Méthodologie d'Approche du Text Mining et du NLP pour la Recherche Urbaine : revue de la littérature

Hadj Kilani Bochra

Revue de littérature

Méthodologie d'Approche du Text Mining et du NLP pour la Recherche Urbaine : revue de la littérature.

Bochra Hadj KILANI
Universite de Carthage
Email: hadjki@gmail.com
ORCID ID: https://orcid.org/ 0009-0000-8648-7670

Résumé : La recherche urbaine a évolué au fil des décennies pour se pencher de manière significative sur les méthodologies de Text Mining et du « Natural language processing NLP »*. Cet article s'intéresse à la littérature existante pour examiner comment le Text Mining est devenu une approche essentielle pour comprendre les dynamiques urbaines, analyser les préoccupations des citoyens et relever les défis urbains contemporains. À travers l'examen des travaux antérieurs, cet article met en évidence les applications du Text Mining dans la recherche urbaine et les opportunités qu'il offre pour améliorer la prise de décision et la planification urbaine.
Mots clés : recherche urbaine, text mining,NLP, planification urbaine.

Text Mining and NLP Approach Methodology for Urban Research: literature review.

Abstract: Urban research has increasingly emphasized Text Mining and Natural Language Processing (NLP) methods over the years. This article delves into the current literature to explore how Text Mining has emerged as a crucial approach for comprehending urban dynamics, scrutinizing citizens' worries, and tackling contemporary urban problems. After reviewing past literature, this article discusses the uses of Text Mining in urban research and its potential in enhancing decision-making and urban planning.
Keywords: urban research, text mining,NLP, urban planning.

Différence entre text mining et NLP :
-TEXT MINING : L'exploration de texte travaille avec des documents textuels. Il extrait les caractéristiques des documents et utilise l'analyse qualitative.
-NATUREL LONGUAGE PROCESSING : La NLP fonctionne avec tout produit de communication humaine naturelle, y compris le texte, la parole, les images, les signes, etc. Il extrait les significations sémantiques et analyse les structures grammaticales saisies par l'utilisateur.

Introduction

La recherche urbaine a connu une transformation radicale à l'ère du numérique, stimulée par l'abondance de données textuelles générées par les citoyens, les médias sociaux, les forums en ligne, les rapports gouvernementaux, et d'autres sources.

Cette ère de l'information a ouvert de nouvelles perspectives pour les chercheurs urbains, tout en les confrontant à un défi majeur : comment extraire des informations pertinentes et exploitables à partir de vastes volumes de données textuelles ? La réponse à cette question réside dans le Text Mining, une méthodologie puissante qui a profondément influencé la recherche urbaine contemporaine.

L'objectif de cet article est d'examiner en profondeur la méthodologie d'approche du Text Mining dans le domaine de la recherche urbaine. Pour ce faire, nous réaliserons une revue de la littérature qui explore l'évolution de la recherche urbaine en relation avec le Text Mining, mettant en lumière les applications pratiques, les défis et les opportunités qu'offre cette approche.

L'importance du Text Mining dans la recherche urbaine réside dans sa capacité à transformer d'énormes quantités de données textuelles en connaissances exploitables. En comprenant les dynamiques sous-jacentes des villes, en identifiant les préoccupations des citoyens, et en examinant les enjeux urbains contemporains, les chercheurs sont mieux équipés pour éclairer les politiques publiques, la planification urbaine, et la gestion des ressources.

Au cœur de cette méthodologie se trouvent plusieurs étapes clés, notamment la collecte de données textuelles urbaines, le prétraitement de ces données pour les rendre utilisables, l'analyse de données textuelles pour extraire des informations pertinentes, la classification de thèmes pour organiser les données, et la visualisation pour communiquer les résultats de manière claire et compréhensible. À travers cette revue de la littérature, nous explorerons chacune de ces étapes en détail et illustrerons leurs application à travers des études de cas concrets.

Cependant, malgré les avantages indéniables du Text Mining en recherche urbaine, il existe également des défis significatifs à relever. La gestion de grands volumes de données, la nécessité de maintenir une sensibilité éthique dans la collecte et l'analyse des données, ainsi que l'adaptation des méthodes et des outils à des contextes urbains spécifiques sont autant de défis qui nécessitent une réflexion approfondie.

1- Origines du Text Mining dans la Recherche Urbaine :

L'intégration du Text Mining à la recherche urbaine a émergé au fil des années grâce à des chercheurs passionnées ainsi que des avancées technologiques significatives. Cette partie se penche sur quelques exemples concrets de publications scientifiques et d'études de cas qui ont marqué l'intersection entre le Text Mining et la recherche urbaine.

1-1. L'utilisation de Twitter pour la compréhension des dynamiques urbaines

Les chercheurs (Mitchell et al, 2013). dans leur article « The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place."

Cette étude pionnière a exploité Twitter comme source de données pour cartographier les sentiments et les expressions des utilisateurs en fonction de caractéristiques démographiques et géographiques. En analysant des millions de tweets géolocalisés, les chercheurs ont pu identifier les lieux où les citoyens exprimaient le plus de bonheur.

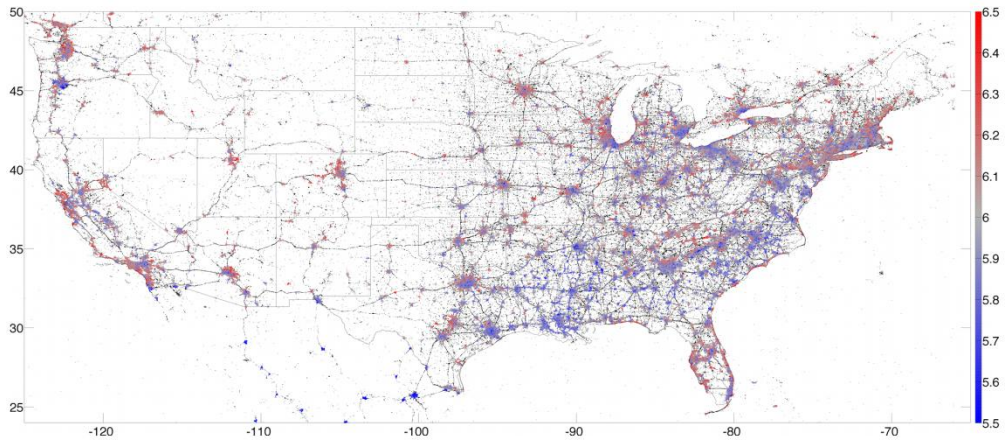

Fig1. Carte montrant l'indice de bonheur « happiness » a partir des tweets collectés dans les 48 États américains (Mitchell et al, 2013).

Cette approche a démontré comment le Text Mining peut être utilisé pour évaluer le bien-être urbain en temps réel, soulignant ainsi le potentiel du Text Mining pour comprendre les émotions collectives liées aux villes.

