Leveraging AI Models for Automated Pattern Detection in Citizen Participation Data

Amal Marzouki[a*], *Boutheyna* Nouri[b], *Sehl* Mellouli[c]

[a]Department of Information Systems, Université du Québec à Rimouski, Campus de Lévis 1595 Boulevard Alphonse-Desjardins, Lévis, QC G6V 0A6, Canada; amal_marzouki@uqar.ca; ORCID:0009-0002-9931-1564

[b]Department of Information Systems, Université du Québec à Rimouski, Campus de Lévis 1595 Boulevard Alphonse-Desjardins, Lévis, QC G6V 0A6, Canada; Boutheyna.Nouri@uqar.ca; ORCID: 0009-0007-1339-1564

[c]Department of Information Systems, Université Laval, 2325, rue de la Terrasse, Québec, QC G1V 0A6, Canada; sehl.mellouli@fsa.ulaval.ca; ORCID: 0000-0003-0977-4603

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 10 July 2025

Abstract. Citizen Participation (CP) is essential for urban projects, traditionally done through face-to-face meetings. Information Technologies (IT) have introduced electronic participation (e-Participation), enhancing inclusivity and engagement. CPPs generate valuable data for decision-making, but processing large volumes of unstructured data is challenging. Traditional methods are inefficient. AI algorithms can improve data analysis, automate classification, detect patterns, and extract relevant information, reducing the workload for decision-makers. This research explores how AI can detect and classify patterns in CPP data, contributing to best practices in applying AI to government operations.

Keywords. Citizen participation, Artificial Intelligence models, Data analysis, Pattern detection, Semantic, spatial and temporal analysis.
Poster, DOI: https://doi.org/10.59490/dgo.2025.1068

1. Introduction

Citizen Participation (CP) integrates citizens into decision-making for urban and community projects. Historically, CP was facilitated through face-to-face meetings, allowing stakeholders to share their perspectives. The advent of Information Technologies (IT) has introduced electronic participation (e-Participation), leveraging tools like blogs and social networks to enhance inclusivity and engagement. CPPs generate valuable data that must be meticulously analyzed to aid decision-making and improve service delivery (Marzouki et al., 2022b). However, processing large volumes of unstructured data is challenging. Traditional manual methods are inefficient and limit scalability. AI algorithms can automate the classification of citizen posts, detect patterns, and extract relevant information, reducing the workload for facilitators and edecision-makers (Bonsón et al., 2015; Marzouki et al., 2022a). This research explores how AI can enhance CPP data analysis, contributing to better decision-making and government service delivery. By leveraging AI, governments can gain deeper insights into citizens' needs and preferences, ensuring that actions are aligned with community demands (Burgess-Allen & Owen-Smith, 2010; Sanford & Rose, 2007).

2. Theoretical background

To integrate Artificial Intelligence (AI) into participatory platforms, we focus on the analysis of participation data derived from these platforms. At the core of this conceptualization, we propose a systemic approach in which the human aspect is represented by citizens, facilitators, and decision-makers. This model integrates

AI techniques for faster, more efficient, and responsive interactions to citizens' needs.
Key concepts include textual data analysis (TDA) and the semantic, spatial, and temporal contextualization of participation data. TDA examines textual data to extract meaningful patterns using natural language processing (NLP) techniques like Named Entity Recognition (NER) and text classification. AI models such as LDA, CNN, LSTM, and RBFN are considered for text classification. NER identifies key entities in text, classifying results into spatial, semantic, and temporal dimensions (Marrero et al., 2013).
The anticipated results aim to extract knowledge about citizens' living contexts, classify this knowledge into distinct patterns, and inform decision-makers. This approach promotes better decisions aligned with citizens' needs, generating valuable insights to refine policies, improve services, and foster stronger engagement between decision-makers and communities. Anchoring decisions in citizen-centered data strengthens trust and governance effectiveness.

