sémantiques

CHAPITRE 5

LES DONNÉES

5.1 Introduction

Les messages courts issus de la base de données Succeed Together sont d'une part des données issues des ressources humaines internes et d'autre part des verbatims clients concernant des produits ou des services des entreprises. Les métiers traités par l'entreprise peuvent être segmentés par secteur d'activité :

— Agroalimentaire

— Association et Fédération

— Banque et Assurance

— Cabinet de Conseil

— Distribution

— Énergie et Environnement

— Ingénierie et Infrastructure

— Service et Utilities

— Transport et Tourisme etc.

Bien que les données proviennent de secteurs d'activés différents, elles ont des points communs. En effet, les sujets traités concernent généralement les ressources humaines, le management et les problématiques de transformation.

5.2 Les différents types des données

Les données proviennent de trois solutions, s'inscrivant dans des moments de management [1] importants identifiés par l'entreprise : Pulsation pour des enquêtes en ligne s'appuyant sur des questionnements ouverts, SucceedMeeting pour des réunions interactives et SucceedData pour le traitement de données textes de masse.

1. Période de réunions entre la direction et les collaborateurs

5.2.1 Données provenant de la solution SucceedMeeting

Succeed Meeting est une solution qui permet de rendre interactive une réunion professionnelle ou un séminaire. Pendant l'évènement, plusieurs questions sont adressées aux participants, idéalement répartis par tables de 4 à 8 personnes. Une tablette est disposée sur chacune des tables pour impliquer tous les participants et les mettre dans une posture active. Cela initie et guide les échanges pour encourager la collaboration. Les participants saisissent au fur et à mesure leurs réponses sur la tablette durant le temps imparti (généralement 7 minutes par question), sans consensus. Les réponses sont anonymes et écrites, chacun peut s'exprimer sans craindre le regard des autres Toutes les idées sont prises en compte, il n'y a pas de bonne ou mauvaise réponse. Pour chaque question, les participants peuvent ajouter autant de réponses qu'ils le souhaitent. En moyenne une centaine de réponses par question est recupérée.

Meeting Software® produit le regroupement sémantique des réponses par question à la fin du temps imparti. La synthèse des réponses pour chaque question est alors diffusée en temps réel, on peut facilement identifier les idées fortes et prendre connaissance des faiblesses. Les intervenants peuvent réagir à chaud aux résultats et créer un dialogue avec la salle. C'est là que l'interactivité commence, à travers l'échange et le débat entre les intervenants et les collaborateurs.

Il faut noter que l'outil est piloté par un expert humain, qui corrige les éventuelles erreurs du système (réponses mal classées). Une fois l'événement terminé, un rapport contenant l'ensemble de la production des participants ainsi que le regroupement sémantique revu par le pilote humain est remis au client.

5.2.2 Données provenant de la solution Pulsation

Pulsation est une enquête en ligne mise à la disposition du client sur une plateforme Succeed Together. Accessible durant 2 à 3 semaines, ce protocole de questionnement ouvert est utilisé pour :

— Réaliser une enquête isolée ou sous forme d'abonnement en soutien d'une pratique managériale régulière pour maintenir l'organisation sous tension positive permanente.

— Recueillir du contenu en amont du prochain séminaire et les thèmes que les participants souhaitent voir aborder.

— Connaître simplement leur ressenti sur une thématique ou mesurer l'impact d'une action.

— Lancer un plan stratégique.

Les participants invités se connectent anonymement d'un simple clic et donner leurs avis sur des questions variées. Comme pour la solution SucceedMeeting, il est possible de donner plusieurs avis par question, chaque réponse est limitéé à 255 caractères. Grâce au regroupement

sémantique de Meeting Software®, une synthèse riche et facilement exploitable est produite rapidement à partir de ces avis. Cette synthèse correspond au regroupement sémantique, revu et corrigé par le pilote humain, des avis émanant des participants, qui peut être suivi progressivement par l'entreprise cliente en temps réel via la plateforme dédiée. Le rapport final est fourni dans un délai de 2 jours ouvrés maximum.

5.2.3 Données provenant de la solution SucceedData

SucceedData est une solution qui permet d'analyser sémantiquement des volumes importants de données textes disponibles auprès d'une entreprise cliente. Les objectifs sont les suivants :

— Améliorer l'expérience client et valoriser les données des ressources humaines ;

— Dégager rapidement une vision d'ensemble tout en prenant en compte chaque personne ;

— Extraire des informations pertinentes.

Les données de cette solution proviennent de mails clients, d'études marketing, de questions ouvertes dans un sondage, de retranscriptions, de baromètres de satisfaction, etc. Succeed Together intervient pour fournir un regroupement sémantique sur cette masse de données importante que l'entreprise cliente n'arrive pas à exploiter. Les données sont envoyées à Succeed Together sous format excel, la quantité de données récoltée (par thématique, sujet ou question) est de l'ordre du millier. À la différence des autres solutions, la taille des données n'est pas limitée. Chaque cellule du fichier excel peut prendre la forme d'un paragraphe avec plusieures idées ou une liste d'idées séparées par des tirets.

Pour une analyse sémantique toujours plus efficace, Succeed Together effectue le fracking [2] de l'ensemble des données texte envoyées. Ce premier algorithme permet la séparation des thèmes présents dans chacune des cellules excel afin d'isoler des idées fortes dans les paragraphes. Le regroupement sémantique est mis à la disposition de l'entreprise cliente dans un délai de 2 jours ouvrés maximum.

La table 5.1 synthétise les caractéristiques de chaque solution et met en évidence les différences entre les types de données obtenus.

2. Segmentation des paragraphes en phrases

Solutions	**Qté de données**	**T. commentaires**	**Prétraiment**	**Srce de données**	**temps im-parti**
SucceedMeeting	1-100	255 caractères	non	tablette	7 minutes
Pulsation	< 1000	255 caractères	non	plateforme	2 jours
SucceedData	> 1000	pas limité	oui	Fichier excel	2 jours

TABLE 5.1 – Différence entre les données provenant des solutions

5.3 Structure des données et quelques chiffres

Lors d'un évenement plusieurs questions, ou sujets, peuvent être abordés. Les réponses obtenues pour les trois solutions sont stockées dans des fichiers XML. Le format XML permet facilement de stocker des données structurées et de les explorer. La figure 5.1 fournit la structure des fichiers XML stockés dans la base de données de l'entreprise.

```
<xml>
   <meta>
   </meta>

      <participants>
      </participants>

         <questions>

            <question q_id = q_id1>
                  <label> titre de la question </label>

                     <group id = g_id1>
                        <answer ans_id = ans_id1> reponse 1</answer>
                        <answer ans_id = ans_id1> reponse 2</answer>
                        .
                        .
                        .
                     </group>
                        .
                        .
                        .
             </question>
                  .
                  .
                  .
        </questions>
</xml>
```

FIGURE 5.1 – Structure du fichier XML

Comme l'illustre la Figure 5.1, les réponses à chaque question sont catégorisées par l'outil

en groupes sémantiques. Ces derniers correspondent à des ensembles de réponses sémantiquement liées. Par exemple, les réponses *améliorer la qualité de vie au travail*, *des bons moments de partages entre collègues*, *moment de détente au bureau* sont catégorisées dans un même groupe "bien être au travail".

Le tableau 5.2 illustre les métadonnées chiffrées des réponses obtenues pour les 2329 questions disponibles dans la base de l'entreprise.

Nombre de questions	2329
Nombre total de réponses	427828
Nombre minimum de formes graphiques (mots) dans une réponse	0
Nombre maximum de formes graphiques (mots) dans une réponse	372
Minimum de réponses dans une question	1
Maximum de réponses dans une question	12630
Nombre de groupes	20216
Minimum de réponses dans un groupe	1
Maximum de réponses dans un groupe	2856

TABLE 5.2 – Chiffres sur les réponses obtenues pour les questions disponibles

Nombre de réponses vides	1774
Nombre de réponses de taille 1 mot	33382
Taille moyenne des réponses	9
Taille minimale des réponses	0
Taille maximale des réponses	323

TABLE 5.3 – Quelques chiffre sur la taille de réponses traitées par l'outil

Nombre de formes graphiques (mots)	64340
Nombre dans le vocabulaire après filtrage de stopwords	63979
Nombre dans le vocabulaire après filtrage de stopwords et lemmatisation	49691
Nombre dans le vocabulaire après filtrage de stopwords et stemmatisation	38146

TABLE 5.4 – Quelques statistiques sur le vocabulaire global

5.4 Conclusion

L'historique des données Succeed Together constitue une base de connaissance importante pour la compréhension du type de données traitées et une richesse sur laquelle nous nous appuyons pour améliorer le regroupement sémantique de l'outil Meeting Software. Les données utilisées proviennent des trois solutions SucceedMeeting, Pulsation, SucceedData et de secteurs d'activité différents.

Cependant, les données imposent un certain nombre de défis à l'outil Meeting Software :

— Certaines questions n'obtiennent que peu ou très peu de réponses (cf. Table 5.2),

— De même certains groupes ne contiennent que peu ou très peu de réponses,

— Certaines réponses sont parfois très courtes (un seul mot), voire vides.

Ces problèmes à différents niveaux, auxquels s'ajoutent les fautes d'orthographes et les abréviations, rendent difficile la classification. Les algorithmes de classification supervisée (et non supervisée) sont ainsi moins précis.

Les groupes sémantiques ainsi formés sont ensuite revus et corrigés par des experts humains (employés du pôle S&D et R&D). Ils sont utilisés pour la construction des ressources sémantiques et considérés comme des "gold standard" pour nos évaluations. Ces processus seront détaillés dans le chapitre 7.

CHAPITRE 6

LA DÉMARCHE PROPOSÉE

6.1 Introduction

La prise en compte de l'aspect sémantique des données textuelles lors de la tâche de classification s'est imposée comme un réel défi ces dix dernières années. Cette difficulté vient s'ajouter au fait que la plupart des données disponibles sur les réseaux sociaux sont des textes courts (moins denses en mots). Ces derniers sont généralement représentés par des vecteurs très clairsemés (contenant beaucoup de zero) dans un espace où les dimensions sont des mots. Dans ce cas, les techniques traditionnelles de classification (supervisée et non supervisée) basées sur le calcul de similarité entre textes se confrontent au problème des mesures très proches de zéro, car les messages courts, même les très similaires, ont peu ou pas de termes en commun.

la plupart des solutions proposées ont pour objectif d'enrichir la représentation de ces messages courts en injectant de la sémantique. Cette dernière peut provenir, d'une part de la collection de messages courts traitée [27] et d'autre part d'une base de connaissance importante externe comme des ontologies [44], Wikipedia [45] [3] et WordNet [79] . La première approche d'extraction de la sémantique requière peu de traitements et techniques, tandis que la deuxième est exigeante en quantité de données appropriées aux messages courts traités. L'utilisation des ressources externes peut comporter des risques. Si les données ne sont pas appropriées, cela peut conduire à un ajout excessif d'information ou un ajout de "bruit".

L'approche proposée dans ce projet de recherche est différente des approches proposées dans les travaux antérieurs sur l'enrichissement des messages courts et ce pour trois raisons. Tout d'abord, nous n'utilisons pas des bases de connaissances externes comme Wikipedia parce que généralement les messages courts qui sont traités par l'entreprise proveniennent des domaines spécifiques. Dans notre cas, c'est l'historique de données de l'entreprise qui est utilisé. Deuxièment, les données à traitées ne sont pas utilisées pour la constitution de ressources à cause du fonctionnement de l'outil. En effet, les données sont traitées par paquet dans l'outil. Cela ne permet pas pour une étape de traitement d'avoir une masse de données conséquente pour l'extraction de motifs. Troisièment, à notre connaissance il n'existe pas des travaux d'une part qui exploitent des données structurées comme celles de l'entreprise pour constituer des ressources sémantiques, et d'autre part qui mesurent l'impact de l'enrichissement sur un système

interactif de regroupement de flux de textes. La démarche proposée s'appuye sur la construction de ressources sémantiques, en utilisant des techniques de fouille de données séquentielles pour extraire des motifs émergents. Ces derniers ont déjà été utilisés dans [61] pour caractériser les genres de textes. Ainsi nous extrayons des motifs émergents pour enrichir les messages courts dans le cadre de la classification (supervisée et non supervisée). Nous montrons que le classification est améliorée grâce à cet enrichissement.

Les ressources sont construites en exploitant des historiques des partitions des messages courts. Nos expérimentations montrent que l'on peut utiliser un historique des partitions des messages courts d'un sujet bien défini pour améliorer la qualité de la classification des nouveaux messages courts traitant le même sujet. Notre approche consiste à valoriser l'historique des partitions, d'un sujet, en les transformant en lexique. Ce dernier représente une cartographie du sujet qui permettra d'enrichir des nouveaux messages courts afin de mettre en évidence leurs sens communs. Les partitions sont des ensembles des groupes de messages courts, chaque groupe étant une collection des messages courts sémantiquement proches.

Ce chapitre décrit les étapes de la construction des ressources du système proposé (section 6.2), détaillées en cinq étapes : constitution du corpus (section 6.2.1), extraction des motifs fréquents (section 6.2.2), sélection des motifs émergents (section 6.2.3), validation de ressource (section 6.2.4) et la sérialisation des ressources (section 6.2.5). Le processus d'enrichissement permettant d'intégrer la sémantique dans les messages courts est présenté dans la section 6.3.

6.2 L'approche de construction de ressources

M partitions [1] de messages courts sont proposées au processus d'extraction de ressources. Grâce à une étape d'annotation manuelle, un corpus composé de N groupes de messages courts est construit. Chaque groupe représente un ensemble de messages courts sémantiquement proches. Chaque message court est tout d'abord pré-traité (tokenisation, lemmatisation et/ou stemmatisation et l'utilisation d'une liste de stopwords). Les motifs séquentiels (fréquents) sont extraits pour chacun des groupes de messages courts. Ensuite des motifs émergents sont calculés à partir des N collections de motifs fréquents extraits. La figure 6.1 illustre les différentes étapes du processus d'extraction de ressource mise en place.

— P_i : partition des messages courts, un ensemble des groupes de messgaes courts.

— G_j : groupe des messages construit en fusionant les groupes des partitions traitant les mêmes sujets.

— MF_j : la collection des motifs fréquents correspondante au groupe G_j.

— ME_j : la collection des motifs emergents correspondante au groupe G_j.

1. Une partition est composée des groupes de messages courts

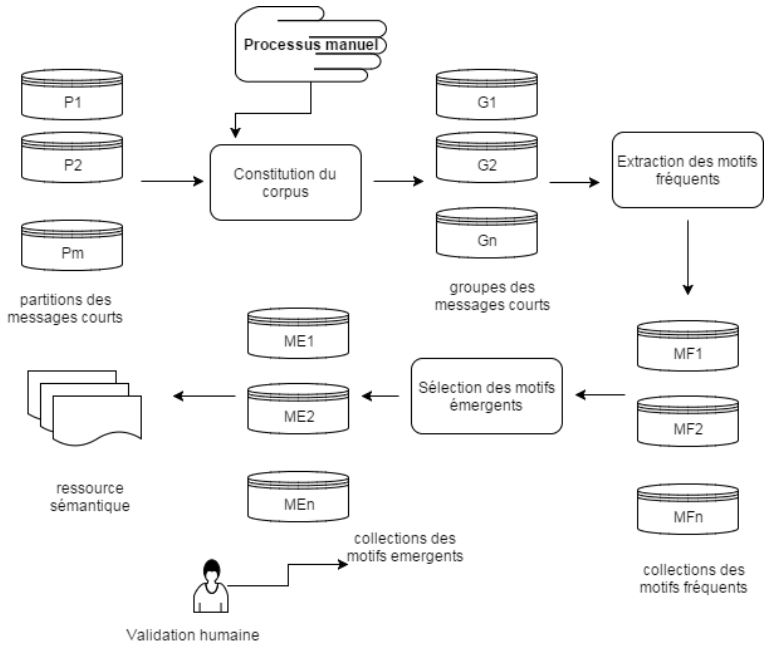

FIGURE 6.1 – Vue générale du processus de constitution de ressources

L'approche de construction de ressources proposée est composée de quatre étapes :

— Constitution du corpus

— Extraction des motifs fréquents

— Sélection des motifs émergents

— Formalisation des ressources

6.2.1 Constitution du corpus

Cette étape consiste à construire un corpus, composé des N groupes de messages courts, à partir des M partitions proposées au processus de constitution de ressources. Il s'agit de constuire une partition type dont chacun des groupes est un groupe ou une union des groupes de messages courts similaires de M partitions.

La similarité des groupes de messages courts est déterminée manuellement par les ingénieurs du pôle recherche et développement. Elle consiste à annoter les groupes de messages courts de M partitions par des étiquettes liées aux thématiques sous-jacentes des groupes. Par exemple, pour un groupe traitant de la thématique « cohésion des équipes » l'étiquette serait « cohesion_equipe ». Les groupes de messages courts ayant la même étiquette sont considérés comme similaires. Les tables 6.1 et 6.2 montrent deux groupes de messages courts de deux questions traitant le même sujet. Ces deux groupes peuvent être associés à l'étiquette "transition-numérique" par exemple.

groupe n°3 de Q_1
A quand google ?
Google du nucleaire
Intégrer le numérique pour accéder plus facilement aux docs
Avoir des outils de recherche conviviaux
Ergonomie de l 'Outils informatique documentaire
Une application unique pour saisir les travaux.
Avoir un vrai moteur de recherche
Avoir un moteur de recherche type google pour simplifier les recherches documentaires
reduire le nomnbre d application informatiques
La mobilité numérique
L'accompagnement des innovations
Homogénéisation des outils informatiques
diminuer le nombre d 'applications informatiques
Utilisation de supports numeriques pour utilisation documents et logiciels type AIC etc ...
Limiter le nombre d 'outils du SI
Interface informatique unique qui simplifie l 'acces aux Si
Simplification des outils informatiques.
Les outils informatiques
Le reseau informatique.

TABLE 6.1 – Les messages courts du groupe n°3 de la question Q_1

groupe n°2 de Q_2
Developper les outils numeriques avec pertinence (p ex REX a disposition sur tablette)
Base informatique
La wifi pour tous
transition numérique
Bases de donnees et bases documentaires
Optimiser le portail et l 'arborescence du reseau
Numerisations des formulaires
Systeme informatique optimisė
Moins de papiers
Develloper la visioconference
Simplification des applications et des outils informatiques
Outil de demande de travaux unique (pilotimo,di,Dt,epsilon,totem etc...)

TABLE 6.2 – Les messages courts du groupe n°2 de la question Q_2

6.2.2 Extraction des motifs fréquents

Cette étape consiste à extraire des motifs fréquents par groupe de messages courts du corpus construit à l'étape précédente. Un motif dans un groupe de message courts est fréquent si son support est supérieur à un seuil fixé (voir section 2.2). À la sortie de cette étape, on obtient une collection des motifs fréquents.

Exemple 1

Supposons que le corpus construit à l'étape 1 est composé de deux groupes de messages courts :

- **fideliser-client**, composé des messages courts suivants :
  - il faut fideliser les clients ;
  - faire des promotions ;
  - être proche du client.
  - fidéliser clients
- **responsabiliser-collaborateur**, composé des messages courts suivants :
  - responsabiliser salariés et clients ;
  - responsabiliser ses salariés, c'est avoir des salariés autonomes ;
  - salarié autonome ;
  - permettre aux clients de faire des achats seuls.

En choisissant comme paramètres $supportmotifs = 2$, 2 comme taille maximale des motifs et 0 comme gap, on obtient les collections suivantes (les messages courts ont subi comme prétraitements : tokenisation, lemmatisation et filtrage de stopwords) :

fideliser-client	**responsabiliser-collaborateur**
client, 3	autonome, 2
fidéliser, 2	client, 2
fidéliser client, 2	responsabiliser, 2
	salarié, 3
	responsabiliser salarié, 2
	salarié autonome, 2

TABLE 6.3 – Collection des motifs fréquents

Un élément d'une collection est un motif suivi de son support, le motif $client$ a 3 comme support.

6.2.3 Sélection des motifs émergents

Dans [61], les auteurs ont détaillé la démarche qu'ils ont suivie pour extraire les motifs émergents. Un motif (choisi après l'étape précédente) dans un groupe est émergent si sa fréquence d'apparition dans ce groupe est supérieure à celle du même motif dans n'importe quel autre groupe (voir section 2.3). On obtient alors des collection des motifs emergents.

Exemple 2

Prenons les groupes de messages courts de l'exemple précédent. En considérant 1.5 comme seuil de sélection des motifs emergents, on obtient les collections de la Table 6.4. Nous choisissons un seuil entre 1 et deux en raison de faible quantité de données manipuler. On constate que le motif ***client*** n'est pas un motif discriminant.

fideliser-client	**responsabiliser-collaborateur**
fidéliser	autonome
fidéliser client	responsabiliser
	salarié
	responsabiliser salarié
	salarié autonome

TABLE 6.4 – Collection des motifs emergents

6.2.4 Validation de ressources

La validation de ressources consiste à supprimer des motifs "bruits" dans chacune des collections de motifs emergents extraites dans l'étape précédente. Cette tâche est réalisée manuellement par des experts humains via un consensus. Ces experts humains sont des employés du pôle Recherche et Développement.

Exemple 3

Pour valider les collections de motifs emergents extraits dans l'exmple 2, on vérifie s'il n'y a pas la présence des motifs ambigus ou non pertinents. On remarque que le motif ***salarié*** de la collection ***responsabiliser-collaborateur*** n'est pas pertinent pour caractériser ladite collection. On obtient alors les collections suivantes après validation :

6.2.5 Sérialisation des ressources

Après les étapes précédentes, on obtient T collections de motifs emergents validés, chaque collection est représentée par un concept (le nom de la thématique associé à la collection). Dans la suite, nous appelerons collections les **classes sémantiques**. Les classes sémantiques

fideliser-client	responsabiliser-collaborateur
fidéliser	autonome
fidéliser client	responsabiliser
	responsabiliser salarié
	salarié autonome

TABLE 6.5 – Collection des motifs emergents après validation

constituent ce que nous appelons ressource sémantique.En considérant l'exemple 3, les motifs *fidéliser* et *fidéliser client* appartiennent à la classe sémantique ***fideliser-client***.

