Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Etudes sur la géolocalisation de tweets

Thibaud Martin[1]
(1) Institut National des Sciences Appliquées de Lyon, 7 Av. Jean Capelle O, 69100 Villeurbanne, France
thibaud.martin@insa-lyon.fr

RÉSUMÉ

La géolocalisation de textes non structurés est un problème de recherche consistant à extraire un contexte géographique d'un texte court. Sa résolution passe typiquement par une recherche de termes spatiaux et de la désambiguïsation.

Dans cet article, nous proposons une analyse du problème, ainsi que deux méthodes d'inférence pour déterminer le lieu dont traite un texte :

1. Comparaison de termes spatiaux à un index géographique
2. Géolocalisation de textes sans information géographique à partir d'un graphe de co-occurrence de termes (avec et sans composante temporelle)

Nos recherches sont basées sur un corpus de 10 millions de tweets traitant de lieux français, dont 57 830 possèdent une coordonnée géographique.

ABSTRACT

Unstructured text geoparsing

Unstructured text geoparsing is a search problem in which we want to infer a geographical context from a short text. This usually involves the search of spatial terms as well as their disambiguation in order to obtain a set of coordinates.

In this article, we propose an analysis of the aforementioned problem, as well as two inference methods to solve it :

1. Comparing spatial terms to a gazetteer
2. Geoparsing text which does not contain geographical terms (e.g : cities, regions, countries, etc.) with a term co-occurrence graph (with and without temporal contextualisation)

Our research is based on a dataset of 10 million tweets pertaining to french places, in which 57 830 have a set of geographical coordinates.

MOTS-CLÉS : TALN, géolocalisation de textes, fouille de données, Twitter.

KEYWORDS: NLP, text geolocation, data mining, Twitter.

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 114–130.

1 Introduction

Depuis plus d'une décennie, le partage d'informations médiatiques a évolué pour comporter de nouveaux canaux de communication. Historiquement, un article de presse était rédigé par un journaliste qui consultait directement des personnes qui avaient un lien avec le sujet en question. Aujourd'hui, la présence de réseaux sociaux tels que Twitter a ajouté un nouvel intermédiaire à cette communication : l'information est transmise par une petite quantité d'utilisateurs, puis est ensuite diffusée par effet boule de neige (likes, retweets, etc.), jusqu'à ce qu'elle soit suffisamment populaire pour être retranscrite dans un article.

Cette dynamique de communication fait l'objet de plus d'une décennie de recherches dans les domaines de la détection d'évènements, du suivi de catastrophes (ex : pandémies, phénomènes météorologiques, etc.), et de la géolocalisation d'utilisateurs. Le sujet commun de ces travaux est l'extraction de contextes géographiques à partir d'une agrégation de sources. Il s'agit d'un processus complexe, car les informations sur des localisations sont rares - par exemple, environ 1% à 3% des tweets mondiaux contiennent au moins une méta-donnée à caractère géographique (Qazi *et al.*, 2020) - et sont présentées à des échelles hétérogènes (ex : pays, région, ville, point d'intérêt).

La géolocalisation de textes est un problème de recherche dont l'objectif est d'extraire, à partir d'un texte et d'une quantité minimale de métadonnées, le ou les lieux qui sont cités, afin d'inférer une zone géographique qui pourra ensuite être traitée en aval par des processus de fouille de données. Cette méthode est décomposée en deux étapes :

— Une extraction de termes à caractère géographiques et / ou un rapprochement de certains termes communs à un évènement dynamique pour en inférer une localisation approximative
— Une désambiguïsation des différents termes géographiques en coordonnées (latitude / longitude) ou contours géographiques. Au-delà d'un simple géocodage, il faut ici retrouver la localisation exacte en fonction du contexte du message. Par exemple, la phrase « Je suis en train de visiter Paris » contient le terme *Paris*, mais 11 villes possèdent ce nom dans le monde.

Dans cet article, nous proposons une première analyse du problème de la géolocalisation de textes non structurés à partir d'un corpus de tweets que nous avons collecté et sélectionné. pour ce faire, nous avons appliqué plusieurs méthodes inspirées des constats relevés dans d'autres articles de recherche (cf. références), afin de donner plus de clarté sur les difficultés à surmonter pour détecter avec une forte précision la localisation d'un texte.

Tout d'abord, nous introduisons le sujet avec un état de l'art sur les jeux de données et méthodes utilisées par des travaux de recherche récents. Dans la section 3, nous présentons le jeu de tweets que nous avons généré et utilisé pour les différentes méthodes présentées dans la section 5. Enfin, nous apportons dans les parties 6 et 7 les résultats de nos expérimentations et ce que nous avons pu en tirer.

2 Etat de l'art

La géolocalisation basée sur des textes est principalement basée sur des jeux de données provenant de réseaux sociaux. Twitter est la source la plus fréquente dans les travaux de recherches actuels, puisqu'il permet d'avoir accès à une grande quantité de posts avec de la géolocalisation, même si leur quantité reste limitée face à au nombre total de tweets (Cheng *et al.*, 2010; Qazi *et al.*, 2020).

Prévue pour la détection de localisations d'utilisateurs de réseaux sociaux à partir de leurs propriétés et contenus, la géolocalisation de textes est maintenant utilisée dans d'autres domaines tels que la détection d'évènements (Hui *et al.*, 2021), d'interaction entre groupes (Kumar *et al.*, 2019), ou encore du suivi de comportements dynamiquement temporels comme des pandémies (Qazi *et al.*, 2020).

Parmi les méthodes proposées pour résoudre ce problème, de premières approches se basent sur l'usage de méta-données disponibles dans les tweets, telles que les coordonnées GPS ou les localisations déclarées par les utilisateurs dans leurs posts et leur profil (Zohar, 2021; Qazi *et al.*, 2020). Ces informations deviennent cependant limitées en raison de changements dans le fonctionnement de Twitter (Zhang *et al.*, 2022; Kruspe *et al.*, 2021), qui ne permettent plus de déclarer une position précise sans mettre à jour un post en utilisant l'API.

Une méthode des études réalisées sur le traitement de contenus textuels est l'extraction toponymique (Qazi *et al.*, 2020), qui consiste à récupérer des termes pertinents (ex : villes, régions) en comparant chaque terme d'un message à un index géographique suite à un filtrage des données non-pertinentes.