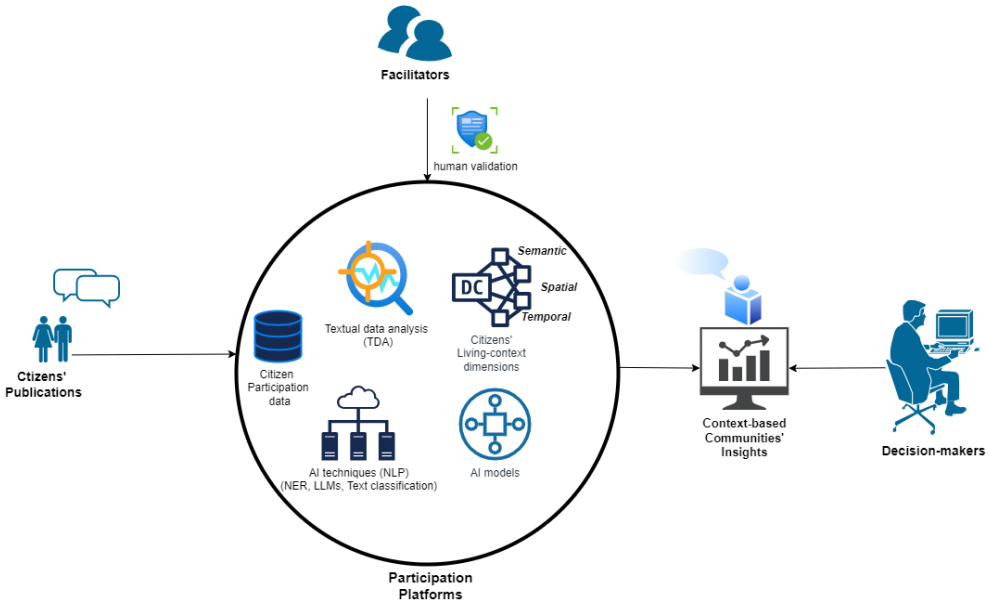

Fig. 1 – A conceptual model for AI models integration in citizens' living context identification in citizen participation data (authors main contribution)

2.1. Textual Data Analysis (TDA)

TDA systematically examines textual data to extract meaningful patterns, isnsights, or classifications using techniques like natural language processing (NLP), text mining, sentiment analysis, and topic modeling. In our study, TDA allows qualitative assessment and automatic classification of citizens' posts, reducing the workload for facilitators and improving decision-making (Feldman & Sanger, 2007; Mohit, 2014).

2.2. Semantic, Spatial, and Temporal Contextualization

After collecting citizens' opinions and feedback, data analysis and information extraction processes are necessary. Patterns capturing stakeholders' living contexts are classified into three dimensions (Marzouki et al., 2022a; Meersman, 1997):

1. **Semantic Dimension:** Represents the meaning of information provided by citizens, including issues, suggestions, lived experiences, metrics, governing entities, references, questions, compliments, links, tags, hashtags, and emoticons (Marzouki et al., 2022a).
2. **Spatial Dimension:** Addresses the question "Where," representing objects in geographical space, including internal and external spatial entities, similar organizations, hypothetical and approximated positions, cities, provinces, countries, hashtags, and location stamps (Lafrance et al., 2019; Marzouki et al., 2022a).
3. **Temporal Dimension:** Addresses the question "When," including past and future calendar expressions, deictic and anaphoric expressions, founded calendar expressions, and temporal hashtags (Lafrance et al., 2019; Marzouki et al., 2022a).

2.3. Named Entity Recognition (NER)

NER identifies and categorizes important names and proper nouns in text, enabling information extraction of Named Entities (NE) (Marrero et al., 2013). We will use pre-trained models due to their high accuracy:

1. **Spacy:** Utilizes deep learning for NLP, processing text from tokenization to annotation (Dasagrandhi, 2021; Partalidou et al., 2019).
2. **Stanford CoreNLP:** Java-based tool with feature extractors for named entities, using a sequence of annotations (Bondielli et al., 2018).
3. **OpenNLP:** Contains components for a complete NLP system, including sentence detection, tokenization, and parsing (Mohanan & Samuel, 2016).
4. **Gate:** Open-source tool with multilingual support, providing reusable processing resources for NLP tasks (Cunningham et al., 2014; Elsherif et al., 2019).
5. **NLTK:** Python library for NLP tasks, analyzing data programmatically (Siva Rama Rao et al., 2022).

This approach aims to extract knowledge about citizens' living contexts, classifying this knowledge into distinct patterns, and informing decision-makers in their decision-making processes. By anchoring decisions in citizen-centered data, this approach strengthens trust and the overall effectiveness of governance .