Les ressources sémantiques sont sérialisées pour être facilement utilisables ou stockées dans des tableurs, une colonne par classe sémantique. La première celulle de chacune des colonnes correspond au concept associé à la classe sémantique. La figure ci-dessous représente un extrait d'une ressource sémantique :

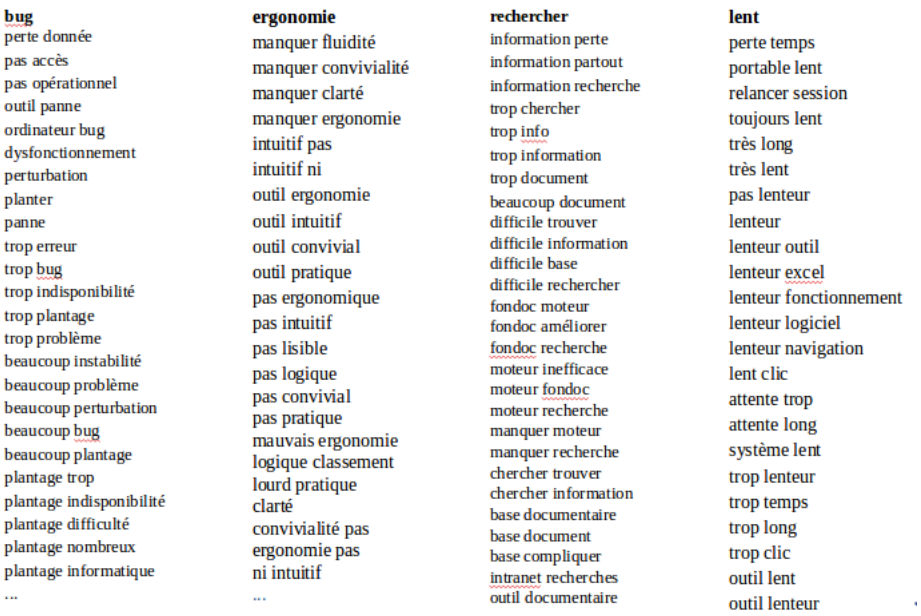

FIGURE 6.2 – Extrait d'une ressource sémantique

La Figure 6.2 montre par exemple que les motifs *perte donné*, *outil panne*, *dysfonctionnement* appartiennent à la classe sémantique ***bug*** et les motifs *manque fluidité*, *manque convivialité*, *pas intuitif* appartiennant à la classe sémantique ***ergonomie***.

6.3 Les processus d'enrichissement

L'enrichissement consiste à utiliser la ressource comme étant un vecteur de champs sémantiques. Il consiste à transformer un message court dans le but de l'enrichir. Cette transformation est liée soit à un ajout d'information dans le message court soit à un remplacement.

Le processus d'enrichissement permet de mettre en évidence le sens commun des messages courts grâce aux concepts associés aux motifs les composant, afin d'améliorer leurs coefficients de similarité. En effet lorsque les messages courts traités partagent peu ou quasiment pas de mots en commun, les techniques de similarité utilisées par les algorithmes de classifications (supervisée et non supervisée) fournisent des scores proches de zero. Cela rend difficile leur classification dans des groupes homogènes. Qautre processus d'enrichissement ont été mis en place :

— Add Concept (**AC**)

— Add Frequent Concept (**AFC**)

— Replace Pattern by Concept (**RPC**)

— Replace short Message by Frequent Concept (**RMFC**)

Nous nous sommes inspirés des travauax réalisés dans [8] pour mettre en place ces processus d'enrichissement. Les auteurs montrent qu'on peut concidérablement améliorer la précision du clustering de textes en leurs ajoutant des concepts provenant de Wikipedia comme ressource externe. La figure 6.3 présente la vue générale d'un processus d'enrichissement.

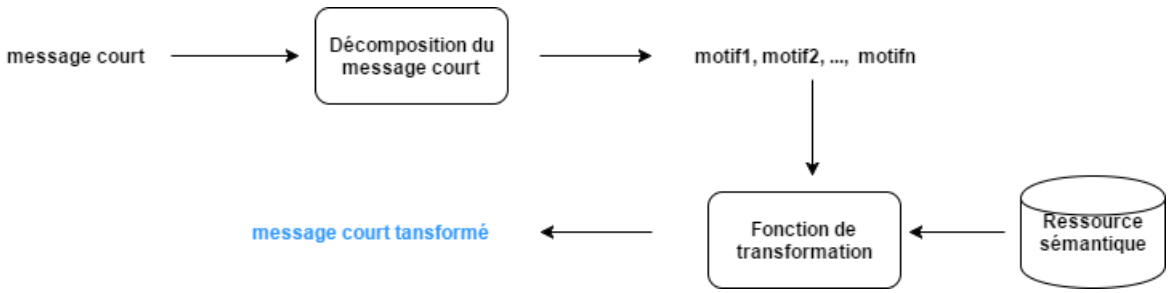

FIGURE 6.3 – Vue générale du processus d'enrichissement

Soit $R = \{T_1, T_2, ..., T_n\}$ une ressource sémantique composée de n collections des motifs emergents $T_i = \{me_1, me_2, ..., me_{n_i}\}$ où me_k avec $k = 1...n_i$ représentant les motifs émergents de la collection T_i. Soient $Concept = \{C_1, C_2, ..., C_n\}$, l'ensemble des concepts associés aux n classes sémantiques et $M = \{m_1, m_2, ..., m_j\}$ un message court à enrichir.

Process AC

```
initialisation d'un message court vide M1

Pour chaque motif m dans M
   si m est defini dans R
       insérer le concept C associé à m dans M1;

retourner M + M1
```

Exemple 4 : processus d'enrichissement *Add Concept*

Soit une ressource sémantique contenant la classe sémantique concept **ergonomie** composé des motifs : `pas fluide; pas ergonomique; manque convivialité; intuitif`. À partir de deux messages courts suivants : `mon système manque de convivialité` et `logiciel pas fluide`, Le processus **Add Concept** permet de les enrichir (en faisant un ajout d'information) et d'obtenir : `mon système manque de convivialité` **`ergonomie`** et `logiciel pas fluide` **`ergonomie`**.

Process AFC

```
initialisation d'un message court vide M1

Pour chaque motif m dans M
   si m est defini dans R
       insérer le concept C associé à m dans M1;

emergent_pattern = concept ayant la frequence maximale d'apparition  dans M1

si la longeur de emergent_pattern est égale à 1
   retourner M + emergent_pattern
sinon
   retourner M
```

Exemple 5 : processus d'enrichissement *Add Frequent Concept*

En prenant la ressource et les deux messages courts de l'exemple 4, le processus **Add Frequent Concept** permet d'obtenir le même résultat que le processus **Add Concept** car le concept le plus fréquent ajouté aux deux messages est **ergonomie**. Dans le cas où on obtient plusieurs concepts ayant une même fréquence, le message court n'est pas enrichi.

Process RPC

```
Pour chaque motif m dans M
   si m est defini dans R
       remplacer m par le concept C associé;
   else
       laisser m dans M;

retourner M
```

Exemple 6 : processus d'enrichissement *Replace Pattern by Concept*

En prenant la ressource et les deux messages courts de l'exemple 4, le processus **Replace Pattern by Concept** permet d'obtenir : `mon système` **`ergonomie`** et `logiciel` **`ergonomie`**.

Process RMFC

```
initialisation d'un message court vide M1

Pour chaque motif m dans M
   si m est defini dans R
       insérer le concept C associé à m dans M1;

emergent_pattern = concept ayant la frequence mawimale d'apparition  dans M1

si la longeur de emergent_pattern est égale à 1
    retourner emergent_pattern
sinon
   retourner M;
```

Exemple 7 : processus d'enrichissement *Replace Message by Frequent Concept*

En prenant la ressource et les deux messages courts de l'exemple 4, le processus **Replace Message by Frequent Concept** permet d'obtenir : `mon système` **`ergonomie`** et `logiciel` **`ergonomie`**.

6.4 Conclusion

Les ressources construites selon l'approche proposée font parties des ressources autonomes, plus précisement des thesaurus. Pour une partition (ou des partitions) de messages courts, l'approche construit un ensemble de collections de motifs sémantiquement proches. Les collections sont comme des synsets, des composantes atomiques sur lesquelles repose WordNet [25]. Chaque collection est représentée par un concept qui est lié par un lien d'hyponymie avec des motifs qui le caractérisent. Les ressources sont utilisées pour enrichir les messages courts via

les processus d'enrichissement mis en place (section 6.3). Les points faibles de l'approche se trouvent au niveau des étapes de la construction du corpus et de la validation. En effet, ces étapes sont des étapes manuelles qui nécessitent une expertise métier. Il est important pour ces étapes d'impliquer plusieurs experts métiers afin de calculer un coefficient d'accord (coefficient Kappa par exemple).

TROISIÈME PARTIE

Expérimentation et bilan

CHAPITRE 7

MISE EN PLACE D'UN BANC DE TEST

7.1 Introduction

Avant de s'intéresser aux les résultats obtenus lors de nos expérimentations, nous allons présenter dans ce chapitre le système mis en place pour évaluer l'impact de notre approche. Nous décrivons les chaînes de traitements comparées, à savoir les chaînes de traitements de l'outil Meeting Software avant et après l'intégration de notre approche. Nous détaillons également dans ce chapitre les jeux de données tests sur lesquels les chaînes de traitement sont comparées ainsi que les mesures de qualité utilisées.

7.2 Les jeux de données tests

Le traitement réalisé sur les messages courts dans Meeting Software est synchrone (en temps réel). Soient T le temps imparti pour réaliser le regroupement de n messages courts et t_0 le temps où les premières données sont disponibles.

Nous considérons $t_i = t_{i-1} + dt$, le $i^{\text{ème}}$ temps avec dt une unité de temps variable tel que $\sum_i t_i = T$. Une classification non supervisée est appliquée aux messages courts arrivant au temps t_0, puis corrigée manuellement par le pilote humain. Cette dernière consiste à déplacer des messages courts mal classés dans des groupes existants ou dans des nouveaux groupes (création). On obtient alors des groupes des messages courts cohérents. Les données arrivant au temps t_1 sont ajoutées aux groupes construits précédement moyennant un algorithme de classification supervisée. Le pilote intervient aussi pour effectuer une correction manuelle. Les données arrivant au temps t_i subissent le même traitement que celles reçu par l'outil au temps t_1. À la fin de chaque temps t_i une partition $GS2_i$ (regroupement sémantique de niveau 2 : système + une correction humaine) des messages courts est stockée, qui correspond au regroupement des messages courts reçus par Meeting Software jusqu'au temps t_i. La dernière partition $GS2_T$ qui correspond au regroupement de l'ensemble de messages courts reçues par Meeting Software, peut être soumise à une dernière correction par le pilote avant sa mise à la disposition au client. C'est ce que nous appelons GS3 (regroupement sémantique de niveau trois). Chaque partition est sérialisée en utilisant la libraire Python *Pickle*.

Exemple d'un jeu de données

En 2016, un groupe spécialisé dans la réparation d'automobiles a réuni ses collaborateurs afin d'échanger sur les nouvelles transformations. Lors de cet évenement, la question suivante a été posée aux participants : ***Comment concrétiser le principe de responsabilisation dans notre pratique de tous les jours ?***

Les Tables 7.1 à 7.5 décrivent les différentes étapes du traitement des 32 réponses à la question reçues par l'outil. L'étape 0 correspond à une classification non supervisée (clustering) et les autres étapes à des classifications supervisées. Un code couleur montre la coordination entre le système et le pilote humain :

— En rouge : Les nouvelles réponses traitées par le système (classification supervisée) et validées par le pilote humain.

— En bleu : Les réponses de l'étape précédente deplacées par le pilote humain.

— En magenta : La création d'un nouveau groupe par le pilote humain.

Groupes	**Réponses**
groupe n°1	
groupe n°2	
	Aller Jusqu'à l'execution
	Reformuler ce que chacun attend
groupe n°3	
	Notion de confiance
	Confiance, partage, écoute
	Faire confiance aux BUs qui connaissent leur marché
	Authenticité
groupe n°4	
	Monter des projets !
groupe n°5	
	Points managériaux réguliers
groupe n°6	
	Savoir déléguer et donc responsabiliser nos collaborateurs
	Valoriser les initiatives même si elles échouent

TABLE 7.1 – Etape 0 : clustering

Groupes	Réponses
groupe n°1	
groupe n°2	
	Aller Jusqu'à l'execution
	Reformuler ce que chacun attend
groupe n°3	
	Notion de confiance
	Confiance, partage, écoute
	Faire confiance aux BUs qui connaissent leur marché
	Authenticité
	Plus de bienveillance dans les visites des pays chez nous en France
	Un homme = une mission = des moyens = un objectif
	Se dire les choses
groupe n°4	
	Monter des projets !
groupe n°5	
	Points managériaux réguliers
groupe n°6	
	Savoir déléguer et donc responsabiliser nos collaborateurs
	Valoriser les initiatives même si elles échouent
groupe n°7	
	Accompagner, former nos collaborateurs

TABLE 7.2 – Etape 1 : classification supervisée

Groupes	Réponses
groupe n°1	
	Aller Jusqu'à l'execution
	Monter des projets !
groupe n°2	
	Donner de la visibilité et du sens
	Render visible les talents et les identifier
groupe n°3	
	Notion de confiance
	Confiance, partage, écoute
	Faire confiance aux BUs qui connaissent leur marché
	Authenticité
	Plus de bienveillance dans les visites des pays chez nous en France
	Un homme = une mission = des moyens = un objectif
	Se dire les choses
groupe n°4	
	Transmission des savoirs être et faire par l'échange et le parrainage (accompagnement)
groupe n°5	
	Points managériaux réguliers
groupe n°6	
	Savoir déléguer et donc responsabiliser nos collaborateurs
	Valoriser les initiatives même si elles échouent
	Laisser reellememt les collaborateur prendre les decisions
	Prendre conscience régulièrement des décisions que nous avons prises
	Reformuler ce que chacun attend
groupe n°7	
	Accompagner, former nos collaborateurs

TABLE 7.3 – Etape 2 : classification supervisée

Groupes	Réponses
groupe n°1	
	Aller Jusqu'à l'execution
	Monter des projets !
	Privilégier la personnalité,le savoir être,le savoir faire pour les postes clefs
groupe n°2	
	Donner de la visibilité et du sens
	Render visible les talents et les identifier
	Reformuler ce que chacun attend
groupe n°3	
	Notion de confiance
	Confiance, partage, écoute
	Faire confiance aux BUs qui connaissent leur marché
	Authenticité
	Plus de bienveillance dans les visites des pays chez nous en France
	Un homme = une mission = des moyens = un objectif
	Se dire les choses
	Donner le droit à l'erreur
	La confiance
groupe n°4	
	Points managériaux réguliers
	C'est une méthode apprenante
	Transmission des savoirs être et faire par l'échange et le parrainage (accompagnement)
	Accompagner, former nos collaborateurs
groupe n°5	
	Savoir déléguer et donc responsabiliser nos collaborateurs
	Valoriser les initiatives même si elles échouent
	Laisser reellememt les collaborateur prendre les decisions
	Prendre conscience régulièrement des décisions que nous avons prises
	Le rendre compte

TABLE 7.4 – Etape 3 : classification supervisée

Groupes	Réponses
groupe n°1	
	Il vaut mieux faire son cash
	Ouvrir la participation aux Codir
	Privilégier la personnalité,le savoir être,le savoir faire pour les postes clefs
groupe n°2	
	Donner de la visibilité et du sens
	Render visible les talents et les identifier
	Reformuler ce que chacun attend
groupe n°3	
	Notion de confiance
	Confiance, partage, écoute
	Faire confiance aux BUs qui connaissent leur marché
	Authenticité
	Plus de bienveillance dans les visites des pays chez nous en France
	Un homme = une mission = des moyens = un objectif
	Se dire les choses
	Donner le droit à l'erreur
	La confiance
	Accepter les erreurs
	Faire confiance aux hommes
groupe n°4	
	Points managériaux réguliers
	C'est une méthode apprenante
	Transmission des savoirs être et faire par l'échange et le parrainage (accompagnement)
	Accompagner, former nos collaborateurs
groupe n°5	
	Savoir déléguer et donc responsabiliser nos collaborateurs
	Valoriser les initiatives même si elles échouent
	Laisser reellememt les collaborateur prendre les decisions
	Prendre conscience régulièrement des décisions que nous avons prises
	Le rendre compte
	Encourager les initiatives
	Appliquer le principe de subsidiiarité
	Comment récompenser ou sanctionner cette responsabilité ?
	Associer les experts pour valider sans passer par la hierarchie
groupe n°6	
	Aller Jusqu'à l'execution
	Monter des projets !

TABLE 7.5 – Etape 4 : classification supervisée

La Table 7.6 fournit des statistiques sur les jeux de données tests utilisés, que nous appelons *benchmarks*. Ces derniers sont un mélange de données provenant des trois solutions proposées par l'entreprise (section 5.2) et sont regroupés par type (chaque type est caractérisé par le nombre de messages courts). Chaque benchmark correspond à un regroupement des données lié à une question ou un sujet traité. Pour chaque type de benchmarks, on fournit les informations suivantes :

— Le nombre minimum et maximum de messages courts dans les benchmarks (NB-MC),

— Le nombre minimum, le nombre moyen et le nombre maximum des étapes correspondantes aux classifications supervisées (NB-CS),

— Des chiffres concernant la première, deuxième et dernière étapes de la classification supervisée (Etap 1, Etap 2 et Etap N) :

  — Taille de vocabulaire minimale, moyenne et maximale (vb),

  — Nombre de groupe minimum, moyen et maximum (group).

Benchmarks	**NB-MC**	**NB-CS**	**Etape 1**	**Etape 2**	**Etape N**
type 1	10-39	0-5-17	vb : 1-14-29 group : 3-5-9	vb : 4-22-48 group : 3-5-8	vb : 8-52-110 group : 3-6-9
type 2	40-100	4-7-13	vb : 5-52-121 group : 3-6-9	vb : 10-71-187 group : 3-6-9	vb : 84-137-234 group : 5-7-13
type 3	101-300	11-18-29	vb : 9-41-73 group : 4-7-12	vb : 22-59-111 group : 4-8-13	vb : 6-10-13 group : 6-10-13
type 4	500-1000	12-27-47	vb : 58-93-134 group : 3-5-9	vb : 89-135-205 group : 3-5-8	vb : 744-1084-1693 group : 3-6-9
type 5	1000-1600	9-23-60	vb : 102-268-738 group : 5-7-8	vb : 183-382-939 group : 6-8-9	vb : 1185-1780-2525 group : 11-15-20

TABLE 7.6 – Caractéristiques des Benchmarks par type constituant les jeux de données test utilisés

7.3 Les chaînes de traitement comparées

La chaîne de traitement de l'outil Meeting Software est l'ensemble des processus et algorithmes qui permettent à l'outil de produire un regroupement de messages courts. Dans cette section, nous présentons la chaîne de traitement de base (CTB) et la chaîne de traitement intégrant l'enrichissement sémantique (CTE).

7.3.1 Chaîne de Traitement de Base (CTB)

La Chaîne de Traitement de Base de l'outil Meeting Software est composée des éléments suivants :

— Pré-traitement : tokenisation, filtrage de stopwords, racinisation (lemmatisation ou stemmatisation ou lemmatisation et stemmatisation)

— Vectorisation : tf-idf de la librairie Python *Sklearn*. Les paramètres utilisés sont :

  — **min_df** : 1. Nous ne considérons dans le vocabulaire construit que les termes (mots ou combinaisons de mots) dont la fréquence d'apparition est supérieure à 1.

  — **ngram_range** : (1, 2). Ce paramètre limite la taille des n-grammes considérés. Dans notre cas les n-grammes sont de taille minimale 1 et maximale 2. Exemple : ***collaborateur***, ***accompager collaborateur***.

— Classification non supervisée : Ward [55], algorithme de clustering hiérarchique de la librairie Python *Sklearn*. Le nombre de clusters est prédéfini.

— Classification supervisée : Extratrees, algorithme de classification supervisée (de la famille des arbres aléatoires) de la librairie Python *Sklearn*. Les paramètres utilisés sont :

  — **n_estimetors** : 100, qui est le nombre d'arbres dans la forêt.

  — **random_state** : 0, pour annuler l'effet aléatoire de l'algorithme.

7.3.2 Chaîne de Traitement Enrich (CTE)

La Chaîne de Traitement Enrich est composée des éléments ci-dessous. Cette chaîne utilise les mêmes paramètres que la chaîne de base auxquels nous ajoutons l'enrichissement lors de l'étape de pré-traitement :

— Pré-traitement : tokenisation, filtrage de stopwords, racinisation (lemmatisation ou stemmatisation ou lemmatisation et stemmatisation), enrichissement.

— Vectorisation : tf-idf de la librairie Python *Scikit-learn*.

— Classification non supervisée : *Ward* [55], algorithme de clustering hiérarchique de la librairie Python *sklearn*. Le nombre de clusters est prédéfini.

— Classification supervisée : *Extratrees*, algorithme de classification supervisée (de la famille des arbres aléatoires) de la librairie Python *Sklearn*.

Quatre chaînes de traitement **CTE** seront comparées avec la chaîne de traitement de base **CTB** :

— CTE + AC : BOW utilisant le processus d'enrichissement Add Concept.

— CTE + AFC : BOW utilisant le processus d'enrichissement Add Frequent Concept process.

— CTE + RPC : BOW utilisant le processus d'enrichissement Replace Pattern by Concept process.

— CTE + RMFC : BOW utilisant le processus d'enrichissement Replace Message by Frequent Concept process.

Lors de deux premières années de thèses, seuls les regroupements finaux des données (**GS3**) ont été utilisés pour tester notre approche (Expérimentations 1 et 2). Par la suite, les données ont été stockées par étape ($GS2_i$ avec $i = 1, ..., T$, cf. section 7.2), ce qui nous a permis de tester les données à toutes les étapes (Expérimentation 3).

Pour les expérimentations 1 et 2, des chaînes de traitement réduites sont comparées :

— **CTBR**, chaîne de traitement déduite de la chaîne **CTB** en supprimant la classification supervisée.

— **CTER**, chaîne de traitement déduite de la chaîne **CTE** en supprimant la classification supervisée.

La Figure 7.1 montre comment les éléments des chaînes de traitement sont coordonnés pour traiter les données. La différence entre les deux chaînes comparées se trouve au niveau de l'enrichissement qui n'est pas présent dans la chaîne de base.

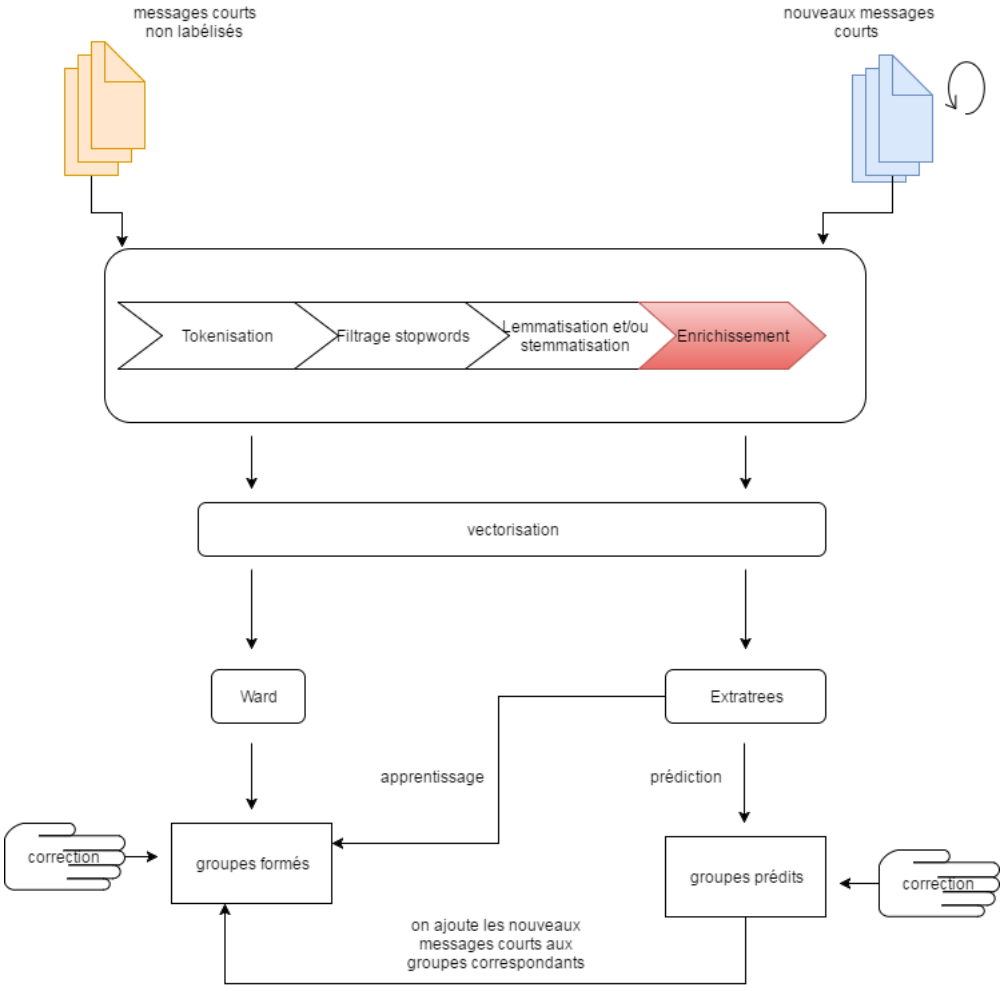

FIGURE 7.1 – Chaîne de traitement Meeting Software

7.4 Evaluation

Nous utilisons le test de Student (test t apparié) [69] pour évaluer la chaîne de traitement de l'outil Meeting Software sur plusieurs benchmarks. Le test de Student permet de comparer les moyennes de deux séries de valeurs présentant un lien. Voici des exemples d'application :

— Observations avant et après sur des mêmes sujets (par exemple, test de diagnostic des résultats d'élèves avant et après un module ou un cours particulier).

