

Analyse multidimensionnelle de contenus des réseaux sociaux : stratégies, calcul et visualisation d'indicateurs pour le tourisme

Multidimensional analysis of social network content: strategies, calculation and visualisation of indicators for tourism

Maxime Masson^{1,2}, Rodrigo Agerri², Christian Sallaberry¹, Marie-Noelle Bessagnet¹, Philippe Roose¹, Annig Le Parc Lacayrelle¹

¹ LIUPPA, E2S, Université de Pau et des Pays de l'Adour (UPPA)

² Centre HiTZ – Ixa, Université du Pays Basque EHU/UPV

RÉSUMÉ. L'influence grandissante des réseaux sociaux dans le domaine du tourisme souligne le besoin d'approches efficaces en traitement automatique du langage naturel (TALN) pour exploiter cette ressource. Toutefois, transformer des textes multilingues, informels et non structurés en connaissances structurées reste un défi, notamment à cause de la nécessité de données annotées pour l'entraînement des modèles. Cet article examine, dans un premier temps, différents techniques et modèles de TALN basés sur l'apprentissage pour optimiser les performances tout en réduisant le besoin de données annotées manuellement. Un nouveau jeu de données multilingues (français, anglais, espagnol) spécifique au tourisme a été créé, se concentrant sur la région du Pays Basque. Ce jeu de données inclut des tweets avec des annotations manuelles sur les entités nommées spatiales, les concepts thématiques touristiques et les sentiments. Une comparaison des méthodes de fine-tuning et d'apprentissage few-shot avec des modèles multilingues indique que les techniques few-shot peuvent produire de bons résultats avec peu d'exemples annotés. Les expérimentations menées sur ce jeu de données suggèrent la possibilité d'appliquer les méthodes de TALN à base d'apprentissage à divers domaines, tout en réduisant le besoin d'annotations manuelles et évitant les complexités des méthodes basées sur des règles. Dans un second temps, nous présentons TextBI, un tableau de bord générique multimodal conçu pour présenter des analyses d'annotations textuelles multidimensionnelles sur de grands volumes de données de médias sociaux multilingues. Cet outil se concentre sur plusieurs dimensions : spatiale, temporelle, thématique et personnelle, et prend également en charge des données d'enrichissement supplémentaires telles que le sentiment et l'engagement. Notre approche, TextBI, représente une avancée significative dans le domaine de la visualisation de résultats d'annotations TALN en intégrant et en mélangeant les caractéristiques d'une variété d'outils de Business Intelligence, de systèmes d'information géographique et de TALN.

ABSTRACT. The growing influence of social networks in the field of tourism highlights the need for effective automatic natural language processing (NLP) approaches to exploit this resource. However, transforming multilingual, informal and unstructured texts into structured knowledge remains a challenge, not least because of the need for annotated data for model training. This paper first examines different learning-based NLP techniques and models to optimise performance while reducing the need for manually annotated data. A new multilingual dataset (French, English, Spanish) specific to tourism has been created, focusing on the Basque Country region. This dataset includes tweets with manual annotations on spatial named entities, tourism thematic concepts and sentiments. A comparison of fine-tuning and few-shot learning methods with multilingual models indicates that few-shot techniques can produce good results with few annotated examples. Experiments conducted on this dataset suggest the possibility of applying learning-based NLP methods to various domains, while reducing the need for manual annotations and avoiding the complexities of rule-based methods. Secondly, we present TextBI, a generic multimodal dashboard designed to present multidimensional text annotation analysis on large volumes of multilingual social media data. The tool focuses on several dimensions: spatial, temporal, thematic and personal, and also supports additional enrichment data such as sentiment and engagement. Our approach, TextBI, represents a significant advance in the field of NLP annotation results visualisation by integrating and blending features from a variety of Business Intelligence, Geographic Information Systems and NLP tools.

MOTS-CLÉS. Apprentissage Few-Shot, Modèle de Langage Masqué (MLM), Multilinguisme, Science Sociale Informatique, Traitement Automatique du Langage Naturel, Visualisation, Multidimensionnel, Tourisme.

KEYWORDS. Few-Shot Learning, Masked Language Model (MLM), Multilingualism, Social Science Computing, Natural Language Processing, Visualisation, Multidimensional, Tourism.

1. Introduction

De nos jours, les réseaux sociaux se sont imposés comme des moyens de communication essentiels pour le partage d'opinions et d'expériences dans une variété de domaines, devenant ainsi une ressource précieuse pour les professionnels du tourisme tels que les offices de tourisme et les agences de voyage [ZEN 14]. Toutefois, l'analyse de grandes quantités de données issues des réseaux sociaux représente un défi majeur [MAY 12], en particulier pour extraire des connaissances structurées de textes non structurés. Ainsi, les acteurs du tourisme se tournent souvent vers les informaticiens et les linguistes pour l'extraction de connaissances, qui utilisent alors des techniques de Traitement Automatique du Langage Naturel (TALN).

Le projet APs (Augmented Proxemic Services) [MAS 24], illustré figure 1, vise l'analyse de trajectoires d'utilisateurs des réseaux sociaux : par exemple, l'activité d'un utilisateur ou de cohortes d'utilisateurs dans l'espace touristique du réseau social Twitter/X. Cette figure illustre les principaux objectifs du projet : (1) une bibliothèque de domaines d'application (tels que le tourisme, l'éducation, la santé, etc.) pour lesquels sont définis des vocabulaires contenus dans des ressources sémantiques (de simples dictionnaires, jusqu'à des structures sémantiques plus complexes) ; (2) le domaine d'application actif : il s'agit du domaine d'intérêt utilisé ; (3) une boîte à outils génériques pour traiter et analyser les données issues de médias sociaux ; (4-5) une interface de visualisation d'indicateurs dédiée à des utilisateurs finaux, qui ne sont pas nécessairement des informaticiens.

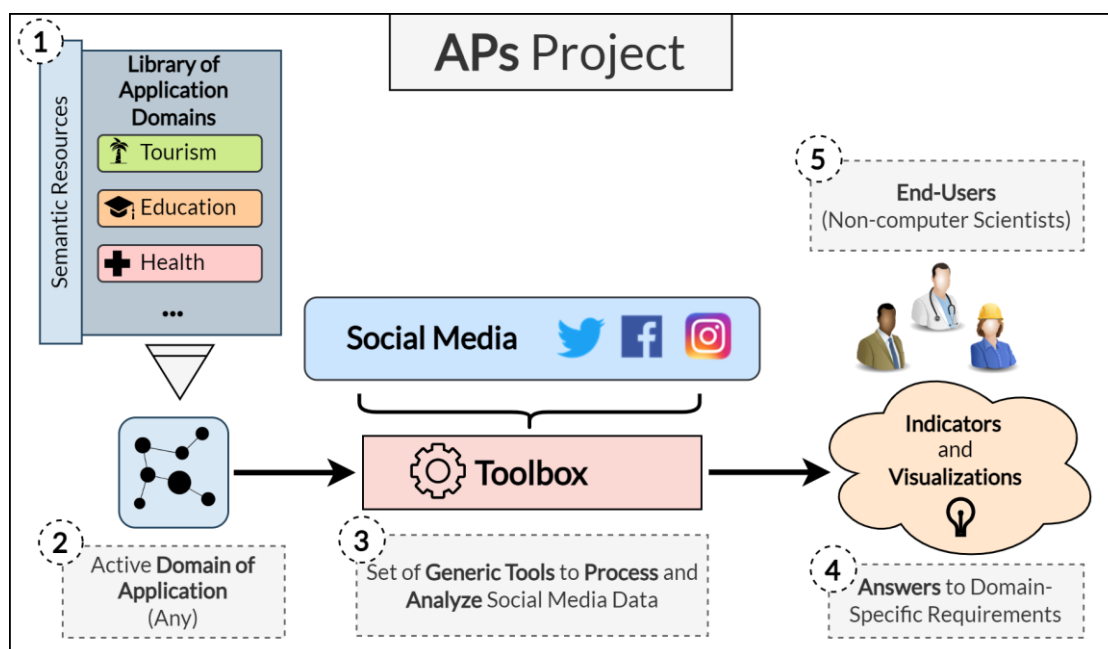


Figure 1. Le projet APs (Augmented Proxemic Services)

Nous allons donner deux éclairages relatifs aux points 3 et 4 de la figure 1, respectivement. Dans un premier temps, le travail sera centré sur le traitement de données issues de médias sociaux pour lequel les outils de TALN sont essentiels. En effet, le TALN est un outil puissant pour traiter et analyser les données textuelles, souvent employé pour automatiser des tâches telles que la détection de sentiments, la reconnaissance d'entités nommées spatiales et l'extraction de concepts thématiques fins [ROS 15]. Les progrès récents dans ce domaine, notamment avec l'émergence de l'apprentissage profond et le développement des Modèles de Langage Masqués (MLM), offrent des avantages significatifs par rapport aux méthodes traditionnelles basées sur des règles. Ces nouvelles techniques d'apprentissage automatique, adaptées aux variations linguistiques [MIN 21], permettent une analyse plus dynamique et adaptative, en opposition aux approches basées sur des règles, souvent ad hoc et plutôt statiques. Toutefois, pour obtenir des résultats optimaux dans des domaines d'application spécifiques, les modèles d'apprentissage doivent préalablement subir une étape dite de *fine-tuning*, c'est-à-dire un enrichissement avec des exemples annotés fortement liés au domaine concerné. Cela soulève deux

questions récurrentes pour les chercheurs : (1) quelles sont les stratégies et modèles d'apprentissage les plus appropriés pour un domaine d'application et une tâche donnée, et (2) combien d'exemples annotés spécifiques au domaine sont nécessaires pour obtenir des résultats satisfaisants. L'annotation manuelle d'exemples est un processus souvent coûteux, fastidieux et chronophage, la majorité des chercheurs visent donc à obtenir les meilleurs résultats possibles tout en minimisant au maximum la quantité d'exemples annotés nécessaire.

Ainsi, nous présentons une étude comparative sur les besoins en annotations pour obtenir de bonnes performances sur trois tâches d'extraction de connaissance communes appliquées au **domaine du tourisme**. Plus précisément, nous cherchons à savoir quelles stratégies d'apprentissage sont les meilleures pour minimiser le processus d'annotation manuelle des données et éviter les approches basées sur des règles, peu dynamiques. Nous supposons que parmi les modèles de langage masqués et les techniques d'entraînement existants, certains seront mieux adaptés à ce domaine précis. En effet, les messages des réseaux sociaux sont caractérisés par des textes informels courts, des erreurs grammaticales fréquentes et la présence d'emojis et hashtags. L'objectif de ce travail est d'analyser des collections de posts, figure 1, selon un domaine d'application générique (le tourisme dans cette expérimentation), des lieux et des sentiments émis par les utilisateurs de réseaux sociaux. Par conséquent, nous nous concentrons sur les trois tâches d'extraction de connaissances suivantes : la classification de la polarité des sentiments (classification de textes), la reconnaissance d'entités nommées spatiales et l'extraction de concepts thématiques fins (classification de tokens). Des indicateurs visuels issus de ces 3 catégories d'annotations aideront le décideur, expert du domaine ciblé, dans son analyse..

De plus, face à l'ampleur des données disponibles sur les plateformes sociales et à la multiplicité des annotations possibles, il est impératif de disposer d'outils adaptés permettant une visualisation et une interprétation optimisées, spécifiquement pour des utilisateurs tels que les décideurs d'un domaine donné (par exemple : le tourisme, les politiques publiques, etc.). Aussi, dans un second temps, nous nous focalisons sur la visualisation interactive de résultats d'analyses multidimensionnelles des médias sociaux. Nous faisons l'hypothèse qu'un tableau de bord générique et interactif, centré sur quatre dimensions (spatiale, temporelle, thématique et personnelle), correspondrait à une plateforme conviviale et adaptable permettant aux non-informaticiens d'obtenir facilement des informations clef sur les réseaux sociaux. Après avoir passé en revue des outils de visualisation dédiés à l'aide à la décision dans la partie des travaux connexes, nous présenterons le tableau de bord générique TextBI que nous illustrerons par des exemples, dans le contexte applicatif du tourisme. Cependant, cet environnement a été conçu pour être générique et adaptable à de multiples domaines.