La popularisation de l'apprentissage supervisé a permis de réaliser des avancées sur la précision des méthodes historiques, avec en particulier l'utilisation de réseaux de neurones type CNN (Mahajan & Mansotra, 2021), RNN (Kumar *et al.*, 2019) et Bi-LSTM (Mahajan & Mansotra, 2021; Lau *et al.*, 2017), et plus récemment de transformers (Li *et al.*, 2022). Ces modèles permettent d'atteindre dans certains cas une précision au niveau du point d'intérêt, contrairement à l'identification de villes au mieux précédemment.

D'autres méthodes se basent sur la co-occurrence de termes dans des textes (Ozdikis *et al.*, 2018). L'idée est de pouvoir retrouver dans un texte, pour un ensemble de termes liés à des évènements dynamiques (ex : hôpital, théâtre, etc.), des localisations ayant été citées dans d'autres textes où ils apparaissent. Cela amène à l'utilisation de graphes de connaissance (Rossi *et al.*, 2020; Hui *et al.*, 2021).

3 Jeu de données

Notre jeu de données contient une agrégation de 10 millions de tweets avec l'ensemble de leurs métadonnées collectés depuis début 2020. Initialement centré sur le sujet du vélo dans la ville de Lyon, il fut ensuite étendu à l'ensemble du territoire français afin de pouvoir l'utiliser dans d'autres projets.

Twitter propose dans les métadonnées de ses messages deux propriétés à caractère géographique :

— coordinates : paire de coordonnées (latitude / longitude) définissable par l'utilisateur en mettant à jour un post par le biais de l'API de Twitter
— place : lieu nommé que l'utilisateur peut choisir à partir d'une liste fournie par Twitter lors de la création d'un tweet. En fonction de son type, l'objet retourné contient le nom du lieu, son pays et son contour (bounding box)

Cependant, l'usage de ces propriétés est très marginal par rapport au nombre total de tweets présents sur la plateforme. Sur notre jeu de données, seuls 57 000 tweets sont précisément géolocalisés (~0,5%) et ~500 000 tweets ont un champ « place » non nul (~5%).

Au-delà des considérations quantitatives, il est impossible de savoir quelle information un utilisateur

souhaite communiquer en utilisant l'une de ces métadonnées : il pourrait s'agir du sujet du message, ou bien de la localisation de la personne au moment où elle crée son tweet. Dans la figure 1, nous avons répertorié l'ensemble des 57 830 tweets de notre jeu de données possédant la propriété coordinates (retweets exclus). Les points les plus foncés sur la carte représentent les lieux avec la plus grande quantité de tweets.

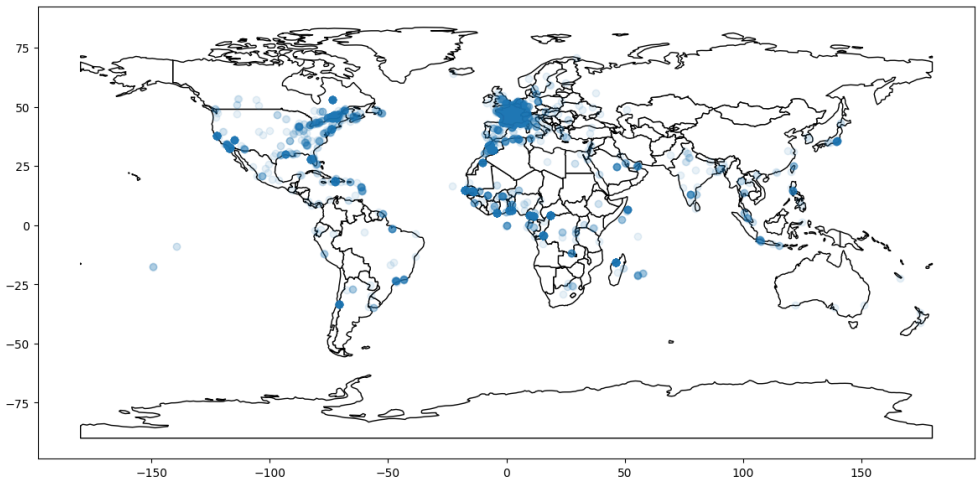

FIGURE 1 – Localisation des tweets dans notre jeu de données (coordonnées uniquement)

Nous avons ainsi noté que, malgré le fait que la grande majorité des tweets de notre jeu de données traitent de lieux en France métropolitaine, il est possible de remarquer des géolocalisations à l'étranger. Par exemple, le post https://twitter.com/PhotosAlain/status/1587845842182053890 traite du Moulin Rouge à Paris, alors que la géolocalisation indique la ville de Madrid en Espagne.

wEnfin, nous avons remarqué que la taille moyenne des tweets a évolué dans l'histoire, suite à de réguliers changements sur leur taille maximale. Historiquement fixée à 140 caractères, la limite fut augmentée en 2017 pour atteindre 280 caractères, et d'ici le printemps 2023 certains utilisateurs pourront aller jusqu'à 4000 [1]. Néanmoins, la figure 2 montre que la majorité des tweets se situent environ entre 8 et 22 mots. Cela peut poser un problème dans le cadre de la géolocalisation de ces textes, puisque la probabilité qu'un mot à caractère géographique apparaisse est plus faible.

4 Pré-traitement du corpus

Notre première version de jeu de données est contenu dans un fichier JSON unique, avec un objet par tweet. Pour nos premières expérimentations, nous n'avons gardé que les objets possédant une propriété *coordinates* non nulle, soit au total 57 830 lignes. Une première analyse des messages montre qu'il y a une grande quantité de duplications (en réalité, des retweets), que nous enlevons pour garder une instance unique par message :

1. https://twitter.com/elonmusk/status/1627388350612004865

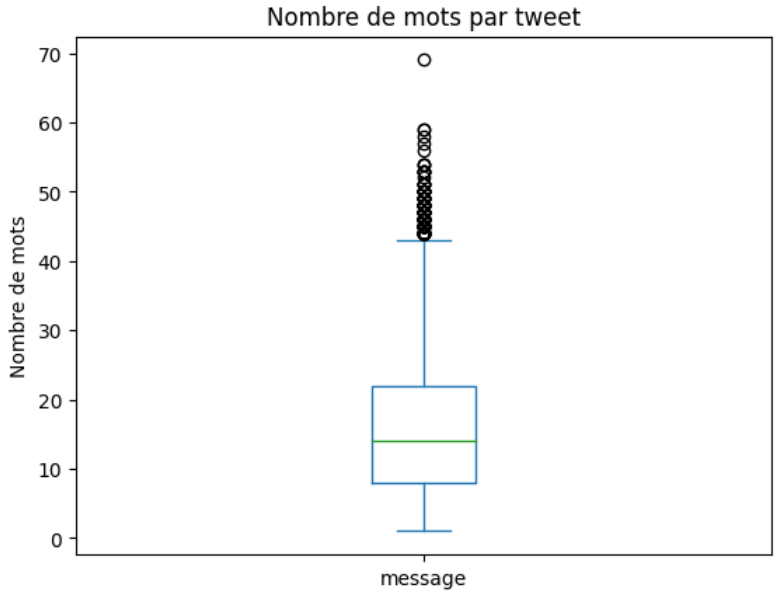

FIGURE 2 – Nombre de mots par tweet dans notre jeu de données

Nombre de tweets	tweets uniques	Message le plus fréquent	Fréquence
57 830	36 638	Vient de publier une photo à Paris France	3174