— Comparaison de deux méthodes ou traitements différents appliqués sur des mêmes sujets.

Dans notre cas, il s'agira du deuxième exemple d'application. Nous comparons les performances de la Chaîne de Traitement de Base (CTB) et la Chaîne de Traitement Enrich (CTE) sur les mêmes jeux de données.

7.4.1 Procédure du test

Soient $X = [x_1, x_2, ..., x_n]$ et $Y = [y_1, y_2, ..., y_n]$ les vecteurs contenant les nombres de déplacement en utilisant respectivement le premier et le deuxième pré-traitement. Pour tester l'hypothèse nulle selon laquelle la différence moyenne réelle est égale à zéro, la procédure suivante est utilisée :

— Calcul des différences $d_i = (x_i - y_i)$ avec $i = 1, ..., n$

— Calcul de la moyenne des différences d_i, $\tilde{d}$

— Calcul de l'écart-type des différences d_i, S_d

— Calcul de statistique de test $T_{stat} = \frac{\tilde{d}}{S_d}$ sous l'hypothèse nulle. Cette statistique suit une distribution de student à $n - 1$ degré de liberté.

7.4.2 Interprétation

L'interprétation du test t est fournie par la lecture de la table de la distribution de Student à $n - 1$ degré de liberté. On compare la statistique T_{stat} à la valeur équivalente t_{n-1} avec un seuil α généralement égal à 5% dans la table :

— Si T_{stat} est inférieur t_{n-1} au seuil α, les moyennes ne diffèrent pas significativement. On obtient alors un intervalle de confiance contenant 0.

— Si T_{stat} est supérieur à t_{n-1} au seuil α, les moyennes diffèrent significativement. On obtient un intervalle de confiance positif si les valeurs du premier vecteur en entrée sont supérieures à celles du deuxième vecteur (et inversement lorsque l'intervalle est négatif).

7.4.3 Application sur les données de l'entreprise

Les données utilisées pour effectuer les expérimentations ont une structure particulière. Lors d'un séminaire, pour chaque question, les données arrivent par vague dans Meeting Software. Soit l'ensemble de données de référence $data_{ref} = \{v_1, v_2, ..., v_n\}$ où v_i correspond aux données envoyées à l'étape i. v_0 correspond aux données utilisées lors du clustering et v_i avec $i >= 1$ aux données à classer en utilisant comme base d'apprentissage le regroupement fait à l'étape $i - 1$. Nous distinguons deux types de tests :

— **Test Global**

  — La constitution de l'échantillon : Soit m données de référence, l'échantillon T est constitué de toutes les étapes provenant des m données de référence.

  — Les entrées du t test : $Vect_1$ et $Vect_2$, deux vecteurs représentant les résultats avant et après l'application d'un traitement sur l'échantillon T. Le $j^{ième}$ élément du $Vect_1$ (respectivement du $Vect_2$) correspond au pourcentage du nombre de messages déplacés par rapport aux données du $j^{ième}$ élément de l'échantillon T.

  — La sortie : l'intervalle de confiance des différences ($Vect_1[k]$ - $Vect_2[k]$)

— **Test par étape**

  — La constitution de l'échantillon : Soient m données de référence et $N = \{N_1, N_2, ..., N_m\}$ où N_i correspond au nombre d'étapes associées aux données de référence i. On choisit $n = min(N)$ pour la constitution des échantillons $T = \{t_1, t_2, ..., t_n\}$ où t_j est constitué de toutes les étapes j des m données de référence.

  — Les entrées du test t : Pour chaque étape j le test t prend en entrée $Vect_1$ et $Vect_2$, deux vecteurs représentant les résultats avant et après l'application d'un traitement sur l'échantillon t_j.

  — La sortie : Pour chaque étape, la sortie est un intervalle de confiance des différences ($Vect_1[k]$ - $Vect_2[k]$)

7.5 Conclusion

Le banc de tests mis en place permet de mesurer l'impact de l'enrichissement sur l'outil Meeting Software. Il s'agit de comparer la chaîne de traitement de base avec la chaîne de traitement intégrant l'enrichissement. 58 jeux de données tests ont été utilisés pour l'évaluation, chaque jeu de données étant composé des données pour chaque étape de traitement de l'outil. Nous utilisons le test t afin de comparer les deux chaînes de traitement.

La stucture de jeux de données tests décrite à la section 7.2 a été mise en place récemment par notre équipe du pôle R&D avec pour objectif de décrire la réalité du fonctionnement de

l'outil. Lors de deux premières années de thèse, seuls les jeux de données issus de la dernière étape de traitement de l'outil (GS3) ont été utilisés pour tester l'impact de l'enrichissement. Ces expérimentations sont décrites dans le chapitre qui suit.

CHAPITRE 8

EXPÉRIMENTATIONS ET ÉVALUATIONS

8.1 Introduction

Ce chapitre a pour objectif de répondre à la question suivante : l'utilisation de ressources externes basées sur les partitions de données historiques d'un sujet peuvent-elles améliorer la performance de la classification (supervisée et non supervisée) de nouveaux messages courts du même sujet ? Au cours de ce chapitre, nous décrivons trois expérimentations réalisées afin d'évaluer l'impact de notre approche sur la chaîne de traitement de l'outil Meeting Software.

Pour les deux premières expériences, les messages courts proviennent des secteurs d'activité bien identifiés et correspondent à des réponses liées à des questions spécifiques. Pour chaque expérience, on construit une ressource sémantique sur un historique de messages courts pour enrichir des nouveaux messages dans le but d'améliorer leurs regroupements. Le jeu de données test utilisé diffère de ceux présentés dans la section 7.3. En effet, le système de recupération de messages courts à chaque étape de traitement de Meeting Software n'avait pas été mise en place. Le test est donc réalisé sur le regroupement final fournit au client c'est-à-dire le GS3.

L'objectif de la troisième expérience est de savoir si l'on peut mettre en place une ressource générique, c'est-à-dire une ressource qui améliorerait la performance de Meeting Software quelque soit le secteur d'activité, le sujet (ou la question) et la source de messages courts (la solution). L'impact de l'enrichissement a été testé sur les jeux de données tests présentés dans la section 7.3.

8.2 Expérimentation 1

8.2.1 Jeux de données utilisées

L'expérience est réalisée sur les données d'une enquête repetitive où une entreprise du domaine conseil demande à ses collaborateurs leurs points de vues sur ses "points critiques". Nous disposons de quatre jeux de données correspondant à la même enquête réalisée à des périodes différentes. Ces données correspondent aux données de la solution Pulsation (voir section 5.2.3). L'objectif de cette expérience est de montrer qu'une ressource peut être construite sur les enquêtes $E1$, $E2$, $E4$ pour améliorer le la classification non supervisée de messages courts de

l'enquête $E3$. La table 8.1 fournit des informations sur les quatres enquêtes utilisées.

Enquêtes	**Nombre de participants**	**Nombre de contribution**	**Période de collecte**
$E1$	328	970	25/02/2016
$E2$	80	274	26/04/2016
$E3$	39	121	27/06/2016
$E4$	151	502	05/09/2016

TABLE 8.1 – Informations sur les quatres enquêtes utilisées

8.2.2 La ressource sémantique extraite

La ressource sémantique extraite est composée de 15 classes sémantiques. La Table 8.2 fournit pour chacune des classes sémantiques le nombre de motifs qui la composent. La Table 8.3 montre un extrait des motifs composant les classes sémantiques ***encourager-initiative*** et ***parcours-professionnel***. On peut voir par exemple que les motifs *idée*, *expression-idée*, *hésiter-pas* appartiennent à la classe sémantique ***encourager-initiative*** et les motifs *expatriation*, *recrutement*, *compétence* appartenant à la classe sémantique ***parcours-professionnel***.

classes sémantiques	**nombre de motifs**
clarifier-strategie	119
encourager-initiative	113
voloriser-reconnaissance	112
valoriser-collectif	87
outils-adapte	32
collectif-comex	160
communication-adapte	43
stabiliser-organisation	17
meilleure-prioriser	72
responsabilite-manager	94
respect-regle	124
evolution-professionnel	135
parcours-professionnel	21
simplifier-processus	103
deleger-responsabiliser	108

TABLE 8.2 – Répartition des motifs par classe sémantique

encourager-initiative	**parcours-professionnel**
émotivité-entreprise	expatriation
manquer-marge	carriere
idée	rh
rapidement-hésiter	formation
initiatif	recruter
inhibant-initiative	expertise
demander-oser	évolution
echec-plus	contrat
hésiter-pas	opportunité
décision-prendre	emploi
initiative-personnel	carrière
idée-nouvel	embauchére
soutenir-accompagner	accompagnement
crainte-pas	parcours
expression-idée	départ
autonomie-individuel	recrutement
priser-initiative	étranger
idée	compétence
rapidement-hésiter	personnel
niveaux-capacité	postes
environnement-percevoir	mobilité
nouvel-valoriséer	
pouvoir-constituer	
capacité-comprendre	
manquer-envie	
conformité-reproduction	
idée-nouvel	
soutenir-accompagner	
encourager	
initiative	

TABLE 8.3 – Extrait des motifs de deux classes sémantiques

8.2.3 Effet de l'enrichissement en utilisant notre approche

Nous utilisons comme mesures d'évaluation (voir section ...), *adjusted Rand Index* (ARI) et la mesure developpée en interne (Le nombre de déplacements). On compare la chaîne de traitement réduite de base **CTBR** avec les chaînes de traitement réduites **CTER**. La Figure 8.1 et la Table 8.4 fournissent les résultats obtenus lors de cette expérimentation.

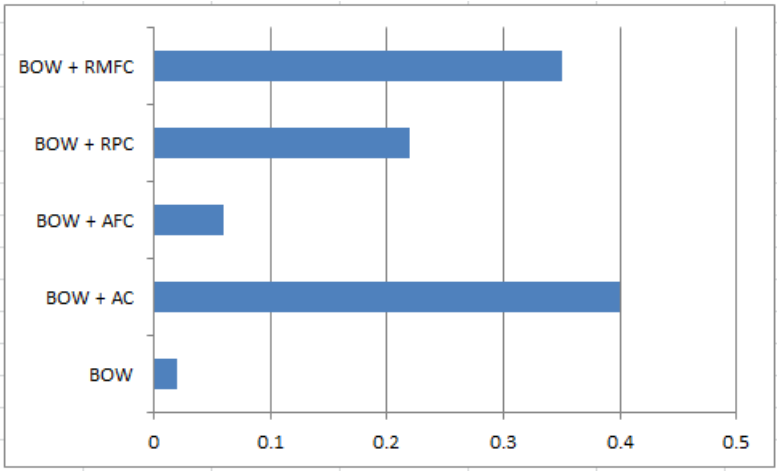

FIGURE 8.1 – Effet d'enrichissement sur l'enquête $E3$

	CTBR	CTER + AC	CTER + AFC	CTER + RPC	CTER + RMFC
Adjusted Rand Index	0.02	0.4	0.06	0.22	0.35

TABLE 8.4 – Effet d'enrichissement sur l'enquête $E3$

Les résultats obtenus montrent que les chaînes de traitement **CTER** fournissent de meilleurs résultats comparer à la chaîne de traitement **CTBR**. On peut voir aussi que l'amélioration varie en fonction de la stratégie d'enrichissement utilisée. On obtient par exemple un ARI (Adjusted Rand Index) de 0.4 avec la chaîne **CTER + AC** et 0.06 avec **CTER + AFC**. Il est important d'effectuer plusieurs expérimentations afin de choisir le processus d'enrichissement adapté aux messages courts traités.

8.3 Expérimentation 2

8.3.1 Jeux de données utilisés

L'expérience est réalisée sur les données d'une enquête répétitive où une entreprise du domaine bancaire demande à ses collaborateurs leurs ressentis sur la qualité de l'environnement technique mis à leur disposition. Suite à chaque enquête, l'entreprise fournit à Succeed Together un fichier excel contenant les verbatims afin d'obtenir un regroupement sémantique de ces données. Ces dernières correspondent aux données de la solution Pulsation (voir section 5.2.2). L'objectif de cette expérience est de montrer qu'une ressource peut être construite sur l'enquête $E2$ pour améliorer la classification non supervisée de messages courts de l'enquête $E1$.

8.3.2 La ressource sémantique extraite

La ressource sémantique extraite est composée de 13 classes sémantiques. La Table 8.6 fournit pour chacune des classes sémantiques le nombre de motifs qui la composent. La Table

Enquêtes	**Nombre de participants**	**Nombre de contributions** [1]	**Période de collecte**
$E1$	39	121	27/06/2016
$E2$	151	502	05/09/2016

TABLE 8.5 – Informations sur les quatres enquêtes utilisées

8.7 montre un extrait des motifs composant les classes sémantiques ***encourager-initiative*** et ***parcours-professionnel***. On peut voir par exemple que les motifs *idée*, *expression-idée*, *hésiter-pas* appartiennent à la classe sémantique ***encourager-initiative*** et les motifs *expatriation*, *recrutement*, *compétence* appartenant à la classe sémantique ***parcours-professionnel***.

classes sémantique	**nombre de motifs**
outils_pas_convivial	113
système_compliqué	46
manque_formation	11
trop_disfonctionnement	160
problème_imprimantes	4
base_document_incomplet	175
outil_pas_adapté	20
quand_outil_unique	59
écran_trop_petit	6
moyen_age	18
resolution_incident	20
outil_trop_lent	96
gestion_mots_passe_lourde	68

TABLE 8.6 – Répartition des motifs par classe sémantique

outils-pas-convivial	**trop_disfonctionnement**
ergonomique	trop-plantage
manquer-convivialité	anomalie
pas-logique	stres
ergonomie-outil	souvent-indisponible
pratique-pas	redémarrer
trop-lien	erreur-grave
intranet-manquer	redémarrage
ergonomie-poste	récurrent
clair	obliger-relancer
raccourci	beaucoup-perturbation
lourd-peu	nombreux-bug
pas-accessible	beug
plus-intuitif	bloquer
clarté	indisponibilité
non-intuitif	problème-récurrent
illogique	ouverture-poste
intuitif-outil	plant
conviviaux	souci
illisible	dysfonctionnement-bug
manquer-clarté	jour-problème
plusieur-clic	problème-connexion
visibilité	instabilité
logique	fonctionnel
manquer-ergonomie	perturbation
environnement-travail	déconnecter
peu-ergonomique	souvent-bug
intuitivité	instabilité-système
ordre-consultation	perte-donnée
logique-classement	fermeture
plus-simple	plantage

TABLE 8.7 – Extrait des motifs de deux classes sémantiques

8.3.3 Effet de l'enrichissement en utilisant notre approche

Comme dans l'expérimentation précédente, nous utilisons comme mesures d'évaluation (voir section ...), Adjusted Rand Index (ARI) et la mesure developpée en interne (le nombre de déplacements). On compare la chaîne de traitement réduite de base **CTBR** avec les chaînes de traitement réduites **CTER**. La Figure 8.2 et la Table 8.8 fournissent les résultats obtenus.

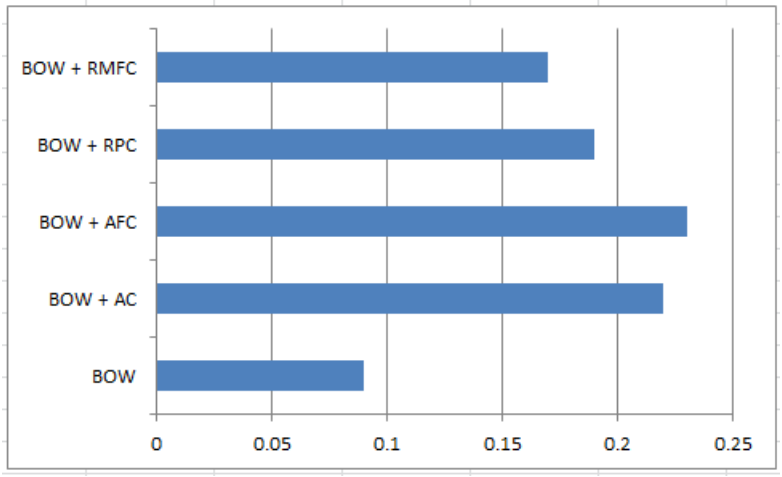

FIGURE 8.2 – Effet d'enrichissement sur l'enquête $E1$

	CTBR	CTER + AC	CTER + AFC	CTER + RPC	CTER + RMFC
Adjusted Rand Index	0.09	0.22	0.23	0.19	0.17

TABLE 8.8 – Effet d'enrichissement sur l'enquête $E1$

Les observations sont les mêmes que celles de l'expérimentation précédente. Les chaînes de traitement avec enrichissement fournissent les meilleurs résultats. On constate aussi la variation de l'impact d'enrichissement suivant la stratégie d'enrichissement utilisée.

8.4 Expérimentation 3

8.4.1 Jeux de données utilisés

Nous avons construit un corpus sur les données du premier trimestre de l'année 2015. On a extrait la ressource pour mesurer l'impact sur les données de l'année 2016/2017. Ici, on fait pas attention aux sujets des questions (les deux autres expériences : les données sur lesquelles les ressources sont construites et les données de références traitent les mêmes sujets). De plus le format du jeu de données test n'est pas le même, les données de reférence recentes retracent l'activité du pilote : pour une question, on recupère les regroupements effectués par étape de traitement. Il ya plus de 50 données de reférences. On constate que l'enrichissment peut aider tout comme il peut dégrader.

8.4.2 La ressource sémantique extraite

La ressource sémantique extraite est composée de 32 classes sémantiques. Nous appelons cette ressource sémantique **RSG0**. La table 8.9 fournit pour chacune des classes sémantiques le nombre de motifs qui la composent. La table 8.10 montre un extrait des motifs composant les

classes sémantiques ***thema-bureautique*** et ***thema-cohesion-equipe***. On peut voir par exemple que les motifs *action-simplif*, *centralis*, *processu-lourd* appartiennent à la classe sémantique ***thema-bureautique*** et les motifs *travail-équipe*, *confianc-reciproqu*, *consider-comm* appartenant à la classe sémantique ***thema-cohesion-equipe***.

Répartition des motifs par classe sémantique	
classe sémantique	**nombre de motifs**
thema-compétitivité	78
thema-organisation	8
thema-responsabilités	27
thema-reactivite	82
thema-investissement	7
thema-RAS	25
thema-cohesion-equipe	30
thema-positif	11
thema-evenementiel	96
thema-qualite-produit	14
thema-bureaucratie	128
thema-valorisation	111
thema-nouvelles-technologies	167
thema-echanges-horizontaux	113
thema-formation	79
thema-experience-client	16
thema-feedback	2
thema-echanges-verticaux	53
thema-exemplarite	18
thema-outils	17
thema-management	9
thema-charge-travail	25
thema-relations-partenaires	1
thema-communication-externe	260
thema-ressources-humaines	31
thema-innovation	32
thema-performance-financière	5
thema-concurrence	13
thema-professionalisme	22
thema-motivation	13
thema-négatif	15

TABLE 8.9 – Répartition des motifs par classe sémantique

thema-bureaucratie	**thema-cohesion-equipe**
action-simplif	travail-équip
pert-efficac	cohes-équip
différent-entit	confianc-agent
organis-processu	climat
centralis	plus-confianc
processu-lourd	climat-confianc
lourdeur-complex	confianc-réciproqu
lourdeur-processu	confianc-réseau
mod-fonction	object-commun
proces-souscript	consider-comm
lourdeur-administr	solidarit
contrôl-trop	cohes-equip
fluidifi-process	confianc-compagn
fluidifi	action-cohes
souscript-plus	communaut
démarch-simplif	projet-commun
lourdeur-system	intérêt-commun
homogénéis	rassembl
hierarch	relat-confianc
simplifi-processu	esprit-équip
manqu-lisibil	travail-collect
procédur	
reduir-interfac	
réduir-nombr	
processu-décis	
plus-clair	
amélior-processu	
lisibil-organis	
simplif-proc	
complex-organis	

TABLE 8.10 – Extrait des motifs de deux classes sémantiques

8.4.3 Effet de l'enrichissement en utilisant notre approche

Nous utilisons *t-test* de Student pour évaluer le gain en nombre de déplacements obtenus en utilisant notre approche. On compare la chaîne de traitement de base **CTB** avec les chaînes de traitement **CTE**. Dans cette expérimentation, nous essayons aussi de voir l'impact des certaines classes sémantiques dans la ressource **RSG0** sur la chaîne de traitement de l'outil Meeting Software. L'objectif est d'évaluer aussi l'impact des classes sémantiques sur les résultats du système. Pour cela, nous deduisons de la ressource **RSG0** deux ressources sémantiques **RSG1** (composée de 16 classes sémantiques) et **RSG2** (composée de 14 classes sémantiques). La table 8.11 fournit les classes sémantiques composant les deux ressources déduites de **RSG0**

RSG1	**RSG2**
thema-compétitivité	thema-organisation
thema-organisation	thema-reactivite
thema-reactivite	thema-investissement
thema-investissement	thema-cohesion-equipe
thema-cohesion-equipe	thema-echanges-verticaux
thema-echanges-verticaux	thema-bureaucratie
thema-bureaucratie	thema-valorisation
thema-valorisation	thema-formation
thema-formation	thema-echanges-horizontaux
thema-echanges-horizontaux	thema-nouvelles-technologies
thema-nouvelles-technologies	thema-management
thema-management	thema-communication-externe
thema-communication-externe	thema-ressources-humaines
thema-ressources-humaines	thema-innovation
thema-innovation	
thema-concurrence	

TABLE 8.11 – classes sémantiques composant les ressources **RSG1** et **RSG2**

Les tables 8.12 à 8.14 fournissent les résultats obtenus. Pour chaque type de jeux de données test, un intervalle sur le gain du nombre de déplacement en utilisant une chaîne de traitement *enrich* par rapport à la chaîne de traitement de base. Les résultats sont organisés par ressource utilisée pour l'enrichissement.