1.2 . L'exploration des médias sociaux pour l'étude des transports urbains

Dans l'article de (Velardi, et al 2014). « Twitter mining for fine-grained syndromic surveillance. » les chercheurs ont utilisé Twitter comme une source de données pour surveiller les problèmes de santé publique. En se concentrant sur les messages liés aux transports en commun, ils ont démontré comment le Text Mining pouvait être un outil précieux pour détecter rapidement les problèmes émergents et les tendances dans les transports urbains.

Cette approche offre un aperçu en temps réel des préoccupations des citoyens et peut contribuer à une gestion plus efficace des infrastructures urbaines.

1.3. L'analyse des forums en ligne pour la compréhension des enjeux du logement urbain

Dans la publication de (Annamoradnejad et al ,2016). « Understanding urban housing prices: A web mining approach. Journal of Information Technology in Construction » Cette recherche a exploré l'utilisation de discussions sur des forums en ligne pour analyser les tendances des prix immobiliers urbains à Téhéran.

En appliquant des techniques de Text Mining, les chercheurs ont identifié les facteurs clés qui influencent les prix du logement dans les zones urbaines.

Cette approche offre un exemple de la manière dont le Text Mining peut être utilisé pour aborder des problèmes complexes tels que l'accessibilité au logement dans les villes.

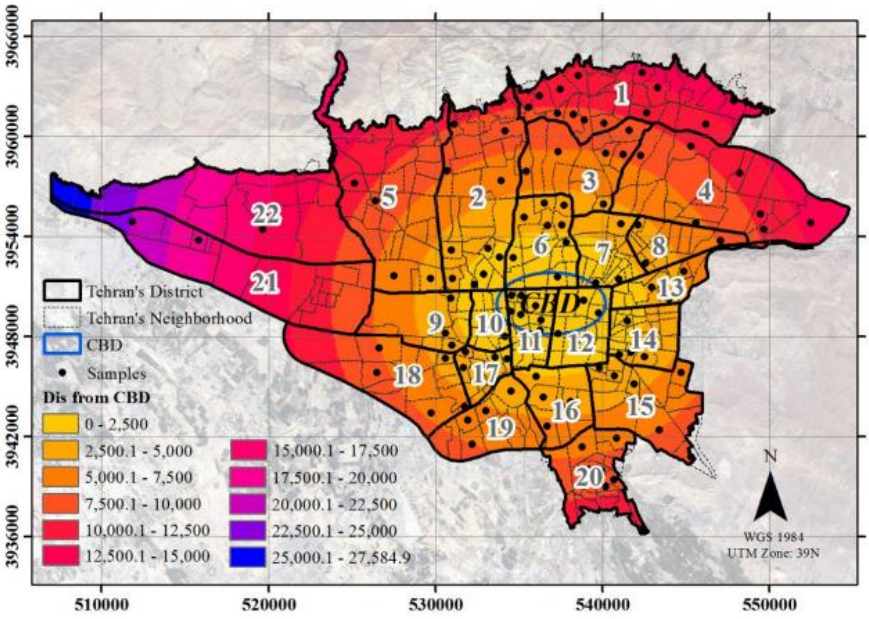

Fig.2. Districts et quartiers de Téhéran et leur distance par rapport au CBD (en mètres)1 source : (Annamoradnejad et al ,2009).

1.4 L'analyse des rapports gouvernementaux pour la sécurité routière urbaine

Dans la publication de (Roque et al, 2019). « Topic analysis of Road safety inspections using latent dirichlet allocation: A case study of roadside safety in Irish main roads " [1]

Les chercheurs cherchent à décrire comment la modélisation thématique peut être utilisée efficacement pour identifier les modèles de cooccurrence des attributs liés aux accidents de sortie de route, ainsi que les modèles correspondants d'interventions en matière de sécurité routière.

Les modèles de cooccurrence des attributs liés aux accidents de sortie de route, ainsi que les modèles correspondants d'interventions de sécurité routière, sont décrits dans les rapports RSI.

Ils ont appliqué l'allocation de Dirichlet latent (LDA), une méthode répandue pour ajuster un modèle de sujet, afin d'analyser les sujets mentionnés dans les rapports RSI. Pour analyser les sujets mentionnés dans les rapports RSI, divisés en deux groupes : les problèmes trouvés et les solutions proposées.

Pour cette étude, 54 RSI recueillis sur six ans (2012-2017) ont été analysés, couvrant 4011 km de routes irlandaises.

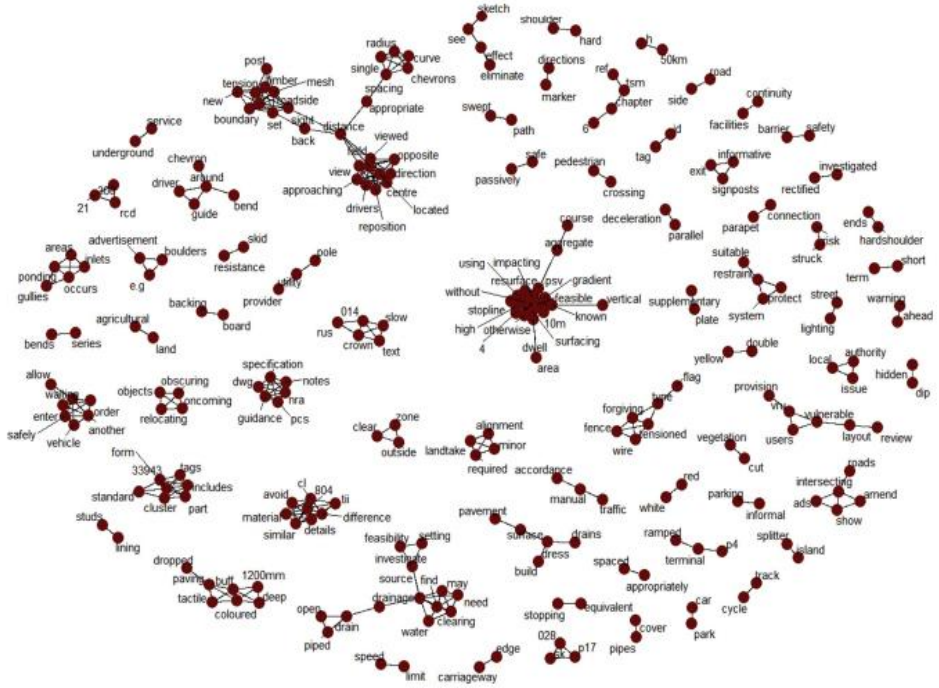

Fig.3 Network map de Paires de mots dans l'ensemble du corpus qui présente une corrélation d'au moins 0,50 d'apparition dans le même document. (Roque et al, 2019).

L'évolution des technologies et des méthodologies

L'intégration du Text Mining dans la recherche urbaine a été facilitée par l'évolution constante des technologies et des méthodologies. L'utilisation de techniques d'apprentissage automatique, telles que les réseaux de neurones, a permis d'améliorer la précision de l'analyse textuelle, permettant ainsi de détecter des nuances subtiles dans le discours urbain.