TABLE 1 – Analyse statistique des messages du jeu de données

Nous filtrons ensuite les métadonnées que nous n'avons pas considérées comme utiles pour notre expérimentation. Restent alors :

- Le message du tweet
- La date de création
- La propriété *coordinates*
- La propriété place (si elle existe)

Les messages du jeu de données sont principalement en français et en anglais, et contiennent des termes et des caractères qui ne sont pas pertinents dans le cadre d'une géolocalisation de texte. Nous avons choisi d'enlever les éléments suivants dans nos tweets :

- URL de redirection vers une page tierce (https ://t.co/[...])
- Emojis et caractères unicodes spéciaux
- Ponctuation (`!"#$%&'()*+,-./:;<=>?@[]^_`{}\~`)
  - Nous avions initialement décidé de garder les caractères ' et -, car beaucoup de villes françaises en font usage. Cela a changé suite aux implémentations de l'index géographique.
- Retour chariot (`\n` et `\r`)
- Mots vides (en anglais : *stop words*) en français et anglais
- tweets dont les coordonnées données ne sont pas en France

Après l'ensemble des étapes de pré-traitement, le jeu de données final est composé de 32 670 tweets. Par la suite, nous souhaitons aborder les hashtags, dont le format est plus compliqué à traiter puisque les mots sont attachés ensembles. Pour l'instant, nous les considérons comme des mots, ce qui permet tout d'analyser les termes "simples" à un mot (ex : #lyon, #paris).

5 Méthodologie

5.1 Extraction des termes géographiques

Dans cette partie, nous avons cherché à déterminer le nombre de tweets de notre jeu de données qui contiennent au moins un terme géographique (ex : nom d'une ville, d'un pays, d'un point d'intérêt, etc.). Dans le problème de la géolocalisation de texte non structuré, il s'agit d'instances plus simples que les autres, car nous pouvons directement appliquer des méthodes d'inférences sur ces termes qui auraient été extraits au préalable.

Pour ce faire, nous avons utilisé le module SpaCy [2] qui permet, entre autres, d'utiliser des méthodes de TALN classiques telles que la reconnaissance d'entités nommées. Notre objectif était d'extraire les entités de type *LOC* qui correspondent à une localisation géographique, à l'aide du modèle *fr_core_news_sm* [3] de SpaCy qui est entraîné sur des textes en français. Sur nos **32 670** tweets pré-traités, **18 601** contiennent au moins une entité *LOC*, soit environ 57% du jeu de données.

Certaines entités *LOC* ne sont cependant pas faciles à désambiguïser dans leur entièreté (ex : adresse exacte, combinaison de lieux comme « Paris Île-de-France France »). Ainsi, nous avons aussi tokenisé ces entités afin que, s'il n'est pas possible de géocoder l'entité composée de plusieurs mots, un essai soit réalisé sur les mots individuels.

Pour pallier aux 43% de tweets sans entité *LOC*, nous analysons aussi le message des tweets mot par mot. La structure des tweets étant libre grammaticalement, il est possible que SpaCy ait du mal à détecter certains termes comme étant des lieux. Au final, la figure 3 résume les étapes suivies afin d'extraire les termes géographiques de nos tweets.

5.2 Index géographique

Pour désambiguïser les termes géographiques que nous avons extraits dans la partie précédente, nous utilisons une base de données Elasticsearch [4] qui fait correspondre un terme à des coordonnées géographiques. Nous disposons au total de 13 070 831 noms de lieux qui proviennent de deux sources :

— Geonames [5], un dictionnaire gratuit couvrant 253 régions pour un total de 12 355 522 localisations
— Toutes les entités de DBpedia [6] contenant au moins un nom et une paire latitude / longitude, soit 715 309 lieux

La nature des tweets fait que les fautes d'orthographe sont courantes. En effet, les utilisateurs sont libres d'écrire ce qu'ils souhaitent, ce qui peut poser problème lors de l'identification de lieux. Nous requêtons ainsi notre base de données de deux manières :

— Recherche de terme exact (sensible à la casse)
— Recherche floue avec une distance d'édition variable en fonction de la taille de la chaîne de caractères en entrée (entre 3 et 6)

2. https://spacy.io/
3. https://spacy.io/models/fr#fr_core_news_sm
4. https://www.elastic.co/
5. https://www.geonames.org/
6. https://www.dbpedia.org/

