Stratégies optimales pour l'analyse multidimensionnelle de contenus multilingues issus des réseaux sociaux

Maxime Masson[1,2], Rodrigo Agerri[2], Christian Sallaberry[1], Marie-Noelle Bessagnet[1], Philippe Roose[1], Annig Le Parc Lacayrelle[1]

1. LIUPPA, E2S, Université de Pau et des Pays de l'Adour (UPPA)
2. Centre HiTZ – Ixa, Université du Pays Basque EHU/UPV

RESUME. 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 différentes 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.

Mots-clés : Tourisme, Apprentissage Few-Shot, Modèle de Langage Masqué (MLM), Multilinguisme, Science Sociale Informatique, Traitement Automatique du Langage Naturel

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 (Zeng *et al*., 2014). Toutefois, l'analyse de grandes quantités de données issues des réseaux sociaux représente un défi majeur (Maynard *et al*., 2012), 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 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 (Rosenthal *et al*., 2015). 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, plus adaptatives aux variations linguistiques (Min *et al.,* 2021), permettent une analyse plus dynamique et adaptative, en opposition aux approches basées sur des règles, souvent ad hoc et rigides. 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.

Dans cet article, 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. 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*).

Nos principales contributions sont les suivantes : (1) nous proposons un **nouveau jeu de données de tweets touristiques**. 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 loisirs*[1] 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 *et al.,* 2022) avec pour objectif d'établir quelle méthode est la plus efficace pour chaque tâche d'extraction de connaissances ; (3) finalement, nous expérimentons avec différentes méthodes d'échantillonnage de données pour déterminer **combien d'exemples annotés sont réellement**

[1] https://www.e-unwto.org/doi/book/10.18111/9789284404551

nécessaires pour obtenir des résultats compétitifs sur chacune des trois tâches. L'objectif est d'éviter aux chercheurs d'annoter manuellement une trop grande quantité de données par rapport à leurs besoins.

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. La section 3 couvre le processus de construction et d'annotation de notre jeu de données. La section 4 décrit la configuration expérimentale de notre étude. Dans la section 5, nous présentons une analyse comparative des différentes approches de TALN pour chacune des trois tâches décrites précédemment. Les résultats et les limitations sont discutés dans la section 6. Enfin, la section 7 présente les perspectives futures.

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 et al., 2021). 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) (Devlin et al., 2018) et XLM-RoBERTa (Conneau et al., 2019) 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 (Toporkov et Agerri, 2023).

2.1. Travaux antérieurs sur les réseaux sociaux dans le domaine du tourisme

Une approche fréquemment utilisée consiste à ***fine-tuner*** (apprentissage par transfert) les modèles de langage masqués pour des tâches spécifiques au domaine (Sun *et al.,* 2019). Cela se fait en altérant les poids du modèle pour l'adapter à la nouvelle tâche, via un enrichissant 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 la détection de *spam* dans les avis d'hôtels (Crawford et Khoshgoftaar, 2021) ou encore l'analyse des sentiments dans les avis touristiques (Enríquez *et al.,* 2022) ou dans le domaine du transport durable (Serna *et al.,* 2021). 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 (Cheng *et al.,* 2020). 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 (Chantrapornchai et Tunsakul, 2021). L'un des principaux défauts de l'approche *fine-tuning* est qu'elle nécessite généralement un volume conséquent d'exemples annotés pour être efficace. Cela génère une charge de travail d'annotation importante pour les chercheurs.

2.2. Aborder le manque de données annotées spécifiques au domaine

Face au défi des données annotées limitées, 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), utiles lorsque les données abondantes font défaut.

Le Pattern-Exploiting Training (PET) est un exemple d'apprentissage *few-shot* dédié à la classification de texte, où le modèle est guidé par des descriptions de tâches en langage naturel et des phrases à trous (Schick et Schütze, 2020). 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 *et al.*, 2022) 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.

2.3. Ressources annotées existantes

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** (Galliano *et al.*, 2006) est une collection complète de transcriptions de radio françaises ; **AnCora** (Taulé *et al*., 2008) 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** (Derczynski *et al.*, 2016) comprend des annotations sur les lieux, personnes et organisations, tandis que **Sentiment140** (Go *et al.*, 2009), **STS-Gold** (Saif *et al.*, 2013), et de nombreux autres jeux de données développés dans le cadre de tâches d'évaluation partagées comme **SemEval** (Rosenthal *et al.*, 2015), sont utilisés pour la classification de la polarité des sentiments. D'autres corpus incluent le jeu de données de dialogue **MultiWOZ** (Budzianowski *et al.*, 2018), le jeu de données **Stanford NLI** (Bowman *et al.*, 2015) 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 annotés 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.