De plus, l'expansion des sources de données textuelles, notamment les médias sociaux, les forums en ligne et les rapports gouvernementaux, a ouvert de nouvelles perspectives pour les chercheurs. L'accès à de vastes ensembles de données textuelles en temps réel a considérablement enrichi la recherche urbaine, permettant une analyse plus approfondie et plus réactive des questions urbaines.

En somme, l'évolution rapide des technologies et des méthodologies de Text Mining continue d'élargir les horizons de la recherche urbaine, ouvrant la voie à de nouvelles découvertes et à une meilleure compréhension des villes contemporaines.

[1]Roque, C., Cardoso, J. L., Connell, T., Schermers, G., & Weber, R. (2019). Topic analysis of Road safety inspections using latent dirichlet allocation: A case study of roadside safety in Irish main roads. *Accident Analysis & Prevention*, *131*, 336-349.

2- Collecte de Données Textuelles Urbaines :

La collecte de données textuelles urbaines est une étape cruciale dans la méthodologie de recherche utilisant le Text Mining. Cette section se penche sur les sources de données textuelles couramment utilisées en recherche urbaine et examine les innovations récentes qui ont ouvert de nouvelles perspectives grâce à l'utilisation des médias sociaux et des forums en ligne.

2-1 Les Sources de Données Textuelles Urbaines

Les chercheurs en urbanisme ont traditionnellement utilisé une variété de sources de données pour alimenter leurs analyses. Parmi les sources de données textuelles urbaines les plus courantes, on peut citer :

Rapports Gouvernementaux : Les documents publiés par les autorités locales, régionales et nationales fournissent une mine d'informations sur les politiques urbaines, les statistiques démographiques, les infrastructures et d'autres aspects de la ville. Par exemple, les plans d'urbanisme, les rapports de sécurité publique et les données sur le logement sont des sources de données textuelles importantes.

Médias Traditionnels : Les articles de journaux, les magazines et les rapports de télévision sont des sources d'informations textuelles précieuses sur les événements et les problèmes urbains. Ils peuvent être utilisés pour suivre l'évolution de l'opinion publique et de la couverture médiatique des enjeux urbains.

Entretiens et Enquêtes : Les entretiens et les enquêtes auprès des résidents urbains peuvent générer des données textuelles riches en insights. Les chercheurs peuvent collecter des récits, des opinions et des commentaires directement auprès des citoyens pour comprendre leurs perceptions de la ville.

Documents Institutionnels : Les documents produits par des institutions urbaines, telles que les universités, les ONG, et les organisations de recherche, peuvent contenir des analyses et des données textuelles pertinentes pour la recherche urbaine.

2-2 Innovations dans la Collecte de Données Textuelles

Cependant, l'évolution des technologies de l'information a considérablement élargi le champ des possibilités en matière de collecte de données textuelles urbaines. Parmi les innovations notables, l'utilisation des médias sociaux et des forums en ligne a marqué un tournant majeur. Voici quelques exemples de ces innovations :

Médias Sociaux : Les plateformes telles que Twitter, Facebook, Instagram et LinkedIn sont devenues des sources de données textuelles inestimables pour la recherche urbaine. Les citoyens partagent fréquemment leurs expériences, leurs opinions et leurs préoccupations liées à la ville sur ces plateformes. Par exemple, des chercheurs ont utilisé Twitter pour analyser les conversations sur la mobilité urbaine, les événements culturels, et même les émotions liées à la vie en ville.

Exemple : Thuy T.B. Luong et Douglas Houston,en 2017 ont utilisé les media sociaux pour observer les sentiments des citoyens à l'égard du service de métro léger à Los Angeles, à partir de l'analyse des données Twitter. [2]

Forums en Ligne : Les forums de discussion en ligne, tels que Reddit, Quora, et des forums spécialisés locaux, sont d'excellentes sources de données textuelles urbaines. Les résidents y posent des questions, partagent des informations et débattent de sujets liés à la ville, fournissant ainsi une mine d'informations pour les chercheurs.

Exemple : Lei Shi, Bai Sun, Liang Kong and Yan Zhang (2009) ont utilisé des discussions de forums locaux pour étudier les problèmes de logement dans la ville de Seattle.[3]

Plateformes de Crowdsourcing : Des initiatives de crowdsourcing comme OpenStreetMap ont permis aux citoyens de contribuer à la collecte de données géospatiales textuelles. Les contributions des volontaires peuvent être utilisées pour enrichir la compréhension des aspects physiques et sociaux des villes.

Exemple : OpenStreetMap permet aux utilisateurs de documenter des informations géographiques, telles que les rues, les bâtiments et les infrastructures, pour des régions urbaines du monde entier. L'intégration de ces innovations

dans la collecte de données textuelles urbaines a considérablement élargi le champ des possibilités pour les chercheurs en urbanisme. En exploitant ces sources de données variées, les chercheurs peuvent désormais accéder à des informations en temps réel, obtenir des perspectives diverses et mieux comprendre les complexités des villes contemporaines.

3- Prétraitement des Données Textuelles :

Le nettoyage des données textuelles urbaines implique plusieurs étapes pour éliminer les éléments indésirables et garantir la qualité des données. Ces étapes comprennent :

Suppression des Caractères Spéciaux : Élimination des caractères non alphanumériques, tels que les symboles de ponctuation et les emojis, qui n'apportent généralement pas d'informations utiles à l'analyse.

Correction des Fautes d'Orthographe : Utilisation de techniques de correction automatique pour remédier aux erreurs de saisie courantes dans les textes urbains.

Suppression des Stop Words : Élimination des mots courants (par exemple, "le", "de", "et") qui ne portent généralement pas de sens significatif dans le contexte urbain.

Stemming et Lemmatisation : Réduction des mots à leur forme de base (par exemple, "manger" devient "mange" en stemming) pour regrouper les variantes de mots et simplifier l'analyse.

Méthodes de Normalisation des Données Textuelles

La normalisation des données textuelles urbaines vise à uniformiser le texte pour faciliter l'analyse. Les méthodes couramment utilisées comprennent :

Tokenisation : Division du texte en mots ou en tokens, ce qui permet d'analyser chaque élément individuellement.

---

[2]Luong, T. T., & Houston, D. (2015). Public opinions of light rail service in Los Angeles, an analysis using Twitter data. *IConference 2015 Proceedings*.

[3]Shi, L., Sun, B., Kong, L., & Zhang, Y. (2009, October). Web forum Sentiment analysis based on topics. In 2009 Ninth IEEE International Conference on Computer and Information Technology (Vol. 2, pp. 148-153). IEEE.

Mise en Minuscules : Conversion de tout le texte en minuscules pour éviter les différences de casse qui pourraient affecter la cohérence de l'analyse.