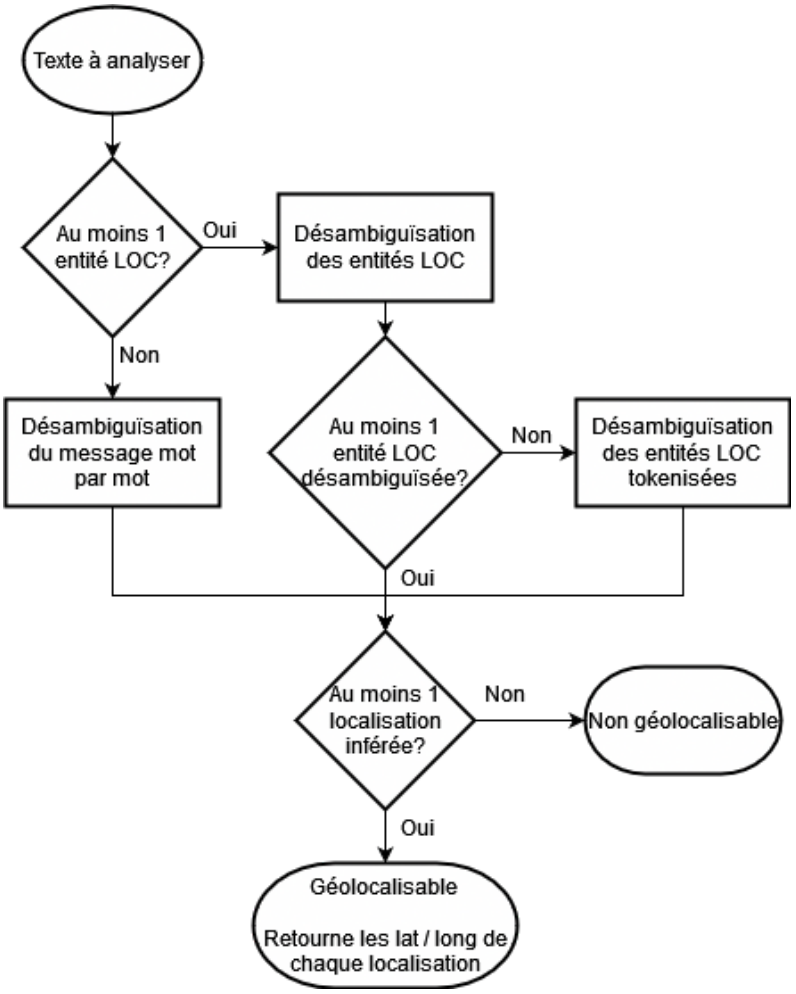

FIGURE 3 – Diagramme du processus d'extraction des entités géographiques

Chaque résultat retourné par Elasticsearch contient un score de confiance que nous utilisons pour choisir la localisation la plus pertinente. Nous appliquons un *boost* de 2 aux résultats retournés par la recherche de terme exacte afin qu'ils puissent apparaître au-dessus des autres.

Dans le cas où plusieurs termes seraient géolocalisés dans un tweet, nous retournons le centroïde composé des différentes paires de coordonnées données par l'index géographique. De plus, nous calculons un nouveau score de confiance afin de donner une indication sur la proximité de chaque point. En effet, plus les points sont dispersés, et moins le centroïde donné en résultat sera pertinent.

$$\text{confiance} = \max\left(1 - \frac{\sum_{i=0}^{n} \text{geodesic}(\text{centroid}, p_i)}{\|p\| - 1000}, 0\right) \tag{1}$$

Avec :

- **p** : ensemble de vecteurs de position (lat, long) inférées à partir du texte donné en entrée
- **centroid** : vecteur de position (lat, long) déterminé à partir des positions **p**
- **geodesic(p1, p2)** : fonction de calcul d'une distance géodésique entre 2 vecteurs de position (lat / long)
- Une distance maximale est utilisée pour avoir une borne supérieure de notre confiance. Pour toute distance ≥ 1000km, le score est de 0. Cette distance arbitraire correspond à la longueur

moyenne de la France, et sera revue dans la suite de nos travaux pour qu'elle ait plus de cohérence.

Par la suite, nous souhaitons remplacer le calcul de centroïdes par des calculs évitant les valeurs extrêmes, afin d'éviter que des positions aberrantes par rapport aux autres ne viennent fausser le résultat final.

5.3 Graphe de co-occurrence de termes

Nous avons constaté dans la partie sur l'extraction de termes géographiques qu'environ 43% des tweets de notre jeu de données ne contiennent pas d'entité *LOC* (localisation) identifiables par SpaCy. Au-delà des limitations du module, cela est aussi explicable par le fait que beaucoup de textes ne contiennent pas de termes géographiques.

Pour tenter de géolocaliser ces textes, nous proposons l'usage d'un graphe qui répertorie des co-occurrences entre termes, indépendamment de leur type (géographique ou non). Cette co-occurrence est définie par la présence de deux termes distincts dans un même texte. L'objectif est de trouver des liens entre des termes non géographiques et des lieux qui pourraient être cités dans des textes différents de celui que nous cherchons à géolocaliser.

Pour ce faire, nous avons alimenté un graphe Neo4J [7] en extrayant les noms et noms propres dans chaque texte de notre jeu de données. Pour cela, nous avons utilisé la fonctionnalité de POS-Tagging de SpaCy (codes POS : NOUN et PROPN). Ensuite, nous utilisons notre index géographique pour déterminer quels termes correspondent à une localisation. Enfin, nous générons le graphe avec deux types de sommets contenant des propriétés différentes :

— Entité géographique : nom, latitude, longitude
— Entité non géographique : nom

Nous créons ensuite une arête par co-occurrence. Puisque deux termes peuvent être co-occurrents dans plusieurs textes différents, nous identifions chaque arête par la date de création du tweet concerné.

Au final, notre graphe de co-occurrence comporte :

— 795 entités géographiques
— 24 580 entités non géographiques
— 11 141 arêtes entre entités géographiques
— 553 018 arêtes entre entités non géographiques
— 122 123 arêtes d'une entité géographique à une non géographique

5.3.1 Approche avec le graphe global

Dans cette situation, nous ne prenons pas en compte les différentes dates si plusieurs co-occurrences sont présentes entre deux sommets. Dans le cas où le terme donné serait identifié comme une entité géographique, nous retournons directement ses coordonnées (latitude, longitude).

Au contraire, si le terme est une entité non géographique, nous utilisons un *BFS* (Breadth First Search) avec une profondeur maximale de 5 pour explorer son voisinage. L'algorithme s'arrête lorsqu'une

7. https://neo4j.com/

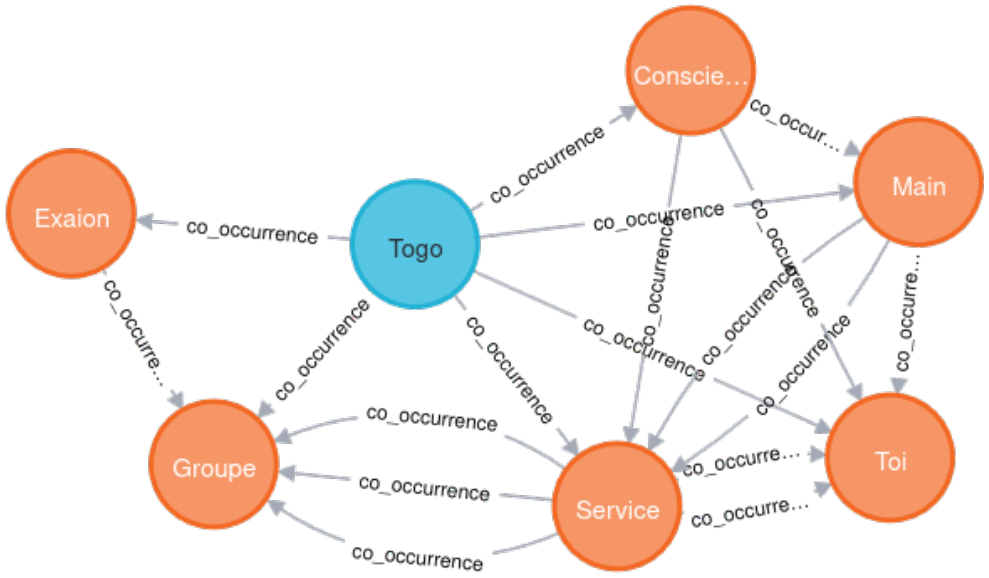

FIGURE 4 – Exemple de graphe de co-occurrence de termes (Bleu = Entité géographique, Rouge = Entité non géographique)

première entité géographique est trouvée, puis retourne ses coordonnées géographiques et le nombre de sommets traversés pour estimer une confiance sur le résultat :

$$\text{confiance} = 1 - \frac{\text{profondeur}}{\text{PROFONDEUR\_MAX} - 1} \tag{2}$$