3. Construction du jeu de données et processus d'annotation

Dans cette section, 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 Academic[2] 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 Masson *et al*., 2022 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 **2 961 tweets** provenant de 624 utilisateurs ont été sélectionnés pour l'annotation et l'utilisation dans nos expérimentations. Ces tweets spécifiques (2 961) 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 1 montre la répartition linguistique du jeu de données et la subdivision des tweets pour l'apprentissage (60% pour l'entraînement, 20% pour le développement, 20% pour le test). Cette division maintient un équilibre entre le nombre d'utilisateurs et de langues dans chaque jeu de données.

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

Tableau 1. Répartition du jeu de données collecté par langue – Tweets (Utilisateurs)

	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)

3.1 Annotation du sentiment

Le processus d'annotation des 2 961 tweets a été effectué **semi-automatiquement**, comme illustré dans la Figure 1. 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.

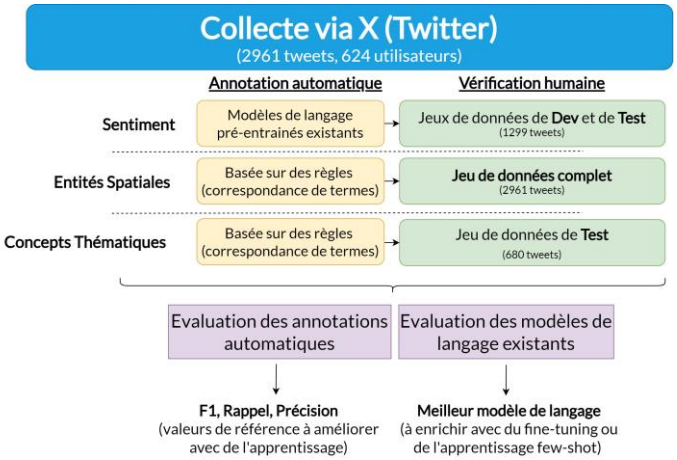

Figure 1. Processus de construction et d'annotation du jeu de données

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. Le Tableau 2 montre que le modèle XLM-T Sentiment (Barbieri *et al.*, 2022) 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.

Tableau 2. Précision de modèles de prédiction du sentiment existants (jeu de test)

Langue des tweet	**Barbieri *et al.*, 2020**	**Perez *et al.*, 2021**	**Seethal *et al.*, 2023**	**Hartmann *et al.*, 2023**	**Barbieri *et al.*, 2022**
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**

3.2 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 avions 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 jeu de *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 jeu de 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.

Tableau 3. Performance de la méthode d'annotation par correspondance de termes

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

Les performances de la méthode par correspondance de termes appliquée au jeu de test sont rapportées dans le Tableau 3. Ces résultats serviront de référence pour comparer avec les différentes approches par apprentissage dans la section expérimentale. 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.

4. Protocole expérimental

L'expérimentation se concentre sur les trois tâches présentées précédemment : classification de la polarité des sentiments (section 4.1), reconnaissance d'entités nommées spatiales et extraction de concepts thématiques fins (section 4.2) afin de déterminer quelles approches sont les plus efficaces et avec quelle quantité de données d'entraînement. La Figure 2 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'entrainement 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 jeu de données d'entrainement, 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).

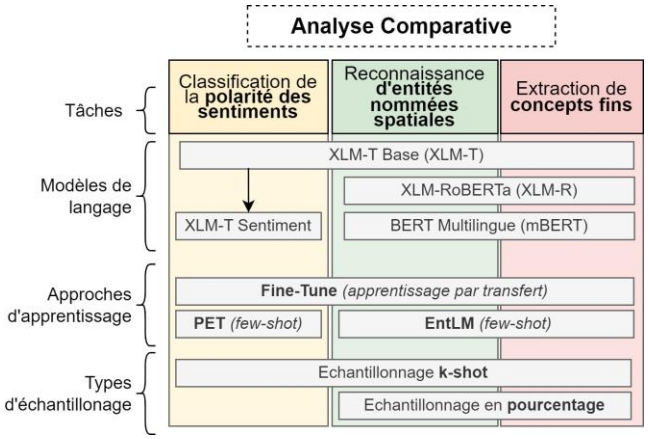

Figure 2. Configuration expérimentale de l'étude

4.1 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 (Barbieri *et al.*, 2022) pour nos expérimentations sur la classification de la polarité des sentiments. XLM-T est basé sur XLM-RoBERTa (Conneau *et al.*, 2019), 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 (Barbieri *et al.*, 2022), 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** (dénoté 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 requêtage *few-shot* avec plusieurs échantillons de notre jeu de données d'entrainement 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)** (voir Schick et Schütze, 2020) : 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.