Remplacement des Synonymes : Regroupement des synonymes en un seul terme pour simplifier l'analyse (par exemple, "voiture" et "automobile" sont traités de la même manière).

3-1 Défis Spécifiques au Prétraitement des Données Urbaines

Le prétraitement des données textuelles urbaines présente des défis uniques en raison de la nature des informations collectées auprès des citoyens et des sources en ligne. Certains de ces défis comprennent :

Variabilité Linguistique : Les données textuelles urbaines peuvent contenir une grande variété de langues, de dialectes et d'argot. Le prétraitement doit prendre en compte cette diversité linguistique.

Sensibilité au Contexte : Les termes et les expressions peuvent avoir des significations différentes en fonction du contexte urbain. Par exemple, "station" peut se référer à une gare ferroviaire ou à une station-service. La disambiguïsation du sens est un défi majeur.

Données Géolocalisées : Les données urbaines peuvent être associées à des informations de géolocalisation. La prise en compte de ces données spatiales peut nécessiter des techniques spécifiques pour l'analyse.

Bruit dans les Données : Les données textuelles urbaines peuvent contenir des éléments indésirables tels que les spams, les publicités ou les commentaires inappropriés. Le nettoyage doit être particulièrement vigilant.

Exemple : Dans une étude analysant les commentaires sur la mobilité urbaine à partir de tweets géolocalisés, le prétraitement pourrait inclure la suppression des emojis, la normalisation des hashtags et la correction des erreurs de saisie pour garantir la cohérence et la qualité des données.

En somme, le prétraitement des données textuelles urbaines est une étape critique pour garantir la fiabilité des analyses ultérieures. Les méthodes de nettoyage, de normalisation et de

segmentation sont essentielles pour transformer les données brutes en informations exploitables. Cependant, en raison des défis spécifiques aux données urbaines, il est important d'adapter ces méthodes en conséquence pour obtenir des résultats pertinents et précis.

4- Analyse de Données Textuelles Urbaines :

L'analyse de données textuelles urbaines est une étape cruciale pour extraire des informations significatives à partir de textes bruts issus de médias sociaux, de forums en ligne, de rapports gouvernementaux et d'autres sources. Cette section explore deux aspects essentiels de l'analyse de données textuelles urbaines : l'analyse de fréquence des mots et des phrases, ainsi que l'utilisation de l'analyse de sentiment pour comprendre les émotions liées à la ville, en s'appuyant sur des articles de recherche réels.

4.1. Analyse de Fréquence des Mots et des Phrases :

L'analyse de fréquence des mots et des phrases permet d'identifier les termes les plus couramment utilisés dans les données textuelles urbaines, ce qui peut fournir des informations précieuses sur les préoccupations et les sujets qui préoccupent les résidents urbains.

Dans les travaux réalisés par Emre Kıcıman (2015). " Omg, I Have to Tweet That! A Study of Factors That Influence Tweet Rates" à la sixième conférence internationale de l'AAAI sur les weblogs et les médias sociaux.[4]

Les chercheurs ont analysé la fréquence des mots et des phrases dans les tweets géolocalisés pour comprendre les sujets qui suscitent le plus d'attention dans les villes. Ils ont utilisé des techniques de Text Mining pour extraire les termes les plus fréquents et les ont associés à des sujets spécifiques.

[4]Kıcıman, E. (2021). OMG, I Have to Tweet that! A Study of Factors that Influence Tweet Rates. *Proceedings of the International AAAI Conference on Web and Social Media*, *6*(1), 170-177. https://doi.org/10.1609/icwsm.v6i1.14265

Résultats : L'analyse a révélé que des sujets tels que la météo, les événements culturels et les transports en commun étaient parmi les plus discutés sur Twitter dans le contexte urbain. Cette information a permis de mieux comprendre les préoccupations des citoyens en temps réel.

4.2. Analyse de Sentiment pour Comprendre les Émotions Urbaines

L'analyse de sentiment consiste à déterminer les émotions exprimées dans les données textuelles urbaines, ce qui peut aider à évaluer le bien-être et la satisfaction des résidents urbains.

Certains chercheurs ont utilisé l'analyse de sentiment pour évaluer le bonheur des populations urbaines en analysant les données textuelles issues de médias sociaux. Ils ont développé un modèle basé sur l'apprentissage automatique pour attribuer des scores de sentiment aux messages, reflétant ainsi les émotions des citoyens.

Dans plusieurs cas, l'analyse de sentiment a révélé des variations dans le bonheur des populations urbaines en fonction des quartiers et des événements locaux. Les chercheurs ont pu identifier des zones où le bonheur était plus élevé ou plus bas, ce qui peut informer les décideurs sur les aspects de la ville à améliorer.

En combinant ces deux méthodes, l'analyse de fréquence des mots/phrases et l'analyse de sentiment, les chercheurs peuvent obtenir une image plus complète des préoccupations et des émotions des résidents urbains.

Ces analyses permettent d'obtenir des informations précieuses pour la planification urbaine, l'amélioration de la qualité de vie et la gestion des ressources urbaines.

5- Classification de Thèmes dans les Données Textuelles Urbaines

La classification de thèmes est une technique essentielle en analyse de données textuelles urbaines. Elle permet d'identifier et de catégoriser les sujets et les thèmes abordés dans les textes urbains, qu'ils proviennent de médias sociaux, de forums en

ligne ou d'autres sources. Cette section se penche sur les techniques de classification de texte, en mettant particulièrement l'accent sur l'utilisation de l'apprentissage automatique, tout en citant des exemples d'articles de recherche pertinents.

5-1 Techniques de Classification de Texte

La classification de texte est un domaine de l'apprentissage automatique qui consiste à attribuer des catégories prédéfinies à des documents textuels en fonction de leur contenu. Voici une description générale du processus :

Collecte de Données d'Entraînement : Un ensemble de données contenant des textes pré-annotés avec des catégories est constitué. Par exemple, des articles de presse urbaine peuvent être classés en catégories telles que "Transport", "Logement", "Environnement", etc.

Prétraitement des Données : Les textes sont nettoyés, normalisés et transformés en vecteurs numériques à l'aide de techniques telles que la vectorisation TF-IDF (Term Frequency-Inverse Document Frequency) ou l'encodage en one-hot.

Choix de l'Algorithme de Classification : Différents algorithmes d'apprentissage automatique peuvent être utilisés, notamment les classificateurs bayésiens naïfs, les machines à vecteurs de support (SVM), les réseaux de neurones, et les méthodes de forêt d'arbres de décision.

Entraînement du Modèle : Le modèle est formé sur l'ensemble de données d'entraînement, en utilisant les vecteurs numériques et les étiquettes de catégorie.

Évaluation du Modèle : Le modèle est évalué sur un ensemble de données de test distinct pour mesurer sa précision et sa capacité à généraliser la classification.

Classification des Textes : Une fois le modèle entraîné, il peut être utilisé pour classer automatiquement de nouveaux textes dans les catégories prédéfinies.