On constate que la plupart des intervalles du gain de déplacements contiennent 0, ce qui signifie qu'on peut gagner et perdre suivant l'étape de traitement. Globalement le gain est faiblement positif pour les stratégies d'enrichissement **AC**, **AFC**, **RPC**. On peut gagner jusqu'à $1,45$ déplacements ou faire $0,45$ déplacement en plus. Par contre, le gain est négatif pour la stratégie **RMFC** quelque soit la ressource utilisée.

On observe aussi la variation du résultat en fonction de la ressource. En effet, l'enrichissement en utilisant la ressource **RSG1** sur bench4 avec la stratégie **AFC** fournit un meilleur résultat par rapport à l'enrichissement en utilisant la ressource **RSG0** avec la même stratégie. On gagne en moyenne par étape de traitement en utilisant la ressource sémantique **RSG1** sur le **bench4** $1,6$ déplacements par rapport $1,15$ déplacements en utilisant la ressource sémantique **RSG0**. C'est le cas aussi du **bench5**, on gagne en moyenne par étape de traitement 1 déplacement en utilisant la ressource **RSG0** avec la stratégie d'enrichissement **AC** contre $1,18$ en utilisant la ressource **RSG1** avec la même stratégie d'enrichissement.

	bench1	**bench2**	**bench3**	**bench4**	**bench5**	**global**
AC	[-10.46 ; 5.46]	[-2.44 ; 4.1]	[-2.03 ; 0.83]	[-0.13 ; 2.48]	[-0.36 ; 2.36]	[-0.32 ; 1.22]
AFC	[-20.91 ; 10.91]	[-4.21 ; 5.87]	[-1.73 ; 1.1]	[-0.04 ; 2.34]	[-0.46 ; 1.58]	[-0.33 ; 1.09]
RPC	[-12.99 ; 12.99]	[-4.21 ; 5.87]	[-1.31 ; 1.83]	[-0.36 ; 2.4]	[-1.23 ; 1.0]	[-0.4 ; 1.16]
RMFC	[-12.99 ; 12.99]	[-17.95 ; 4.61]	[-10.83 ; -6.21]	[-6.43 ; -2.19]	[-9.49 ; -6.02]	[-8.1 ; -5.68]

TABLE 8.12 – Effet de l'enrichissement en utilisant **RSG0**

	bench1	**bench2**	**bench3**	**bench4**	**bench5**	**global**
AC	[0.0 ; 0.0]	[-3.41 ; 5.08]	[-1.55 ; 1.1]	[-0.09 ; 2.66]	[0.09 ; 2.27]	[-0.01 ; 1.41]
AFC	[0.0 ; 0.0]	[-2.71 ; 2.71]	[-1.41 ; 1.3]	[0.34 ; 2.83]	[-0.3 ; 2.07]	[0.02 ; 1.43]
RPC	[-5.46 ; 10.46]	[-3.41 ; 5.08]	[-1.46 ; 1.41]	[-0.58 ; 2.35]	[-0.78 ; 1.21]	[-0.37 ; 1.11]
RMFC	[-5.46 ; 10.46]	[-6.18 ; 5.56]	[-9.52 ; -5.09]	[-5.76 ; -2.08]	[-9.59 ; -6.0]	[-7.26 ; -5.0]

TABLE 8.13 – Effet de l'enrichissement en utilisant **RSG1**

	bench1	**bench2**	**bench3**	**bench4**	**bench5**	**global**
AC	[0.0 ; 0.0]	[-5.42 ; 5.42]	[-1.69 ; 0.93]	[0.1 ; 2.61]	[-0.24 ; 2.07]	[-0.15 ; 1.26]
AFC	[-10.46 ; 5.46]	[-6.23 ; 2.89]	[-1.72 ; 1.3]	[-0.5 ; 2.01]	[-0.35 ; 1.75]	[-0.45 ; 1.02]
RPC	[0.0 ; 0.0]	[-2.0 ; 5.34]	[-0.92 ; 1.67]	[-0.23 ; 2.61]	[-0.37 ; 1.6]	[0.03 ; 1.42]
RMFC	[0.0 ; 0.0]	[-8.37 ; 6.09]	[-8.71 ; -4.5]	[-5.79 ; -2.41]	[-8.56 ; -5.51]	[-6.79 ; -4.72]

TABLE 8.14 – Effet de l'enrichissement en utilisant **RSG2**

Pour voir concretement l'impact de type de ressource sémantique sur les jeux de données tests, nous avons utilisé la stratégie **AC** et évaluer pour chaque type de jeux de données tests :

— la proportion de données tests pour lesquelles l'enrichissement a été positif **(oui)**

— la proportion de données tests pour lesquelles l'enrichissement a été négatif **(non)**

— la proportion de données tests pour lesquelles l'enrichissement n'a produit aucun n'effet **(neutre)**

La figure 8.3 montre que la ressource initiale **RSG0** fournit des meilleurs résultats sur le **bench1** avec 10,3% des données pour lesquelles l'enrichissement est négatif contre 24,1% pour **RSG1** et 27,6% pour **RSG2**. Pour le jeu de données **bench3**, c'est la ressource déduite **RSG1** qui fournit les meilleurs résultats (figure 8.5) avec 30% des données pour lesquelles l'enrichissement est négatif contre 66,7% pour **RSG0** et 40% pour **RSG2**.

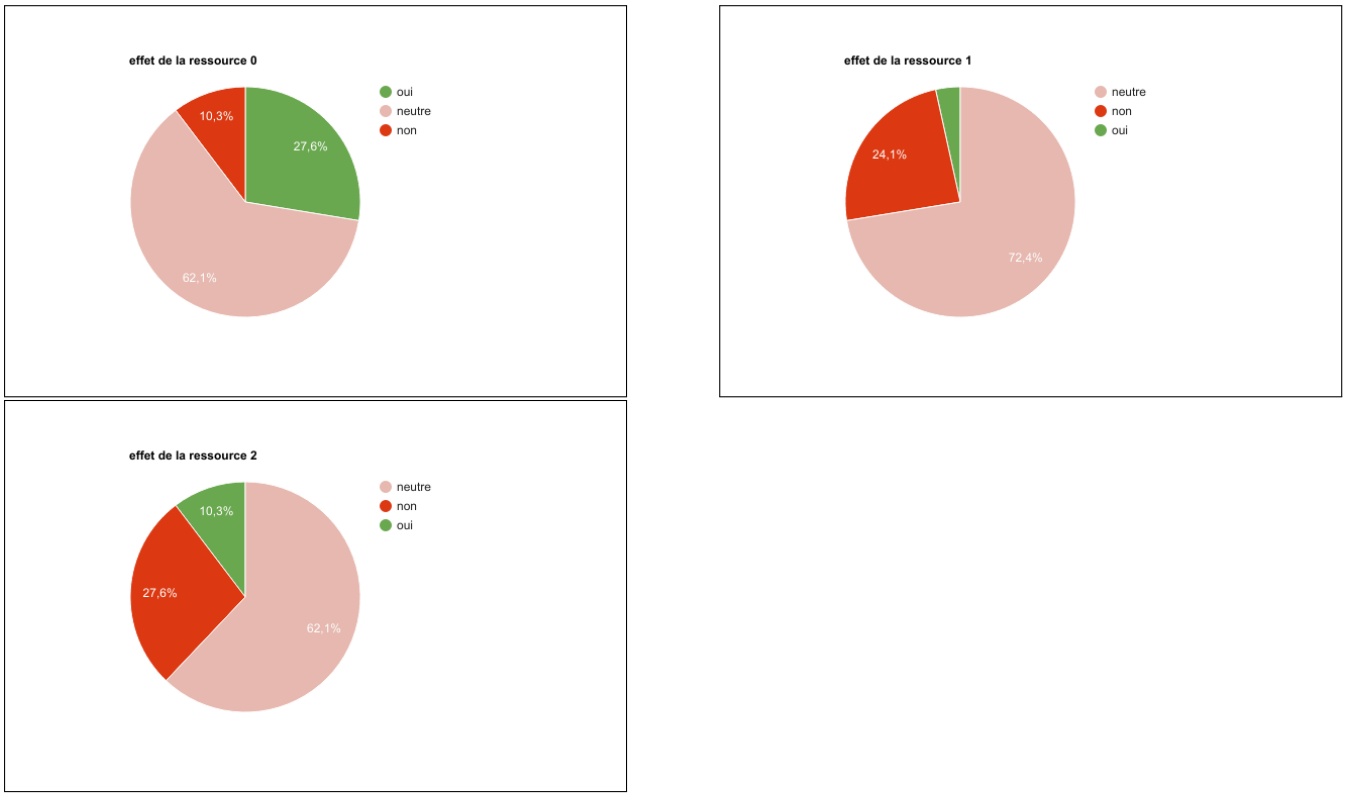

FIGURE 8.3 – effet de l'enrichissement sur le bench1

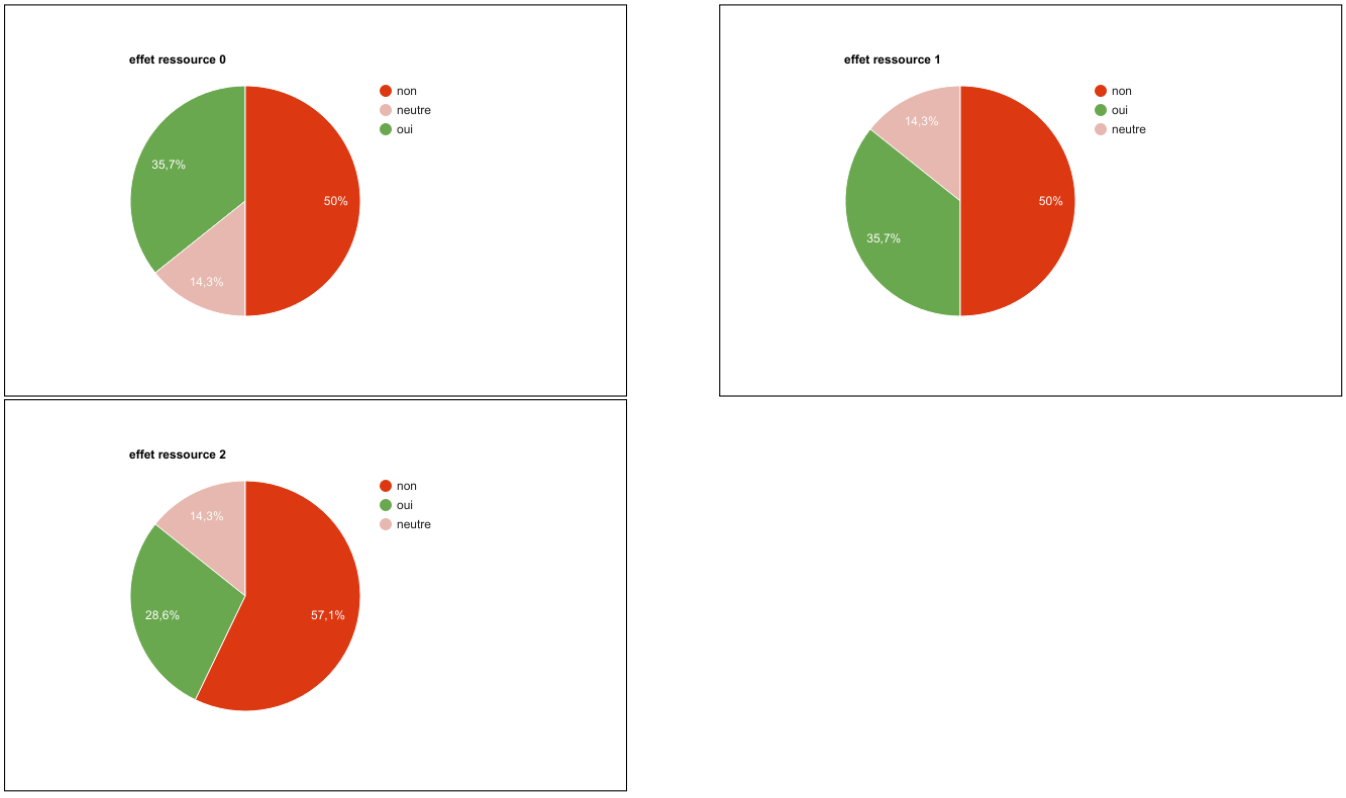

FIGURE 8.4 – effet de l'enrichissement sur le bench2

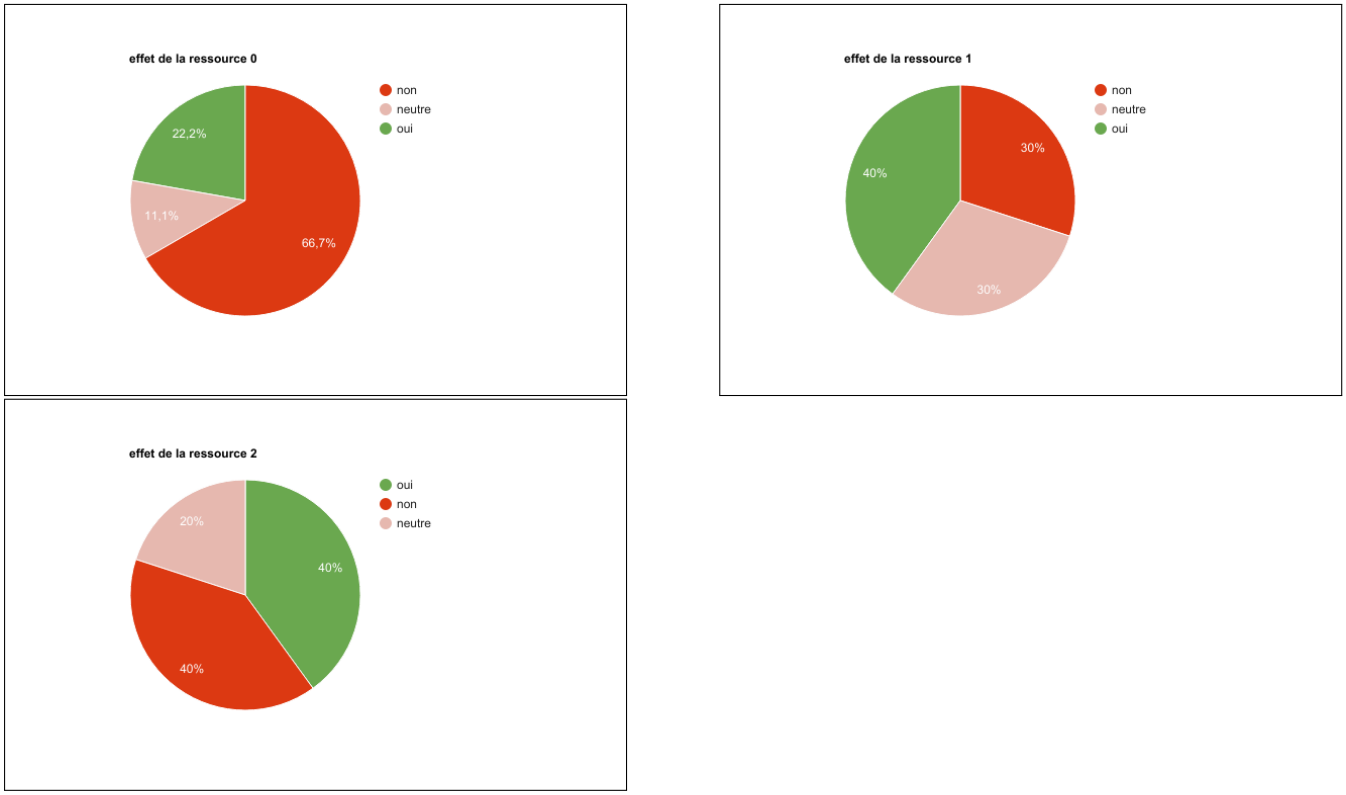

FIGURE 8.5 – effet de l'enrichissement sur le bench3

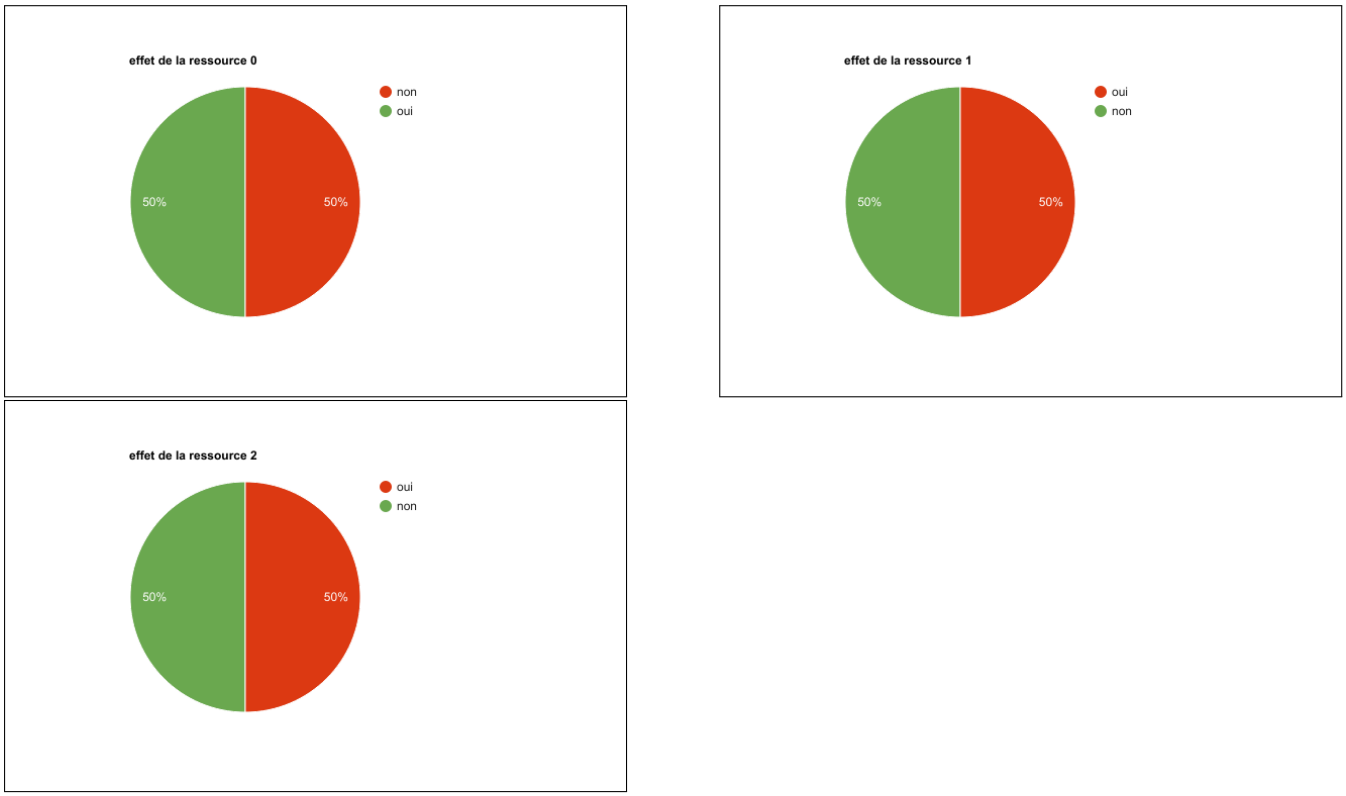

FIGURE 8.6 – effet de l'enrichissement sur le bench4

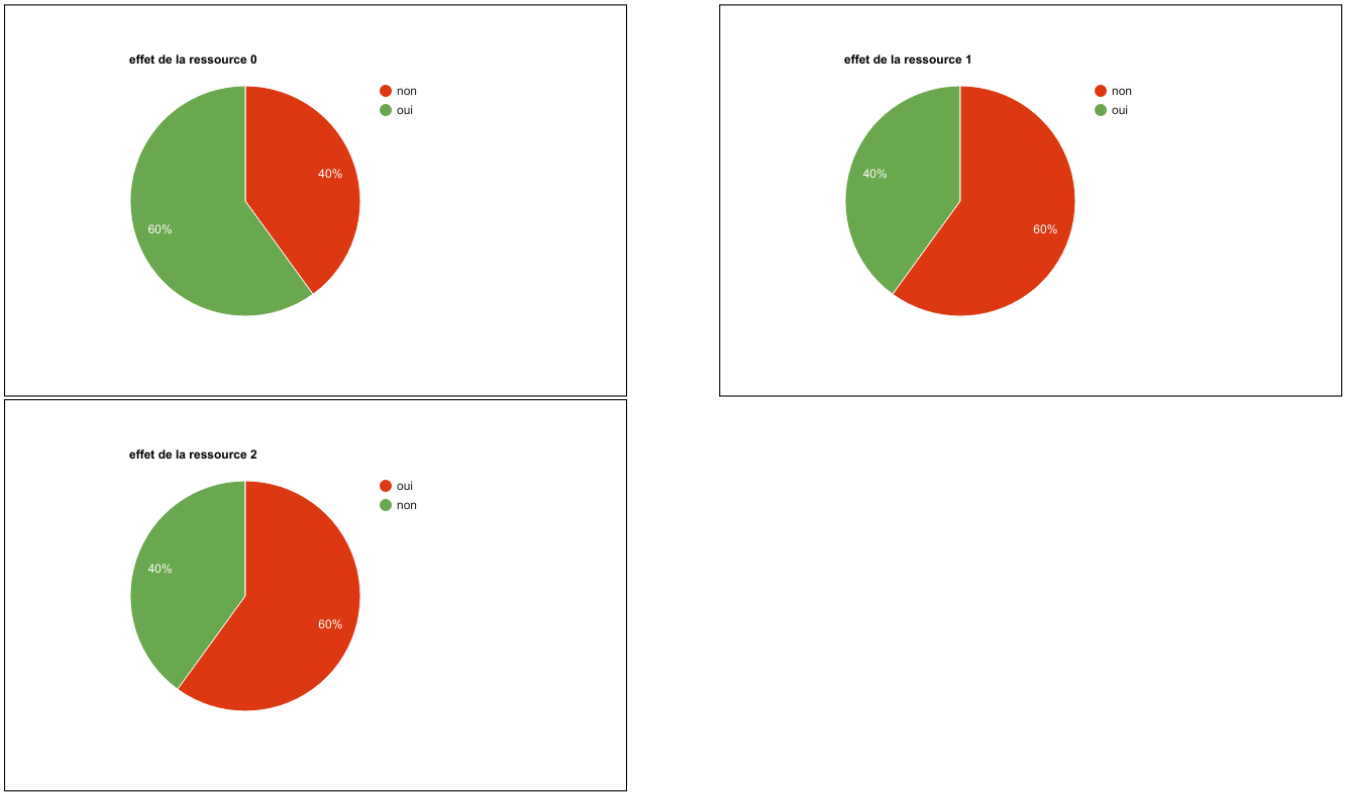

FIGURE 8.7 – effet de l'enrichissement sur le bench5

CHAPITRE 9

CONCLUSIONS GÉNÉRALE

9.1 Conclusion

Cette thèse se positionne dans la construction et l'exploitation des ressources linguistiques en vue d'améliorer la classification (supervisée et non supervisée) des messages courts, des tâches réalisées par l'outil Meeting Software. Cet outil se heurte à deux défis :

— **Défi imposé par la nature de données traitées** : Les données traitées par Meeting Softawre sont des messages courts, des données moins riches en contexte. En plus, les données sont traitées en synchrone en des intervalles de temps irréguliers que nous appelons phases. Les quantités des données traitées pendant les phases sont généralement inégales. Les caractéristiques des données sont resumées par les points suivants :

  — messages très courts : ne contenant pas de forme graphique ou une seule

  — messages contenant des bruits, abréviations, des fautes d'orthographes ;

  — question ou sujet traité contenant un seul ou peu de messages courts ;

  — question ou sujet avec des groupe de messages courts non homogènes. Les groupes peuvent contenir qu'un seul messages court et un autre une disaine. Cela peut dégrader la qualité de la classification supervisée.