5.3.2 Approche avec composante temporelle

Pour rendre les résultats de l'approche précédente plus pertinente, nous souhaitons faire le lien avec le domaine de la détection d'évènement. Ainsi, nous introduisons une méthode basée sur une proximité temporelle, qui favorise les co-occurrences de termes les plus « fraîches ». Notre intuition se base sur le fait que certains lieux dépendent d'évènements : par exemple, la « Fête des lumières » est très fréquemment associée à la ville de Lyon puisqu'elle se déroule dans cette ville.

Pour obtenir des résultats plus précis, nous envisageons, pour chaque terme à analyser dans le graphe, à réaliser une découpe pour ne garder que les sommets et les arêtes avec une date comprise entre t_0 (date de création du tweet correspondant) et $t_0 - x$ (x étant un paramètre). En complément, nous souhaitons aussi assigner à chaque arête restante un poids de « fraîcheur » afin de pouvoir appliquer l'algorithme de recherche de Dijkstra et privilégier les termes avec une co-ocurrence proche du tweet à analyser. Le calcul de ces poids pourrait s'apparenter à une gaussienne définie dans l'intervalle $[t_0 - x, t_0]$.

5.4 Approche incrémentale

Malgré le pourcentage de textes géolocalisables très élevé pour chaque proposition (cf. résultats), il peut être dans certains cas nécessaire d'avoir un rappel maximal, en dépit d'une précision plus faible. Pour ce cas de figure, nous proposons une approche incrémentale qui consiste, pour chaque

tweet, à exécuter toutes les méthodes citées précédemment une par une, jusqu'à ce qu'au moins une localisation soit extraite dans une paire type d'entité / méthode donnée. L'ordre est décidé par les précisions mesurées individuellement au préalable.

```
def inferer_localisation(texte, liste_methodes)
    localisations = []

    for methode in liste_methodes:
        for type_entite in ["LOC", "LOC_TOKENS", "TEXT_TOKENS"]:
            entites = texte[type_entite]
            if len(entites) > 0:
                for entite in entites:
                    localisation_inferee = methode(entite)
                    if localisation_inferee is not None:
                        localisations.append(localisation_inferee)

            if len(localisations) > 0:
                break

        if len(localisations) > 0:
            break
```

Algorithme 1 – Code Python de l'approche incrémentale

6 Evaluation - Résultats

Les programmes correspondants aux méthodes présentées ci-dessus ont été exécutés sur une machine avec un Intel® Core™ i5-8600K et 24 GiB de RAM. Pour les 32 670 tweets de notre jeu de données, le temps d'exécution moyen est de 5 min 30 pour les différents types d'index géographique. Pour le graphe de co-occurrences, ce temps est proportionnel à sa taille :

— Pour un petit graphe (5125 sommets, 18 118 arêtes) : ~5 minutes
— Pour un graphe de plus grande taille (75 953 sommets, 1 333 974 arêtes) : ~2 heures

Pour le processus d'évaluation, nous avons utilisé les métriques suivantes (toutes les distances sont exprimées en km) :

— % Géolocalisable : pourcentage de tweets où au moins une localisation a pu être extraite du texte
— Dist. moyenne : distance moyenne entre la localisation inférée et la vérité (coordonnées du tweet)
— Q(0.1), ..., Q(0.5) : quantiles des distances entre la localisation inférée et la vérité (coordonnées du tweet)
— Acc @k km : pourcentage de tweets dont la distance entre la localisation inférée et la vérité (coordonnées du tweet) est inférieure ou égale à k. En général, une Acc @10 km ou moins équivaut à la précision d'une ville, et une Acc entre 50 et 100km correspond à un département.

Dans cette partie, les méthodes suivront le nommage suivant :

— **Index géo** : index géographique
  — *DBPedia* : utilisation des localisations de DBpedia dans l'index
  — *Geonames* : utilisation des localisations de Geonames dans l'index

    - *Exact* : Recherche de terme exact (casse, orthographe)
    - *Fuzzy* : Recherche de terme floue avec une distance d'édition variable en fonction de la taille de la chaîne de caractères en entrée (entre 3 et 6)
- **Graphe de co-occurrences** (**uniquement la version sans composante temporelle**)
- **Approche incrémentale**
    - *Entités LOC* : extraction des termes géographiques sur les entités *LOC* (localisation) détectés par SpaCy dans le texte
    - *Texte tokenisé* : extraction des termes géographiques sur le texte mot par mot
    - *LOC tokenisé* : extraction des termes géographiques sur les entités *LOC* mot par mot

6.1 Résultats des méthodes

Les tableaux 2 et 3 présentent les résultats des méthodes citées précédemment. L'approche incrémentale est divisée pour montrer la précision qui peut être obtenue pour chaque type d'entité utilisé.

Méthode	% Géolocalisable	Dist. moyenne	Q(0.1)	Q(0.25)	Q(0.5)	Distance max.
Index géo, DBpedia, Exact	0,958	**175,841**	**0,105**	**0,798**	**3,642**	17 016,5
Index géo, DBpedia + Geonames, Exact	0,998	1814,32	0,174	15,535	1107,81	18 916,3
Index géo, DBpedia + Geonames, Fuzzy	**0,999**	1956,97	0,174	104,044	1438,21	19 031,3
Graphe de co-occurrences	0,954	294,115	0,207	3,592	96,697	**10 907,7**
Approche incrémentale (Entités *LOC*)	0,4	**283,939**	**0**	**0**	**0**	19 031,3
Approche incrémentale (Texte tokenisé)	**1**	388,026	0,955	3,133	69,36	**17 016,5**
Approche incrémentale (*LOC* + *LOC* tokenisé)	0,565	354,502	**0**	**0**	0,105	19 031,3
Approche incrémentale (tous types)	**1**	341,202	0,105	1,251	4,284	**17 016,5**

TABLE 2 – Statistiques des distances par rapport à la vérité (distances en km)