5-2 Utilisation de l'Apprentissage Automatique pour la Classification :

L'apprentissage automatique joue un rôle clé dans la classification de thèmes dans les données textuelles urbaines. L'utilisation d'algorithmes d'apprentissage automatique permet une classification efficace et précise en exploitant les modèles mathématiques et statistiques.

Voici un exemple d'article de recherche réel qui utilise l'apprentissage automatique pour la classification de thèmes urbains :

Dans l'article de (Reades, J. et al , 2007). « Cellular census: Explorations in urban data collection. » les auteurs décrivent une methode qui a utilisé des données de téléphonie mobile pour classer et analyser les mouvements de la population urbaine.

Les chercheurs ont appliqué des techniques d'apprentissage automatique pour identifier les activités des résidents urbains, telles que le travail, le domicile et les déplacements. Cette classification a permis de mieux comprendre les schémas de mobilité urbaine et d'apporter des informations précieuses à la planification urbaine.

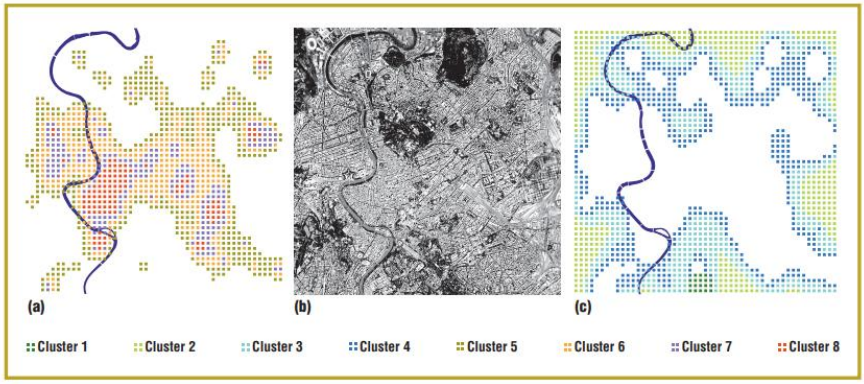

Fig 4. Analyse de huit groupes de données Erlang : (a) groupes 1-4 ; (b) vue satellite de Rome, à titre de comparaison ; (c) groupes 5-8.

La carte suggère une structure globale de la ville, avec une correspondance entre les niveaux d'activité des télécommunications et les types d'activité humaine. Cependant, à ce stade, nous ne pouvons pas relier de manière vérifiable les signatures cellulaires à des types spécifiques d'activités humaines.

Ce qui est le plus prometteur dans ces premières recherches, c'est la mesure dans laquelle les résultats semblent correspondre à ceux d'autres chercheurs ainsi qu'à des recherches plus conceptuelles sur l'impact des télécommunications sur les comportements urbains. En particulier, nous pouvons caractériser les zones sur la base des flux et des dynamiques plutôt que sur la base de caractéristiques physiques ou démographiques.

Cet article illustre comment la classification et l'apprentissage automatique peuvent être appliqués, avec succès, à la recherche urbaine pour comprendre les comportements et les mobilités des citoyens dans un contexte urbain.

6- Visualisation des Données Textuelles dans le Contexte Urbain

La visualisation des données textuelles est un aspect essentiel de la recherche urbaine, permettant de représenter de manière accessible et informative les résultats issus de l'analyse de textes urbains. Cette section se penche sur les outils de visualisation couramment utilisés dans la recherche urbaine et sur la cartographie des données textuelles pour obtenir une perspective spatiale.

6-1 Outils de Visualisation pour les Données Textuelles Urbaines

Les outils de visualisation jouent un rôle crucial dans la communication des résultats de recherche et dans la compréhension des tendances et des modèles présents dans les données textuelles urbaines. Voici quelques outils couramment utilisés :

Nuages de Mots (Word Clouds) : Les nuages de mots affichent les mots les plus fréquemment utilisés dans un corpus de texte sous forme de graphique, où la taille des mots est proportionnelle à leur fréquence. Ils sont utiles pour identifier rapidement les termes clés liés à un thème urbain spécifique.

Graphiques et Diagrammes : Les graphiques, tels que les diagrammes en barres, les camemberts et les graphiques en courbes, sont utilisés pour représenter visuellement les statistiques et les tendances des données urbaines. Par

exemple, un diagramme en barres peut montrer la fréquence de certaines catégories de thèmes urbains.

Réseaux de Cooccurrence : Les réseaux de cooccurrence mettent en évidence les relations entre les mots ou les concepts en montrant comment ils sont liés dans le texte. Ces visualisations permettent de comprendre les associations entre les termes urbains.

Graphiques Temporels : Si les données textuelles contiennent des informations temporelles, les graphiques temporels peuvent illustrer comment les tendances urbaines évoluent au fil du temps. Ils sont particulièrement utiles pour l'analyse des tendances à long terme.

6-2 Cartographie des Données Textuelles pour une Perspective Spatiale

La cartographie des données textuelles consiste à associer des informations géo spatiales aux données urbaines, ce qui permet d'obtenir une perspective spatiale. Voici comment cela peut être réalisé :

Si les données textuelles incluent des informations de géo localisation (par exemple, des coordonnées GPS ou des noms de lieux), ces données peuvent être utilisées pour cartographier les textes sur une carte géographique. Cela permet de visualiser où les discussions, les événements ou les problèmes urbains se produisent.

En associant des informations de géo localisation aux données textuelles, il est possible de créer des cartes de sentiment qui montrent comment les émotions varient dans différentes parties de la ville. Cela peut aider à identifier les zones où les résidents sont plus ou moins satisfaits.

Les cartes thématiques affichent des informations sur des thèmes urbains spécifiques, tels que la répartition des crimes, la densité de la population ou les problèmes environnementaux. Ces cartes permettent de mieux comprendre la distribution spatiale de ces thèmes.

Les cartes d'infographie combinent des données textuelles avec des éléments visuels sur une carte pour créer des représentations visuelles plus riches des problèmes urbains.

Par exemple, une carte d'infographie peut montrer les zones à risque pour un problème de santé urbaine spécifique. La cartographie des données textuelles dans le contexte urbain offre une perspective spatiale qui peut aider les chercheurs, les urbanistes et les décideurs à prendre des décisions informées et à mieux comprendre les dynamiques urbaines. Elle permet également de présenter visuellement les résultats de la recherche de manière accessible au public.

7- Études de Cas Illustrants l'Utilisation du Text Mining dans la Recherche Urbaine

Le Text Mining a été largement utilisé pour aborder des problèmes urbains concrets en exploitant les données textuelles disponibles. Cette section présente quelques études de cas illustrant ces applications, ainsi que les avantages et les défis rencontrés.