— **L'interaction pilote humain et Meeting Software**

  — nombre d'étapes important : on rencontre des jeux de données ayant peu de messages courts mais beaucoup d'étapes correspondantes à des classifications supervisées.

  — le pilote peut à n'importe quelle étape modifier ou créer des groupes de messages courts. La modification d'un groupe consiste à supprimer un groupe ou à déplacer des messages courts de ce groupe.

La principale piste de recherche s'appuie sur la constitution de ressources sémantiques qui permettront de mieux caractériser les messages courts. Ces ressources sémantiques exploitent les données historiques de l'entreprise. Les messages courts traités par Meeting Software sont généralement des données métiers, construire des réssources sémantiques sur des données autres que métiers pourrait dégrader la performance de l'outil. La structure de données de l'entreprise nous a guidé à mettre en place notre approche. Les données sont composées

des questions (ou des sujets) dans lesquelles les messages courts sont régroupés par classes. Notre objectif est de valoriser cettre structure qui resprésente une richesse car validée par nos consultants et nos clients. La plupart des méthodes existantes d'extraction de ressources exploitent des données non structurées (un corpus brut de messages courts). L'approche est complémentaire aux approches qui s'appuient uniquement sur la puissance des calculs statistiques sans prendre en compte les traits sémantiques des unités linguistiques. Elle consiste pour un historique de données de l'entreprise à extraire des classes sémantiques représentées par des concepts. Chaque classe sémantique est composée des motifs (combinaison des formes graphiques) qui sont sémantiquement liés. Quatre processus d'enrichissement ont été mis en place pour enrichir des nouveaux messages courts (voir section 6.3). Pour évaluer l'impact de l'approche sur l'outil Meeting Software un banc de tests a été mis en place (voir chapitre 7)

Les expérimentations réalisées montrent des meilleurs résultats quant il s'agit d'utiliser des ressources construites sur un sujet bien identifié pour enrichir des nouveau messages du mêmes sujet (Expérimentations 1 et 2). Ces expérimentations ont été réalisées sur des jeu de données non dynamiques. Elle consistent à comparer le clustering des jeux de données avant et après l'enrichissement de données. L'expérimentation n°3 est celle qui reproduit le fonctionnement de l'outil Meeting Software. Les jeux de données sont récupérés par phase de traitement (Voir section 7.2) et contiennent les interactions du pilote humain avec l'outil. Pour cette expérience une ressource générique a été mise en place, en faisant abstraction de la thématique des données et des secteurs d'activité. Les résultats sont faiblement positifs lorsque l'on compare les chaînes de traitement enrichies (sauf la chaîne **RMFC**) avec la chaîne de traitement de base.

9.2 Autres contributions

Cette section contient mes contributions sur des sujets connexes à ma thèse qui ont fait l'objet d'une transcription dans le document CIR de l'entreprise.

9.2.1 CBC

La méthode Clustering Based on Clustered data (CBC) est une méthode permettant pour un sujet donné d'entrainer un classifieur sur un historique des données afin d'utiliser le modèle entrainé pour faire effectuer un regroupement non supervisé des nouveaux messages courts du même sujet. Cette méthode est plus adaptée aux données issues de la solution **SucceedData**. Le principe de la méthode est de supposer la redondance d'un certain nombre de thématiques lorsque qu'on effectue une enquête répétitive sur un même sujet.

Cette approche pose des problèmes aux pilotes humains pour qui il est plus simple de vérifier et valider de faible quantité de données. En effet, cette méthode produit pour une masse

de données automatiquement des groupes. Une solution proposée est de mettre en place une méthode de détection des nouvelles thématiques.

9.2.2 Algorithme Baobab

L'objectif de cette méthode est de classer des messages courts dans un ou plusieurs groupes, en se basant sur les motifs émergents (section 2.3). Elle est composée de deux fonctions :

— Apprentissage : l'apprentissage consiste à extraire les motifs émergents de chacun des groupes d'apprentissage.

— Prédiction : pour un message court, le groupe prédit est celui avec lequel il partage le plus de motifs.

Cette méthode à l'avantage de faciliter l'interprétation des résultats de la prédiction, ce qui n'est pas le cas avec d'autres algorithmes de classification supervisée. Cependant, l'algorithme Extratrees utilisé actuellement est plus performant que l'algorithme **Baobab**.

9.3 Perspectives

L'approche proposée pour prendre en compte l'aspect sémantique de messages courts nécessite des tâches manuelles. Ces dernières se situent au niveau de l'étape de la préparation du corpus (section 6.2.1) et l'étape validation de ressources (section 6.2.4). Ces tâches ont été globalement réalisées par deux personnes du pôle R&D. Une mesure de qualité n'a pas été mise en place pour évaluer ces tâhes. Il est important d'utiliser par exemple le coefficient kappa afin de mesurer l'accord entre les personnes réalisant ces tâches. Nous avons vu que l'amélioration est plus nette quand il s'agit des données, d'un sujet bien identifié, spécifiques à des secteurs d'activité. Ce résultat ouvre une perspective d'amélioration de l'outil Meeting Software : les données de l'entreprise peuvent être segmentées en secteur d'activité puis par sujet afin de construire des ressources spécifiques.

Des améliorations peuvent être apportées sur le choix de l'algorithme de classification. Actuellement, c'est l'algorithme extratrees (un algorithme de la famille des forêts aléatoires) qui est utilisée dans la chaîne de traitement. Des études et des publications récentes montrent la suppremacie des réseaux de neurones sur les algorithmes de classification classiques utilisées. Une autre piste d'amélioration est donc d'utiliser les réseaux de neurônes pour la classification supervisée lorque le jeu de données est important. Pour intégrer de la sémantique, on peut faire recours à l'utilisation du modèle word2vec pour la représentation des messages courts.

BIBLIOGRAPHIE

[1] Rakesh AGRAWAL et Ramakrishnan SRIKANT. « Mining sequential patterns ». *in* : *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*. IEEE. 1995, p. 3–14.

[2] Rakesh AGRAWAL, Ramakrishnan SRIKANT et al. « Fast algorithms for mining association rules ». *in* : *Proc. 20th int. conf. very large data bases, VLDB*. T. 1215. 1994, p. 487–499.

[3] Iyad ALAGHA et Rami NAFEE. « An Efficient Approach For Semantically-Enhanced Document Clustering By Using Wikipedia Link Structure ». *in* : *International Journal of Artificial Intelligence & Applications* 5.6 (2014), p. 53.

[4] Khaled ALSABTI, Sanjay RANKA et Vineet SINGH. « An efficient k-means clustering algorithm ». *in* : (1997).

[5] Massih-Reza AMINI. *Apprentissage machine : de la théorie à la pratique*. Editions Eyrolles, 2015.

[6] Ismail BADACHE. « Recherche d'information sociale : exploitation des signaux sociaux pour améliorer la recherche d'information ». Thèse de doct. Université de Toulouse, Université Toulouse III-Paul Sabatier, 2016.

[7] Johan BALTIÉ, SCIA SPECIALISATION et Responsable M ADJAOUTE. « DataMining : ID3 et C4. 5 ». *in* : *Epita SCIA* (2002).

[8] Somnath BANERJEE, Krishnan RAMANATHAN et Ajay GUPTA. « Clustering short texts using wikipedia ». *in* : *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2007, p. 787–788.

[9] Yves BASTIDE. « Data mining : algorithmes par niveau, techniques d'implantation et applications ». Thèse de doct. 2000.

[10] Yves BASTIDE et al. « Pascal : un algorithme d'extraction des motifs fréquents ». *in* : *Techniques et Sciences Informatiquess* 21.*1* (2002), p. 65–95.

[11] Roberto J BAYARDO JR. « Efficiently mining long patterns from databases ». *in* : *ACM Sigmod Record* 27.*2* (1998), p. 85–93.

[12] Nicolas BÉCHET et al. « SDMC : un outil en ligne d'extraction de motifs séquentiels pour la fouille de textes ». *in* : *Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'13)*. 2013.

[13] Nicolas BÉCHET et al. « Sequence mining under multiple constraints ». *in* : *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. ACM. 2015, p. 908–914.

[14] N BELACEL. « Multicriteria classification methods : methodology and medical applications ». Thèse de doct. PhD thesis, Free University of Brussels, Belgium, 1999.

[15] Pavel BERKHIN. « A survey of clustering data mining techniques ». *in* : *Grouping multidimensional data*. Springer, 2006, p. 25–71.

[16] Guillaume BOUCHARD. « Les modèles génératifs en classification supervisée et applications à la catégorisation d'images et à la fiabilité industrielle ». Thèse de doct. Université Joseph-Fourier-Grenoble I, 2005.

[17] Didier BOURIGAULT et Jean CHARLET. « Construction d'un index thématique de l'Ingénierie des connaissances ». *in* : *Conférence ingénierie des connaissances*. 1999, p. 107–118.

[18] Leo BREIMAN. « Random forests ». *in* : *Machine learning* 45.*1* (2001), p. 5–32.

[19] Leo BREIMAN et al. *Classification and regression trees*. CRC press, 1984.

[20] Chris BUCKLEY, James ALLAN et G SALTON. « Automatic retrieval with locality information using SMART ». *in* : *Proceedings of the First Text REtrieval Conference TREC-1*. 1993, p. 59–72.

[21] Gilles CELEUX et Jean DIEBOLT. « Une version de type recuit simulé de l'algorithme EM ». Thèse de doct. INRIA, 1989.

[22] Gilles CELEUX et Gérard GOVAERT. « A classification EM algorithm for clustering and two stochastic versions ». *in* : *Computational statistics & Data analysis* 14.*3* (1992), p. 315–332.

[23] Peggy CELLIER et Thierry CHARNOIS. « Fouille de données séquentielle d'itemsets pour l'apprentissage de patrons linguistiques ». *in* : *17e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*. 2010, 6–pages.

[24] Thierry CHARNOIS et al. « Fouille de données séquentielles pour l'extraction d'information dans les textes ». *in* : *Traitement Automatique des Langues* (2009), pp59–87.

[25] François-Régis CHAUMARTIN. « WordNet et son écosystème : un ensemble de ressources linguistiques de large couverture ». *in* : *Colloque BD lexicales*. 2007.

[26] Corinna CORTES et Vladimir VAPNIK. « Support-vector networks ». *in* : *Machine learning* 20.*3* (1995), p. 273–297.

[27] Zichao DAI, Aixin SUN et Xu-Ying LIU. « Crest : Cluster-based representation enrichment for short text classification ». *in* : *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2013, p. 256–267.

[28] David L DAVIES et Donald W BOULDIN. « A cluster separation measure ». *in* : *IEEE transactions on pattern analysis and machine intelligence 2* (1979), p. 224–227.

[29] Joseph C DUNN. « Well-separated clusters and optimal fuzzy partitions ». *in* : *Journal of cybernetics* 4.*1* (1974), p. 95–104.

[30] Vincent DUQUENNE. « Latticial structures in data analysis ». *in* : *Theoretical Computer Science* 217.*2* (1999), p. 407–436.

[31] Ali EL AKADI. « Contribution à la séllection de variables pertinentes en classification supervisée : Application à la sélection des gènes pour les puces à ADN et des caractéristiques faciales ». *in* : (2012).

[32] Jérôme EUZENAT, Pavel SHVAIKO et al. *Ontology matching*. T. 18. Springer, 2007.

[33] UM FAYYAD et R UTHURUSAMY. « Efficient algorithms for discovering association rules ». *in* : *AAAI Workshop on KDD, Eds*, p. 181–192.

[34] Usama FAYYAD, Gregory PIATETSKY-SHAPIRO et Padhraic SMYTH. « From data mining to knowledge discovery in databases ». *in* : *AI magazine* 17.*3* (1996), p. 37.

[35] Germain FORESTIER. « Connaissances et clustering collaboratif d'objets complexes multisources ». Thèse de doct. Université de Strasbourg, 2010.

[36] Bernhard GANTER et Rudolf WILLE. *Formal concept analysis : mathematical foundations*. Springer Science & Business Media, 2012.

[37] Nizar GHOULA, Gilles FALQUET et Jacques GUYOT. « Modèle d'entrepôt de ressources hétérogènes pour le traitement sémantique des documents ». *in* : *Document numérique* 13.*2* (2010), p. 97–124.

[38] Stéphane GOSSELIN. « Recherche de motifs fréquents dans une base de cartes combinatoires ». Thèse de doct. Université Claude Bernard-Lyon I, 2011.

[39] Thomas R GRUBER. « Toward principles for the design of ontologies used for knowledge sharing ? » *in* : *International journal of human-computer studies* 43.*5-6* (1995), p. 907–928.

[40] Jiawei HAN et al. « DBMiner : A System for Mining Knowledge in Large Relational Databases. » *in* : *KDD*. T. 96. 1996, p. 250–255.

[41] Jiawei HAN et al. « Prefixspan : Mining sequential patterns efficiently by prefix-projected pattern growth ». *in* : *proceedings of the 17th international conference on data engineering*. 2001, p. 215–224.

[42] J HAN et al. « Frequent Pattern Projected Sequential Pattern Mining ». *in* : *Proceedings of international Conference on KDD, Boston (August 2000)*.

[43] Hui HE et al. « Short text feature extraction and clustering for web topic mining ». *in* : *Semantics, Knowledge and Grid, Third International Conference on*. IEEE. 2007, p. 382–385.

[44] Andreas HOTHO, Steffen STAAB et Gerd STUMME. « Ontologies improve text document clustering ». *in* : *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE. 2003, p. 541–544.

[45] Anna HUANG et al. « Clustering documents using a wikipedia-based concept representation ». *in* : *Advances in Knowledge Discovery and Data Mining* (2009), p. 628–636.

[46] Lawrence HUBERT et Phipps ARABIE. « Comparing partitions ». *in* : *Journal of classification* 2.*1* (1985), p. 193–218.

[47] Tapas KANUNGO et al. « An efficient k-means clustering algorithm : Analysis and implementation ». *in* : *IEEE transactions on pattern analysis and machine intelligence* 24.*7* (2002), p. 881–892.

[48] Atanas KIRYAKOV et al. « Semantic annotation, indexing, and retrieval ». *in* : *Web Semantics : Science, Services and Agents on the World Wide Web* 2.*1* (2004), p. 49–79.

[49] Sotiris B KOTSIANTIS, Ioannis D ZAHARAKIS et Panayiotis E PINTELAS. « Machine learning : a review of classification and combining techniques ». *in* : *Artificial Intelligence Review* 26.*3* (2006), p. 159–190.

[50] Siwei LAI et al. « Recurrent Convolutional Neural Networks for Text Classification. » *in* : *AAAI*. T. 333. 2015, p. 2267–2273.

[51] Thi Hoang Diem LE. « Utilisation de ressources externes dans un modèle Bayésien de Recherche d'Information. Application à la recherche d'information multilingue avec UMLS. » Thèse de doct. Université Joseph-Fourier-Grenoble I, 2009.

[52] Christopher D MANNING, Prabhakar RAGHAVAN, Hinrich SCHÜTZE et al. *Introduction to information retrieval*. T. 1. 1. Cambridge university press Cambridge, 2008.

[53] Alice MARASCU. « Extraction de motifs séquentiels dans les flux de données ». Thèse de doct. Université Nice Sophia Antipolis, 2009.

[54] Satoshi MORINAGA et al. « Mining product reputations on the web ». *in* : *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2002, p. 341–349.

[55] Fionn MURTAGH et Pierre LEGENDRE. « Ward's hierarchical clustering method : clustering criterion and agglomerative algorithm ». *in* : *arXiv preprint arXiv :1111.6285* (2011).

[56] Nicolas PASQUIER. « Data mining : algorithmes d'extraction et de réduction des règles d'association dans les bases de données ». Thèse de doct. Université Blaise Pascal-Clermont-Ferrand II, 2000.

[57] Nicolas PASQUIER et al. « Efficient mining of association rules using closed itemset lattices ». *in* : *Information systems* 24.*1* (1999), p. 25–46.

[58] Nicolas PASQUIER et al. « Pruning closed itemset lattices for association rules ». *in* : *BDA'1998 international conference on Advanced Databases*. 1998, p. 177–196.

[59] Gregory PIATETSKY-SHAPIRO. « Knowledge discovery in databases : 10 years after ». *in* : *ACM SIGKDD Explorations Newsletter* 1.*2* (2000), p. 59–61.

[60] Martin F PORTER. « Snowball : A language for stemming algorithms. October 2001 ». *in* : *Retrieved March* 1 (2001), p. 2014.

[61] Solen QUINIOU et al. « What about sequential data mining techniques to identify linguistic patterns for stylistics ? » *in* : *Computational linguistics and intelligent text processing* (2012), p. 166–177.

[62] J Ross QUINLAN et al. « Bagging, boosting, and C4. 5 ». *in* : *AAAI/IAAI, Vol. 1*. 1996, p. 725–730.

[63] Julien RABATEL. « Extraction de motifs contextuels : Enjeux et applications dans les données séquentielles ». *in* : *Theses, Université MontpellierII Sciences et Techniques du Languedoc* (2011).

[64] William M RAND. « Objective criteria for the evaluation of clustering methods ». *in* : *Journal of the American Statistical association* 66.*336* (1971), p. 846–850.

[65] Andrew ROSENBERG et Julia HIRSCHBERG. « V-Measure : A Conditional Entropy-Based External Cluster Evaluation Measure. » *in* : *EMNLP-CoNLL*. T. 7. 2007, p. 410–420.

[66] Peter J ROUSSEEUW. « Silhouettes : a graphical aid to the interpretation and validation of cluster analysis ». *in* : *Journal of computational and applied mathematics* 20 (1987), p. 53–65.

[67] Gerard SALTON. « Automatic text processing : The transformation, analysis, and retrieval of ». *in* : *Reading : Addison-Wesley* (1989).

[68] Gerard SALTON, Anita WONG et Chung-Shu YANG. « A vector space model for automatic indexing ». *in* : *Communications of the ACM* 18.*11* (1975), p. 613–620.

[69] Gilbert SAPORTA. *Probabilités, analyse des données et statistique*. Editions Technip, 2006.

[70] Fabrizio SEBASTIANI. « Machine learning in automated text categorization ». *in* : *ACM computing surveys (CSUR)* 34.*1* (2002), p. 1–47.

[71] Radek SILHAVY et al. *Artificial Intelligence Perspectives and Applications : Proceedings of the 4th Computer Science On-line Conference 2015 (CSOC2015), Vol 1 : Artificial Intelligence Perspectives and Applications*. T. 347. Springer, 2015.

[72] Tony C SMITH et Eibe FRANK. « Introducing machine learning concepts with WEKA ». *in* : *Statistical Genomics : Methods and Protocols* (2016), p. 353–378.

[73] Ge SONG et al. « Short Text Classification : A Survey. » *in* : *Journal of Multimedia* 9.*5* (2014), p. 635–643.

[74] Alexander STREHL et Joydeep GHOSH. « Cluster ensembles—a knowledge reuse framework for combining multiple partitions ». *in* : *Journal of machine learning research* 3.*Dec* (2002), p. 583–617.

[75] Maguelonne TEISSEIRE. « Autour et alentours des motifs séquentiels ». Thèse de doct. Université Montpellier II-Sciences et Techniques du Languedoc, 2007.

[76] Juan-Manuel TORRES-MORENO. *Automatic text summarization*. John Wiley & Sons, 2014.

[77] TM TRAN, Marine TRANCART et Domitille SERVENT. « Littéracie, SMS et troubles spécifiques du langage écrit ». *in* : *Congrès Mondial de Linguistique Française*. EDP Sciences. 2008, p. 168.

[78] Vladimir Naumovich VAPNIK et Vlamimir VAPNIK. *Statistical learning theory*. T. 1. Wiley New York, 1998.

[79] Tingting WEI et al. « A semantic approach for text clustering using WordNet and lexical chains ». *in* : *Expert Systems with Applications* 42.*4* (2015), p. 2264–2275.

[80] SM WEISS et CA KULIKOWSKI. « Computer Systems that Learn Morgan Kaufman Publishers ». *in* : *San Mateo* (1991).

[81] Max WELLING. « Fisher linear discriminant analysis ». *in* : *Department of Computer Science, University of Toronto* 3 (2005), p. 1–4.

[82] Sue Ellen WRIGHT et Gerhard BUDIN. *Handbook of terminology management : application-oriented terminology management*. T. 2. John Benjamins Publishing, 2001.

[83] Cheng-Lin YANG, Nuttakorn BENJAMASUTIN et Yun-Heh CHEN-BURGER. « Mining hidden concepts : Using short text clustering and wikipedia knowledge ». *in* : *Advanced Information Networking and Applications Workshops (WAINA), 2014 28th International Conference on*. IEEE. 2014, p. 675–680.

[84] Ka Yee YEUNG et Walter L RUZZO. « Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data ». *in* : *Bioinformatics* 17.*9* (2001), p. 763–774.

[85] Osmar R ZAIANE. « Principles of knowledge discovery in databases ». *in* : *Department of Computing Science, University of Alberta* (1999).

ANNEXE

9.4 Présentation de Meeting Software (MS)

L'outil central de notre offre, Meeting Software®, vise à collecter, analyser et présenter les idées de chaque participant pendant une réunion. Les réunions sont le plus souvent des séminaires d'entreprise pouvant accueillir comités de direction, managers, ou exécutants. Succeed Together®, afin de permettre une utilisation plus pertinente de Meeting Software®, préconise lors de ces séminaires d'entreprise l'utilisation de temps courts et la présence d'un animateur.

Meeting Software® a la capacité de poser des questions à une assemblée ; les questions qui sont posées doivent suivre un déroulement pédagogique, c'est-à-dire dans un certain ordre permettant une avancée dans le questionnement et dans la problématique posée par l'entreprise. De manière plus fonctionnelle, cela se présente sous la forme d'une suite de questions alternant questions dites fermées (est proposé au participant un ensemble de choix), questions dites ouvertes (le participant répond ce qu'il souhaite avec ses propres mots), et d'autres types de questions (questions imagées, questions n'attendant qu'un seul mot, possibilité de donner un commentaire, etc.).

Le but étant d'animer des séminaires de manière dynamique et pédagogique, tous les participants doivent répondre aux questions et ce sans consensus, ainsi on obtient la pensée de toute l'assemblée.

Le temps donné pour répondre à une question est généralement court afin de rester dans une dynamique forte, de plus l'animateur doit aussi rythmer l'enchaînement des questions.

Les synthèses des résultats des questions sont rendues et remontées en temps réel par le logiciel. Enfin pour rester dans cette logique dynamique et pédagogique, l'animateur commente les résultats devant l'assemblée, cette dernière peut aussi commenter les résultats et un débat peut s'ouvrir.

Pour ce qui est des questions ouvertes, Meeting Software® est capable de regrouper de manière sémantique en contributions de même thème l'ensemble des contributions des participants, quel que soit le nombre de participants. Ainsi il est aisé de faire ressortir les idées principales exprimées par l'assemblée répondante.