Méthode	Acc @1km	Acc @5km	Acc @10km	Acc @50km	Acc @100km
Index géo, DBpedia, Exact	**0,27**	**0,573**	**0,62**	**0,694**	**0,73**
Index géo, DBpedia + Geonames, Exact	0,137	0,22	0,23	0,261	0,275
Index géo, DBpedia + Geonames, Fuzzy	0,13	0,198	0,208	0,234	0,245
Graphe de co-occurrences	0,136	0,301	0,326	0,41	0,505
Approche incrémentale (Entités *LOC*)	**0,738**	**0,806**	**0,816**	**0,847**	**0,853**
Approche incrémentale (Texte tokenisé)	0,105	0,37	0,414	0,476	0,551
Approche incrémentale (*LOC* + *LOC* tokenisé)	0,606	0,744	0,763	0,802	0,816
Approche incrémentale (tous types)	0,228	0,531	0,579	0,653	0,688

TABLE 3 – Précision des méthodes

Nous pouvons constater que la méthode d'index géographique avec une recherche de termes exacts et l'usage de DBpedia comme unique source de localisations produit les meilleures précisions de manière générale. L'ajout de Geonames permet de géolocaliser plus de tweets, mais la distance entre

la vérité et les termes inférés est trop élevée pour être pertinente : dans le tableau 3, seulement 27,5% des tweets inférés par une recherche de termes exacte possèdent une précision inférieure à 100km (Acc @100km), tandis que la recherche de termes floue atteint 24,5%.

La méthode permettant d'identifier le plus de localisations dans les tweets est l'approche incrémentale (% géolocalisable = 100 %). Malgré une précision sensiblement plus faible que notre meilleure méthode, les résultats restent assez comparables (avec une Acc @5km égale à ~53%, contrairement à ~57% pour l'index géographique), ce qui permet d'avoir un bon équilibre entre les deux critères.

En comparaison avec les différents types de termes tokenisés, les entités *LOC* donnent les meilleurs résultats avec ~74% de précision $\leq$ 1km, ce qui laisse à croire que le NER de SpaCy garantit d'extraire des termes de qualité. Cependant, nous constatons que cela ne concerne que 40% du jeu de données.

Les localisations inférées se concentrent aussi sur une liste restreinte de villes en France. Par exemple, pour la méthode *Index géo, DBpedia, Exact* qui retourne 42 649 lieux, les cinq les plus inférés sont :

— Paris (24 593, soit ~58%)
— Lyon (5046, soit ~12%)
— France (4644, soit ~11%)
— Bordeaux (2047, soit ~5%)
— Publier (761, soit ~2%), qui est potentiellement un faux positif : Publier est une ville situé à la frontière franco-suisse, mais aussi un terme non-géographique très commun dans le contexte d'un tweet

Cela signifie que ~88% des localisations inférées font partie de cette liste.

Nous nous sommes aussi intéressés au score de confiance calculé à partir du calcul de centroïde pour les tweets avec ≥ 2 localisations identifiées, afin de connaître l'effet de plusieurs critères sur la qualité des inférences.

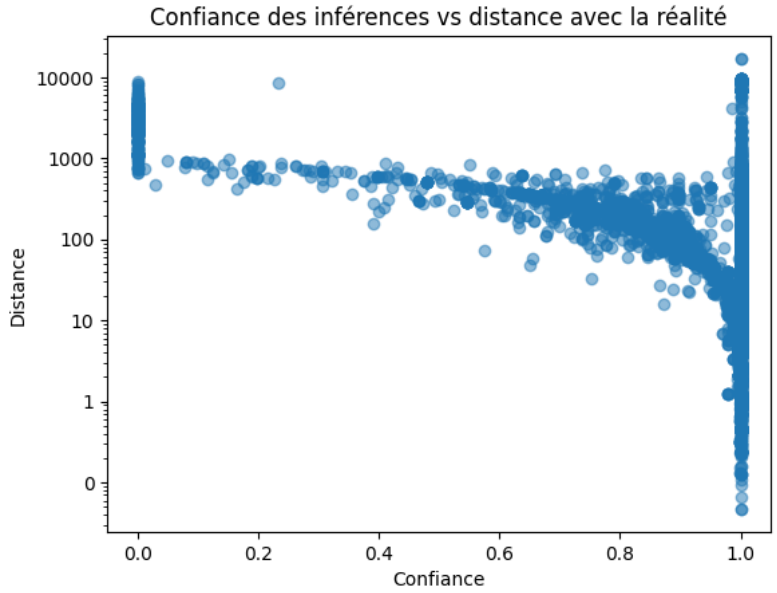

FIGURE 5 – Comparaison entre la confiance des inférences et la distance avec la réalité

La figure 5 montre une corrélation bien observable entre le score de confiance obtenu et la distance

entre la localisation inférée d'un tweet et la vérité. Nous pouvons cependant observer qu'aux deux extrémités de l'intervalle, nous retrouvons les tweets avec les distances les plus élevées. Cela peut s'expliquer de deux manières :

— Les inférences avec une confiance de 0 sont les moins cohérentes, puisque les localisations utilisées pour calculer le centroïde sont très éparpillées
— Les inférences avec une confiance de 1 et une distance très éloignée avec la vérité sont des erreurs de localisations. Un travail supplémentaire sur le calcul du score de confiance pourrait diminuer cet effet

Afin de pousser cette analyse, nous avons souhaité savoir si le score de confiance était impacté par le nombre de localisations que nous avons pu extraire dans un tweet.

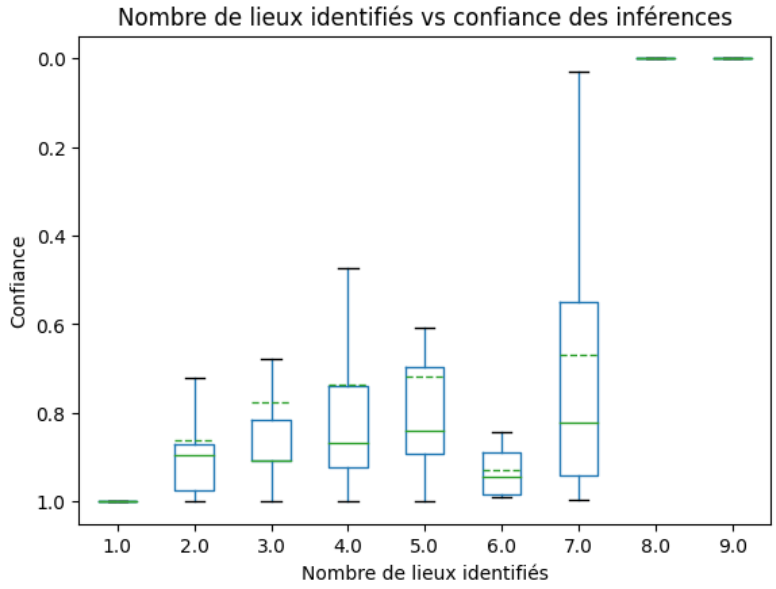

FIGURE 6 – Représentation de la confiance des inférences en fonction du nombre de lieux identifiés dans un texte

La figure 6 confirme que plus nous avons de localisations extraites dans un tweet donné, plus le centroïde calculé pour l'inférence a des chances d'être peu précis. Cela peut être causé par l'apparition plus probable de positions aberrantes (ex : [« Paris », « Musée du Louvre », « Jardin des Tuilleries », **« New York »**] qui auront tendance à fausser le résultat.