Étude de Cas 1 : Analyse des Sentiments dans les Transports en Commun

Dans l'article « Recherche d'information et analyse de sentiments dans des tweets concernant les transports en Île de France »[5], (Paroubek, P et al, 2018). A partir d'un corpus de tweets, quatre tâches ont été proposées : identifier les tweets sur la thématique des transports, puis parmi ces derniers, identifier la polarité (négatif, neutre, positif, mixte), identifier les marqueurs de sentiment et la cible, et enfin, annoter complètement chaque tweet en source et cible des sentiments exprimés.

Douze équipes ont participé, majoritairement sur les deux premières tâches. Sur l'identification de la thématique des transports urbains.

Étude de Cas 2 : Prévision des Crimes Urbains

Dans leur article "Crime Prediction Using Machine Learning and Deep Learning" (Varun M et al, 2023)[6] les auteurs examinent plus de 150 articles pour explorer les différents algorithmes d'apprentissage automatique et d'apprentissage profond appliqués à la prédiction de la criminalité.

Fig 5. Wordcloud * du corpus de 150 articles pour dégager les mots clés

L'étude fournit un accès aux ensembles de données utilisés pour la prédiction de la criminalité par les chercheurs et analyse les approches les plus prometteuses appliquées dans les algorithmes d'apprentissage automatique et d'apprentissage profond pour prédire la criminalité, offrant des informations sur différentes tendances et facteurs liés aux activités criminelles.

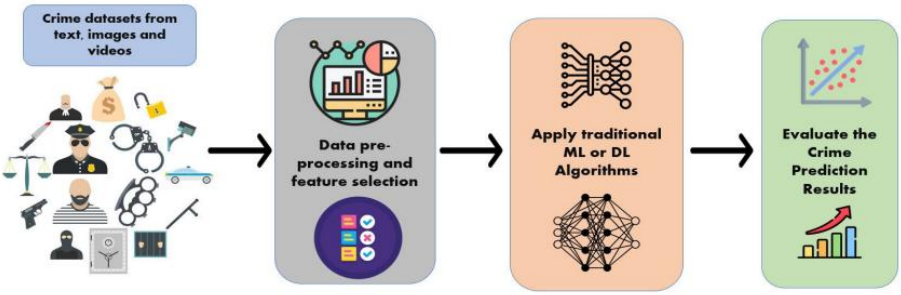

Fig 7. Flux structurel de prévision de la criminalité

Dans l'ensemble, ces ensembles de données fournissent des informations précieuses aux chercheurs pour élaborer des modèles de prédiction de la criminalité

Les modèles Random Forest (RF) sont ici utilisés pour analyser un large éventail de caractéristiques et de faire des prédictions sur les schémas de criminalité.

---

[5]Patrick Paroubek, Cyril Grouin, Patrice Bellot, Vincent Claveau, Iris Eshkol-Taravella, et al.. DEFT2018 : Recherche d'information et analyse de sentiments dans des tweets concernant les transports en Île de France. DEFT 2018 - 14ème atelier Défi Fouille de Texte, May 2018, Rennes, France. pp.1-11. ffhal-01839407

*Wordcloud :Un nuage de mots - Word Cloud en anglais - est une visualisation graphique des mots contenus dans un texte ou un site internet classés par nombre d'occurrences

En plus de ces techniques, les modèles traditionnels d'apprentissage automatique peuvent également être utilisés pour la détection d'anomalies et l'analyse de dans les données criminelles. En identifiant des modèles inhabituels ou des dans les données, les services répressifs peuvent détecter des activités criminelles potentielles et prendre des mesures pour les prévenir.

Étude de Cas 3 : Planification Urbaine Participative

Dans leurs articles “ From tweets to semantic trajectories: mining anomalous urban mobility patterns ” (Gabrielli,,et al , 2013)[7] Cet article propose et expérimente de nouvelles techniques pour détecter les modèles de mobilité urbaine et les anomalies en analysant les trajectoires extraites des traces de médias sociaux géolocalisés accessibles au public laissées par les citoyens (notamment Twitter).

En collectant un large ensemble de tweets géolocalisés caractérisant une zone urbaine spécifique au fil du temps, ils ont enrichi sémantiquement les tweets disponibles avec des informations sur leur auteur - c'est-à-dire un résident ou un touriste - et l'objet du mouvement - c'est-à-dire l'activité réalisée dans chaque lieu. (fig.8)

Fig8.les tweets géolocalisés en associant les lieux/catégories

[6]Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions. *IEEE Access*.

Étude de Cas 4 : Analyse des performances du tourisme

Dans leurs articles ”What do hotel customers complain about? Text analysis using structural topic model ’ (Hu, et al, 2019)[8] analysent La capacité à comprendre les causes des plaintes des clients est essentielle pour les hôtels afin d'améliorer leur qualité de service, la satisfaction de la clientèle et leurs revenus. Cette étude adopte une méthode novatrice d'analyse de texte basée sur un modèle de sujet structurel pour analyser 27 864 avis d'hôtels à New York City, et démontre qu'elle permet une meilleure compréhension de l'insatisfaction des consommateurs.

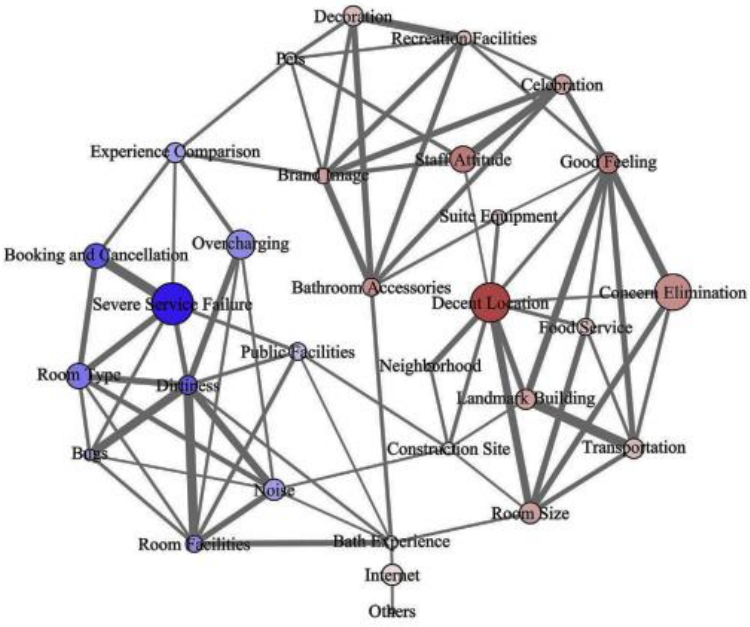

Fig9. Network map des thèmes de corrélations (Hu ,et al , 2019)

De plus, les chercheurs ont examiné comment les plaintes des clients varient en fonction de la catégorie de l'hôtel. Les résultats indiquent que les plaintes des clients concernant les hôtels haut de gamme sont principalement liées à des problèmes de service, tandis que les clients des hôtels économiques sont fréquemment dérangés par des problèmes liés aux installations.