Meeting Software® répond donc de manière instantanée à un problème qui aurait pu prendre plusieurs jours pour récolter puis trier l'information.

Seule l'utilisation de Meeting Software® permet de récupérer sur un temps restreint un aussi grand nombre de contributions. Pour une entreprise le temps économisé est important, et donc d'une grande valeur.

Exemple d'utilisation de MS

Un séminaire d'entreprise est organisé, durant lequel les participants sont regroupés par tables de 6 à 8 collaborateurs. Chaque table dispose d'une tablette présentant les produits retenus pour le cas considéré.

Prenons l'exemple d'un séminaire d'une banque de détail comprenant une centaine de participants répartis sur 12 tables de 8 participants. Chaque table contient une tablette reliée à un serveur Meeting Software®. La salle est animée par un facilitateur.

Au début de l'intervention de SGC, une question est envoyée à chaque tablette, précédemment définie en commun par l'entreprise et le chef de projet de Succeed Together suivant le sujet du séminaire, dont voici un exemple :

9.5 Ressource n°1

n°1	clarifier-strategie	confiance-orientation ; prise-conscience ; transition-énergétique ; stratégie-pas ; projet-entreprise ; comme-objectif ; concret ; conquête ; clarté-orientation ; service-public ; projet-industriel ; perspective ; objectif-incohérent ; objectif-parfois ; expliciter ; nouveau-nucléaire ; exploitation ; communication-descendant ; manquer-lisibilité ; enjeux-objectif ; réduire-taille ; stratégie-claire ; transition ; soutien ; communication-projet ; défi ; donner-sen ; situation-financier ; manquer-souffle ; stratégique ; stratégique-entreprise ; avenir-filier ; engagement-initial ; enjeux ; feuille-claire ; santé-entreprise ; planning ; vente ; projet-stratégique ; stratégie ; rapprochement-areva ; ambition ; septicisme ; sen-stratégie ; objectif-réaliste ; incertitude-avenir ; absence-vision ; visibilité-stratégie ; avenir-entreprise ; énergétique ; prévision ; assurer-cohérence ; nucléaire-france ; compréhension ; comme-irréaliste ; visibilité-avenir ; moyen-terme ; objectif-plus ; plan-action ; projet-stratégie ; préciser ; filier-nucléaire ; stabilité-objectif ; transparence ; conscience ; manquer-communication ; filier ; opérationnel ; rachat ; arme-international ; décision-projet ; orientation-entreprise ; vision ; partenariat ; ni-conquête ; orientation-stratégique ; bilan ; démontrer-objectif ; afficher-clairement ; faisabilité-technique ; expliquer-clairement ; exploit ; objectif-projet ; pas-stratégie ; méconnaissance ; visibilité-terme ; rêver-pas ; vente-arme ; orientation ; compétitivité ; destin ; déclinaison ; marcher ; inquiétude ; fixer-objectif ; incohérent ; adhésion ; faisabilité ; cap ; stratégie-entreprise ; malaise-vente ; crise ; explication ; plans-actions ; planning-réaliste ; vision-stratégique ; perspectif ; manquer-vision ; managérialement ; avenir ; atteignable ; clareté ; jalon ; choix-stratégique ; cohérence ; futur ; conquête-manquer ; avenir-nucléaire ; nucléaire-français
n°2	outils-pas-adapte	réactivité-demander ; informatique-logiciel ; dispostion-outil ; siècle ; pas-stable ; outil-pilotage ; obsolète ; outil-fonction ; informatique-concevoir ; outil-retard ; travail-distance ; dépassére ; numérique ; industriel-vieillir ; trop-lourd ; outil-adaptére ; information-spécialiste ; informatique ; documentaire-trop ; outil ; longu ; trop-outil ; ergonomie ; matériel ; outil-standard ; modernité ; logiciel ; informatique-adaptére ; smartphone ; ingénieur ; inefficacité-outil ; irritant-informatique

n°3	encourager-initiative	émotivité-entreprise ; manquer-marge ; idée ; rapidement-hésiter ; initiatif ; inhibant-initiative ; demander-oser ; droit ; echec-plus ; hésiter-pas ; décision-prendre ; initiative-personnel ; valoriser-initiative ; dissuadée-echec ; manquer-innovation ; initiative-collaborateur ; valoriser-initiatif ; idée-nouvel ; soutenir-accompagner ; crainte-pas ; expression-idée ; autonomie-individuel ; priser-initiative ; niveaux-capacité ; environnement-percevoir ; nouvel-valoriséer ; pouvoir-constituer ; capacité-comprendre ; manquer-envie ; conformité-reproduction ; innovation-passer ; livrable-conforme ; personnel-bridée ; libérer-davantage ; temps-innover ; dévalorisant-succès ; proactivité-management ; reconnaitre-personnel ; restreindre ; davantage-innovation ; prendre-risque ; constituer-frein ; nouvel-idée ; zone-confort ; reproduction-norme ; pas-incitation ; proactivité ; pas-inscriter ; cloisonner-personnels ; défavorable-engagement ; inscriter-adn ; libre-proactivité ; davantage-confiance ; priser-risque ; libérer ; favorable-innovation ; encourager ; innover ; trop-restreindre ; oser-pas ; espace-libre ; notion-risque ; entrepreneriat-développer ; statique-défavorable ; favoriser-innovation ; pensée-unique ; hésiter ; start-up ; manque-initiative ; initiative ; conséquence-décision ; innovation-freinée ; procédure-inhibant ; capacité-initiative ; erreur ; bridage-initiatif ; place-initiative ; organisation-statique ; encouragement ; esprit-start ; favoriser-esprit ; besoin-changer ; management-opérationnel ; critique-systématique ; frein-initiative ; manquer-initiative ; risque-entrepreneriat ; échec-résultat ; manquer-créativité ; erreur-notion ; oser ; innovation-domaine ; plus-dévalorisant ; trop-centraliser ; innovation ; risque-expression ; envie-croire ; personnel-innovation ; reconnaissance-initiative ; collaborateur-motiver ; sanction-échec ; contexte-favorable ; passer-droit ; libre-arbitre ; moindre-risque ; initiative-jugéer ; risque-prendre ; créer ; engagement-innovation ; individuel-encouragéer ; peur-sanction ; incitation-prendre ; autonomie-libre
n°4	parcours-professionnel	expatriation ; carriere ; rh ; formation ; recruter ; expertise ; évolution ; contrat ; opportunité ; emploi ; carrière ; embauchére ; accompagnement ; parcours ; départ ; recrutement ; étranger ; compétence ; personnel ; postes ; mobilité

n°5	voloriser-reconnaissance	reconnaître ; valoriser-personnel ; améliorer-reconnaissance ; gratification ; individuel-performance ; progression ; meilleure-reconnaissance ; rétribution ; monter-prime ; reconnaissance-métier ; reconnaissance-travail ; valorisation-performance ; enveloppe-prime ; politique-salariale ; salaire ; récompenser ; individuel-collectif ; projet-difficulté ; féliciter ; rémunération-variable ; reconnaissance-maille ; avantage ; reconnaissance-résultat ; système-rémunération ; pas-valorisation ; favoriser-reconnaissance ; valorisation ; meilleur-valorisation ; manquer-valorisation ; plus-marquéer ; bénéficiaire ; performance-groupe ; reconnaissance-sanction ; mieux-valoriser ; marquer-différence ; évaluation-performance ; reconnaissance-individuel ; remuneration ; système-reconnaissance ; terme-rémunération ; valorisé ; levier-reconnaissance ; reconnaitre ; valorisation-travail ; rémunération-dessous ; rémunéré ; politique-reconnaissance ; reconnaissance-salariale ; résultat-collectif ; égalitaire ; valeur-travail ; différenciée ; pas-reconnaître ; salariale ; prime ; grilles ; rémunération-individuel ; valorisation-personnel ; résultat-équipe ; méritant ; investissement-personnel ; equité ; investissement-individuel ; évaluation ; performance-individuel ; valoriser-succès ; prime-performance ; différenciation ; traitement ; moyen-reconnaissance ; risque-sanction ; rémunération-devoir ; reconnaissance ; levier ; reconnaissance-performance ; salarié ; valoriser ; discriminer ; résultat-individuel ; collectif-individuel ; rémunération-plus ; rénumération ; proposer-prime ; absence-reconnaissance ; équipe-succès ; mieux-reconnaître ; reconnaissance-insuffisant ; rémunération-performance ; politique-rémunération ; valorisation-insuffisant ; différencier ; mérite ; valoriser-réussite ; pas-reconnaissance ; mériter ; équité ; reconnaissance-contribution ; cadeau ; motiver ; reconnaissance-adaptéer ; personnel-contribuer ; pas-suffisamment ; rémunération ; récompense ; donner-prime ; reconnaissance-financier ; effort-individuel ; part-variable ; augmentation ; variable-rémunération ; travail-réaliser ; responsabilité-engagement

n°6	valoriser-collectif	balkanisation ; actions-transverse ; esprit-équipe ; inter-unité ; transverse-assurer ; travail-silo ; transverse-objectif ; matriciel-freiner ; local-pas ; objectif-individuel ; équipe ; silot ; collaboratif ; manquer-esprit ; intelligence-individuel ; travailler-transverse ; fonctionnement-projet ; transverse-pertiner ; travail-collectif ; multisite ; hierarchique-transverse ; projet-equipe ; unité-service ; pluridisciplinaire-responsabilite ; performance-collectif ; persistance-silo ; collectif-pas ; mutualiséer-multisite ; engagement-collectif ; mutualiser ; transverse ; démanteler-esprit ; pas-transversal ; collaboration ; collaborer ; équipe-local ; objectif-commun ; collectif ; transversal ; plusieur-entité ; fonctionnement-silo ; échange ; travail-équipe ; renforcer-collectif ; individualisme ; travailler-ensemble ; plateforme-equipt ; favoriser-échange ; équipe-localement ; supprimer-cloisonnement ; individuel-pas ; transversalité ; cloisonnement-plateforme ; equipt ; cloisonnement-entité ; equipe ; objectif-collectif ; décloisonner ; individuel-objectif ; silo ; cloisonnement ; esprit-collectif ; intelligence-collectif ; pluridisciplinaire ; partage ; coordination ; différent-entité ; esprit ; objectif-contradictoire ; réussite-collectif ; démanteler ; action-mutualiséer ; projet-équipe ; cloisonnée ; trop-cloisonnement ; dépassement-collectif ; commun ; contradictoire-collaborateur ; fédérer ; individualité ; equipe-pluridisciplinaire ; difficulté-travailler ; individualisation ; localement-pas ; mutualiséer ; esprit-territorial ; commun-plusieur
n°7	communication-adapte	compte-idée ; pas-prise ; écouter ; fiche-doléance ; considération-équipe ; écouter-base ; équipe-visite ; insuffisant-avis ; priser-compte ; calage-planning ; projet-écouter ; prise-compte ; écoute-insuffisant ; communication-positif ; écouter-problème ; écouter-mauvais ; considération-agent ; ecoute-métier ; connaître-besoin ; mauvais-écouter ; ecoute ; mauvais-communication ; écouter-niveaux ; compte-remontées ; déconnection-décision ; sentiment-appartenance ; visite-direction ; avis ; manque-écouter ; écouter-métier ; direction-écouter ; participation-décision ; ecoute-actif ; écouter-working ; ecoute-hierarchie ; communiquer ; doléance ; avis-salarié ; remontées-salarié ; compte-avis ; communication ; agent-ecoute ; sentiment-décision

n°8	collectif-comex	venir-étude ; uniformisation-pratique ; implication-haute ; union-exemplarité ; dg-directif ; étude-terrain ; exemplarité-décision ; cohésion-management ; engagement-direction ; écouter-personnalité ; personnel-directement ; solidarité-direction ; comex-organisation ; pouvoir-chapelle ; lutte-interne ; unité-répartition ; toujours-claire ; haute-hiérarchie ; véritable-écouter ; parisien-décision ; comex-insuffisamment ; discours-très ; interne-direction ; besoin-dg ; engagement-managérial ; pratiques-dcn ; traumatisme-message ; diverse-direction ; exemplarité-comex ; sentiment-idée ; conscience-avenir ; jouer-collectif ; cohésion-direction ; communication-corporate ; image-négatif ; critique-dg ; directif-contradictoire ; niveau-comex ; guerre-pouvoir ; ségrégation-diverse ; dg ; aligner-plutôt ; cohérence-objectif ; contradiction-niveau ; extérieur-groupe ; avenir-personnel ; instabilité-lutte ; direction-insuffisament ; cohérence-management ; forte-centralisation ; idée-pas ; conflit-direction ; directement-lier ; tête-dcn ; managérial-niveau ; din-pgr ; trop-confu ; manquer-cohérence ; besoin-uniformisation ; clareté-message ; pas-preuve ; conflit-inter ; manquer-cohésion ; solidaire-aligner ; doute-cohésion ; donner-naissance ; message-entreprise ; identité-site ; compétitivité-parfois ; comex-divergence ; global-prendre ; seulement-tour ; dirigeants-pas ; sentiment-comex ; dcn-donner ; direction-général ; important-visible ; dg-entité ; rivalité-unité ; site-pdg ; réglement-conflit ; activité-pas ; pas-compte ; dcn-union ; objectif-différent ; climat-instabilité ; insuffisamment-collectif ; centralisation-parisien ; cohésion-comex ; pas-toujours ; solidarité-sein ; malsain-divisions ; trop-ségrégation ; trop-global ; conflit ; entité-démotivant ; defiance-sein ; compte-différence ; parfois-malsain ; contradictoire-compétitivité ; direction-programme ; cadrage-venir ; ecoute-existe ; equipe-direction ; naissance-règles ; forte-présenter ; négatif-comex ; inter-divisions ; comex-jouer ; preuve-unicité ; pas-passére ; pratique-besoin ; résultat-entreprise ; uniformisation-pratiques ; passére-niveaux ; message-parfois ; cloisonnement-service ; divisions ; gouvernance ; cadres-dirigeant ; comex-quotidien ; dg-adjoindre ; comex-sentir ; pas-écoutéer ; écouter-direction ; gouvernance-forte ; tension-sein ; sein-dg ; message-vou ; entreprise-pas ; impression-manquer ; message-constance ; discorde-plus ; venir-extérieur ; divergence-important ; programme-envers ; plutôt-guerre ; très-critique ; sentiment-discorde ; sein-comex ; trop-forte ; sentir-quotidien

n°9	meilleure-prioriser	multiplicité-initiatif ; augmenter-disponibilité ; sentiment-anticipation ; dossier-parallèle ; permanent-priorité ; changement-fréquent ; charge-actions ; actions-évolution ; claire-priorité ; priser-décision ; sujet-prioritaire ; parallèle-priorité ; beaucoup-energie ; arbitrage-mieux ; court-terme ; nombreux-dossier ; décision-arbitrage ; mieux-gérer ; priorité-multiple ; absence-gestion ; manquer-arbitrage ; contradictoire-prioritaire ; energie-consommée ; papillonnage-priorité ; plus-priorité ; malgré-priorisation ; capacité-prioriser ; priorisation-actions ; ressources-priorisation ; prioriser-niveau ; arbitrer ; afficher-priorité ; arbitrage-priorité ; priorisation-urgence ; beaucoup-décision ; ni-hiérarchisation ; arbitrage-falloir ; fréquenter-ressources ; changement-permanent ; ambition-parallèle ; fréquent-priorité ; trop-priorité ; prioritaire ; arbitrage-insuffisant ; priorisation ; pas-arbitrage ; prioriser ; priorité ; priorisation-stratégie ; urgence ; impliquére-plusieur ; incapacité-mener ; plusieur-projet ; priorisation-falloir ; arbitrage-ni ; gérer-urgence ; prioriser-renoncer ; trop-projet ; plupart-projet ; manquer-priorisation ; priorité-trop ; disponibilité-collaborateur ; changement-priorité ; priorité-plus ; devoir-prioriser ; parallèle ; travail-priorisation ; trop-actions ; hiérarchisation-priorité ; priser-court ; réaffectation-trop ; turnover-réaffectation
n°10	stabiliser-organisation	changement-organisation ; confusion-organisation ; stabilité-organisation ; organisation ; stabilité ; organiser ; organisation-lourd ; organisationnel ; lisibilité-organisation ; organisationnelle ; complexité-organisation ; organisation-savoir ; organisation-complexe ; organisation-trop ; manquer-stabilité ; organisation-processu ; organisation-lisible

n°11	responsabilite-manager	ligne-managériale ; managerial ; manager-regard ; cascading ; programme ; chef ; parole-direction ; exemplarité-managériale ; dialogue-social ; alignement ; exemplaire ; management-manager ; développer-courage ; former-premier ; pas-manager ; manquer-management ; décision-priser ; hiérarchie ; manquer-courage ; management-très ; manquer-exemplarité ; manquer-prise ; construire-ligne ; direction-trop ; management-objectif ; exemplarité-management ; management-terrain ; responsable ; terrain ; pas-responsable ; partage-information ; confiance-équipe ; ligne ; complexe-management ; maintien-manager ; intermédiaire ; manquer-continuité ; prise-décision ; renforcer ; paix-social ; courage ; comprendre-rôle ; manager ; progrès-courage ; bse-exclusif ; justifier ; comportement-cible ; identifier ; entrainement ; encadrement ; niveau-chef ; écouter-terrain ; présence-managériale ; managérial ; comportement ; exemplarité ; information-équipe ; définir-comportement ; suivi ; managériale-niveau ; niveau-management ; manager-pas ; message-managériaux ; hiérarchie-faciale ; management-montant ; manager-postes ; chef-équipe ; entrainer ; responsable-partage ; autorité ; dimension-managériale ; faciale ; critique ; drastiquement-choix ; manager-fonction ; leader ; respect-parole ; postes-management ; dirigeant ; incapacité ; nommer ; management-trop ; syndicalisme ; insuffisant-management ; managériale ; décider ; proximité ; volonté ; management-intermédiaire ; choix-manager ; exemplarité-manager ; management ; managériaux ; remettre-cause

n°12	respect-regle	régle ; respect-règles ; importer-règles ; pas-appliquére ; contournement-réussir ; zèlée-audit ; parcours-pas ; regle-métier ; responsabilité-établir ; appliquer-règles ; quasi-systématiquement ; régle-métier ; sanction ; cause-personnel ; direction-arbitrage ; décision-comex ; généraliser-toute ; systématiquement-contestéer ; règles-métier ; absence-partage ; évoluer-collaborateur ; respect-décision ; suivre-décision ; trop-importer ; digne-acccompagnement ; nouveau-arrivant ; dcn-nouveau ; fonctionnement-inter ; pas-discipline ; très-formel ; règles-valeur ; toute-organisation ; exécution-optionnelle ; application-régle ; règles-compliance ; règles-insuffisant ; clarté-suivre ; fonctionnement-entreprise ; discipline-plus ; manquer-rigueur ; inconsciemment-exécution ; social-méconnure ; lisibilité-régle ; établir-objectif ; audit-règles ; règles-établier ; règles-appliquer ; rappel-règles ; initiative-encourager ; décision-quasi ; davantage-respect ; formel-pas ; excessif-zèlée ; pas-gpec ; appliquer-référentiel ; respect-engagement ; référentiel-très ; optionnelle-manquer ; partage-culture ; métier-accord ; encourager-contournement ; dense-pas ; pas-respectéer ; manque-respect ; clarté-règles ; connaissance-régle ; règles-contraint ; manquer-clareté ; entreprise-port ; nombre-trop ; organisation-très ; très-dense ; trop-détournée ; application-processu ; contournement-règles ; respect-application ; esprit-remise ; sanctionner ; règle ; méconnaissance-regle ; sanctionner-respect ; arbitrage-mou ; sport-national ; changeant-pragmatique ; inter-direction ; pouvoir-remise ; sentiment-impunité ; remise-cause ; métier-dcn ; respect-engagements ; accord-social ; mauvais-connaissance ; règles-fonctionnement ; règles-manquer ; régle-évoluer ; discipline ; engagements-règles ; respect-objectif ; pas-lisibilité ; règles ; limiter-prise ; national-contournement ; decision-pouvoir ; détournée-pas ; mou-respectére ; pas-rappel ; gpec-digne ; respect-méconnaissance ; engagement-généraliser ; pas-cultiver ; priser-application ; port-epi ; application-excessif ; acccompagnement-parcours ; contraint-limiter ; manquer-engagement ; référentiel ; prise-initiative ; personnel-pas ; manquer-discipline ; règles-changeant ; contournement ; rigueur-discipline ; culture-engagement

n°13	evolution-professionnel	évolution-professionnel ; professionnel-difficile ; replacement-expatriation ; difficile-mettre ; élaborer-salarier ; frein-mobilité ; maîtrise-parcours ; gestion-ressources ; salarier-parcours ; motiver-troupe ; personnalisé-moyen ; nouveau-profil ; mobilité-inter ; écart-salaire ; avenir-professionnel ; adhésion-objectif ; réel-visbilité ; moyens-objectif ; développer-compétence ; attent-personnel ; professionnel-mobilité ; visibilité-évolution ; évolution-terme ; professionnel-pas ; saupoudrage-moyens ; manquer-ressources ; mobilité-étranger ; construire-autant ; politique-rh ; mobilité-afficher ; visibilité-parcours ; nouveau-recrutement ; gestion-carriere ; plans-carrière ; professionnel-personnalisé ; ressources-besoin ; juste-job ; moyens-motiver ; parcours-professionnel ; évolution-carrière ; nouveau-arrivants ; mobilisére-évolution ; favoriser-mobilité ; offre-pas ; rh-développement ; valoriser-expertise ; perspectif-évolution ; manquer-accompagnement ; moyens-formation ; fonction-support ; person-pas ; ambitions-moyens ; privilégie-privéer ; cadres-mobilisére ; professionnel-individuel ; objectif-personnel ; pourquoi-engager ; ressources-homme ; postes-plus ; collaborateur-expérience ; ressources-sujet ; faible-turn ; charte-expatriation ; mobilité-personnel ; équité-côtation ; homme-efficace ; formation-nuire ; monter-capacité ; manquer-équité ; individu-projet ; politique-formation ; manquer-mobilité ; rapport-situation ; visibilité-rh ; utilisation-ressources ; vision-parcours ; départ-retraite ; compétence-jugéer ; personnel-groupe ; activité-individu ; mobilité-difficile ; recrutement-étranger ; vécuer-collaborateur ; moyens-valoriser ; situation-personnel ; manquer-moyens ; carrière-technique ; absence-objectif ; embauche-externe ; cotation-écart ; confort-actuel ; casting-parfois ; efficace-gestion ; véritable-parcours ; côtation-postes ; formation-cohérent ; collaborateur-pourquoi ; développement-collaborateur ; social-homme ; étranger-meilleure ; gestion-mobilité ; expérience-dehors ; responsabilité-similiaire ; proner-parcours ; profil-caste ; compétence-automatisme ; individuel-personnel ; trop-comptable ; valorisation-compétence ; arrivants-ralenti ; afficher-cadre ; retraite-personnel ; contrat-moral ; recruter-prix ; obtenir-recrutement ; accompagnement-prise ; gestion-social ; difficile-construire ; entreprise-person ; pas-accompagnement ; diriger-postes ; réel-compétence ; mobilité-pratiquéer ; ouvrir-embauche ; recrutement-monter ; perte-savoir ; ressources-opérationnelle ; durées-contrat ; compatibilité-moyens ; visbilité-