Nos principales contributions sont les suivantes : (1) nous proposons un **nouveau jeu de données composé de tweets touristiques** (disponible à la demande, à des fins de recherche). Ce jeu de données est multilingue (français, anglais et espagnol) et a été manuellement annoté au niveau du texte avec le sentiment (positif, négatif et neutre) et au niveau du token avec les *lieux* et les *concepts thématiques*. Ces derniers sont liés au thésaurus du tourisme et des loisirs¹ de l'Organisation Mondiale du Tourisme (OMT) ; (2) nous réalisons une **analyse comparative** entre des techniques de TALN basées sur des règles, le *fine-tuning* et l'apprentissage *few-shot* [MA 22] avec pour objectif d'établir quelle méthode est la plus efficace pour chaque tâche d'extraction de connaissances ; (3) nous expérimentons différentes méthodes d'échantillonnage de données pour déterminer **combien d'exemples annotés sont réellement nécessaires** afin d'obtenir des résultats compétitifs sur chacune des trois tâches. L'objectif est d'éviter aux experts du domaine d'application d'annoter manuellement une trop grande quantité de données par rapport à leurs besoins ; (4) finalement, nous proposons une interface de type *Business Intelligence* pour visualiser les indicateurs résultant d'analyses des données issues de ces traitements TALN. Une vidéo de démonstration est disponible à l'adresse suivante : maxime-masson.github.io/TextBI.

¹ <https://www.e-unwto.org/doi/book/10.18111/9789284404551>

Ces contributions se veulent génériques, indépendantes d'un domaine d'activité. A des fins d'expérimentation, nous avons travaillé sur deux domaines : nous avons collaboré avec (1) les offices de tourisme du Pays Basque et du Béarn et (2) des experts en management des collectivités territoriales (Chaire Optima rattachée au laboratoire LIREM de l'UPPA).

Cet article est structuré comme suit. Dans la section 2, nous passons en revue les techniques de TALN basées sur l'apprentissage profond les plus communément utilisés dans le domaine du tourisme et nous présentons différentes familles d'outils de visualisation d'indicateurs dédiés à la prise de décision. La section 3 développe le processus de construction et d'annotation semi-automatique de notre propre jeu de données. La section 4 détaille les expérimentations menées sur notre jeu de données afin de déterminer une stratégie optimale ; aussi, nous y présentons la configuration expérimentale de notre étude comparative, l'analyse des résultats obtenus dans le cadre des différentes approches de TALN, pour chacune des trois tâches : annotation du sentiment, des lieux et des concepts thématiques. La section 5 est dédiée à la présentation de notre tableau de bord universel TextBI et à son expérimentation sur le jeu de données. La conclusion et les perspectives sont développées dans la section 6.

2. Travaux connexes

Dans le secteur en évolution constante du traitement automatique du langage naturel (TALN), l'une des avancées les plus significatives a été l'avènement des modèles de langage masqués (*Masked Language Model* ou MLM). Ces modèles, entraînés sur de vastes corpus textuels, capturent un large éventail de structures linguistiques, de nuances et de connaissances [MIN 21]. Ils offrent ainsi une amélioration significative des performances dans de nombreuses tâches de TALN, allant de la classification de textes ou de tokens aux systèmes de questions-réponses. Des modèles de langage masqués tels que BERT (*Bidirectional Encoder Representations from Transformers*) [DEV 18] et XLM-RoBERTa [CON 19] sont entraînés sur d'importantes quantités de données textuelles, englobant parfois des téraoctets d'informations issues de sources diverses telles que des livres, des articles et des sites Web. Grâce à ce processus d'entraînement étendu, ces modèles sont capables de comprendre les relations complexes entre mots, expressions et constructions textuelles. Ils capturent des signaux sémantiques, syntaxiques et contextuels, leur conférant la capacité de générer et de comprendre le langage naturel [TOP 23].

Parallèlement à ces travaux, nous nous intéressons aux outils de visualisation de résultats d'analyse de données en général et issues de réseaux sociaux, en particulier. Nous allons notamment étudier le potentiel des Systèmes d'information géographique (SIG), des outils de *Business Intelligence* (BI), ou encore, celui des outils de TALN.

Nous allons faire un tour d'horizon des opportunités liées aux modèles de langages (section 2.1) et aux outils de visualisation de données (section 2.2), dans un objectif de réinvestissement dans un domaine applicatif spécifique (par exemple, le tourisme) et dans le contexte des réseaux sociaux.

2.1. Les modèles de langage masqués au service du TALN

Une approche fréquemment utilisée consiste à *fine-tuner* (apprentissage par transfert) les modèles de langage masqués pour des tâches spécifiques au domaine [SUN 19]. Cela se fait en altérant les poids du modèle pour l'adapter à la nouvelle tâche, via un enrichissement de ce dernier avec des exemples annotés spécifiques au domaine étudié. Afin d'améliorer la précision, les modèles de langage masqués ont été affinés pour des tâches de classification de textes notamment, comme, par exemple, la détection de spam dans les avis d'hôtels [CRA 21] ou encore l'analyse des sentiments dans les avis touristiques [ENR 22] ou dans le domaine du transport durable [SER 21]. En REN (Reconnaissance d'Entités Nommées), le *fine-tuning* de modèles de langage masqués a été employé pour extraire des informations de localisation de corpus touristiques [CHE 20]. Enfin, les modèles de langage masqués ont démontré des résultats prometteurs dans l'extraction de concepts thématiques, tels que l'identification de thèmes

et sujets liés aux voyages dans des textes touristiques [CHA 21]. L'un des principaux défauts de l'approche *fine-tuning* est qu'elle nécessite généralement un volume conséquent d'exemples annotés pour être efficace (plusieurs centaines d'exemples sont nécessaires, au minimum, pour obtenir des résultats non aléatoires). Cela génère une charge de travail d'annotation importante pour les experts du domaine d'application.

Face au défi des données annotées disponibles en nombre limité, les techniques d'apprentissage *few-shot* ont gagné en popularité. Elles permettent aux modèles de langage masqués d'apprendre avec peu d'exemples annotés (de l'ordre de la dizaine), lorsque les données abondantes font défaut.

Le *Pattern-Exploiting Training* (PET) est un exemple d'apprentissage *few-shot* dédié à la classification de texte, dans lequel le modèle est guidé par des descriptions de tâches en langage naturel et des phrases à trous [SCH 20]. Par exemple, pour classer les avis de films en fonction du sentiment prédominant qu'ils expriment, le modèle sera requêté (prompt) avec l'avis du film et la classification souhaitée : « Le film était {?} ». Le modèle essaierait alors de prédire le {?}, en choisissant parmi des options telles que « génial » ou « décevant ». Plusieurs travaux récents ont également appliqué ce principe à la classification de tokens, par exemple, EntLM [MA 22] pour la reconnaissance d'entités nommées. Dans les deux cas, les approches *few-shot* ont démontré leur efficacité pour obtenir des résultats satisfaisants avec peu d'exemples. D'autres techniques mélangent *Fine-Tuning* et apprentissage *Few-shot* comme SetFit [TUN 22], par exemple

Récemment, de nouveaux grands modèles de langages (LLM) sont exploités pour la recherche d'information. Contrairement aux modèles masqués, présentés plus haut, ces modèles ne visent pas la prédiction d'un masque mais la génération d'un texte cohérent. Nous pouvons citer les modèles de la famille des GPT comme GPT 3 [BRO 20] et GPT 4 [ACH 23] qui ne sont pas *open source*. Il y a également des modèles *open source* tels que Mistral [JIA23], LLaMA 2 [TOU 23] et FlanT5 [CHU 24]. Enfin, d'autres LLM sont spécialisés dans la recherche d'information, comme Gollie [SAI 23], par exemple.

2.2. Différentes catégories d'outils au service de la visualisation de données

Les Systèmes d'information géographique (SIG) tels que QGIS [MEN 16] et ArcGIS [BOO 01] offrent une grande variété de fonctionnalités pour répondre aux différents besoins des utilisateurs géographes. Bien que ces systèmes soient utiles pour visualiser en profondeur les données géospatiales, leur principal intérêt réside dans l'aspect spatial des données, ce qui limite leur efficacité dans des contextes de données non spatiales. Bien que les outils SIG soient capables d'afficher des données thématiques [MUR 03], elles doivent être associées à une zone spatiale spécifique.

En ce qui concerne les outils de *Business Intelligence* (BI), Tableau Public [DAT 18], Power BI [FER 16] et Qlik [SHU 16] sont reconnus pour leur capacité à supporter les processus de prise de décision [HAN 10] [ORL 20] grâce à l'exploration interactive de données et de tableaux de bord conviviaux. Cependant, ces outils sont principalement conçus pour gérer des données numériques bien structurées, ce qui pose des problèmes lorsqu'il s'agit de travailler avec des données nativement textuelles. Bien que certains efforts aient été faits pour intégrer des processus de TALN dans les outils BI [VAS 20] [DES 21], ils manquent d'efficacité pour présenter des trajectoires et pour établir des connexions entre les annotations textuelles de différentes dimensions.

Concernant les outils de traitement automatique de la langue (TALN), ils se scindent en deux principales catégories. Certains, comme SpaCy [CHA 21], TextRazor [RAJ 22], GATE [MAY 00], et Gensim [REH 11], se concentrent notamment sur le traitement des données, offrant des capacités de visualisation limitées comme les nuages de mots, les graphes sémantiques et les vues d'annotations basées sur le texte. D'autres outils, tels que IRaMuTeQ [LOU 14], Voyant [WEL 14], VOSviewer [VAN 13] et SentimentViz [HEA 22], offrent une gamme plus large d'options de visualisation, mais se concentrent généralement sur une seule dimension (comme le sentiment) ou des analyses statistiques

basées sur les mots. De plus, certains de ces outils peuvent être très complexes pour des utilisateurs non-informaticiens.

Notre projet de plateforme interactive, TextBI, a pour objectif la visualisation des annotations TALN et des indicateurs issus d'analyse proxémique [MAS 24] en intégrant et combinant les fonctionnalités de ces diverses catégories d'outils. TextBI combinera des fonctionnalités des outils BI, SIG et TALN (voir tableau 1) au service de la visualisation de corpus annotés issus de réseaux sociaux. Des outils BI, nous emprunterons le design interactif, les interfaces conviviales et la synchronisation visuelle, en adaptant ces fonctionnalités pour leur utilisation avec des données textuelles annotées plutôt qu'avec des données numériques uniquement.

Composant visualisation	Systèmes d'information géographique (SIG)	Outils de BI	Outils de TALN
Exemples	QGIS, ArcGIS	Tableau Public, PowerBI, Qlik	SpaCy, GATE, TextRazor, IRaMuTeQ, VOSViewer, SentimentViz
Points forts (+)	Analyse spatiale poussée ; Granularités spatiales multiples	Exploration de données multidimensionnelles ; Simple d'utilisation ; Interactivité inter et intra visuel ; Filtrages combinés ; Synchronisation des visuels	Statistiques avancées (cooccurrence, fréquences, etc.) ; Gestion native du texte ; Visuels adaptés au texte : graphes sémantiques, nuages de mot, etc.
Faiblesses (-)	Difficile de visualiser des données sans composante spatiale ; Parfois complexe pour des utilisateurs néophytes (non-géographes)	Nécessite des données structurées (e.g., base de données relationnelle) ; Dédiés aux données numériques ; Difficile de visualiser des données séquentielles	Dédiés aux utilisateurs spécialistes du TALN ; Analyse au niveau des mots ; Focus sur une seule dimension (ex. Sentiment)

Tableau 1. Comparaison des points forts et faiblesses entre plusieurs catégories d'outils ayant une composante visualisation

Nous adoptons les vues spatiales détaillées des outils SIG, reconnaissant que les données des réseaux sociaux contiennent souvent un aspect spatial. Cependant, nous visons une approche intégralement multidimensionnelle. Des outils TALN, nous prenons les aspects analytiques, comme l'analyse de cooccurrence et de fréquence, mais nous allons au-delà pour agréger des entités comme des concepts thématiques, des lieux ou encore des entités temporelles.

Ces outils ne peuvent exister sans données, aussi, nous allons présenter le processus de création du jeu de données utilisé dans le cadre de nos expérimentations.

3. Construction du jeu de données et annotation semi-automatique

En vue de l'expérimentation (section 4), nous construisons notre propre jeu de données et un référentiel d'annotations de contenu de type spatial, thématique et sentiment.

