Aide à l'extraction de connaissances d'entretiens semi-dirigés

Ghinevra Comiti*, Louiza Belarif*
Paul-Antoine Bisgambiglia*, Nathalie Lameta *, Paul Bisgambiglia*

* Université de Corse
comiti_g@univ-corse.fr, belariflouiza@gmail.com

1 Introduction et Contexte

La qualité de vie est une préoccupation clé des politiques publiques, mais sa définition reste complexe en raison de sa nature subjective et multidimensionnelle, dépendante des contextes locaux. En Corse, un territoire marqué par ses spécificités géographiques et sociales, nous souhaitons développer un indicateur de qualité de vie à l'échelle communale, combinant données quantitatives et qualitatives pour refléter au mieux les réalités locales et soutenir les décideurs.

Cet article se concentre sur l'analyse des données qualitatives recueillies via des entretiens individuels semi-directifs.

Traditionnellement, ces analyses reposent sur des méthodes manuelles, riches, mais chronophages. Pour en optimiser l'efficacité et soutenir ces démarches, nous intégrons des outils de traitement automatique du langage naturel (TALN), tels que RoBERTa (Liu et al., 2019) ou VADER (Hutto et Gilbert, 2014).

L'objectif est de permettre à nos modèles de TALN d'identifier, à partir des entretiens, les thèmes abordés par les enquêtés ainsi que l'opinion exprimée (positive, neutre ou négative). Ces modèles visent à accompagner le chercheur en identifiant rapidement les thèmes principaux et les sentiments exprimés, tout en réduisant certains biais interprétatifs.

Nous explorons ainsi comment ces outils peuvent enrichir et accélérer le travail et discutons de leurs implications pour la recherche sur la qualité de vie.

L'article aborde successivement les méthodes adoptées et les résultats obtenus, puis les perspectives futures.

2 Méthode

Un état de l'art nous a permis de définir les thèmes principaux de la qualité de vie. Cela permet à nos modèles de retrouver ces thèmes dans les comptes rendus d'entretien. Nous en avons identifié onze : logement, revenus, environnement, ressources, santé, sécurité, politique, communauté et relations, attractivité et éducation (Pesta et al., 2010; Group, 1993).

Notre travail s'est déroulé en plusieurs étapes : [1] Le texte des entretiens a été collecté ; [2] Nous avons effectué une phase de prétraitement : classification des interventions selon leur type (interventions de la chercheuse ou du participant), suppression des "stop-words", traduction en anglais (à cause de problèmes de performances des modèles francophones testés). [3] Nous

avons procédé à deux types d'analyses textuelles : une analyse des sentiments, qui classifie les phrases du répondant comme négatives, positives ou neutres à l'aide des modèles VADER et RoBERTa, et une analyse thématique, qui relie chaque phrase à un thème de la qualité de vie précédemment identifié, réalisée à l'aide d'un modèle de type Transformer ainsi que de l'approche LDA. [4] Une matrice synthétisant les thèmes abordés dans les entretiens et les sentiments exprimés a été élaborée.

3 Résultats et discussions

Pour évaluer les résultats de la matrice, nous les avons croisés avec le compte rendu de la chercheuse ayant réalisé les entretiens.

Nous retrouvons des éléments communs dans l'analyse automatique et dans l'analyse humaine. Par exemple, l'analyse automatique attribue un score négatif au terme de "santé", et on retrouve cela dans le compte rendu humain : les participants qui mentionnent cet aspect déplorent l'éloignement des services de santé ou leur difficulté d'accès. Notre analyse montre aussi un sentiment plutôt positif lié au thème "revenus". Cela peut paraître contre-intuitif et s'explique à la lecture des entretiens, car le sujet est peu mentionné, sauf dans 3 entretiens dans lesquels les répondants expliquent comment ils ont réussi à créer leurs entreprises. Ainsi, notre analyse a permis de révéler un biais dans nos entretiens en attirant notre attention sur un élément inattendu.

Pour corroborer nos résultats, nous avons fait tourner deux modèles distincts, pour chaque tâche. Les résultats de la matrice générée avec ce modèle étaient quasiment identiques.

En conclusion, ces modèles combinent analyse thématique et de sentiments pour restituer les grandes tendances des opinions des habitants, offrant une vision synthétique qui facilite l'élaboration de politiques publiques adaptées. Nous proposons donc une approche innovante d'analyse qualitative automatisée des entretiens individuels via des modèles de TALN, permettant d'extraire rapidement thèmes et sentiments tout en identifiant les éléments divergents à prioriser.Les perspectives sont nombreuses, ce qui nous conduit à ouvrir un agenda de recherche pour améliorer cet outil, et en explorer davantage les implications possibles.

Références

Group, W. (1993). Study protocol for the World Health Organization project to develop a Quality of Life assessment instrument (WHOQOL). *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation 2*(2), 153–159.

Hutto, C. et E. Gilbert (2014). VADER : A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media 8*(1), 216–225, doi: 10.1609/icwsm.v8i1.14550. Number : 1.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, et V. Stoyanov (2019). RoBERTa : A Robustly Optimized BERT Pretraining Approach. arXiv :1907.11692.

Pesta, B. J., M. A. McDaniel, et S. Bertsch (2010). Toward an index of well-being for the fifty U.S. states. *Intelligence 38*(1), 160–168, doi: 10.1016/j.intell.2009.09.006.