A cet effet, les observations sur les figures 5 et 6 nous confortent dans l'idée de remplacer le calcul d'un centroïde par l'utilisation de clusters pour les inférences avec ≥ 2 localisations identifiées, afin de concentrer le résultat à un ensemble de lieux qui sont proches.

6.2 Résultats de l'extraction d'entités géographiques

Afin d'appréhender l'efficacité de notre méthode d'extraction des entités géographiques, nous avons souhaité connaître le nombre de tweets géolocalisables par type d'entité et méthode d'inférence sur l'approche incrémentale. Le tableau 4 résume une exécution sur nos 32 670 tweets :

— chaque case indique le nombre de tweets dont au moins une localisation a été identifiée pour la méthode (ligne) et le type d'entité (colonne) testés

— il n'y a pas de remise. Lorsqu'une localisation est identifiée pour un tweet, le type d'entité et la méthode sont retenues, et l'exécution ne passe pas à la suite. Le tableau se lit de gauche à droite, puis de haut en bas (cf. code de l'approche incrémentale).

Méthode	Entité *LOC*	*LOC* + *LOC* tokenisé	Texte tokenisé
Index géo, DBpedia, Exact	9515	6609	15 167
Index géo, DBpedia + Geonames, Exact	52	158	1121
Index géo, DBpedia + Geonames, Fuzzy	0	0	32
Graphe de co-occurrences	0	0	16

TABLE 4 – Nombre de tweets géolocalisables par méthode et type d'entité (sur un corpus de 32 670 textes)

Nous remarquons que le nombre de tweets dont au moins une entité *LOC* extraite par SpaCy est géolocalisable est assez faible comparé aux autres types de termes, ce qui est cohérent avec les 40% de textes géolocalisables constatés ci-dessus. Nous nous sommes donc demandés si cela est causé par un manque de performance de SpaCy ou un problème venant de nos méthodes.

Au total, le nombre de tweets sans entité *LOC* est de 14 069, soit ~43% du jeu de données pré-traité. Parmi ces textes, le pourcentage de géolocalisation dépend de la méthode utilisée (cf. résultats des méthodes). Le tableau 5 résume le nombre d'entités *LOC* qui ont pu être géolocalisées par type de méthode (exécutée individuellement).

Méthode	Entités *LOC* avec géolocalisation	Entités *LOC* sans géolocalisation
Index géo, DBpedia, Exact	219	5424
Index géo, DBpedia + Geonames, Exact	702	4941
Index géo, DBpedia + Geonames, Fuzzy	1253	4390
Graphe de co-occurrences	940	4703

TABLE 5 – Nombre d'entités *LOC* géolocalisables ou non par méthode

Ces résultats démontrent que, malgré la grande quantité d'entités *LOC* uniques disponibles, peu d'entre elles sont véritablement géolocalisées par nos méthodes. Voici quelques exemples non-exhaustifs :

— Hôtel Novotel Paris Tour Eiffel
— Palais Tokyo Ville Paris
— Stade Groupama Lyon
— Palais Luxembourg invalides

Ce qui indique que des localisations « complexes » ne sont pas géolocalisables par notre index géographique actuel. Ceci est une des raisons pourquoi nous utiliserons un index plus performant par la suite (cf. index géographique).

7 Discussion

7.1 Limites de notre jeu de données

L'agrégation des tweets que nous avons employé pour générer notre jeu de données apporte des biais sur le fonctionnement de nos méthodes et les résultats de nos méthodes. En effet, la majorité des textes parlent uniquement de lieux en France, et les inférences réalisées par nos différentes méthodes se concentrent sur un petit sous-ensemble de régions et villes (cf. résultats).

De plus, les hashtags ne peuvent pas être pris en compte en l'état actuel. En effet, nous utilisons soit SpaCy soit une fonction de tokénisation pour traiter les différents termes dans notre corpus, or elles ne sont pas capables de traiter un ensemble de mots attachés, dont la casse est généralement aléatoire (souvent en titlecase ou minuscule). Nous avons pour objectif de nous concentrer sur le sujet dans la suite de nos travaux.

Par ailleurs, nous avons pu montrer que le NER de SpaCy possède des limites sur notre jeu de données qui sont causées par le contexte multilingue (anglais + français), mais aussi par la nature des tweets qui possèdent une grammaire et une orthographe très libre. D'autres NER comme Flair [8] et TwitterNER [9] doivent être explorés afin de potentiellement pouvoir améliorer l'extraction des entités géographiques.

Enfin, le fait de n'avoir que des paires de coordonnées géographiques nous empêche de mieux décrire les différents niveaux de précision de géolocalisation qui sont possibles. En effet, il est possible qu'un tweet parle de la France en général, donc les coordonnées GPS de la vérité et de la localisation inférée par nos méthodes ait une grande distance, ce qui fausse nos métriques actuelles. Nous souhaitons donc utiliser, où cela est possible, différents niveaux de précision (ex : ville, région, pays) qui peuvent être évaluées avec des contours géographiques, comme la propriété *place* proposée dans certains tweets (cf. les questionnements sur les données utilisables et la partie sur l'index géographique)

7.2 Limites du graphe de co-occurrences

Nous avons pu remarquer dans la partie résultats des méthodes que le graphe de co-occurrence que nous proposons n'atteint pas des performances similaires à notre meilleur type d'index géographique et notre approche incrémentale. Tout d'abord, nous n'avons pas encore pu tester la version avec composante temporelle, qui devrait supposément garantir une meilleure précision en ajoutant plus de localisations inférées à partir de termes non géographiques.