4.2 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 *et al.*, 2022), 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, Devlin et al., 2018), 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'entrainement. 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 (Tjong Kim Sang et Erik, 2002). Tous les résultats rapportés sont la moyenne de trois exécutions initialisées aléatoirement.

5. 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 3. 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 d'exemples annotés utilisés pour l'entrainement, 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 3, (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.

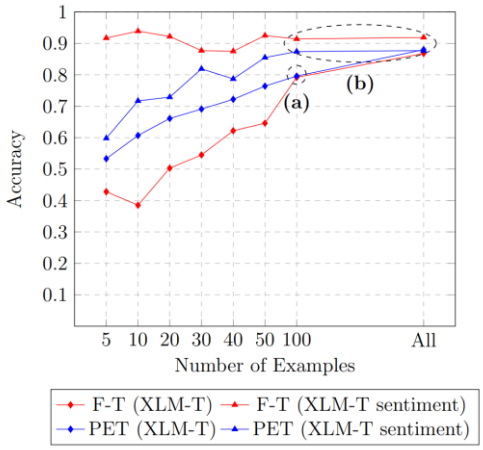

Figure 3. Classification de la polarité des sentiments - Echantillonnage k-shot

Lorsqu'on se concentre sur les approches qui exploitent uniquement nos propres données d'entraînement, l'apprentissage *few-shot* (PET) surclasse le fine-tuning du modèle XLM-T (jusqu'à 100 exemples annotés, voir Figure 3 : (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, 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-entrainé comme XLM-T Sentiment, un jeu de données d'entrainement 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 4 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 4), nous pouvons observer que lors de l'utilisation de toutes les données d'entraînement, les quatre méthodes obtiennent des résultats relativement comparables. Cependant, dans un contexte de faible disponibilité des données, la méthode EntLM nécessite moins d'exemples 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 d'entraînement. 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 d'exemples avec les deux méthodes d'échantillonnage.

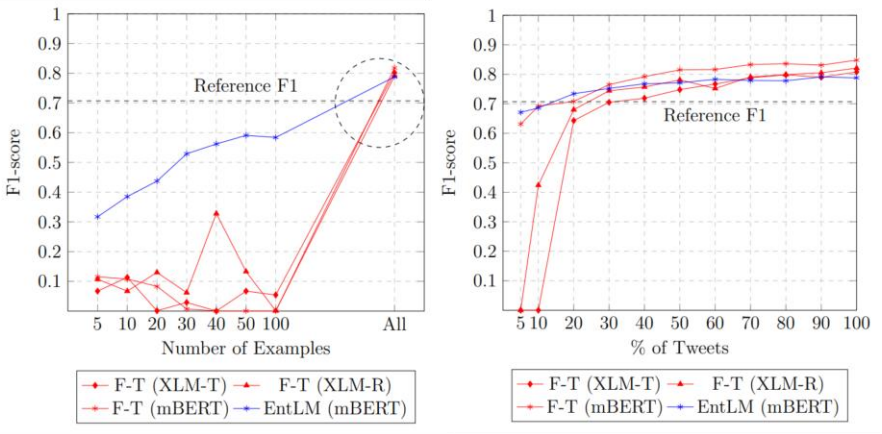

Figure 4. Reconnaissance d'entités nommées spatiales – Echantillonnage des données d'entrainement par k-shot (gauche) et pourcentage (droite)

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 pour l'entraînement (~ 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 5, 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.

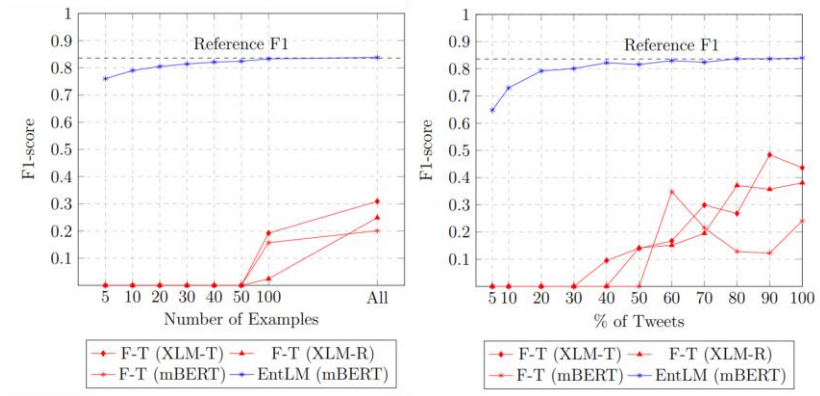

Figure 5. Extraction de concepts thématiques fins – Echantillonnage des données d'entrainement par k-shot (gauche) et pourcentage (droite)

6. Discussion et limitations

Après ces expérimentations, penchons-nous sur les conclusions obtenues pour discuter des principaux éléments ayant émergé. Nous aborderons 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.