2.4. Large Language Models (LLMs)

LLMs are a category of deep learning models designed to understand and generate human language. These models are built on transformer architectures, which enable them to learn complex patterns in text data by processing input sequences in parallel and capturing long-range dependencies between words. LLMs are pre-trained on massive corpora of text and can be fine-tuned for specific tasks such as text classification, named entity recognition (NER), and text generation (Brown et al., 2020). Below, we discuss some of the most prominent LLMs: GPT-4, BERT, RoBERTa, DistilBERT, and T5.

1. **GPT-4 (Generative Pre-trained Transformer 4):** GPT-4 is a state-of-the-art generative language model developed by OpenAI. It has been trained on a vast corpus, which enables it to perform well on a wide range of natural language processing (NLP) tasks (Brown et al., 2020).
2. **BERT (Bidirectional Encoder Representations from Transformers):** BERT, developed by Google, is a transformer-based model designed to understand the contextual relationships between words in a sentence. It is primarily used for tasks like named entity recognition (NER), question answering, and sentence classification (Devlin et al., 2019).
3. **RoBERTa (A Robustly Optimized BERT Pretraining Approach):** RoBERTa is an optimized version of BERT developed by Facebook AI. The improvements result in better performance on a range of NLP tasks, particularly for tasks that require understanding the relationship between words within large text corpora (Liu et al., 2019)
4. **DistilBERT (Distilled Version of BERT):** DistilBERT is a smaller, more efficient version of BERT that retains 97% of BERT's language understanding while being 60% faster and using 60% less memory (Sanh et al., 2019)
5. **T5 (Text-to-Text Transfer Transformer):** T5, developed by Google Research, is a transformer-based model that treats every NLP task as a text-to-text problem. It is highly versatile and can perform a wide variety of NLP tasks by leveraging a unified architecture and large-scale pretraining (Raffel et al., 2020)

2.5. Text Classification

Text classification assigns a document to a predefined category. It can be binary or multi-class, where binary classification can also be used for multi-label classification (Anis, 2022).

2.6. Artificial Intelligence Algorithms

AI algorithms, inspired by human intelligence, perform tasks like learning, reasoning, and problem-solving. They process and analyze complex data, including natural language (Russell & Norvig, 2016). Effective algorithms for text data analysis include:

1. **LDA:** Probabilistic model for discrete data clusters, identifying hidden topics (Sayadi, 2017).
2. **CNN:** Network using minimal preprocessing, useful for detecting named entities (Widiastuti, 2019).
3. **LSTM:** Extends memory of RNNs, storing states of each cell for sequential processing (Schmidhuber & Hochreiter, 1997).
4. **RBFN:** Neural network for curve-fitting in high-dimensional space, useful for classifying participation data into semantic, spatial, or temporal patterns (Zhang et al., 2011).

3. Methodology

This paper explores the integration of AI models for semantic, spatial, and temporal analysis of citizen participation data on digital platforms. The methodology involves several steps:

- **Exploration of Citizen Participation Platforms:** Establish foundational understanding of functionalities and data generated.
- **Framework for CPP Data Contextualization:** Analyze data based on semantic, spatial, and temporal dimensions.
- **Review of Textual Data Analysis Techniques:** Identify key dimensions for analysis.
- **Search for Appropriate AI Architectures:** Design AI models tailored to each dimension.
- **Evaluation of AI Model Designs:** Assess accuracy scores for models like Named Entity Recognition (NER) and Entity Classification.
- **Selection and Integration of Models:** Choose the best models for each dimension and analyze results to understand citizen participation patterns.

3.1. Choice of Pre-trained NER Model:

We selected a pre-trained NER model based on performance, corpus size, and integration environment. After comparing several models, including SpaCy, Stanford Core NLP, OpenNLP, Gate, and NLTK, we excluded Java-based models due to integration constraints. Between SpaCy and NLTK, SpaCy was preferred for its larger corpus size and better performance. After evaluating the available SpaCy models, we chose `fr-core-news-sm` for its smaller size and efficiency in identifying patterns in citizen participation texts.