3.1. Notre jeu de données

Bien que les données annotées disponibles publiquement pour le domaine du tourisme soient extrêmement réduites, il existe plusieurs corpus annotés dans d'autres domaines, qui pourraient être utilisés pour nos expérimentations. Par exemple, le corpus **ESTER** [GAL 06] est une collection complète de transcriptions de radio françaises ; **AnCora** [TAU 08] est un corpus annoté sur plusieurs niveaux (*principalement à partir de journaux*) pour le catalan et l'espagnol. Ces deux ressources sont annotées pour la reconnaissance d'entités nommées. En termes de ressources spécifiques aux réseaux sociaux, le **Broad Twitter Corpus** [DER 16] comprend des annotations sur les lieux, personnes et organisations, tandis que **Sentiment140** [GO 09], **STS-Gold** [SAI 13], et de nombreux autres jeux de données développés dans le cadre de tâches d'évaluation partagées comme **SemEval** [ROS 15], sont utilisés pour la classification de la polarité des sentiments. D'autres corpus incluent le jeu de données de dialogue **MultiWOZ** [BUD 18], le jeu de données **Stanford NLI** [BOW 15] pour l'inférence textuelle, et le corpus **Heldugazte** qui aide à catégoriser les tweets comme formels ou informels.

Ces jeux de données sont vastes mais très généraux et se concentrent souvent uniquement sur l'anglais, manquant ainsi d'informations contextuelles nécessaires pertinentes au domaine du tourisme. Plus important encore, nous n'avons pas trouvé de jeu de données publiques annotées pour l'extraction de concepts thématiques fins dans le domaine du tourisme. Compte tenu de cela, nous avons décidé de construire notre propre jeu de données annoté.

Aussi, nous décrivons le processus de création d'un nouveau jeu de données multilingue composé de tweets liés au tourisme et annotés pour trois tâches d'extraction de connaissance pour des applications dans le domaine du tourisme : (1) la classification de la polarité des sentiments, (2) la reconnaissance des entités nommées spatiales, et (3) l'extraction de concepts thématiques fins (basé sur le Thésaurus de l'OMT).

Le jeu de données a été collecté via X en utilisant l'API Academic² et le processus de collecte a été mis en œuvre en utilisant une méthodologie que nous avons conçue pour la construction de jeux de données. Cette méthodologie est à la fois générique, itérative et incrémentale (voir [MAS 22] pour plus de détails). Plusieurs itérations ont été réalisées, chacune avec un filtrage successif correspondant aux dimensions cibles du jeu de données : spatiale (zone de la côte Basque française, coordonnées et toponymes), temporelle (été 2019, horodatage) et thématique (domaine du tourisme tel que définis par le thésaurus de l'Organisation Mondiale du Tourisme). Pour éviter un bruit excessif, chaque itération a été suivie d'un *feedback* humain pour ajuster et équilibrer les critères de filtrage. De même, nous avons exclu les utilisateurs professionnels et institutionnels car nous sommes principalement intéressés par la compréhension du comportement des touristes en tant qu'individu, et non par l'analyse de contenus promotionnels ou institutionnels.

Le jeu de données final comprend 27 379 tweets, parmi lesquels **2 951 tweets** provenant de 624 utilisateurs ont été sélectionnés pour l'annotation et l'utilisation dans nos expérimentations. Ces tweets spécifiques (2 951) ont été sélectionnés car ils ont été examinés manuellement pour s'assurer qu'ils concernent le tourisme et **sont émis par des touristes** (e.g., nous avons déterminé que seuls 624 utilisateurs du jeu de données initial sont des touristes, représentant environ 10 % des 27 379 tweets initiaux), plutôt que par des professionnels du tourisme ou des médias traitant du tourisme, entre autres. Ce choix privilégie la qualité des tweets par rapport à la quantité. Le jeu de données est multilingue et inclut une variété de tweets en anglais, français et espagnol. Le déséquilibre entre les différentes langues reflète la réalité de l'utilisation des réseaux sociaux sur la côte basque française. Le tableau 2 montre la répartition linguistique du jeu de données et la subdivision des tweets pour l'apprentissage (approximativement 60% pour l'entraînement (*Train*), 20% pour le développement

² <https://developer.twitter.com/en/use-cases/do-research/academic-research> (fin du service en avril 2023)

(*Dev*), 20% pour le test (*Test*)). Cette division maintient un équilibre entre le nombre d'utilisateurs et de langues dans chaque jeu de données.

	Tous	Français	Anglais	Espagnol
Train	1 652 (503)	1 297 (391)	283 (129)	82 (32)
Dev	619 (300)	450 (213)	99 (66)	70 (31)
Test	680 (431)	401 (273)	102 (100)	177 (93)

Tableau 2. Répartition du jeu de données collecté par langue – Tweets (Utilisateurs)

3.2. Annotation du sentiment

Le processus d'annotation des 2 961 tweets a été effectué **semi-automatiquement**, comme illustré dans la figure 2. Pour faciliter le travail des annotateurs humains, 1 299 tweets des divisions *dev* et *test* ont été **pré-annotés automatiquement** (au niveau du texte) en utilisant les 5 modèles de langage dédiés à la prédiction du sentiment décrits dans le tableau 2.

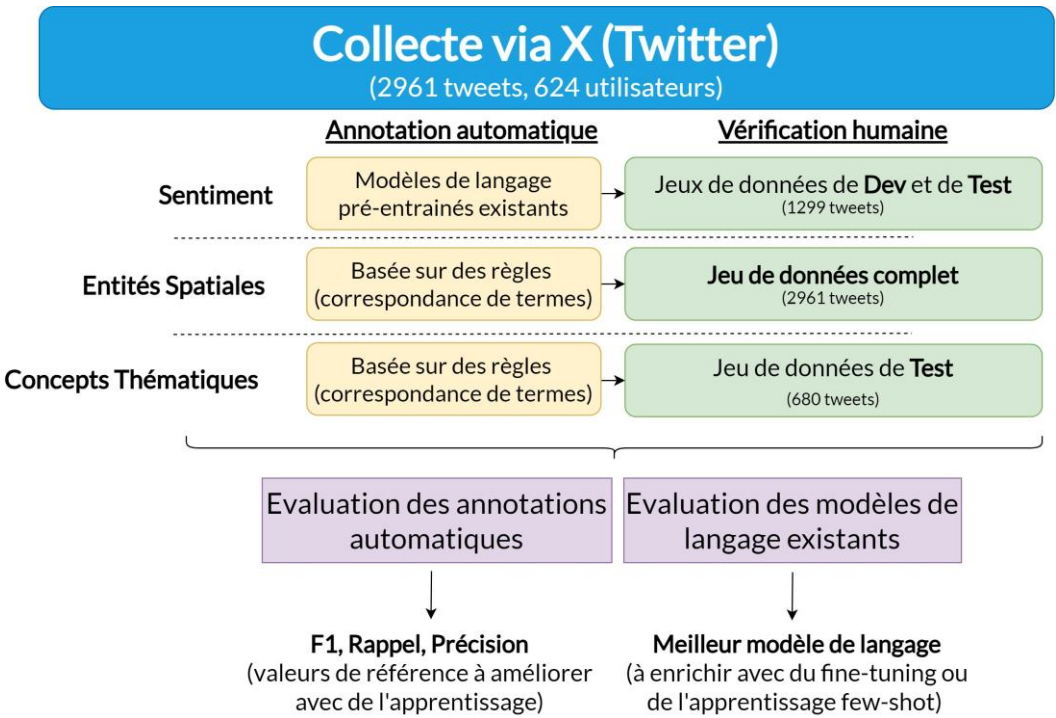


Figure 2. Processus de construction et d'annotation du jeu de données

Ces annotations ont ensuite été révisées manuellement par les annotateurs humains. Chaque tweet a été évalué par deux annotateurs pour mesurer la concordance (via le coefficient *kappa de Cohen*), assurant ainsi la qualité des annotations. Nous avons obtenu $\kappa = 0.79$ pour les tweets en français, $\kappa = 0.75$ pour l'espagnol, et $\kappa = 0.67$ pour l'anglais, indiquant un accord fort. Les divergences ont été résolues par discussion collaborative.

L'étape suivante a consisté à évaluer la performance des 5 modèles de langage choisis pour l'annotation automatique, en les comparant aux annotations révisées par les annotateurs humains sur le sous-jeu de test, *Test*, décrit précédemment. Le tableau 3 montre que le modèle XLM-T Sentiment [BAR 22] qui est déjà *fine-tuné* avec des données multilingues sentimentales provenant de domaines variés, a obtenu les meilleurs résultats en moyenne pour les trois langues. Ce modèle (XLM-T Sentiment) a été utilisé pour annoter automatiquement le jeu de données *train*, que nous avons révisé manuellement pour nos expérimentations.

Langue des tweet	[BAR 20]	[PER 21]	[SEE 23]	[HAR 23]	[BAR 22]
Français	0.56	0.45	0.43	0.47	0.82
Espagnol	0.71	0.64	0.61	0.34	0.83
Anglais	0.81	0.81	0.71	0.66	0.80

Tableau 3. Précision de modèles de prédiction du sentiment existants (sous-jeu de test)

3.3. Annotation des lieux et des concepts thématiques touristiques

Nos tâches suivantes s'intéressent à la reconnaissance d'entités nommées spatiales et à l'extraction de concepts thématiques détaillés dans le domaine du tourisme. Contrairement à la détection de sentiments, qui relève de la classification de textes, ces tâches s'inscrivent dans le cadre de la classification de tokens. Autrement dit, chaque token dans les textes est annoté individuellement.

Avant d'expérimenter des méthodes d'apprentissage automatique, nous avons développé une méthode d'annotation basée sur la correspondance de termes. La précision et le taux de rappel obtenus par cette méthode serviront de référence à améliorer avec des méthodes par apprentissage. Pour détecter les localisations, nous avons utilisé 625 toponymes locaux issus d'Open Street Map, incluant des villes, des points d'intérêt et des repères. Quant aux concepts thématiques, ils ont été identifiés en utilisant leurs étiquettes et synonymes dans le thésaurus de l'Organisation Mondiale du Tourisme, qui recense 1 494 concepts liés au tourisme. Un prétraitement des tweets (lemmatisation, mise en minuscules, suppression des URL, décomposition des hashtags) a été effectué pour faciliter la correspondance des termes. Nous avons appliqué cette méthode pour annoter l'ensemble du jeu de données (*train*, *dev* et *test*). Les annotations générées automatiquement ont par la suite été révisées manuellement par des annotateurs humains. Cette révision a été appliquée sur l'ensemble du jeu de données pour les entités spatiales, tandis que pour les concepts thématiques, elle a été limitée au sous-jeu de données *Test*. Pour les concepts thématiques, la méthode par correspondance de termes a détecté 315 classes de concept unique (sur les 1 494 concepts inclus dans le thésaurus de l'OMT), donnant ainsi une tâche de classification de tokens de granularité très fine. Nous n'avons effectué les révisions manuelles que sur le sous-jeu de données *Test*, car annoter 315 classes de concepts est une tâche complexe et demandant un temps conséquent.

L'accord entre annotateurs a été mesuré sur un échantillon aléatoire de 100 tweets. Concernant les entités spatiales, le coefficient Kappa atteint 0,91 pour les correspondances exactes, c'est-à-dire lorsque tous les tokens constituant une entité correspondent (par exemple, dans le cas de la ville de *New York*, les tokens *New* et *York*). Pour les correspondances partielles, où une entité est identifiée mais présente des tokens manquants ou additionnels (*New* seulement sans le *York*), le coefficient est de 0,93. Ces valeurs témoignent d'un accord quasi parfait entre les annotateurs.

Reconnaissance d'entités nommées spatiales	Rappel	Précision	Mesure F1
Correspondance Exacte	0.692	0.722	0.707
Correspondance Partielle	0.780	0.814	0.796
Extraction de concepts thématiques fins	Rappel	Précision	Mesure F1
Correspondance Exacte	0.746	0.952	0.836
Correspondance Partielle	0.747	0.953	0.837

Tableau 4. Performance de la méthode d'annotation par correspondance de termes

Les performances de la méthode par correspondance de termes appliquée au sous-jeu de données *Test* sont rapportées dans le tableau 4. Ces résultats serviront de référence pour comparer avec les différentes approches par apprentissage dans la section expérimentale (section 4). Nous constatons que, pour les localisations, les performances ne sont pas satisfaisantes, notamment en termes de rappel

(0.692). Cependant, la méthode par correspondance de termes se distingue nettement par sa précision sur l'extraction de concepts thématiques fins (0.952). Bien que cela constitue une valeur de référence solide, le rappel reste relativement faible (0.746), ce qui signifie que de nombreux concepts thématiques ne sont pas détectés. Ainsi, notre principal objectif est désormais de déterminer si les techniques d'apprentissage automatique peuvent égaler ou surpasser cette méthode tout en minimisant la quantité d'annotations manuelles, en particulier pour l'extraction de concepts thématiques fins.