Malgré tout, nous concluons qu'il n'y a pas assez de « chemins » de co-occurrence utiles, ce qui veut dire que nous n'avons pas la capacité d'inférer plus de localisations qu'avec la méthode de l'index géographique. Pour pallier ce problème, nous suggérons de générer le graphe à partir d'un jeu de données différent, composé de textes plus longs (ex : articles de presse, Wikipedia) que des tweets qui sont majoritairement très courts (cf. jeu de données). Cela devrait permettre d'avoir une plus grande richesse de co-occurrences sur des données de meilleure qualité syntaxique.

8. https://huggingface.co/flair/ner-french
9. https ://github.com/napsternxg/TwitterNER

7.3 Questionnement sur les données utilisables

Notre étude du problème de géolocalisation de textes non structurés s'est concentré uniquement sur des tweets, comme pour la majorité des travaux dans le domaine (Cheng *et al.*, 2010; Hui *et al.*, 2021; Kruspe *et al.*, 2021; Lau *et al.*, 2017; Li *et al.*, 2022; Mahajan & Mansotra, 2021; Zohar, 2021; Zhang *et al.*, 2022). Cependant, il serait intéressant de pouvoir géolocaliser d'autres types de sources comme les articles de presse, dont la disponibilité en ligne et leur qualité d'information surpasse la contribution d'utilisateurs de réseaux sociaux.

Les éditions ne proposent généralement pas de système permettant de facilement récupérer les lieux qui sont traités dans leurs articles, ce qui rend pertinent l'usage des mêmes méthodes utilisées sur Twitter pour extraire des localisations. Malheureusement, la quantité de méta-données mises à disposition est moindre : sur Twitter, nous pouvons utiliser les propriétés *coordinates* et *place* afin d'avoir une idée de la localisation des posts qui les possèdent.

A cet effet, notre problème peut être traité de deux manières :

— Développer des méthodes spécifiques à des tweets, en utilisant les méta-données fournies
— Se généraliser à des textes de tailles variables (court ou long), avec un minimum de méta-données disponibles

La première solution nous permettrait d'utiliser des informations supplémentaires pour améliorer la précision des localisations inférées :

— La propriété *place*, qui permet de connaître le type de lieu (point d'intérêt, ville, région, etc.) et fournit un contour géographique (quand cela est applicable) afin de décider si une localisation inférée par nos méthodes se situe bien dans le lieu en question
— Les hashtags (cf. limites de notre jeu de données)
— La localisation de l'utilisateur qui publie le tweet, mais aussi celle des utilisateurs mentionnés dans le texte

8 Conclusion

Dans cet article, nous avons pu évaluer la performance de certaines méthodes communes dans le domaine de la géolocalisation de textes non structurés, à partir d'un jeu de données basé sur des tweets traitant de lieux en France. Les résultats nous encouragent dans la poursuite de leur utilisation, même si des améliorations notables sont à réaliser : par exemple, agréger une plus grande quantité de tweets pour avoir des données plus hétérogènes, et utiliser un index géographique plus complet pour mieux désambiguïser des termes géographiques.

Dans la suite de nos travaux, nous souhaitons tester les méthodes basées sur le principe du plongement de mots et les réseaux de neurones, ainsi que de développer notre propre méthode basée sur un graphe de co-occurrences à composante temporelle.

Références

CAILLAUT G., GRACIANNE C., AUCLAIR S., ABADIE N. & TOUYA G. (2022). Annotation sémantique pour la géolocalisation d'entités spatiales dans des tweets. In *PFIA Résilience et IA*, Saint-Etienne, France. HAL : hal-03682484.

CHENG Z., CAVERLEE J. & LEE K. (2010). You are where you tweet : A content-based approach to geo-locating twitter users. *International Conference on Information and Knowledge Management, Proceedings*, p. 759–768. DOI : 10.1145/1871437.1871535.

HARANG R. & BUSCALDI D. (2021). Apprentissage par transfert avec BERT pour la géolocalisation des Tweets. In *Atelier Deep Learning pour le traitement automatique des langues, EGC 2021*, Montpellier, France. HAL : hal-03166986.

HUI B., CHEN H., YAN D. & KU W.-S. (2021). Edge : Entity-diffusion gaussian ensemble for interpretable tweet geolocation prediction. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, p. 1092–1103. DOI : 10.1109/ICDE51399.2021.00099.

KRUSPE A., HÄBERLE M., HOFFMANN E. J., RODE-HASINGER S., ABDULAHHAD K. & ZHU X. X. (2021). Changes in twitter geolocations : Insights and suggestions for future usage. *W-NUT 2021 - 7th Workshop on Noisy User-Generated Text, Proceedings of the Conference*, p. 212–221. DOI : 10.48550/arxiv.2108.12251.

KUMAR S., ZHANG X. & LESKOVEC J. (2019). Predicting dynamic embedding trajectory in temporal interaction networks. **10**. DOI : 10.1145/3292500.3330895.

LAU J. H., CHI L., TRAN K.-N. & COHN T. (2017). End-to-end network for twitter geolocation prediction and hashing. DOI : 10.48550/arxiv.1710.04802.

LI M., LIM K. H., GUO T. & LIU J. (2022). A transformer-based framework for poi-level social post geolocation. DOI : 10.48550/arxiv.2211.01336.

MAHAJAN R. & MANSOTRA V. (2021). Predicting geolocation of tweets : Using combination of cnn and bilstm. *Data Science and Engineering*, **6**, 402–410. DOI : 10.1007/s41019-021-00165-1.

OZDIKIS O., RAMAMPIARO H. & NØRVÅG K. (2018). Spatial statistics of term co-occurrences for location prediction of tweets. In G. PASI, B. PIWOWARSKI, L. AZZOPARDI & A. HANBURY, Éds., *Advances in Information Retrieval*, p. 494–506, Cham : Springer International Publishing. DOI : 10.1007/978-3-319-76941-7_37.

QAZI U., IMRAN M. & OFLI F. (2020). Geocov19 : A dataset of hundreds of millions of multilingual covid-19 tweets with location information. DOI : 10.48550/ARXIV.2005.11177.

ROSSI E., CHAMBERLAIN B., FABRIZIO T., TWITTER F., TWITTER D. E., MONTI F. & TWITTER M. B. (2020). Temporal graph networks for deep learning on dynamic graphs. DOI : 10.48550/arxiv.2006.10637.

ZHANG J., DELUCIA A. & DREDZE M. (2022). Changes in tweet geolocation over time : A study with carmen 2.0. p. 1–14.

ZOHAR M. (2021). Geolocating tweets via spatial inspection of information inferred from tweet meta-fields. *International Journal of Applied Earth Observation and Geoinformation*, **105**, 102593. DOI : 10.1016/j.jag.2021.102593.