[7]Gabrielli, L., Rinzivillo, S., Ronzano, F., & Villatoro, D. (2013, September). From tweets to semantic trajectories: mining anomalous urban mobility patterns. In *International Workshop on Citizen in Sensor Networks* (pp. 26-35). Cham: Springer International Publishing.

[8]Hu, N., Zhang, T., Gao, B., & Bose, I. (2019). What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management*, *72*, 417-426.

Cette recherche contribue à la littérature sur l'hospitalité en améliorant notre compréhension des aspects de l'insatisfaction des clients d'hôtels grâce à une analyse statistique rigoureuse, en mettant en évidence leurs corrélations et leur importance pour différentes catégories d'hôtels.

8- Limites et Défis du Text Mining en Recherche Urbaine

Malgré les avancées significatives dans le domaine du Text Mining en recherche urbaine, plusieurs limitations et défis subsistent. Cette section explore ces aspects afin de mieux comprendre les limites actuelles et les questions éthiques associées à la collecte et à l'analyse de données textuelles urbaines.

1. Qualité des Données Textuelles

Limites : Les données textuelles urbaines peuvent être de qualité variable, avec des erreurs grammaticales, des abréviations et des langues multiples. Cela rend parfois difficile l'analyse précise, en particulier lorsque les données sont générées par des sources non standardisées comme les médias sociaux.

Défi : Le prétraitement et la normalisation des données textuelles sont essentiels pour obtenir des résultats fiables. L'utilisation de techniques d'apprentissage automatique pour gérer la variabilité linguistique peut aider à surmonter cette limitation.

2. Biais dans les Données

Limites : Les données textuelles urbaines peuvent refléter les biais des auteurs, en particulier sur les médias sociaux. Certains groupes peuvent être sous-représentés, tandis que d'autres peuvent être surreprésentés.

Défi : Les chercheurs doivent être conscients des biais potentiels et prendre des mesures pour les atténuer, notamment en utilisant des techniques de pondération et d'échantillonnage.

3. Confidentialité et Protection de la Vie Privée

Limites : La collecte de données textuelles urbaines peut soulever des préoccupations en matière de confidentialité et de protection de la vie privée. Les informations personnelles peuvent être divulguées involontairement, ce qui soulève des questions éthiques.

Défi : Les chercheurs doivent respecter les normes éthiques et légales en matière de protection de la vie privée. L'anonymisation des données et l'obtention du consentement éclairé des participants sont des mesures importantes pour atténuer ces problèmes.

4. Représentativité des Données

Limites : Les données textuelles peuvent ne pas être représentatives de l'ensemble de la population urbaine. Certaines voix peuvent être exclues, ce qui peut biaiser les résultats.

Défi : Les chercheurs doivent être transparents sur les limites de leurs données et de leurs méthodologies. La combinaison de données textuelles avec d'autres sources de données peut aider à obtenir une image plus complète.

5. Complexité de l'Analyse

Limites : L'analyse de données textuelles urbaines peut être complexe en raison de la richesse du langage naturel. Comprendre le contexte et les nuances peut être un défi.

Défi : L'utilisation de techniques avancées d'analyse de texte, telles que l'analyse de sentiment, la classification de texte et la modélisation de sujets, peut aider à extraire des informations significatives. Cependant, cela nécessite une expertise en traitement automatique du langage naturel (NLP).

Questions Éthiques Liées à la Collecte et à l'Analyse de Données Textuelles Urbaines

La collecte et l'analyse de données textuelles urbaines soulèvent des questions éthiques importantes :

Consentement et Confidentialité : Les individus dont les données sont collectées doivent donner leur consentement éclairé. De plus, la confidentialité des données personnelles doit être préservée.

Transparence : Les chercheurs doivent être transparents quant à leurs méthodes de collecte et d'analyse des données, ainsi qu'en ce qui concerne les sources des données.

Biais : Les chercheurs doivent être conscients des biais potentiels dans les données et prendre des mesures pour les atténuer.

Utilisation Responsable : Les données textuelles urbaines ne doivent pas être utilisées pour nuire aux individus ou à des groupes particuliers. Les résultats de la recherche doivent être utilisés de manière responsable.

Équité : Les résultats de la recherche basée sur des données textuelles ne doivent pas renforcer les inégalités existantes ou marginaliser certains groupes.

Déontologie : Les chercheurs doivent adhérer aux normes éthiques et aux règles de déontologie lors de la collecte, de l'analyse et de la publication des données.

Bien que le Text Mining en recherche urbaine offre d'énormes possibilités, il est essentiel de traiter les limitations et les questions éthiques de manière proactive pour garantir des résultats fiables et éthiques.

9- Discussion et synthèse des aspects clés de la recherche :

Aspect	Points Clés du text mining
Avantages	- Compréhension des dynamiques urbaines.
	- Identification des préoccupations des citoyens.
	- Informations pour les politiques urbaines.
	- Intégration de données multiples.

	- Amélioration de la qualité de vie urbaine.
	- Aide à la prise de décision.
	- Outils de transparence dans la gouvernance urbaine.
Applications	- Analyse des sentiments dans les transports en commun.
	- Prévision des crimes urbains.
	- Planification urbaine participative.
	- Analyse des politiques urbaines.
	- Prévision des tendances urbaines.
Limites	- Qualité variable des données textuelles.
	- Biais potentiels dans les données.
	- Variabilité linguistique.
	- Complexité de l'analyse.
	- Nécessité de normalisation et de nettoyage.
	- Manque de données d'entraînement étiquetées.
Défis	- Gestion des biais dans les données.
	- Confidentialité et protection de la vie privée.
	- Représentativité des données.
	- Utilisation responsable des données.
	- Éthique et déontologie.

© Tableau récapitulatif des aspects clés du text mining pour la recherche urbaine .source : l'auteur 2023 .

Ce tableau offre une vue exhaustive des opportunités qu'offre le text mining dans la compréhension des dynamiques urbaines, allant de l'identification des préoccupations citoyennes à l'amélioration de la qualité de vie par le biais de politiques urbaines éclairées.

La diversité des applications, telles que l'analyse des sentiments dans les transports et la prévision des crimes, souligne la polyvalence de cette méthodologie.