Le jeu de données étant construit, nous pouvons initier les premières expérimentations.

4. Expérimentation : protocole, mise en œuvre et analyse des résultats

Cette section détaille le protocole expérimental mis en place dans l'étude comparative ainsi que l'analyse des résultats obtenus. Ces résultats ont pour but de mettre en évidence la stratégie de TALN optimale pour chacune des tâches d'annotation du sentiment, des lieux et des concepts thématiques, respectivement.

Après la présentation du protocole (section 4.1), l'expérimentation se concentre sur les trois tâches présentées précédemment : classification de la polarité des sentiments (section 4.2), reconnaissance d'entités nommées spatiales et extraction de concepts thématiques fins (section 4.3) avec pour objectif de déterminer quelles approches sont les plus efficaces et avec quelle quantité de données d'entraînement.

4.1. Protocole expérimental

La figure 3 donne un aperçu de notre configuration expérimentale avec (1) les modèles de langage, (2) les méthodes d'échantillonnage et (3) les méthodes d'apprentissage automatique utilisées pour chaque tâche. Les données d'entraînement sont échantillonnées en utilisant deux méthodes :

- **Échantillonnage k-shot** : sélection d'un nombre précis d'exemples pour chaque classe d'annotation. Par exemple, dans le cas de la classification de la polarité des sentiments, si l'on souhaite réaliser un échantillonnage 5-shot, il faudra 15 exemples (5 positifs, 5 négatifs et 5 neutres). Pour nos expérimentations, nous avons utilisé les valeurs de k suivantes : 5, 10, 20, 30, 40, 50, et 100 exemples par classe.
- **Échantillonnage par pourcentage** : utilisation d'un pourcentage précis du jeu de données. Nous avons successivement utilisé 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 90%, et 100% du sous-jeu de données *Train*, tout en essayant de maintenir la distribution originale des classes d'annotation, y compris les étiquettes O (l'étiquette O est attribuée aux tokens qui ne sont ni des lieux ni des entités thématiques).

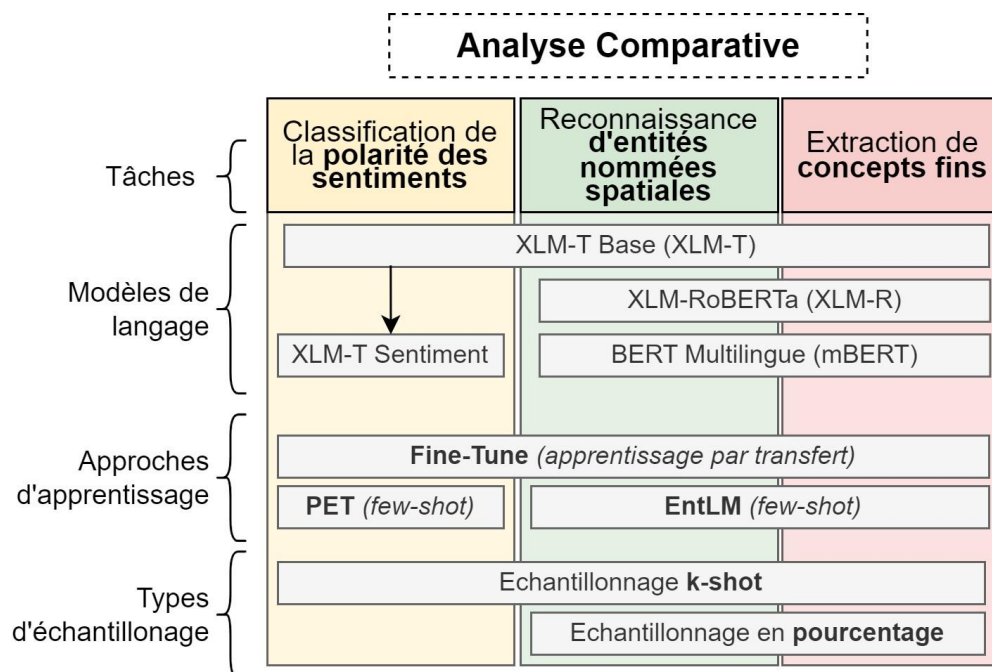


Figure 3. Configuration expérimentale de l'étude

4.2. Classification de textes - Classification de la polarité des sentiments

Sur la base des résultats rapportés par le tableau 2, nous avons choisi le modèle de langage XLM-T [BAR 22] pour nos expérimentations sur la classification de la polarité des sentiments. XLM-T est basé sur XLM-RoBERTa [CON 19], mais entraîné sur un corpus de 198 millions de tweets comprenant 15 langues. Cette version du modèle est spécialement conçue pour gérer les caractéristiques uniques des tweets et plus généralement des publications sur les réseaux sociaux (longueur limitée, langage informel, présence d'emojis, etc.). Nous utilisons 2 variantes :

- La **version de base**, dénommée XLM-T [BAR 22], qui est un modèle de langage masqué (MLM). Elle permet de prédire le prochain mot.
- **XLM-T fine-tuné pour la classification de la polarité des sentiments** (dénote XLM-T Sentiment). Cette variante permettant la prédiction du sentiment a été préalablement *fine-tuné* en utilisant 24 264 tweets couvrant 8 langues différentes (incluant le français, l'anglais, et l'espagnol). Cependant, ces tweets couvrent un large éventail de domaines n'incluant pas le tourisme.

Ces deux modèles vont ensuite être *fine-tunés* ou utilisés dans des approches de requête *few-shot* avec plusieurs échantillons de notre jeu de données d'entraînement générés via les deux méthodes d'échantillonnage décrites précédemment. Nous utilisons les méthodes d'apprentissage suivantes.

- *Fine-Tuning* : les hyperparamètres ont été déterminés en testant toutes les combinaisons possibles (recherche en grille).
- *Pattern-Exploiting Training* (PET) [SCH 20] : il s'agit d'une approche d'apprentissage *few-shot* pour la classification de textes. Elle est basée sur le concept de phrases à trous (*cloze*). Dans notre cas d'utilisation, la requête (prompt) envoyée au modèle est formulée comme suit : « Le sentiment dominant exprimé dans le texte suivant : [Tweet] est {?} ». Le modèle de langage masqué utilisé tentera alors de remplir le {?} avec le sentiment approprié à partir d'une liste d'étiquettes possibles : positif, négatif ou neutre.

En comparant ces deux méthodes d'apprentissage et en évaluant leur efficacité avec différentes quantités d'exemples annotés, nous visons à mieux comprendre les exigences minimales en termes de

données pour obtenir des résultats fiables pour de la classification de la polarité des sentiments dans le domaine du tourisme.

4.3. Classification de tokens – Reconnaissance d'entités nommées spatiales et extraction de concepts thématiques fins

Pour la classification de tokens, englobant à la fois les lieux et les thèmes, nous adoptons une démarche similaire. Toutefois, dans le cadre de l'apprentissage *few-shot*, nous optons pour la méthode EntLM [MA 22], spécifiquement conçue pour la catégorisation de tokens. En complément du modèle XLM-T, nous intégrons également deux autres modèles à notre étude : XLM-RoBERTa (XLM-R) et BERT multilingue (mBERT) [DEV 18], ce dernier étant le modèle par défaut utilisé par EntLM. Pour rappel, l'un des objectifs de cette étude comparative est de déterminer la quantité minimale d'exemples annotés nécessaires pour justifier le passage de méthodes basées sur des règles (comme celle par correspondance de termes) rigides à des méthodes d'apprentissage plus dynamique mais nécessitant des exemples annotés pour l'entraînement. Plus spécifiquement, nous cherchons à déterminer le point de basculement à partir duquel les avantages de l'utilisation des techniques par apprentissage l'emportent sur leurs exigences en matière d'exemples.

Pour la classification de la polarité des sentiments, nous rapportons des résultats de précision, tandis que pour la classification de tokens (lieux et thèmes), nous utilisons la métrique micro-F1 calculée au niveau du segment tel que défini dans la tâche partagée CoNLL 2002 [TJO 02]. Tous les résultats rapportés sont la moyenne de trois exécutions initialisées aléatoirement.

4.4. Résultats

Classification de la polarité des sentiments. Les résultats obtenus pour la tâche de classification de la polarité des sentiments sont rapportés dans la figure 4. Pour rappel, cette tâche consiste à annoter chaque tweet avec une des classes d'annotation suivantes : positif, négatif ou neutre. L'axe x représente le nombre de tweets annotés utilisés pour l'entraînement, tandis que l'axe y indique les scores de précision obtenus en utilisant deux méthodes d'apprentissage différentes : le *fine-tuning* (F-T) et le *Pattern-Exploiting Training* (PET, une méthode de type *few-shot*). Deux modèles de langage sont utilisés XLM-T et XLM-T Sentiment. Les résultats montrent que le *fine-tuning* du modèle XLM-T Sentiment est plus performant que l'apprentissage *few-shot* (PET). Cette observation suggère qu'un *fine-tuning* réalisé sur un vaste ensemble de données multilingues pour la détection de sentiments, même issues de domaines très différents, contribue significativement à l'amélioration des performances sur des données liées au tourisme (figure 4.b). Cela se traduit par une nette amélioration des résultats dans des situations où les données sont peu abondantes. Le modèle XLM-T Sentiment, après *fine-tuning*, atteint une efficacité optimale avec seulement 10 exemples et parvient à égaler les performances obtenues en utilisant l'ensemble des exemples avec un entraînement basé sur 5 exemples seulement.

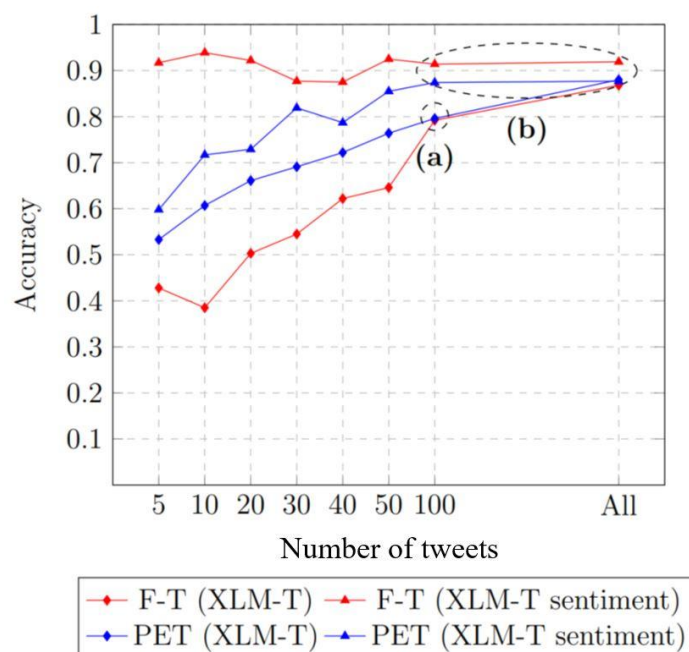


Figure 4. Classification de la polarité des sentiments - Echantillonnage *k-shot*

Lorsqu'on se concentre sur les approches qui exploitent uniquement nos propres données d'entraînement *Train*, l'apprentissage *few-shot* (PET) surclasse le *fine-tuning* du modèle XLM-T (jusqu'à 100 exemples annotés, voir figure 4.a). Cette observation met en évidence l'efficacité de PET, capable d'obtenir une précision importante même avec un nombre très restreint d'exemples. En revanche, figure 4.b, la performance de PET n'atteint pas celle obtenue par le *fine-tuning* du modèle XLM-T Sentiment, qui tire avantage d'un pré-entraînement sur un vaste corpus externe composé de milliers de tweets. En résumé, nous pouvons tirer deux enseignements significatifs de ces résultats: (1) lors de l'utilisation d'un modèle de sentiment déjà pré-entraîné comme XLM-T Sentiment, un jeu de données d'entraînement contenant aussi peu que 10 exemples est suffisant pour obtenir de bonnes performances pour faire de la classification de la polarité des sentiments dans le domaine du tourisme, ajouter plus d'exemples ne semble pas améliorer significativement la précision et (2) lors de l'emploi d'un modèle de langage masqué (MLM) comme XLM-T, PET semble être un choix préférable pour des scénarios à faible disponibilité de données, étant donné qu'une performance quasi optimale peut être atteinte avec 50 exemples.