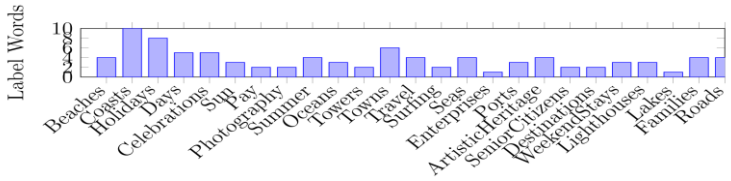

Figure 6. Nombre de mots labels pour les concepts les plus fréquents du corpus

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 6) 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 du tourisme.

Les limitations de cette étude incluent la taille relativement restreinte de notre corpus spécifique au tourisme, qui pourrait influencer la généralisabilité de nos résultats. Bien que notre corpus soit multilingue et annoté avec soin, l'étendre à d'autres langues ou contextes touristiques pourrait fournir des indications supplémentaires sur l'adaptabilité des modèles d'apprentissage dans des contextes variés. De plus, bien que nous ayons concentré nos efforts sur des tâches spécifiques liées au tourisme, les méthodes et conclusions pourraient nécessiter une validation supplémentaire dans d'autres domaines d'application. Notre étude contribue à une meilleure compréhension des stratégies optimales pour l'analyse de données multilingues issues des réseaux sociaux dans le domaine du tourisme. Nos résultats soulignent l'importance de choisir la bonne méthode d'apprentissage en fonction de la spécificité de la tâche, de la disponibilité des données et de la nécessité d'annotations manuelles. Des recherches futures pourraient explorer l'extension de ces techniques à d'autres domaines ou l'intégration de sources de données diversifiées pour enrichir la capacité d'analyse des modèles d'apprentissage profond.

7. 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 l'extraction 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 sera rendu public dans les prochains mois, 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. La prochaine étape du projet est de présenter ces résultats aux acteurs de l'industrie du tourisme, notamment à travers des tableaux de bord dynamiques mettant en évidence les entités extraites des réseaux sociaux, en lien avec des données contextuelles comme le sentiment.

Bibliographie

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.

Barbieri, F., Camacho-Collados, J., Neves, L., Espinosa-Anke, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. arXiv:2010.12421

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.

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.

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.

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.

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.

Crawford M., Khoshgoftaar T.M. (2021). Using inductive transfer learning to improve hotel review spam detection. IRI 2022, IEEE, p. 248-254.

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.

Derczynski L., Bontcheva K., Roberts I. (2016). Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. *COLING 2016: Technical Papers*, p. 1169-1179.

Enríquez M.P., Mencía J.A., Segura-Bedmar I. (2022). Transformers Approach for Sentiment Analysis: Classification of Mexican Tourists Reviews from TripAdvisor.

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 A., Bhayani R., Huang L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report*, Stanford, vol. 1, n° 12, p. 2009.

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

Pérez, J. M., Giudici, J. C., Luque, F. (2021). pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. *arXiv e-prints*, arXiv-2106.

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.

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.

Maynard D., Bontcheva K., Rout D. (2012). Challenges in developing opinion mining tools for social media, *Workshop Programme*, p. 15.

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.

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.

Saif H., Fernandez M., He Y., Alani H. (2013). Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold.

Schick T., Schütze H. (2020). Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint* arXiv:2001.07676.

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.

Seethal (2023). Sentiment analysis generic dataset.. *En ligne le 23 mars 2023.*

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*.

Taulé, M., Martí, M. A., Recasens, M. (2008). Ancora: Multilevel annotated corpora for Catalan and Spanish. *LREC*. Vol. 2008, pp. 96-101.

Tjong Kim Sang, Erik F. (2002). Introduction to the CoNLL-2002 Shared Task. *CoNLL 2002*

Toporkov O., Agerri R. (2023). On the Role of Morphological Information for Contextual Lemmatization. *arXiv preprint*, vol. 2302.00407.

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.