n°14	simplifier-processus	structure-lourd ; rationaliser ; lourdeur ; dispositif-social ; complexe ; trop-rigide ; excessivement ; processu-lourd ; processu-fonctionnement ; excès ; limiter ; administratif-procédure ; rigidité ; trop-rigidité ; multitude ; mille-feuille ; temps-passer ; simplicité ; alléger ; complexe-projet ; reduire-interface ; complexité-outil ; optimiser ; procédure ; souplesse-organisation ; multiplication ; poursuivre-chantier ; alléger-processu ; lourdeur-règles ; simplifier-organisation ; lourdeur-système ; homogénéiser ; lourdeur-circuit ; trop-pesant ; mode-fonctionnement ; réduire-nombre ; décision-trop ; lenteur-décision ; processu-décision ; mail-perte ; processu-décisionnel ; simplifier ; multiplication-canaux ; administratif-complexité ; lourdeur-référentiel ; flexibilté-processu ; actions-simplification ; souplesse ; efficacité ; complexe-dcn ; plus-complexe ; processu-trop ; démarche-simplification ; proces-outil ; redondance ; réduire ; lourdeur-processu ; lisiblement ; processu ; harmoniser ; plus-rapide ; processu-bm ; optimum ; doublon ; complexité-proces ; compliquére ; simplification ; processu-achat ; procédure-trop ; décision-lourd ; lourdeur-mécanisme ; proces ; compliquer ; simplifier-processu ; simplification-organisation ; frein-efficacité ; contrôle-trop ; centralisation ; administration ; circuit-décision ; rigide ; bureaucratique ; lourdeur-dispositif ; multiplier-processu ; management-projet ; processu-instruction ; excès-processu ; fonctionnement-complexe ; trop-frein ; choc-simplification ; processu-niveau ; réorganisation ; simplicité-lisibilité ; dinosaure ; simplification-processu ; trop-contrôle ; reduire ; lourdeur-complexité ; lourdeur-administratif ; chronphage ; complexité-processu ; pléthorique ; diminuer-nombre

n°15	deleger-responsabiliser	bulle-responsabilité ; conscience-enjeux ; responsabilisation-ensemble ; indépendance-liéer ; délégatio-trop ; déresponsabilisation ; responsabiliser-davantage ; liéer-responsabilité ; delegation-pas ; responsabilisation ; plus-pouvoir ; diluation-responsabilité ; acteurs ; responsabilisation-empilement ; fréquent-déresponsabiliser ; autonome ; délégation ; déresponsabiliser ; responsabilité-diluation ; responsabilité-liéer ; nécessaire-laisser ; responsabilisation-écart ; délégation-responsabilité ; responsabilité-différent ; insuffisance-délégation ; responsabilité-diluéer ; centralisation-décision ; stigmatisation-blessant ; répartition-responsabilité ; plus-autonomie ; champ-responsabilité ; désesponsabilisation-agent ; rôle-responsabilité ; manoeuvre-responsabilisation ; autonomie-équipe ; trop-déresponsabilise ; responsabilisation-faible ; responsabilisation-personnel ; trop-matriciel ; manquer-responsabilité ; manquer-responsabilisation ; mélange-rôle ; favoriser-implication ; chef-projet ; prépondérance-siège ; chef-beaucoup ; parole-confisquéer ; responsable-plutôt ; diluéer-commissionnite ; trop-décision ; déresponsabilisation-acteurs ; subsidiarité ; déléguer ; diluéer-management ; responsabiliser-acteurs ; implication-engagement ; réel-délégation ; dissolution-responsabilité ; dilution-responsibilisation ; niveau-délégation ; temps-déresponsabilisation ; ingénieur-cadre ; replacer-ingénieur ; responsabilité-désintérêt ; diluéer ; pas-responsabilisation ; responsibilisation-acteurs ; down-directeur ; directeur-tyrannique ; management-chef ; responsabilité-unité ; comme-délégation ; déresponsabilisation-dilution ; responsabilité-pas ; responsabilité ; hyper-centralisation ; relatif-désinvolture ; responsabilité-opposition ; autonomie-site ; responsabilité-morceléer ; insuffisant-contribution ; dilution ; rôle ; arbitrage ; défaut-responsabilisation ; manquer-delegation ; role-responsabilité ; manquer-dialogue ; intestin-unité ; relle-delegation ; déréssponsabilisation-contributeur ; trouver-responsable ; désinvolture-personnel ; responsabiliser-agent ; trop-chef ; peu-autonomie ; morcellement-responsabilité ; droit-erreur ; marge-manoeuvre ; responsabilisation-individuel ; delegation ; responsabilité-définir ; dilution-responsabilité ; responsabiliser-plus ; responsabilisation-objectif ; responsabiliser ; autonomie ; manque-autonomie

9.6 Ressource n°2

n°1	outils-pas-convivial	retour ; poste-travail ; facilement ; intranet-pas ; guichet ; ergonomique ; manquer-convivialité ; outil-extrassur ; mauvais ; page-accueil ; disponible ; pas-logique ; ergonomie-outil ; gagner ; pratique-pas ; trop-lien ; ascenseur ; iard ; permettre-pas ; fonctionnalité ; menu ; acceder ; agréable ; intranet-manquer ; ergonomie-poste ; clair ; entretien ; fiscalité ; proces ; raccourci ; pas-ergonomique ; ergonomie-pas ; ordre ; permettre ; lourd-peu ; caractère ; pas-accessible ; choses ; partenaire-intranet ; plus-intuitif ; extrasur-pas ; très-ergonomique ; consultation-fiscalité ; clarté ; non-intuitif ; illogique ; intuitif-outil ; amélioration ; conviviaux ; très-intuitif ; droite ; illisible ; manquer-clarté ; option ; code-popix ; plusieur-clic ; globalement ; équinoxe-pas ; visibilité ; tâche ; logique ; insatisfaction ; rubrique ; bas ; convivial-pas ; vision ; très-peu ; portail-equinoxe ; extrasur ; affichage ; insatisfaction-outil ; comparativement ; comme-extrassur ; interface ; manquer-ergonomie ; extrassur-peu ; plus-simple ; forcément-intuitif ; portail ; intuitif ; equinoxe-manquer ; rapide ; télécommande ; traiter ; métier ; ergonomie ; equinoxe-outil ; couleur ; conseiller ; obtenir ; pas-intuitif ; environnement-travail ; pas-plus ; convivialité ; page ; logique-classement ; accessible ; accueil ; mauvais-ergonomie ; ordre-consultation ; falloir-faire ; extrassur-pas ; remplir ; pas-pratique ; cheminement ; consultation ; faciliter ; peu-ergonomique ; nouvel ; peu-intuitif ; prise ; rester ; intuitivité
n°2	système-compliqué	univers-client ; beaucoup-outil ; lourdeur ; devenir ; lourd ; souvent-trop ; trop-complexe ; évolution ; miser ; utilisation ; univers ; simplicité ; modifier ; compliquer ; pénible ; pas-trouver ; parfois-difficile ; très-lourd ; informatique-pas ; difficile-utilisation ; mail ; trop-lourd ; également ; faire-simple ; difficile-trouver ; comprendre ; mise ; pas-facile ; effectuer ; trop-compliquer ; plus-complexe ; alpha ; environnement-pas ; manquer-simplicité ; pas-convivial ; simple ; client-trop ; parfois ; assez-intuitif ; difficile ; complexe ; cisco ; carte ; opération ; tel ; connaître
n°3	manque-formation	fonctionnement-aléatoire ; fonctionnement ; aucune ; aucune-formation ; formation ; outil-intranet ; aléatoire ; erreur ; intranet ; manquer-formation ; manquer

n°4	trop-disfonctionnement	déjà ; stres ; trop-plantage ; anomalie ; revenir ; suite ; souvent-indisponible ; redémarrer ; erreur-grave ; poste ; redémarrage ; récurrent ; obliger-relancer ; beaucoup-perturbation ; popix-très ; technique ; nombreux-bug ; trop-souvent ; beug ; environnement-equinoxe ; bloquer ; nou-pas ; appel ; indisponibilité ; cas ; pas-journée ; accé-information ; minimum ; problème-récurrent ; systématique ; ouverture-poste ; souvent ; plant ; très-régulièrement ; heure ; souci ; souvant ; nombreux ; recurrent ; dysfonctionnement-bug ; reseau ; pas-opérationnel ; jour-problème ; compter ; partenaire-souvent ; pan ; cesse ; problème-connexion ; réseau ; logicile ; midi ; instabilité ; place ; lancer ; fonctionnel ; message ; perturbation ; déconnecter ; manipulation ; souvent-bug ; apre ; deconnexion ; instabilité-système ; jour ; perte-donnée ; bloquére ; déconnexion ; nombreux-pan ; postes ; dernier ; équinoxe ; grave ; obligére ; equinoxe-extrassur ; fermeture ; arret ; opération-nel ; système-plant ; plantage ; semaine ; planter ; redemarrage ; extrassur-souvent ; -cour ; répétition ; impossibilité ; monde ; plant-tous ; ordinateur ; pc-trop ; continuellement ; nombreux-dysfonctionnement ; souvent-relancer ; dysfonctionnement ; rame ; exemple-matin ; seulement ; plusieur-fois ; pb ; bug-popix ; souvent-panne ; beaucoup-problème ; plantage-informatique ; diverse ; plantage-trop ; niveau-pc ; fonctionner-pas ; vou ; relancer ; raison ; informatique-plant ; bug-informatique ; rallumer ; apre-midi ; coup ; beaucoup-bug ; regulier ; arret-intempestif ; correctement ; fois-semaine ; disfonctionnement ; trop-problème ; durer ; poste-régulièrement ; journée ; quotidiennement ; beaucoup-plantage ; coupure ; dysfonctionnement-quotidien ; trop-bug ; régulier ; régulièrement ; panne ; erreur-equinoxe ; journalier ; très-souvent ; moins-fois ; manque ; blocage ; maintenance ; popix-plant ; intempestif ; général ; bug ; démarrer ; popix-faire ; message-erreur ; ferme ; répondre ; nombreux-plantage ; plant-très ; plant-régulièrement ; arrêter ; tous-jour ; test ; probleme ; fois-jour ; transfert ; rib ; génant
n°5	problème-imprimantes	impression ; imprimant ; problème ; très

n°6	base-document-incomplet	information-mal ; difficulté-retrouver ; efficace ; recherche-non ; outil-devoir ; falloir-aller ; insuffisant ; notes ; très-compliquer ; rarement ; trouver-pas ; obsolete ; version ; performant ; recherche-pas ; recherche-mot-clé ; tré ; donner ; bon ; documentaire-devoir ; arborescence ; fournier ; entreprise ; faire-mieux ; documentaire-moteur ; doc-moteur ; très-performant ; recherche-mot ; difficile-rechercher ; type ; tuer-info ; concerner-base ; complexité ; beaucoup-information ; document-pas ; azurcom ; information-pas ; inaccessible ; fonds ; organiséer ; rechercher-document ; recherchéer ; moteur ; documentaire-notes ; temps-chercher ; claire ; documentaire ; simplifier ; retrouver-information ; peu-performant ; moteur-recherche ; nom ; assez-fiable ; impossible ; intranet-base ; peu-efficace ; fonds-doc ; pas-efficace ; tuer ; tre-intuitif ; fond-doc ; toujours-simple ; incomplète ; documentaire-recherche ; intranet-moteur ; pas-jour ; outil-recherche ; marcher ; contenu ; très-contraindre ; trop-info ; base ; impossible-retrouver ; devoir-faire ; trop-information ; avis ; tenir ; recherche-fondoc ; recherches-fondoc ; recherche-très ; directement ; marcher-pas ; hui ; documentaire-très ; info-mal ; mal-organiséer ; très-loin ; trouver-chemin ; rechercher ; toujours-pas ; pas-toujours ; totalement ; renseignement ; recherche-info ; relever ; compliquéer ; fond ; toujours ; quel ; trop-document ; venir-plus ; revoir ; clef ; intranet-trop ; recherche-information ; recherche ; commande ; fonction-recherche ; point ; recherche-totalement ; très-bien ; souvent-périmée ; aller-extrassur ; info-base ; implicit ; aller-chercher ; diffcile ; difficulté-trouver ; collaborateur ; fondoc ; jour-moteur ; information-recherchéer ; périmée ; document ; fastidieux ; non-pertiner ; mots-clé ; base-documentaire ; mot-clef ; documentaire-peu ; toujours-jour ; inefficace ; recherche-base ; très-difficile ; note ; chercher-trouver ; mot-clé ; pertinent ; bon-information ; pertiner ; contraindre ; google ; falloir-toujours ; recherche-intranet ; information-peu ; simplifiéer ; point-négatif ; matricule ; évident ; doc-pas ; info-tuer ; documentaire-inexploitable ; inexploitable ; documentaire-pas ; outil-plus ; actualité ; notes-obsolète ; loin ; trouver-information ; négatif ; recherche-peu ; question ; plus-actualité ; préci ; trouver-info ; venir ; info ; mot-clé ; trouver ; recherche-equinoxe ; améliorer ; clé ; fondoc-moteur ; recherches ; trop-temps
n°7	outil-pas-adapté	service-production ; fonction ; equinoxe-pas ; outil-adapter ; equinoxe ; pas-adapter ; mal ; adapter-service ; service ; siège ; sembler ; adapter ; production ; environnement ; mettre ; transaction ; adaptére ; compte ; productivité ; perte-temps

n°8	quand-outil-unique	difficile-retrouver ; lisa ; systeme-information ; falloir ; connaissance ; domaine ; homogénéité ; gérer ; sharepoint ; progrès ; aucune-homogénéité ; vite ; partenaire-trop ; extrassur-platine ; intégration ; platine-sesame ; outil-différent ; trop-outil ; applicatif-autre ; meme ; utilisére ; permanence ; efficacité ; reliére ; regrouper ; partenaire-code ; passer-applicatif ; trop-application ; différent-applicatif ; outil-partenaire ; double ; bien-mieux ; trop-dispersére ; diver ; toujours-facile ; centraliser ; détail ; lien-pas ; meilleure ; peu-lien ; sesame ; equinoxe-bien ; trop-applicatif ; utile ; outil-aucune ; grand ; info-clients ; souscription ; source ; meme-endroit ; dispersére ; passerelle ; lien ; ligne ; endroit ; partenaire ; seul ; vcc ; trop-onglet
n°9	écran-trop-petit	non ; trop-petit ; ecran ; écran ; ecran-trop ; petit
n°10	moyen-age	outil-informatique ; rapport ; procédure ; ancien ; vieillir ; encore ; ram ; obsolète ; retard-rapport ; rapport-concurrent ; nou-sommer ; retard ; sommer ; modernité ; pc ; travailler ; niveau ; concurrent
n°11	resolution-incident	non-résoudre ; pouvoir ; ibp ; résolution ; faire ; réponse ; fréquent ; cloture ; prendre ; cloture-incident ; déclaration-incident ; trop-fréquent ; incident ; devoir ; déclaration ; client ; résoudre ; incident-trop ; déclarer ; jamais
n°12	outil-trop-lent	moin ; malheureusement ; trop-clic ; relancer-session ; equinoxe-trop ; lenteur-système ; transaction-équinoxe ; plant-pas ; portable ; réponse-trop ; temps-rapport ; nou-permettre ; lent-bug ; transaction-popix ; plus-longu ; problème-lenteur ; mois ; popix-beaucoup ; messagerie-outlook ; trop-lent ; clic-trop ; ralentir ; bugg ; office ; trop-longu ; plus-plus ; récurrer ; vouloir ; nou-faire ; equinoxe-lent ; plus-lent ; beaucoup-lenteur ; inexistant ; perdre-beaucoup ; ralentissement ; lenteur-bug ; faire-perdre ; réponse-long ; très-lent ; reponse ; longu ; clic ; toujours-lent ; lenteur-systeme ; session ; informatique-trop ; rendre ; lenteur-réponse ; lent-plant ; equinox ; bureautique ; trop-lenteur ; outil-bureautique ; secondes ; mum ; souvent-lent ; temps-réponse ; tré-lent ; caisse ; espace ; particulier ; beaucoup-plus ; tre-lent ; beaucoup-trop ; pc-portable ; encore-plus ; temps-attente ; informatique-lenteur ; rapidité ; perdre-temps ; conséquence ; outil-lent ; attente ; délai ; outil-beaucoup ; lent-trop ; lenteur ; lenteur-informatique ; lent-nou ; lent-peu ; très-long ; falloir-temps ; pages ; lenteur-logiciel ; saisir ; informatique-lent ; minutes ; systeme ; lent-souvent ; excel ; gro ; lenteur-navigation ; lent-pas ; suppression ; lent ; gro-problème

n°13	gestion-mots-passe-lourde	code ; connecté ; changer ; autant ; identifiant-différent ; accès-différent ; date ; environnement-différent ; changement ; apogee ; identifier ; créer ; mdp ; code-accès ; début ; trop-code ; différent-bpatl ; acce ; passer ; passer-equinoxe ; authentifier ; éviter ; favorable ; produit ; accé ; deux ; connecter ; multiple ; identifiant-mot ; accéder-différent ; tous-mot ; habilitation ; différent ; accès ; gestion ; multitude-mot ; rapidement ; retenir ; tous-outil ; site ; passer-multiple ; nécessité ; limiter ; platine ; falloir-mot ; moment ; outil-trop ; outil-nécessiter ; mot ; trop-mot ; pourquoi-pas ; différent-fois ; gestion-mot ; bpatl ; limitéer ; log ; commun ; identifiant ; different ; mot-passer ; pourquoi ; passer-différent ; pas-equinoxe ; oublier ; nécessiter ; passer-outil ; différent-logiciel ; multiplicité

9.7 Ressource n°3

n°1	thema-compétitivité	projet-local ; meilleur-rémuner ; revoir-tarif ; tarif-bancair ; cart-mozaic ; attract ; spécif-jeun ; spécif ; plus-competit ; sociétair-jeun ; avantag-mozaic ; plus-soupless ; plus-attract ; offre-assur ; offre-adapté ; specif ; offre-jeun ; promot ; réduct ; rémuner-attract ; avantag-jeun ; préférentiel ; pas-cher ; offre-promotionnel ; rémuner-avantag ; tarifair-produit ; taux-cred ; tarifair ; offre-spécif ; cart-gratuit ; cré-avantag ; plus-interess ; rapport-qualit ; livret ; avantag-spécif ; offertr ; qualit-prix ; taux ; propos-offre ; dévelopd-offre ; offre-plus ; sociétair-offre ; don-avantag ; recom-pens ; term-prix ; reduct ; offre ; prix-servic ; tarif-préférentiel ; avantag-fidelit ; avantag-tarifair ; procur-avantag ; avantag-tarif ; promotionnel ; prix-attract ; prix ; plac-cart ; plus-avantag ; souscript-part ; tarif-march ; offre-compétit ; plac-cin ; interress ; avantag-societair ; avantag-financi ; tarif ; sociétair ; competit ; cart-jeun ; plac-avantag ; offre-specif ; jeun-sociétair ; offrir ; offre-tarifair ; deven-sociétair ; avantag-client ; plus-offre ; avantag
n°2	thema-organisation	stabl-organis ; organis-travail ; spécialis-sit ; organigramm ; coor-din ; reorganis ; entrepris-pouvoir ; strateg
n°3	thema-responsabilités	chef ; augment-deleg ; rôl-mission ; plus-pouvoir ; augment-déleg ; responsabilis ; responsabilit ; déleg-pouvoir ; respons ; indépend ; holding ; manqu-déleg ; respons-dilué ; dilut-respons ; pas-responsabilis ; deleg ; autonom ; pouvoir-inspecteur ; redon-pouvoir ; deleg-agent ; plus-déleg ; manqu-responsabilis ; inspecteur-souscripteur ; plus-deleg ; mission-direct ; déleg ; responsabilis-individuel
n°4	thema-investissement	budget-mba ; fond ; budget ; mba-plus ; coût ; plus-import ; moyen
n°5	thema-RAS	action-cour ; jug-optimis ; répondr-question ; pas-concerné ; pas-répondr ; action-mené ; pas-sujet ; pas-idé ; pas-connaiss ; optimis-action ; mené-pouvoir ; inform-concern ; cour-pas ; savoir-pas ; concern-pas ; pas-pouvoir ; détail-action ; pertin-jug ; pas-pertin ; connaiss-action ; pas-repondr ; con-pas ; connaîtr-pas ; sujet-pas ; pas-détail
n°6	thema-cohesion-equipe	travail-équip ; cohes-équip ; confianc-agent ; climat ; plus-confianc ; confianc-aven ; climat-confianc ; cohes ; confianc-réciproqu ; confianc-réseau ; object-commun ; consider-comm ; profil ; réseau-agent ; heureux ; solidarit ; cohes-equip ; confianc-compagn ; confianc ; action-cohes ; communaut ; projet-commun ; intérêt-commun ; cultur-excellent ; reciproqu ; rassembl ; relat-confianc ; bienveil-reciproqu ; esprit-équip ; travail-collect

n°7	thema-reactivite	engag-del ; réclam ; réactiv-demand ; besoin-client ; réactiv-rappel ; répons ; réactiv ; del-repons ; flexibl ; respect-del ; client-vouloir ; demand-réclam ; engag-rappel ; trait-demand ; rappel-client ; réactiv-engag ; fil-attent ; suit-demand ; rapid-pris ; quelqu-canal ; réactiv-souscripteur ; demand-client ; répondr-demand ; attent-client ; agil ; raccourc-del ; repondr-rapid ; reactivit ; rapidit ; pouvoir-répondr ; repons ; répons-client ; répons-mail ; toujour-plus ; répondr-attent ; répondr-rapid ; del-répons ; execu ; répons-rapid ; réactiv-répons ; tres-réactif ; question-client ; propos-solut ; pris-charg ; demand-rapid ; répondr-client ; del ; court ; reactif-demand ; repons-rapid ; reactif ; reactiv ; client-suit ; immediat ; mail-client ; amélior-réactiv ; répondr-besoin ; reactivit-compagn ; anticip-besoin ; charg-demand ; demand ; bon-moment ; plus-reactivit ; disponibl-client ; raccourc ; demand-répons ; plus-réactiv ; plus-réactif ; apport-répons ; réactif-demand ; réag ; solut-rapid ; plus-rapid ; del-trait ; demand-mail ; reclam ; rapid ; disponibilit ; réduir-del ; flexibil ; rapid-demand ; trait-rapid
n°8	thema-positif	bon-présent ; bon-orient ; bon ; bon-sen ; posit ; espoir ; enthousiasm ; optim ; intéress ; encourag ; interess
n°9	thema-evenementiel	deven-societair ; organis-journ ; even ; moment-convivial ; concour ; forum ; provoqu-rencontr ; écol-univers ; rencontr-réguli ; tabl-rond ; client-réunion ; anim ; don-plac ; expo ; cré-éven ; invit-réunion ; organis-forum ; fest ; sein-agenc ; client-prospect ; visit-sit ; manifest ; plus-rencontr ; festif ; forum-jeun ; rencontr-échang ; port-ouvert ; jeun-particip ; projet-jeun ; ven-client ; dédi-jeun ; them-ag ; rendr-ag ; client-agenc ; them-dédi ; rencontr-annuel ; réunion-caiss ; prospect ; ag-ven ; sport ; lot-gagn ; them-plus ; sportiv ; sportif ; club-sportif ; visit-client ; partenariat-écol ; organis-rencontr ; associ-sportif ; jeun-ag ; attract-jeun ; plac-rencontr ; organis-éven ; plus-present ; niveau-associ ; rencontr-convivial ; invit-client ; anim-ponctuel ; foot ; musiqu ; jeun-associ ; conseil-municipal ; anim-local ; sociétair-organis ; inform-them ; reunion-jeun ; organis-soir ; organis-tabl ; rencontr-them ; développ-présenc ; organis-manifest ; associ-jeun ; développ-partenariat ; réunion-them ; nouveau-sociétair ; local-jeun ; sportif-culturel ; pres-manifest ; rencontr-plus ; reunion-them ; them-client ; petit-déjeun ; cré-conseil ; général-jeun ; ag-jeun ; assemble-general ; rencontr-jeun ; comit-jeun ; davantag-rencontr ; jeun-actif ; partenariat-associ ; plus-fréquent ; administr-jeun ; ag-plus ; organis-port ; invit-jeun