Reconnaissance d'entités nommées spatiales. La figure 5 montre la performance de la reconnaissance d'entités nommées spatiales avec les deux méthodes d'échantillonnage (*k-shot* et pourcentage). Les versions *fine-tunées* de trois modèles : XLM-T, XLM-RoBERTa (XLM-R), et mBERT ont été comparées à EntLM, une méthode d'apprentissage *few-shot* pour les tâches de classification de tokens basée sur un modèle BERT multilingue. En comparant les résultats en utilisant les deux méthodes d'échantillonnage différentes (cf. figure 5), nous pouvons observer que lors de l'utilisation de toutes les données d'entraînement, *Train*, les quatre méthodes obtiennent des résultats relativement comparables. Cependant, dans un contexte de faible disponibilité des données, la méthode EntLM nécessite moins de tweets annotés. En d'autres termes, le BERT multilingue *fine-tuné* ne commence à surpasser EntLM qu'à partir de l'utilisation de 30% des données du sous-jeu de données *Train*. Globalement, le *fine-tuning* n'est compétitif qu'en utilisant l'échantillonnage par pourcentage. Nous pensons que cela pourrait être dû à la faible quantité de tokens n'étant pas des localisations générées par l'échantillonnage *k-shot*. En revanche, EntLM se comporte de manière assez robuste en utilisant un plus petit nombre de tweets avec les deux méthodes d'échantillonnage.

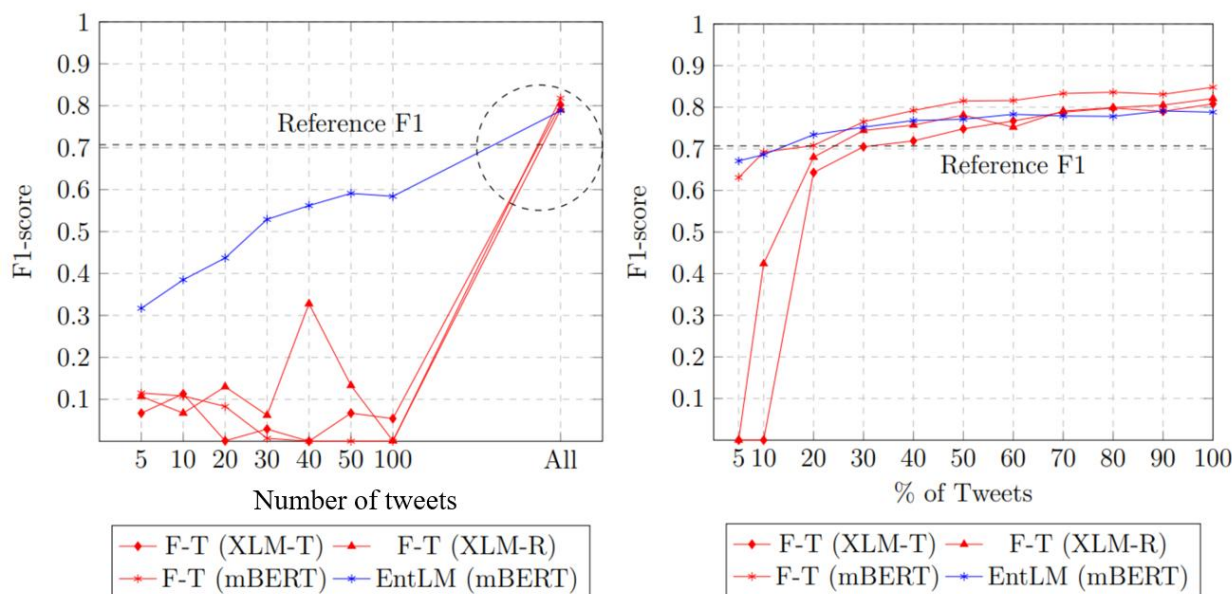


Figure 5. Reconnaissance d'entités nommées spatiales – Echantillonnage des données d'entraînement (sous-jeu Train) par k -shot (gauche) et pourcentage (droite)

Nous constatons également que le BERT multilingue *fine-tuné* et EntLM sont très nettement meilleurs comparés à l'approche basée sur des règles lorsqu'ils utilisent seulement environ 13% des tweets du sous-jeu de données *Train* (~ 200 tweets). De manière surprenante, les modèles généraux (mBERT et XLM-R) obtiennent de meilleures performances que ceux entraînés sur des données X/Twitter (XLM-T). Ce résultat montre que dans certains cas, les modèles généraux tendent à mieux s'appliquer à des domaines spécifiques que des modèles parfois trop spécialisés.

Extraction de concepts thématiques fins. C'est dans l'extraction de concepts thématiques fins, présentée dans la Figure 6, que l'approche d'apprentissage *few-shot* EntLM se distingue le plus. Pour cette tâche, impliquant la catégorisation de tokens dans un inventaire de 315 classes, EntLM se montre très compétitif. Ainsi, avec juste 5 exemples par classe (paramétrage 5-shot), il obtient un score F1 de 0.760. De même, il égale les résultats de l'approche par correspondance de termes avec un entraînement sur seulement 50 exemples. Ces scores indiquent une forte capacité à identifier avec précision les concepts touristiques, comme en témoignent les valeurs de précision élevées allant de 0.80 à 0.913. Bien que les résultats obtenus avec l'approche par correspondance de termes soient similaires, EntLM est légèrement supérieur en termes de rappel tout en étant légèrement moins bon en précision. Néanmoins, la performance d'EntLM est prometteuse pour éviter l'effort d'annotation manuelle ou le développement complexe d'approches basées sur des règles pour des tâches de classification de tokens fines spécifiques à un domaine.

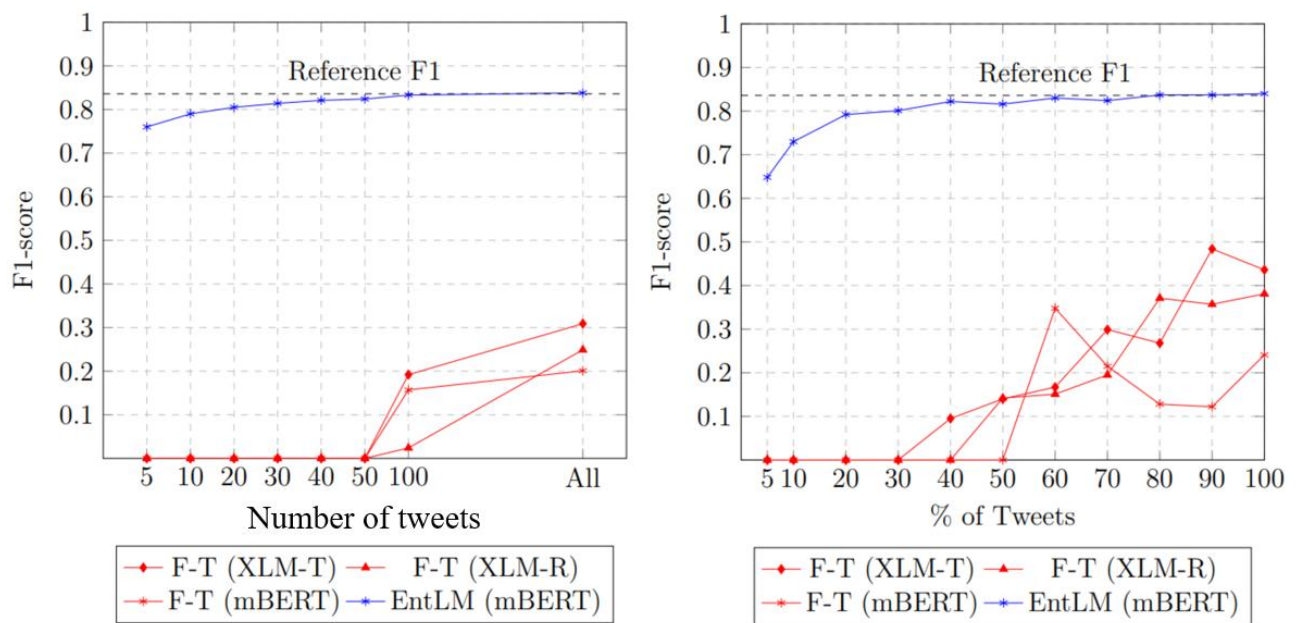


Figure 6. Extraction de concepts thématiques fins – Echantillonnage des données d'entraînement par *k*-shot (gauche) et pourcentage (droite)

4.5. Discussion et limitations

Après ces expérimentations, penchons-nous sur les conclusions obtenues pour discuter des principaux éléments ayant émergé. Nous abordons également les limitations potentielles qui pourraient avoir affecté les résultats. Nos résultats montrent que le *fine-tuning* sur des modèles déjà pré-entraînés, comme XLM-T Sentiment, peut s'avérer très efficace pour la détection de sentiments, même avec un faible volume de données spécifiques au domaine. Cette observation souligne l'importance d'un pré-entraînement riche et varié pour améliorer la performance des modèles dans des contextes de données limitées. Cependant, il est crucial de souligner que l'efficacité de cette approche dépend fortement de la disponibilité de données pré-entraînées pertinentes et de la capacité du modèle à s'adapter au contexte du tourisme.

Pour la tâche de reconnaissance d'entités nommées spatiales (une classe d'annotation *localisation*, mais beaucoup de mots labels associé à cette dernière), nos expérimentations indiquent que les méthodes d'apprentissage *few-shot*, comme EntLM, peuvent rivaliser avec des techniques de *fine-tuning* plus gourmandes en données. Cette observation suggère que des approches d'apprentissage plus légères (comme celles basées sur le principe du *few-shot* comme EntLM) peuvent être suffisantes pour traiter des tâches de NER spécifiques, en particulier dans des contextes où les données annotées sont rares ou coûteuses à obtenir.

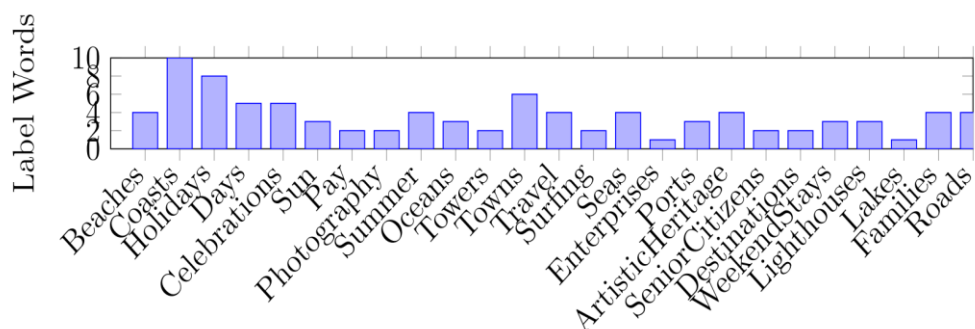


Figure 7. Nombre de mots labels pour les concepts les plus fréquents du corpus

L'extraction de concepts thématiques fins (315 classes correspondant aux concepts du thésaurus de l'OMT, chacun ayant un nombre restreint de mots labels, voir figure 7), en revanche, présente des défis uniques en raison de la granularité et de la spécificité des classes impliquées. Bien que la méthode *few-shot* EntLM ait démontré une efficacité certaine, il est important de reconnaître que la précision de cette approche dépend fortement de la représentativité des exemples d'entraînement et de leur alignement avec les concepts thématiques du domaine d'application, ici le tourisme.

Le jeu de données ayant été créé et validé, les techniques de reconnaissance de sentiments, d'entités nommées de lieux et de concepts thématiques ayant été optimisées, nous avons pu analyser ces résultats et calculer des indicateurs de média sociaux originaux. La présentation de ces calculs dépasse le cadre de cet article. Le lecteur intéressé pourra se reporter à [MAS 24]. Ainsi, dans la section suivante, nous présentons TextBI, le tableau de bord générique multimodal, dédié à la visualisation de ces indicateurs, que nous avons conçu et développé.

5. TextBI

Après avoir présenté la structure du tableau de bord TextBI et ses aspects conceptuels (section 5.1), nous pourrions l'expérimenter sur notre jeu de données dans le domaine du tourisme (section 5.2).