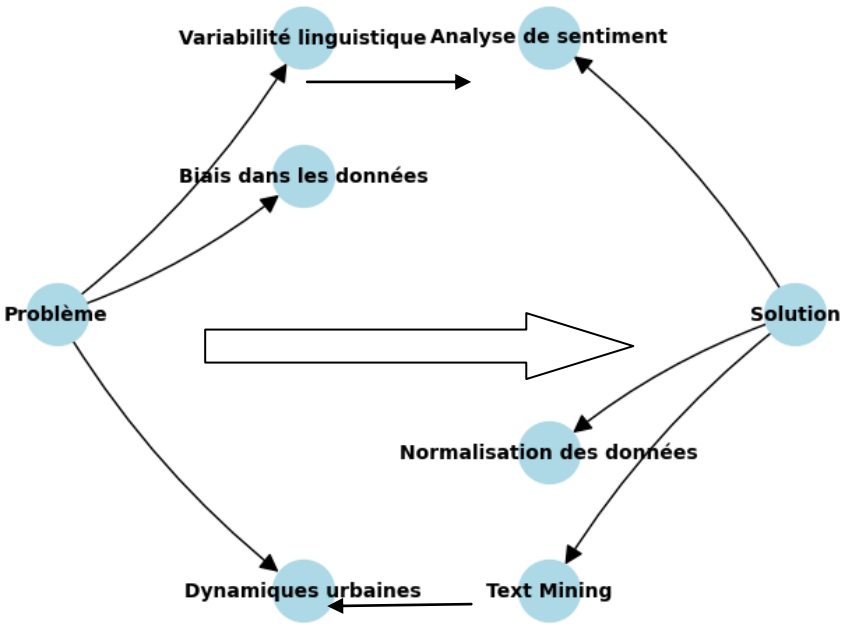

Fig10.diagramme de synthése des composantes du text mining dans la recherche urbaine Problèmes /solutions source : l'auteur 2023

Ce diagramme met en lumière les problèmes potentiels et les solutions envisageables. Il sert d'outil précieux pour la planification stratégique, la communication efficace et la prise de décision éclairée dans le domaine complexe de la recherche urbaine.

De plus, il souligne les défis inhérents, dont la qualité variable des données et les enjeux éthiques, ajoutant une nuance essentielle à l'appréciation de la viabilité du text mining dans un contexte urbain.

Conclusion

La revue de littérature sur le Text Mining en recherche urbaine met en lumière l'importance croissante de cette approche dans la compréhension des dynamiques urbaines, des comportements des citoyens et des politiques urbaines. Voici une synthèse des points clés abordés dans cette revue :

Analyse de Données Textuelles Urbaines dans Le Text Mining permet l'analyse de vastes ensembles de données textuelles urbaines provenant de sources variées telles que les médias sociaux, les forums en ligne, les rapports gouvernementaux et les avis de citoyens. Cette analyse peut fournir des informations précieuses pour la recherche urbaine.

Les avancées technologiques et méthodologiques ont permis de développer des techniques avancées d'analyse de texte, notamment l'analyse de sentiment, la classification de texte et la modélisation de sujets. Ces avancées ont ouvert de nouvelles perspectives pour la recherche urbaine.

La collecte de Données Textuelles Urbaines et Les sources de données textuelles urbaines sont diverses, allant des médias sociaux aux données gouvernementales. Les chercheurs exploitent ces sources pour obtenir des informations sur les préoccupations des citoyens et les tendances urbaines.

Le prétraitement des données textuelles est essentiel pour nettoyer, normaliser et segmenter les données. Les défis spécifiques au prétraitement des données urbaines comprennent la variabilité linguistique et la qualité variable des données.

L'analyse de fréquence des mots, l'analyse de sentiment et la modélisation de sujets sont des techniques couramment utilisées pour extraire des informations significatives des données textuelles urbaines.

L'utilisation de l'apprentissage automatique (Machine Learning) pour la classification de texte permet d'identifier et de catégoriser les thèmes urbains. Cette approche facilite la compréhension des préoccupations urbaines.

Les outils de visualisation, y compris la cartographie des données textuelles, permettent de représenter visuellement les résultats de manière accessible et informative.

Les études de cas ont démontré comment le Text Mining a été appliqué pour aborder des problèmes urbains concrets, tels que l'analyse des sentiments dans les transports en commun et la prévision du crime. Les limites actuelles du Text Mining en recherche urbaine comprennent la qualité des données, les biais potentiels, les questions de confidentialité et les défis liés à la complexité de l'analyse.

Bibliographie

Aas, C., Engen, V., & Øvervåg, K. (2016). Text analytics for open data. Government Information Quarterly, 33(2), 305-314.

Cai, L., Hu, Y., Liu, L., & Sheng, X. (2021). A Topic-Based Text Mining Approach for Analyzing Urban Policies. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM 2021) (pp. 2729-2732).

Hirschberg, J., Manning, C.D., 2015. Advances in natural language processing. Science 349 (6245), 261–266.

Gabrielli, L., Rinzivillo, S., Ronzano, F., & Villatoro, D. (2013, September). From tweets to semantic trajectories: mining anomalous urban mobility patterns. In International Workshop on Citizen in Sensor Networks (pp. 26-35). Cham: Springer International Publishing.

Hong, L., Fu, C., Wu, J., Frias-Martinez, V., 2018. Information needs and communication gaps between citizens and local governments online during natural disasters. Inf. Syst. Front. New York 20 (5), 1027–1039.

Howard, J., Ruder, S., 2018. Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 328–339.

Hu, Y., Deng, C., Zhou, Z., 2019a. A semantic and sentiment analysis on online neighborhood reviews for understanding the perceptions of people toward their living environments. Ann. Assoc. Am. Geogr. 109 (4), 1052–1073.

Hu, Y., Mao, H., McKenzie, G., 2019b. A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. Int. J. Geogr. Inf. Sci. 33 (4), 714–738.

Huang, L., Wu, Y., Zheng, Q., Zheng, Q., Zheng, X., Gan, M., Wang, K., Shahtahmassebi, A., Deng, J., Wang, J., Zhang, J., 2018. Quantifying the spatiotemporal dynamics of industrial land uses through mining free access social datasets in the mega hangzhou bay region, China. Sustainability 10 (10), 3463.

Iaconesi, S., 2015. Emotional landmarks in cities. Sociologica 9 (3), 22.
Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P., 2013. Practical extraction of disaster-relevant information from social media. In: Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion, 1021–1024.

Jang, K.M., Kim, Y., 2019. Crowd-sourced cognitive mapping: a new way of displaying people's cognitive perception of urban space. PloS One 14 (6).

Lai, Y., Kontokosta, C.E., 2019. Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities. Comput. Environ. Urban Syst. 78, 101383

Kadar, F. J., & Adhikari, A. (2019). Crime Prediction in Smart Cities Using Machine Learning Algorithms. In Proceedings of the 2nd International Conference on Communication, Devices and Computing (ICCDC 2019) (pp. 1-5).

Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions. IEEE Access.

Rahman, M. M., & Rafi, M. S. (2019). Analysis of Public Sentiments on Urban Transportation Issues Using Twitter Data: A Case Study of Dhaka City. IEEE Access, 7, 150740-150753.

Ribeiro, F. N. S., Santos, A. D. S., & da Silva, F. S. (2018). A Twitter-Based Approach to Understand Citizens' Perceptions about Urban Mobility. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2018) (pp. 768-775).

Velardi, P., Stilo, G., Tozzi, A. E., & Gesualdo, F. (2014). Twitter mining for fine-grained syndromic surveillance. Artificial intelligence in medicine, 61(3), 153-163.