n°10	thema-qualite-produit	conso ; valeur-mobili ; march-boursi ; gamm ; gestion-valeur ; der-epargn ; connaiss-march ; titr-bours ; prépar-retrait ; banqu-priv ; cred-habitat ; valeur-mobilier ; gestion-patrimoin ; diversif-der
n°11	thema-bureaucratie	pert-efficac ; uniformis ; action-simplif ; différent-entit ; organis-processu ; centralis ; processu-lourd ; chang-organis ; reporting-intern ; tertiair-diffu ; lourdeur-complex ; pyramidal ; cloison-servic ; décisionnair ; processu-trop ; plus-complex ; lourdeur-processu ; décisionnel ; chef-projet ; orga ; simplif-accept ; mod-fonction ; proces-souscript ; decisionnel ; lourdeur-administr ; poursuivr-chanti ; contrôl-trop ; fluidifi-process ; fluidifi ; schem ; souscript-plus ; démarch-simplif ; lourdeur-system ; respons-chacun ; processu-niveau ; projet-unit ; homogénéis ; drastiqu ; hierarch ; simplifi-processu ; manqu-lisibil ; procédur ; reduir-interfac ; réduir-nombr ; processu-décis ; plus-clair ; amélior-processu ; décis-arbitrag ; lisibil-organis ; processus-décis ; soupless-organis ; mail-pert ; plateau-projet ; processu-intern ; complex-processu ; processu-fonction ; jusqu-niveau ; fonction-central ; manag-projet ; processu-associ ; circuit-décis ; nouvel-organis ; choc-simplif ; organis-complex ; nombr-réunion ; processu ; niveau-national ; trop-complex ; organis-fonction ; pris-décis ; pas-simplifi ; workflow-administr ; comit-décis ; complex-projet ; décentralis ; décis-projet ; complex-outil ; organis-interfac ; fonction-entrepris ; organis-trop ; pert-sen ; amélior-proc ; procedur ; simplif-proc ; complex-organis ; projet-méti ; proces ; alleg-processu ; simplifi-organis ; décis-plus ; simplifi-gestion ; canal-inform ; niveau-hiérarch ; interfac-inter ; gestion-agenc ; bureaucrat ; success ; dipnn-unit ; trop-contrôl ; organis-projet ; définit-rôl ; alleg-gestion ; mill-feuill ; activ-pas ; redond ; décis-trop ; tertiair ; process ; temp-pass ; pas-toujour ; simplif-proces ; lign-hiérarch ; décis-pas ; workflow ; réduir-temp ; organisationnel ; simplif-processu ; procédur-intern ; lenteur ; processu-achat ; laiss-marg ; organis-lourd ; stabilis-organis ; rationalis ; simplifi-procédur ; simplifi-proc ; simplif-organis ; proc-décis

n°12	thema-valorisation	meilleur-valoris ; prim-rémuner ; ceu-engag ; prim-perform ; trop-égalitair ; égalitair ; object-individuel ; différenci-agent ; permettr-pas ; responsabilis-acteur ; rémuner-devoir ; responsabilis-plus ; form-manag ; reconnaîtr ; valoris ; découpag-selon ; valoris-perform ; assum ; plac-polit ; reconnaissance ; don-prim ; mesur-perform ; parcour-carri ; succes-équip ; favoris-reconnaiss ; amélior-reconnaiss ; plus-transparent ; reconnaiss-maill ; envelopp-prim ; don-plus ; augment-envelopp ; revoir-system ; levi-rémuner ; réel-déleg ; plus-marqué ; invest-perform ; mettr-progress ; valoris-action ; impliqu-salari ; individuel-perform ; valoris-résultat ; perform-moyen ; résultat-obtenur ; objectiv-reconnaiss ; déleg-responsabilis ; object-annuel ; cadr-rémuner ; conscienc ; marqu-différent ; rémuner-plus ; anciennet ; fonction-perform ; résultat-équip ; rénumer ; maill-équip ; rémuner-reconnaiss ; reconnaiss-invest ; invest-personnel ; perform-collect ; valoris-activ ; acteur-niveau ; reconnaitr-perform ; reconnaiss-salarial ; sanction ; reconnaiss-financi ; rémuner-individuel ; variabl-rémuner ; term-rémuner ; polit-reconnaiss ; reconnaiss-engag ; individuel-prim ; valoris-individuel ; atteint-object ; moyen-reconnaiss ; mont-prim ; contribu-individuel ; fonction-atteint ; part-variabl ; concret-local ; augment-autonom ; carri-moyen ; valoris-succes ; prim-plus ; supprim-system ; associ-résultat ; perform-reconnaiss ; rémuner-trop ; évalu-perform ; vi-diplom ; collect-maill ; levi-reconnaiss ; main-manag ; réduir-écart ; atteint-résultat ; part-individuel ; reconnaiss ; agent-plus ; meilleur-reconnaiss ; responsabilis-agent ; écart-salair ; rémuner-perform ; résultat-collect ; mieux-reconnaîtr ; polit-salarial ; plus-gen ; résultat-individuel ; rémuner-variabl ; contribu-collect ; tourn-perform ; salarial-trop ; polit-valoris
n°13	thema-experience-client	rénov-agenc ; satisfact-client ; pai-téléphon ; pai-mobil ; appliqu-smartphon ; internet-mobil ; pai-cart ; mod-pai ; telephon-portabl ; continu-rénov ; nouveau-mod ; proposit-assur ; pai-smartphon ; dévelopg-pai ; dévelopg-nouveau ; pai-telephon

n°14	thema-nouvelles-technologies	développ-dématérialis ; adapt-nouvel ; impress-recto ; client-pouvoir ; photocopi ; don-possibil ; smartphon ; mobil-client ; noir-blanc ; part-soir ; appareil-veil ; envoi-docu ; dématérialis-contrat ; signatur-électron ; ordin-soir ; piratag ; direct-mail ; lumi-bureau ; imprim-mail ; mis-niveau ; mid-soir ; réalis-contrat ; empreint ; évit-envoi ; fraud-internet ; objet-connect ; valid-réalis ; viseo-mail ; appliqu-mobil ; system-inform ; lier-interact ; client-évolu ; amelior-outil ; téléphon-portabl ; avanc-nouvel ; ordi ; écran-soir ; susceptibl-déménag ; tout-impress ; point-nouvel ; don-client ; not-tach ; collect-final ; dematerialis-total ; photo-cop ; imprim-docu ; sécuris-simplifi ; final-lier ; evit-impress ; développ-nouvel ; des-quitt ; mod-commun ; impress-docu ; présent-compt ; ecran-pc ; appareil ; pas-imprim ; écran ; adapt-évolu ; limit-impress ; viséo ; suivr-évolu ; imprim-systémat ; savoir-adapt ; eteindr-ecran ; outlook ; signatur-electron ; suivr-evolu ; éteindr-imprim ; googl ; lumi-appareil ; dépos-bam ; client-collect ; person-présent ; imprim-pas ; appliqu-informat ; eteindr-lumi ; pc-soir ; vid ; imprim ; scan ; évit-fraud ; pas-laiss ; ecran-soir ; viséo-multicanal ; evit-imprim ; mail-courri ; relat-viséo ; utilis-tablet ; voir-indispens ; interact-parf ; oper-bancair ; impress-papi ; utilis-outil ; evolu-technolog ; outil-technolog ; oper-distanc ; adress-mail ; éteindr-écran ; impress-mail ; mul-timédi ; scann-docu ; outlook-evit ; laiss-veil ; connexion ; évit-impress ; pc ; appli-smartphon ; point-technolog ; eteindr-écran ; tablet ; client-utilis ; outil-exist ; consomm-papi ; plutôt-papi ; technolog-méti ; util-voir ; ordin-nuit ; évit-imprim ; sécur-oper ; pas-prendr ; appli ; soir-part ; impress ; skyp ; nouveau-moyen ; adapt-rapid ; visio-conférent ; quitt-bureau ; envoi-mail ; évit-édit ; mati-nouvel ; tablet-soir ; relat-viseo ; prendr-retard ; scann ; zéro-papi ; numer-outil ; econom-papi ; technolog-pas ; e-rou ; evit-tout ; servic-mobil ; eteindr-lumier ; mettr-veil ; fraud ; signatur-tablet ; email ; tach-demand ; éteindr-lumi ; mail-plutôt ; pdf ; adapt-nouveau ; informat-perform ; recto-verso ; contrat-oper ; sécur-don ; tchat ; édit-papi ; vent-distanc ; pouvoir-valid ; technolog-rest ; rapport-humain ; impress-inutil ; lumi-écran ; écran-ordin ; papi-pas

n°15	thema-echanges-horizontaux	collabor-agenc ; partag-bon ; dialogu ; plus-synerg ; commercial-commun ; unit-sieg ; echang-reunion ; commun-différent ; cré-davantag ; immers-sieg ; sit-agenc ; plus-contact ; plus-échang ; équip-audit ; fili-méti ; bon-pratiqu ; projet-transversal ; march-local ; amélior-commun ; servic-sieg ; gestion-social ; partag-idé ; impliqu-mond ; sein-bureau ; communiqu-cibl ; different-servic ; cre-even ; réflexion ; jumelag ; rencontr-moment ; caf ; problemat ; valoris-synerg ; manqu-commun ; plac-réunion ; practic ; discuss ; cré-plus ; amélior-transversal ; travau ; favoris-echang ; person-différent ; vic-vers ; reunion-conseil ; agent-réseau ; collabor-pas ; visit-agenc ; particip-salari ; déplac-inutil ; compétent-chacun ; sieg-invers ; évit-déplac ; offre-transvers ; difficult-rencontré ; mix-équip ; connaîtr-mieux ; différent-méti ; personnel-sieg ; organis-reunion ; facilit-échang ; synerg-sieg ; différent-marcher ; échang-réguli ; semestriel ; diffus-inform ; sujet-commun ; impliqu-plus ; ag-cl ; territorial ; administr-salar ; collègu-réseau ; trombinoscop ; gen-sieg ; commun-ensembl ; multipli-rencontr ; harmonis-méthod ; reunion-commercial ; tour-tabl ; commun-march ; client-collabor ; transversal ; commun-équip ; meilleur-connaiss ; kpmg-gestion ; conseil-adm ; différent-servic ; sieg-connaîtr ; reunion-commun ; méti-marcher ; commun-sein ; trimestriel ; intern-action ; inter-servic ; porteur-offre ; échang-méti ; servic-support ; action-entrepris ; fonction-support ; reseau-sieg ; rencontr-sieg ; temp-temp ; réunion-mensuel ; bouch-oreil ; visit-servic ; pas-administr ; réunion-commercial ; permettr-collabor ; mélang-équip ; administr-person ; ken-kel ; pas-forc ; davantag-synerg ; sieg-collègu
n°16	thema-charge-travail	action-commercial ; diminu-nombr ; nombr-client ; dégag ; dégag-temp ; dévelopg-commercial ; travail-tres ; occup ; sur-charg ; prioris-renonc ; tach-administr ; surcharg-travail ; affich-priorit ; trop-charg ; temp-collabor ; charg ; manqu-arbitrag ; don-temp ; sent-devoir ; charg-travail ; trop-projet ; plus-temp ; planning ; trop-import ; temp
n°17	thema-relations-partenaires	vi-vi

n°18	thema-formation	mont-compétent ; form-personnel ; form-davantag ; salari-utilis ; form-collabor ; personnel-nouvel ; format-mutual ; savoir-utilis ; renforc-compétent ; form-jeun ; accompagn-collabor ; form-ensembl ; adapt-format ; format-région ; parl-client ; form-equipi ; notion-mutual ; format-nouveau ; jeun-embauch ; plus-format ; form-salari ; assur-plus ; form ; collabor-nouvel ; temp-format ; collabor-utilis ; learning ; form-conseil ; informer ; former ; collabor-mutual ; salari-nouvel ; compétent-salari ; parcour-jeun ; collabor-évolu ; salari-former ; nouvel-techno ; pouvoir-parl ; format-assur ; bien-form ; nou-form ; format-conseil ; plac-format ; form-équip ; adapt-collabor ; niveau-compétent ; format-jeun ; kpmg-academy ; niveau-connaiss ; jeun-embaucher ; altern ; réguli-collabor ; format-collabor ; format-intern ; dévelopg-expertis ; fondamental ; form-équipi ; format ; conduit-chang ; mieux-form ; format-continu ; collabor-former ; inform-salari ; format-salari ; maitris-outil ; format-personnel ; nouveau-embaucher ; developp-competent ; gard-altern ; academy ; apli-gard ; format-pas ; plus-autonom ; parcour-format ; form-évolu ; format-salar ; format-adapte ; consacr-temp ; integr-parcour
n°19	thema-feedback	enquêt-satisfact ; ressent-client
n°20	thema-echanges-verticaux	méti-chacun ; réseau-administr ; impliqu-administr ; ecout-agent ; plus-écout ; actif-ag ; présenc-ag ; administr-cl ; chacun-rendr ; collabor-sieg ; réseau-sieg ; remont ; particip-actif ; transmiss ; conseil-administr ; renforc-synerg ; infos ; administr ; ecout ; remonte ; réflex ; servic-agenc ; remonté ; avis ; salari-conseil ; administr-collabor ; ateli ; rôl-administr ; compt-avis ; pres-ag ; amplifi ; rendr-compt ; synerg-administr ; intervent ; consult ; collabor-particip ; remont-inform ; salari-réseau ; comport ; écout-méti ; journ-immers ; rencontr-salari ; particip-ag ; ten-compt ; sein-conseil ; écout-bas ; écout-problem ; salari-valeur ; échang-administr ; mum-sieg ; administr-réunion ; reporting-commercial ; administr-salari
n°21	thema-exemplarite	plus-fort ; fonction-manag ; don-moyen ; lign-managérial ; courag ; lesquel-manag ; dirig ; encadr ; exemplar ; manag-pas ; exemplar-lign ; présenc-terrain ; confianc-équip ; messag-managérial ; manqu-exemplar ; manag-devoir ; exemplar-managérial ; mpl-mdl
n°22	thema-outils	outil-informat ; report ; disposit ; logiciel ; dispos ; amélior-outil ; outil-travail ; plac-outil ; plus-perform ; outil-fonction ; trop-lourd ; outil ; disposit-outil ; outil-modern ; gagn-temp ; mis-disposit ; outil-numer

n°23	thema-communication-externe	tvl-fal ; campagn-commun ; agenc-client ; plus-sociétariat ; accentu-commun ; fal-tvl ; communiqu ; entrepris-associ ; tv ; temoignag ; communiqu-valeur ; commun-cred ; expliqu-valeur ; commun-offre ; entrepris-local ; somm-banqu ; rajeun-imag ; abord-sujet ; convaincur ; mieux-communiqu ; invit-décideur ; concret-mutual ; commun-assur ; sall-réunion ; sociétair-devoir ; affichag-agenc ; inform-sociétariat ; mutual-plus ; bien-expliqu ; transmettr-valeur ; rappel-ensembl ; adapt-commun ; parl-action ; commun-medi ; parrainag ; connaitr-produit ; radio ; commun-médi ; parl-davantag ; inform-collabor ; parol ; rôl-sociétair ; entourag ; tranch-âge ; expliqu-client ; produit-servic ; mod-multicanal ; campagn-publicitair ; sit-mettr ; cibl-jeun ; communiqu-avantag ; projet-réaliser ; mettr-valeur ; technolog-nouvel ; commun-aupres ; client-mutual ; projet-associ ; recrut-administr ; commun-spécif ; pas-assur ; mutual-cooper ; nouveau-client ; bon-imag ; local-fal ; aupres-proch ; banqu-différent ; pub-tel ; salari-action ; communiqu-davantag ; beaucoup-client ; chef-entrepris ; produit-assur ; plac-commun ; présenc-réseau ; réunion-associ ; campagn-pub ; client-identifi ; commun-agenc ; commun-affich ; accueil-manifest ; commun-national ; ouvr-sall ; expliqu-clair ; davantag-communiqu ; plus-connaitr ; gestion-sinistr ; commun-régional ; communiqu-réseau ; parl-avantag ; communiqu-façon ; réseau-social ; cré-réseau ; affichag-publicitair ; témoignag ; témoign ; sociétariat-avantag ; client-assurer ; prévent ; communiqu-sociétariat ; fort-action ; mettr-plus ; visibl-action ; parl-systémat ; associ-client ; économ-local ; mani-plus ; véhicul-bon ; disposit-sall ; parl-plus ; plus-pub ; nouveau-entrant ; sociétariat-action ; avantag-societariat ; aupres-asso ; term-imag ; publiqu ; jeun-jeun ; societariat-jeun ; sujet-tart ; film ; mutual-aupres ; médi-affich ; chang-imag ; imag-marqu ; plus-publiqu ; banqu-mutual ; acteur-local ; client-pens ; accompagn-client ; promouvoir-societariat ; remis-prix ; aupres-jeun ; assur-client ; commun-local ; client-savoir ; communiqu-press ; commun-plus ; pub-tv ; parl-posit ; action-mis ; communiqu-assur ; pub ; expliqu-jeun ; recommand-aupres ; imag-cred ; façon-plus ; disposit-entrepris ; utilis-banqu ; posit-banqu ; associ-sall ; disposit-associ ; local-associ ; action-réalisé ; valeur-action ; communiqu-client ; banqu-cooper ; communiqu-aupr ; banqu-banqu ; communiqu-aupres ; aupr-client ; developp-commun ; portail ; campagn-affichag ; agenc-action ; associ-secteur ; action-associ ; ensembl-salari ; prendr-temp ; imag-plus ; commun-sociétariat ; agenc-caiss ; expliqu-sociétariat ; aupres-

n°24	thema-management	plan-action ; cultur-résultat ; pouvoir-décis ; plan ; pragmat ; accompagn-chang ; action-concret ; manag-plus ; méthod
n°25	thema-ressources-humaines	manqu-visibil ; object-personnel ; chang-conseil ; perspect ; jeun-administr ; parcour-professionnel ; rh ; diplom ; impress-chang ; opportun ; impliqu-jeun ; pas-uniqu ; recouvr ; valoris-expertis ; ressourc-humain ; cré-vérit ; valoris-compétent ; recrut ; conserv-compétent ; renforc-effect ; post-plus ; diplôm ; contrat-moral ; développ-compétent ; permettr-chacun ; temp-travail ; perspect-évolu ; visibil-parcour ; secrétair ; polit-rh ; mutat
n°26	thema-innovation	liber-initi ; activ-ingénieur ; revolu ; développment ; favoris-esprit ; modern-innov ; encourag-initi ; incub-solut ; favoris-innov ; revolu-numer ; valoris-initi ; temp-innov ; cart-usag ; in-nov ; initi-pris ; pris-risqu ; pai-internet ; adapter-activ ; ima-gin ; financ-projet ; financ-particip ; start-up ; achat-internet ; uniqu-pai ; informat-adapter ; nouveau-servic ; usag-uniqu ; initi-personnel ; solut-start ; smart ; futur ; droit-erreur
n°27	thema-performance-financière	frer-général ; baiss-frer ; opex ; baiss ; temp-réel
n°28	thema-concurrence	démarqu ; rest-lead ; différenci-concurrent ; concurrent ; avantag-banqu ; rapport-concurrent ; rest-compétit ; concurrent-plus ; contr-banqu ; banqu-lign ; concurent ; concurrent-banqu ; lead
n°29	thema-professionalisme	expert-méti ; gestion-risqu ; risqu-client ; efficient ; maîtris-risqu ; expert-sieg ; compétent-réactiv ; bien-coup ; plus-expertis ; achat-immobili ; exemplair ; pertinent ; manqu-compétent ; premi-coup ; bien-premi ; fiabl ; compétent-techniqu ; activit ; expérient ; mod-projet ; maitris-risqu ; conformit
n°30	thema-motivation	vouloir ; envi-affair ; cercl-vertueux ; don-envi ; communiqu-réussit ; besoin-communiqu ; envi ; manqu-motiv ; group-travail ; envi-souscrir ; réussit-projet ; plus-impliqu ; cré-cercl
n°31	thema-négatif	enfumag ; inaccept ; mefianc ; confus ; nul ; inquiet ; bof ; pas-about ; dubit ; trop-commun ; pas-plus ; sceptiqu ; inquietud ; poudr-yeu ; manipul

9.8 Classification basée sur des motifs émergents (CME)

L'objectif de cette méthode est de classer des messages courts dans un ou plusieurs groupes, en se basant sur les motifs émergents, et ensuite de déterminer la performance de la méthode par rapport au classifieur existant (Extratrees). Les messages courts traités par la suite logicielle Meeting Software sont très "sparses", contenant peu des mots. Les méthodes de classification

traditionnelles sont moins performantes sur ce nouveau type de texte. C'est le cas du classifieur Extratrees que nous utilisons, qui fonctionne en exploitant des arbres de décision basés sur les termes d'un vocabulaire. Les décisions sont fortement altérées lorsque les vecteurs de représentation des messages courts sont creux. Pour remédier au problème cité, nous avons pensé mettre en place une technique basée sur la présence ou l'absence de motifs émergents dans les messages courts. Nous appelons motifs des combinaisons de mots contenues dans les messages courts. Cette méthode a l'avantage d'être simple et fournit des résultats facilement exploitable.

9.8.1 Pré-traitement

Les messages courts doivent subir des prétraitements afin de mieux les caractériser. Trois traitements sont utilisés :

— La tokenisation

— Le filtrage des mots vides de sens

— La stemmatisation

— La décomposition des messages courts en motifs

Pour la tokenisation, le filtrage des mots vides de sens et la stemmatisation, on se reportera aux paragraphes précédents. La décomposition des messages courts en motifs consiste à les ramener à des combinaisons des mots. Nous utilisons deux élément pour y parvenir :

— la taille maximale de la combinaison

— la fenêtre maximale pour le choix des mots à combiner

Nous prenons généralement 2 comme taille et 0 comme fenêtre. Supposons que l'on a le message court suivant : synergie aide client. Sa décomposition donnera : synergie, aide, client, synergie aide, aide client.

9.8.2 Principe

En général, les méthodes de classification supervisées sont divisées en deux phases : la phase d'apprentissage et phase de prédiction.

Phase d'apprentissage

Dans la phase d'apprentissage, les groupes des messages courts sont connus. L'objectif est d'apprendre les caractéristiques des groupes afin de prédire les groupes des nouveaux messages. Elle est composée de deux fonctions : une fonction permettant d'extraire les motifs fréquents par groupe et une fonction qui sélectionne par groupe les motifs émergents.

Phase de prédiction

Cette phase consiste à trouver le groupe d'un message court dont le groupe n'est pas connu.