5.1. TextBI : une interface générique de construction de tableaux de bord

TextBI est une plateforme multimodale et générique qui permet de visualiser, de manière simple ou agrégée, les annotations multidimensionnelles présentes dans des corpus extraits des réseaux sociaux. Cette plateforme est multisource et indépendante du domaine. Elle repose sur un modèle de données [MAS 23] capable de représenter n'importe quel corpus annoté selon nos dimensions d'analyse. Ainsi, la plateforme prend la forme d'un tableau de bord reposant sur quatre dimensions centrales d'analyse : la dimension spatiale (données de localisation présentes dans le corpus), temporelle (horodatages et périodes temporelles), thématique (phénomènes décrits dans une ressource sémantique) et personnelle (données et profils associés à l'utilisateur). TextBI prend également en charge des données contextuelles d'enrichissement, telles que les informations liées au sentiment et à l'engagement (par exemple, *likes* et partages). TextBI fonctionne avec tout domaine préalablement modélisé sous la forme d'une ressource sémantique hiérarchisée (dictionnaire, thésaurus, ontologie, etc.).

Cette plateforme s'adresse aux spécialistes d'un domaine d'activité donné en situation de recherche d'informations. Par exemple, dans le secteur du tourisme, les offices de tourisme souhaitent identifier les principales activités touristiques dans une région d'intérêt, les villes fréquemment associées et les sentiments moyens exprimés par les visiteurs. Ces informations peuvent les assister dans la prise de décisions éclairées. Nous faisons l'hypothèse que des spécialistes de domaines d'activité différents peuvent avoir des besoins similaires comme, par exemple, identifier les fréquences, les associations et les séquences d'annotations dans des corpus de réseaux sociaux.

La figure 8 donne un aperçu des différentes fonctionnalités offertes par la plateforme. Nous renvoyons le lecteur à [MAS 24] pour de plus amples détails. Nous précisons que 3 modes de visualisation existent dans la plateforme présentée succinctement ci-après :

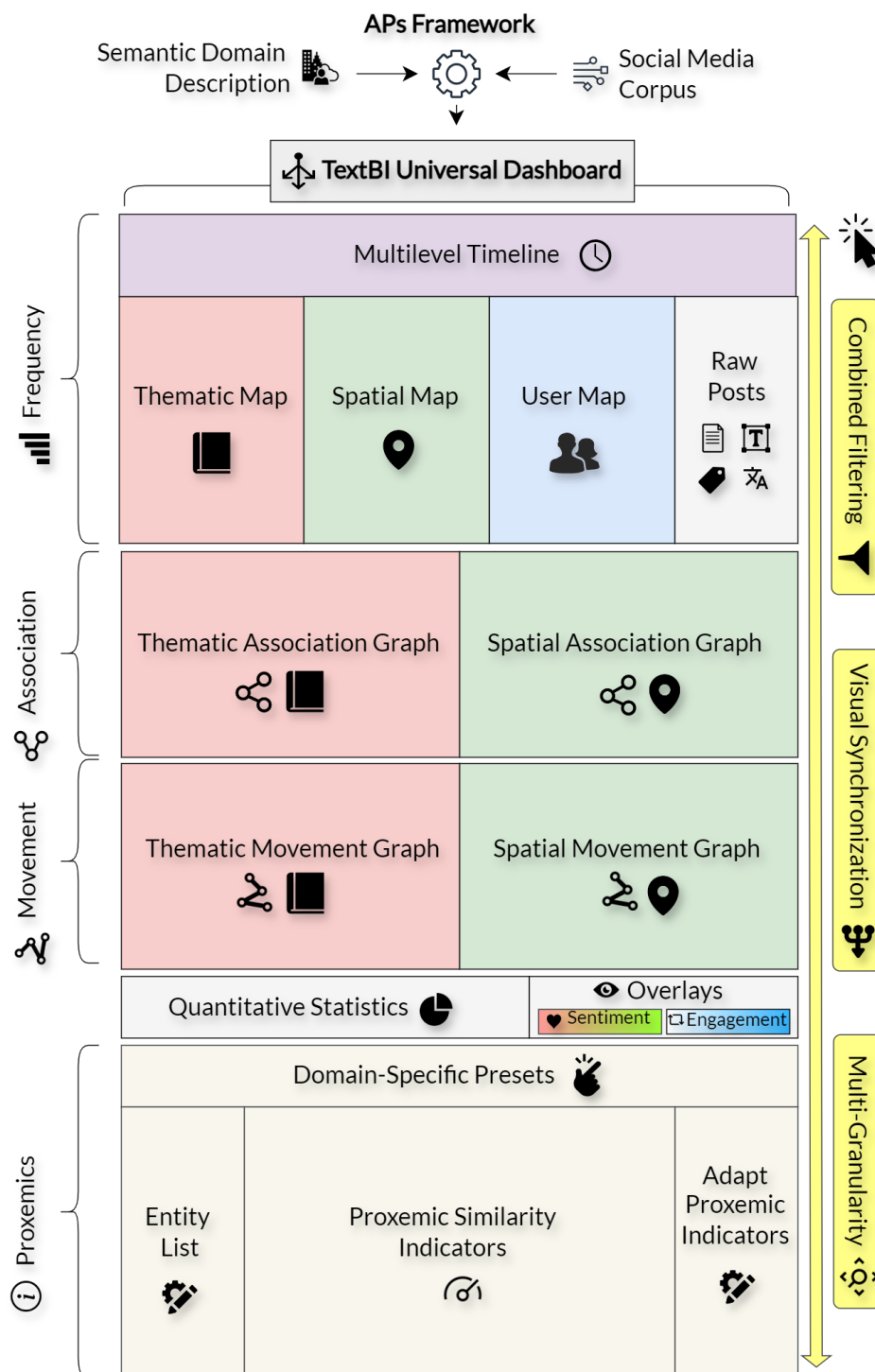


Figure 8. Les fonctionnalités de TextBI

– Le mode « Fréquence » (figure 8 - *Frequency*) sert d'interface principale à la plateforme. Il met en évidence les fréquences spatiales, temporelles et thématiques dans les publications ainsi que la distribution des annotations à travers quatre visualisations majeures : une carte thématique, une carte spatiale, la carte des utilisateurs et la frise chronologique. De plus, les textes des publications sont affichés dans le panneau des publications, sur la droite.

– Les modes « Association » et « Mouvement » : dans le mode « Association » (figure 8 - *Association*), le tableau de bord présente des représentations visuelles illustrant les connexions entre les entités en se basant sur leurs cooccurrences dans les publications. Ce mode peut également se transformer pour afficher les mouvements (figure 8 - *Thematic movement graph* et *Spatial movement graph*), se concentrant sur le séquençement des entités dans les trajectoires spatio-thématiques des utilisateurs, par exemple, la transition d'un concept thématique à un autre.

– Le mode « Proxémique » (figure 8 - *Proxemics*) offre aux utilisateurs la capacité d’analyser des corpus de données de réseaux sociaux via une approche originale de calcul d’indicateurs [HAL 66]. Les utilisateurs peuvent alors sélectionner une entité comme référence (un utilisateur, une cohorte, un concept thématique, un lieu ou une période de temps) et observer la proximité (exprimée via une mesure de similarité proxémique) d’autres entités par rapport à cette référence.

De plus, dans TextBI, les visuels prennent en charge des superpositions (figure 8 - *Overlays*) pour présenter les données contextuelles de sentiment et d’engagement. TextBI gère également des interactions couramment exploitées dans les outils de *Business Intelligence* (BI), garantissant une synchronisation fluide des visualisations. Elle propose des combinaisons de filtrage multidimensionnelles telles que « spatial-temporel », « spatial-thématique » et « utilisateur-temporel ». Elle dispose également d’une vue « Statistiques » (figure 8 - *Quantitative Statistics*) montrant le volume de données associé aux filtres actifs, y compris les décomptes des publications, des utilisateurs, des concepts et des lieux, la plage temporelle actuelle des publications, le niveau total d’engagement, et les sentiments dominants. Ainsi, TextBI sert d’outil d’affichage et d’agrégation de données, fournissant des analyses statistiques et calculant des indicateurs, facilitant ainsi le travail d’analyse de l’utilisateur final.

Nous pouvons désormais expérimenter la plateforme sur un jeu de données lié au Tourisme.

5.2. TextBI : une mise à l’épreuve dans le domaine du Tourisme

Nous rappelons que le jeu de données comprend 2 961 tweets correspondant à 624 utilisateurs (référence section 3). Ainsi, dans les exemples suivants, nous illustrons quelques modes et leurs visuels en chargeant ce corpus de tweets touristiques.

Nous expérimentons le mode Fréquence (voir figure 9), le mode Association (voir figure 10), le mode Proxémique (voir figure 11) et l’interactivité du tableau de bord (voir figure 12) sur l’ensemble des données touristiques.

5.2.1. Le mode Fréquence

Ce mode affiche la distribution des entités dans l’ensemble du corpus, à savoir les thèmes, les lieux, les dates et les utilisateurs.

– La **carte thématique** (figure 9.1) montre que lorsque l’on compare les concepts de tweet au thésaurus des activités de tourisme et de loisirs (Organisation mondiale du tourisme, 2002), on constate que les concepts liés au patrimoine touristique sont les plus fréquents, représentant environ 40 % des concepts découverts (représentés en bleu), dont beaucoup sont liés aux ressources naturelles telles que le littoral et la mer.

– La **carte spatiale** (figure 9.2) est établie au niveau municipalité et sous forme de carte choroplèthe. Nous observons un point chaud de tweets dans trois municipalités de la partie la plus septentrionale de la région. Ces communes proches semblent être populaires auprès des visiteurs.

– La **carte des utilisateurs** (figure 9.3) montre 655 utilisateurs, répartis en 5 langues. Le français est représenté en bleu, l’espagnol en vert, l’anglais en orange, le basque en violet et l’italien en jaune.

– La **frise chronologique** (figure 9.4) offre une visualisation du volume de messages par jour pour l’ensemble des données, segmenté en différents moments de la journée tels que le matin (représenté en cyan), l’après-midi (représenté en orange) et le soir (représenté en violet). Nous observons un pic de tweets entre le 24 et le 28 juillet

– La **liste des messages** (figure 9.5) affiche les messages actuellement chargés. Actuellement, aucun filtrage n’est appliqué, elle contient donc tous les tweets. La **vue Statistiques** (figure 9.6) montre

que ces tweets appartiennent à 655 utilisateurs, couvrant 29 municipalités uniques et 365 concepts uniques du thésaurus du tourisme, au cours de l'été 2019.

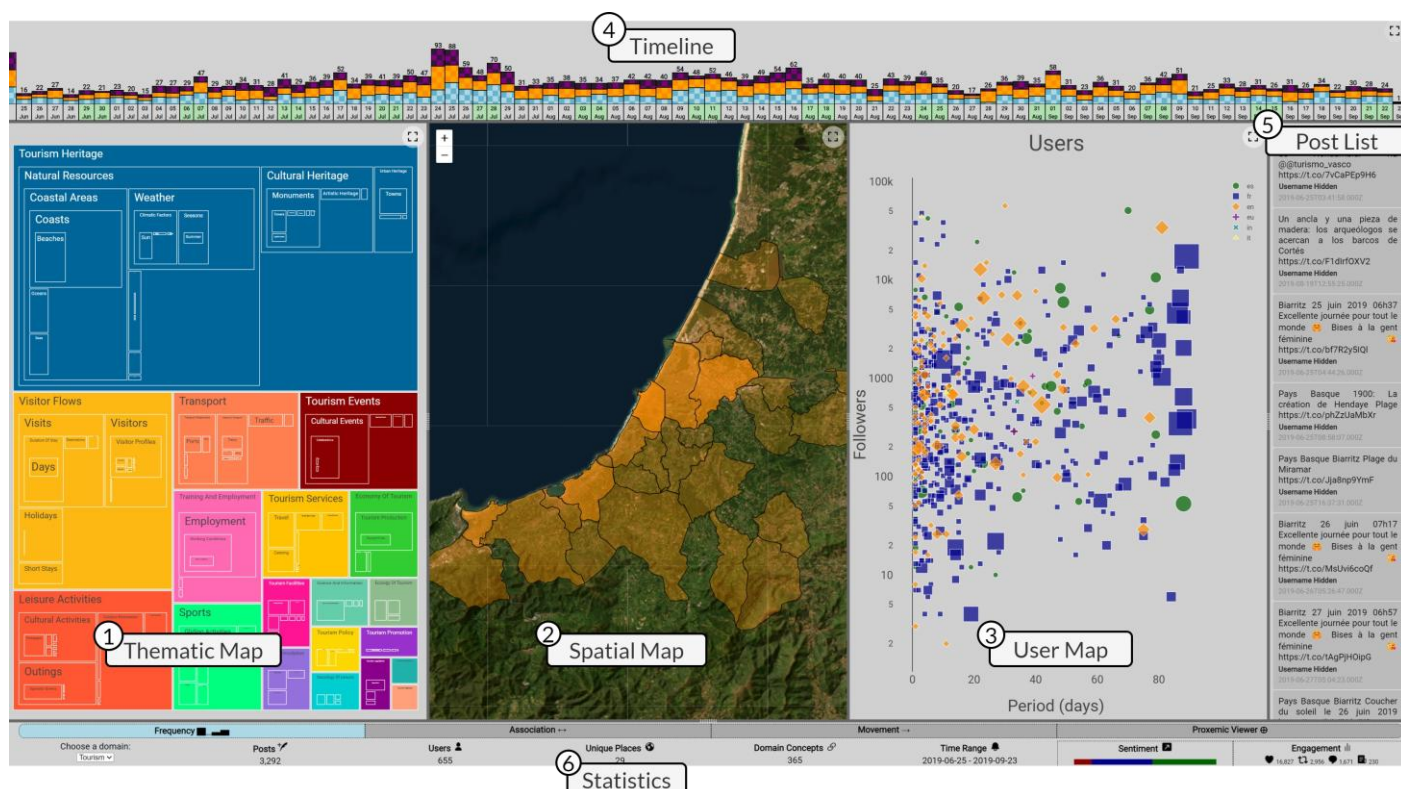


Figure 9. Capture d'écran du mode Fréquence

5.2.2. Le mode Association

La figure 10 présente le mode Association. Ce mode se concentre sur la cooccurrence des thèmes et des lieux pour analyser les entités associées.

Le **graphique d'association thématique** (figure 10.1) montre les concepts thématiques qui sont souvent mentionnés ensemble dans les tweets de notre ensemble de données. Comme on pouvait s'y attendre, le soleil, la plage, le surf, la mer et d'autres concepts côtiers sont fortement liés (liens plus épais).

Le **graphique d'association spatiale** (figure 10.2) montre les municipalités souvent mentionnées ensemble. Ces municipalités sont superposées sur une carte spatiale à des fins de géoréférencement. Comme nous pouvons le voir, Biarritz et Bayonne sont les municipalités les plus associées.

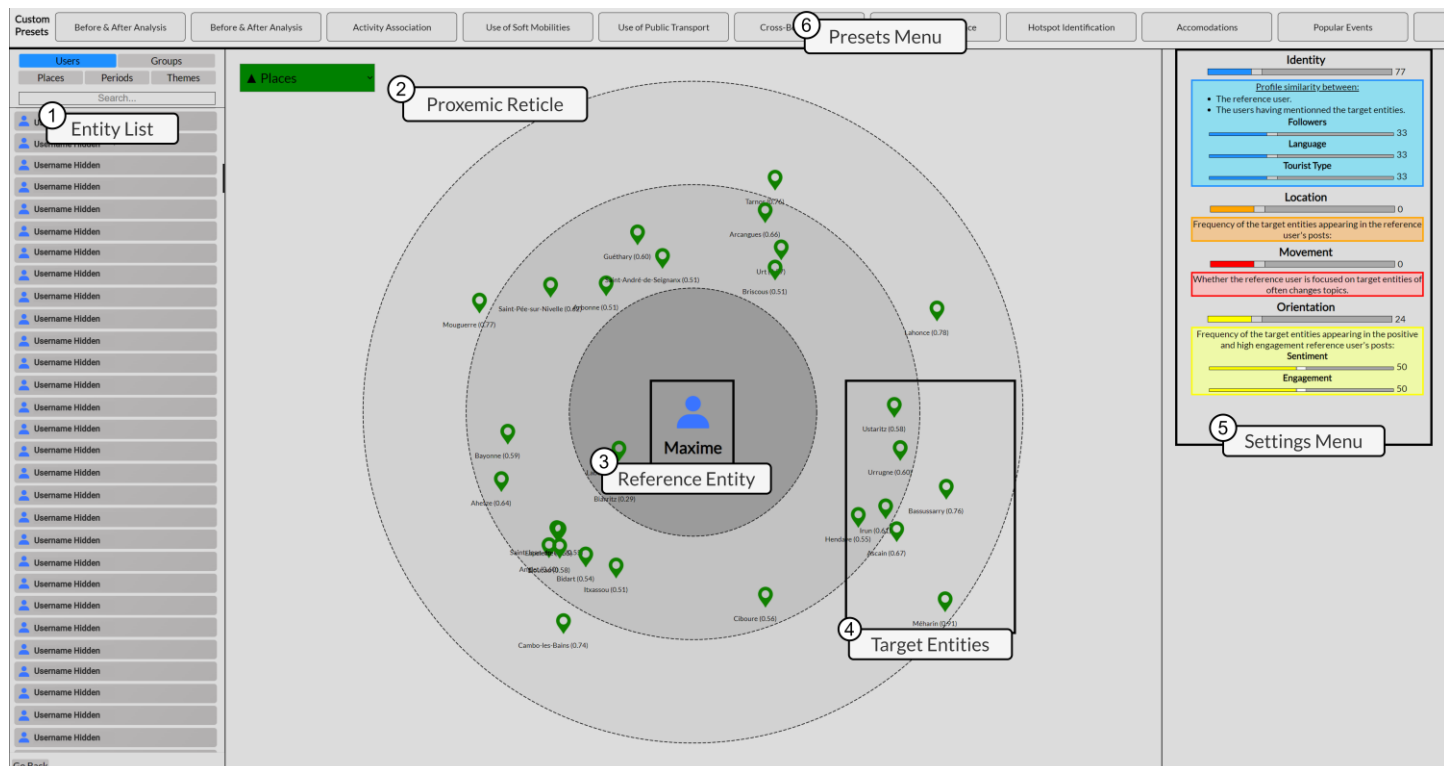


Figure 11. Capture d'écran du mode Proxémique

Cette vision proxémique est particulièrement utile pour plusieurs raisons. La plus importante est qu'elle prend en compte plusieurs critères. Les critères sont basés sur les cinq dimensions proxémiques du modèle DILMO [GRE 11] : Distance (physique, temporelle ou sémantique), Identité (profils des visiteurs), Localisation (lieux, thèmes, dates), Mouvement (trajectoires), et Orientation (sentiment et engagement). Ces dimensions sont interprétées différemment selon les entités comparées (User, Group, Place, Period, Theme) deux à deux. Nous n'avons pas détaillé l'aspect calculatoire pour des raisons de concision. Contrairement aux autres vues de TextBI, qui se concentrent sur des indicateurs spécifiques (par exemple, la fréquence, les associations, etc.), la vue proxémique permet de combiner plusieurs indicateurs dans le même visuel afin de répondre aux exigences complexes de l'utilisateur final. Grâce à la modulation des cinq dimensions proxémiques et au choix des entités de référence et des entités cibles, l'utilisateur final dispose d'un plus grand contrôle sur le visuel qui sera produit par rapport aux autres vues.

5.2.4. Interactions et visuels synchronisés

La dernière capture d'écran (figure 12) illustre l'interactivité du tableau de bord et présente divers types d'interactions que l'on trouve couramment dans les outils de BI. En outre, tous les visuels sont synchronisés et mis à jour de manière dynamique sans qu'il soit nécessaire de les recharger.

Les utilisateurs du tableau de bord peuvent utiliser le filtrage combiné pour restreindre la quantité de données présentées et effectuer des analyses plus fines. Prenons l'exemple de la figure 12 : (1) L'utilisateur sélectionne la période du 24 au 27 juillet dans la ligne de temps ; (2) La carte thématique se met à jour et affiche une plus forte concentration du concept thématique « Célébration » ; (3) En cliquant davantage sur le concept de célébration, la carte spatiale met en évidence un point chaud dans la municipalité de Bayonne ; (4) En examinant la liste des messages, l'utilisateur peut observer que cela coïncide avec les Fêtes de Bayonne, un événement local qui attire plus d'un million de participants.

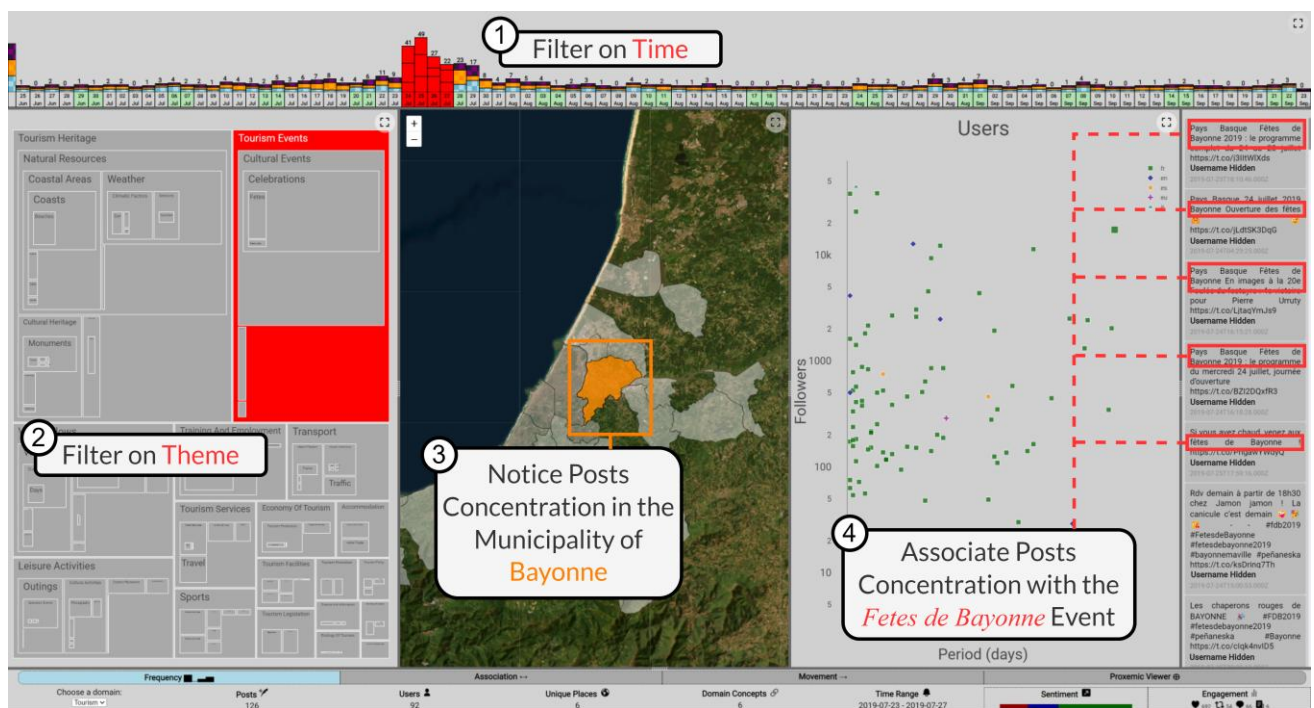


Figure 12. Capture d'écran du mode Proxémique

6. Conclusion et perspectives

Cet article propose une étude comparative de plusieurs techniques d'apprentissage sur 3 tâches d'extraction de connaissance : la classification de la polarité des sentiments, la reconnaissance d'entités nommées spatiales et de concepts thématiques fins dans le domaine du tourisme sur les réseaux sociaux. L'objectif est de déterminer la meilleure stratégie pour obtenir des résultats performants tout en réduisant au maximum les annotations manuelles et le développement de méthodes basées sur des règles, souvent complexe. Pour cela, un nouveau jeu de données multilingue, spécifique au tourisme sur *X/Twitter*, a été créé. Ce jeu de données, qui est disponible à des fins de recherche et à la demande, comprend des annotations au niveau du texte sur le sentiment et des tokens sur les lieux et concepts thématiques. Les résultats de notre étude confirment que l'apprentissage *few-shot* est particulièrement efficace pour ces tâches avec peu d'exemples annotés. Ce résultat est pertinent non seulement pour le développement d'applications spécifiques au tourisme mais aussi pour d'autres domaines nécessitant du TALN. Des recherches supplémentaires sont cependant nécessaires pour valider la généralisabilité de nos résultats dans d'autres domaines d'application.

Nous avons également proposé TextBI (une vidéo de démonstration est disponible à l'adresse suivante : maxime-masson.github.io/TextBI), une interface de type *Business Intelligence* pour visualiser les indicateurs résultant d'analyses des données issues de ces traitements TALN. TextBI a été présenté aux acteurs de l'industrie du tourisme, notamment à travers des tableaux de bord dynamiques mettant en évidence les entités (par exemple, lieux et thèmes) extraites des réseaux sociaux, en lien avec des données contextuelles comme le sentiment. La plateforme a été évaluée de manière positive et s'est révélée plutôt innovante. Une évaluation qualitative du tableau de bord générique a été menée, elle est décrite dans le manuscrit de thèse de Maxime Masson [MAS 24]. Cependant, sur le plan expérimental, il est nécessaire de tester la plateforme avec des ensembles de données massifs afin d'observer ses performances à l'échelle avec des ensembles de données de l'ordre de millions de messages, et potentiellement hétérogènes, c'est-à-dire provenant de plusieurs plateformes de médias sociaux simultanément. Un examen approfondi des schémas de couleurs utilisés dans le tableau de bord est crucial, en particulier en ce qui concerne les gradients de la carte, les visuels de la ligne de temps et la coloration des entités proxémiques. Les commentaires des utilisateurs indiquent que ces fonctionnalités ne sont pas toujours intuitives, ce qui souligne la nécessité de les améliorer. Ainsi, la recherche sur la façon dont les couleurs affectent la perception des données et la justification

de la sélection de la palette de couleurs [AHM 21], [SZA 17] pourrait fournir des indications précieuses pour l'adoption de schémas de couleurs plus adaptés dans TextBI. De plus, nous souhaitons rendre le tableau de bord personnalisable par les utilisateurs finaux. Cela pourrait se faire grâce à une bibliothèque de modules de visualisation (par exemple, des cartes thématiques, des lignes de temps, des graphiques d'association, etc.) qui peuvent être glissés sur le tableau de bord principal pour créer des vues personnalisées.

Bibliographie

- [ACH 23] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- [AHM 21] Ahmad, J., Huynh, E., and Chevalier, F. (2021). When red means good, bad, or canada: exploring people's reasoning for choosing color palettes. In *2021 IEEE Visualization Conference (VIS)*, pages 56–60. IEEE.
- [BAR 20] Barbieri, F., Camacho-Collados, J., Neves, L., Espinosa-Anke, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv:2010.12421*
- [BAR 22] Barbieri F., Espinosa Anke L., Camacho-Collados J. (2022). XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. *LREC 2022*, p. 258-266.
- [BOO 01] BOOTH B., MITCHELL A. (2001). Getting started with ArcGIS. Redlands : ESRI Press, 253 p.
- [BOW 15] Bowman S.R., Angeli G., Potts C., Manning C.D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- [BRO 20] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.
- [BUD 18] Budzianowski P., Wen T.-H., Tseng B.-H., Casanueva I., Ultes S., Ramadan O., Gašić M. (2018). MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. *EMNLP 2018*, novembre 2018, p. 5016-5026.
- [CHA 21] Chantrapornchai C., Tunsakul A. (2021). Information extraction on tourism domain using SpaCy and BERT. *ECTI Transactions on Computer and IT*, vol. 15, n° 1, p. 108-122.
- [CHE 20] Cheng X., Wang W., Bao F., Gao G. (2020). MTNER: A Corpus for Mongolian Tourism Named Entity Recognition. *CCMT 2020*, October 10-12, 2020, Springer, p. 11-23.
- [CHU 24] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70), 1-53.
- [CON 19] Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- [CRA 21] Crawford M., Khoshgoftaar T.M. (2021). Using inductive transfer learning to improve hotel review spam detection. *IRI 2022, IEEE*, p. 248-254.
- [DAT 18] DATIG I., WHITING P. (2018). "Telling your library story: tableau public for data visualization". *Library Hi Tech News*, vol. 35, no 4, p. 6-8.
- [DEV 18] Devlin J., Chang M.-W., Lee K., Toutanova K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [DER 16] Derczynski L., Bontcheva K., Roberts I. (2016). Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. *COLING 2016: Technical Papers*, p. 1169-1179.
- [DES 21] DESAI Z., ANKLESARIA K., BALASUBRAMANIAM H. (2021). "Business intelligence visualization using deep learning based sentiment analysis on Amazon review data". *12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, p. 1-7. IEEE.
- [ENR 22] Enríquez M.P., Mencía J.A., Segura-Bedmar I. (2022). Transformers Approach for Sentiment Analysis: Classification of Mexican Tourists Reviews from TripAdvisor.
- [FER 16] FERRARI A., RUSSO M. (2016). Introducing Microsoft Power BI. Redmond : Microsoft Press.
- [GAL 06] Galliano S., Geoffrois E., Gravier G., Bonastre J.-F., Mostefa D., Choukri K. (2006). Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. *LREC*, Citeseer, p. 139-142.

- [GO 09] Go A., Bhayani R., Huang L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report*, Stanford, vol. 1, n° 12, p. 2009.
- [GRE 11] Greenberg, S., Marquardt, N., Ballendat, T., Diaz-Marino, R., and Wang, M. (2011). Proxemic interactions: the new ubicomp? interactions, 18(1):42–50.
- [HAL 66] Hall, E.T., Hall, E.T.: The hidden dimension, vol. 609. Anchor (1966)
- [HAL 68] Hall, E. T., Birdwhistell, R. L., Bock, B., Bohannon, P., Diebold Jr, A. R., Durbin, M., Edmonson, M. S., Fischer, J., Hymes, D., Kimball, S. T., et al. (1968). Proxemics [and comments and replies]. *Current anthropology*, 9(2/3):83–108.
- [HAN 10] HANSOTI B. (2010). Business Intelligence Dashboard in Decision Making. Purdue e-Pubs, Purdue University.
- [HAR 23] Hartmann, J., Heitmann, M., Siebert, C., Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *Int. Journal of Research in Marketing*, 40(1):75–87
- [HEA 22] HEALEY C. G., RAMASWAMY S. (2022). "Twitter Sentiment Visualization". https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/
- [JIA 23] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L  . R., Lachaux, M., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T. & Sayed, W. E. (2023). Mistral 7B. CoRR, abs/2310.06825. doi: 10.48550/ARXIV.2310.06825
- [LOU 14] LOUB  RE L., RATINAUD P. (2014). Documentation iramuteq 0.6 alpha 3 version 0.1. R  cup  r   sur http://www.iramuteq.org/documentation/fichiers/documentation_19_02_2014.pdf
- [MA 22] Ma R., Zhou X., Gui T., Tan Y., Li L., Zhang Q., Huang X. (2022). Template-free Prompt Tuning for Few-shot NER. *NAACL 2022: Human Language Technologies*, p. 5721-5732.
- [MAS 22] Masson M., Sallaberry C., Agerri R., Bessagnet M.-N., Roose P., Le Parc Lacayrelle A. (2022). A Domain-Independent Method for Thematic Dataset Building from Social Media: The Case of Tourism on Twitter. *Web Information Systems Engineering*, p. 11-20.
- [MAS 23] Masson, M., Roose, P., Sallaberry, C., Agerri, R., Bessagnet, M.-N., and Lacayrelle, A. L. P. (2023). Aps: A proxemic framework for social media interactions modeling and analysis. In *Advances in Intelligent Data Analysis XXI: 21st International Symposium on Intelligent Data Analysis, IDA 2023, Louvain-la-Neuve, Belgium, April 12–14, 2023, Proceedings*, pages 287–299. Springer.
- [MAS 24] M. Masson (2024). Generic Framework for the Multidimensional Processing and Analysis of Social Media Content: A Proxemic Approach. Computer Science Ph.D. Thesis. September 2024
- [MAY 00] MAYNARD D., CUNNINGHAM H., BONTCHEVA K., CATIZONE R., DEMETRIOU G., GAIZAUSKAS R., HAMZA O., HEPPLER M., HERRING P., MITCHELL B., OAKES M., PETERS W., SETZER A., STEVENSON M., TABLAN V., URSU C., WILKS Y. (2000). A survey of uses of GATE. Technical Report CS-00-06, Department of Computer Science, University of Sheffield.
- [MAY 12] Maynard D., Bontcheva K., Rout D. (2012). Challenges in developing opinion mining tools for social media, *Workshop Programme*, p. 15.
- [MEN 16] MENKE K., SMITH R. JR, PIRELLI L., VAN HOESEN J. (2016). Mastering QGIS. Birmingham : Packt Publishing Ltd, 486 p.
- [MIN 21] Min B., Ross H., Sulem E., Veyseh A.P.B., Nguyen T.H., Sainz O., Agirre E., Heinz I., Roth D. (2021). Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.
- [MUR 03] MURTHY K.S. R., AMMINEDU E., VENKATESWARA RAO V. (2003). "Integration of thematic maps through GIS for identification of groundwater potential zones". *Journal of the Indian Society of Remote Sensing*, vol. 31, p. 197-210.
- [ORL 20] ORLOVSKIY D., KOPP A. (2020). "A business intelligence dashboard design approach to improve data analytics and decision making". *CEUR Workshop Proceedings*, vol. 2833
- [PER 21] P  rez, J. M., Giudici, J. C., Luque, F. (2021). pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. *arXiv e-prints*, arXiv-2106.
- [RAJ 22] RAJAONARIVO L., MINE T., ARAKAWA Y. (2022). "Coupling of semantic and syntactic graphs generated via tweets to detect local events". 12th International Congress on Advanced Applied Informatics (IIAI- AAI), IEEE, p. 128-133.
- [REH 11] REHUREK R., SOJKA P. (2011). Gensim – statistical semantics in Python. Poster.

- [ROS 15] Rosenthal S., Nakov P., Kiritchenko S., Mohammad S., Ritter A., Stoyanov V. (2015). SemEval-2015 Task 10: Sentiment Analysis in Twitter. *SemEval 2015*, p. 451-463.
- [SAI 23] Sainz, O., García-Ferrero, I., Agerri, R., de Lacalle, O. L., Rigau, G., & Agirre, E. (2023). Gollie: Annotation guidelines improve zero-shot information-extraction. *arXiv preprint arXiv:2310.03668*.
- [SAI 13] Saif H., Fernandez M., He Y., Alani H. (2013). Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold.
- [SCH 20] Schick T., Schütze H. (2020). Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.
- [SER 21] Serna A., Soroa A., Agerri R. (2021). Applying Deep Learning Techniques for Sentiment Analysis to Assess Sustainable Transport. *Sustainability*, vol. 13, n° 4, article 2397.
- [SEE 23] Seethal (2023). Sentiment analysis generic dataset. *En ligne le 23 mars 2023*.
- [SHU 16] SHUKLA A., DHIR S. (2016). "Tools for data visualization in business intelligence: case study using the tool QlikView". *Information Systems Design and Intelligent Applications: Proceedings of Third International Conference INDIA 2016*, vol. 2, Springer, p. 319-326.
- [SUN 19] Sun, C., Qiu, X., Xu, Y., Huang, X. (2019). How to fine-tune bert for text classification? *Chinese Computational Linguistics: 18th China National Conference, CCL 2019*.
- [SZA 17] Szafir, D. A. (2017). Modeling color difference for visualization design. *IEEE transactions on visualization and computer graphics*, 24(1):392–401.
- [TAU 08] Taulé, M., Martí, M. A., Recasens, M. (2008). Ancora: Multilevel annotated corpora for Catalan and Spanish. *LREC*. Vol. 2008, pp. 96-101.
- [TJO 02] Tjong Kim Sang, Erik F. (2002). Introduction to the CoNLL-2002 Shared Task. *CoNLL 2002*
- [TOP 23] Toporkov O., Agerri R. (2023). On the Role of Morphological Information for Contextual Lemmatization. *arXiv preprint*, vol. 2302.00407.
- [TOU 23] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- [TUN 22] Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.
- [VAN 13] VAN ECK N. J., WALTMAN L. (2013). VOSViewer manual. Univeristeit Leiden, 53 p.
- [VAS 20] VASHISHT V., DHARIA P. (2020). "Integrating chatbot application with qlik sense business intelligence (BI) tool using natural language processing (NLP)". *Micro-Electronics and Telecommunication Engineering: Proceedings of 3rd ICMETE 2019*, Springer.
- [WEL 14] WELSH M. E. (2014). "Review of voyant tools". *Collaborative Librarianship*, vol. 6, no 2, p. 96-98.
- [ZEN 14] Zeng B., Gerritsen R. (2014). What do we know about social media in tourism? A review. *Tourism Management Perspectives*, vol. 10, Elsevier, p. 27-36.