4. Proposed AI Models for Semantic, spatial and temporal Analysis of citizen participation data on digital government platforms

4.1. The proposed Semantic Dimension Model

- **Design of Solution 1 of the Semantic Model**

This solution is based on the pre-trained SpaCy model with custom entity recognition. The goal is to detect and classify semantic patterns from a given text. Semantic patterns include: Question, Governing Entity, Reference, Number/Metric, Link or document/file, Tag, Hashtag, Emoticon.
Figure 2 shows the semantic patterns detected by Solution 1 (using the NER model) and Solution 2 (using the LLM model) in the text.

- **Design of Solution 2 of the Semantic Model**

In this solution, we will employ Large Language Models (LLMs) to enhance the detection and classification of semantic patterns within citizen participation data. The model is pre-trained on vast amounts of textual data and by fine-tuning it, we can detect and classify complex semantic entities which are listed in the solution 1 of the semantic model. By integrating LLMs, the model strengthens the decision-making framework by offering a comprehensive understanding of citizen engagement and by improving the data processing on digital government platforms. The illustrative diagram for this solution is similar to Solution 1, as it shares the same inputs and outputs; however, the model used will be different to detect specific semantic patterns.

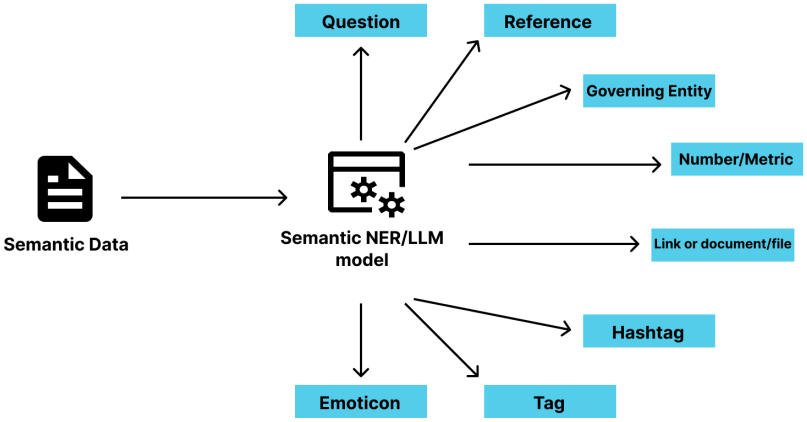

Fig. 2 – Illustrative diagram of the Semantic Dimension Model using NER/LLM

4.2. The proposed Spatial dimension Model

One essential dimension in data analysis is the extraction of spatial data. Two solutions are proposed: one using the SpaCy NER model with a neural classification model, and one using the pre-trained SpaCy NER model combined with a geolocation API and an algorithmic classification process.

- **Design of Solution 1 of the Spatial Model** This solution involves implementing two complementary models:

- Spatial NER Model: Similar to the semantic dimension, input data is prepared, annotated, and divided into training and testing corpora. This model extracts spatial patterns without classifying them. Figure 3 shows the spatial entity that this model is designed to detect in the text. **- Spatial Entity Classification Model:** This

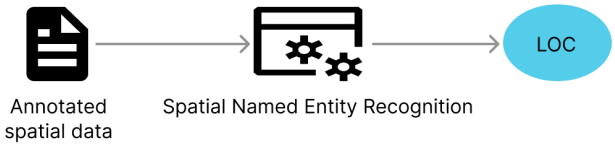

Fig. 3 – Illustrative diagram of the Spatial NER model

model classifies the spatial patterns detected by the first Spatial NER model, processing the annotated corpus and outputting six class categories using a Feed Forward network. Figure 4 shows the spatial patterns that this model is designed to classify.

- **Design of Solution 2 of the Spatial Model**
  This solution uses the pre-trained SpaCy model to identify location and organization entities (Loc, Org), then prepares an NER model to detect classes like Cities_provinces_Countries, Spatial entity approximated position, and Hashtag. The entity results are then classified according to logical constraints. Figure 5 illustrates the second solution as well as the spatial patterns that this solution is designed to classify.

4.3. The proposed Temporal dimension Model

- Design of Solution 1 of the Temporal Model: This solution includes two complementary models: a Temporal NER model, similar to the semantic dimension and the first spatial solution, which detects

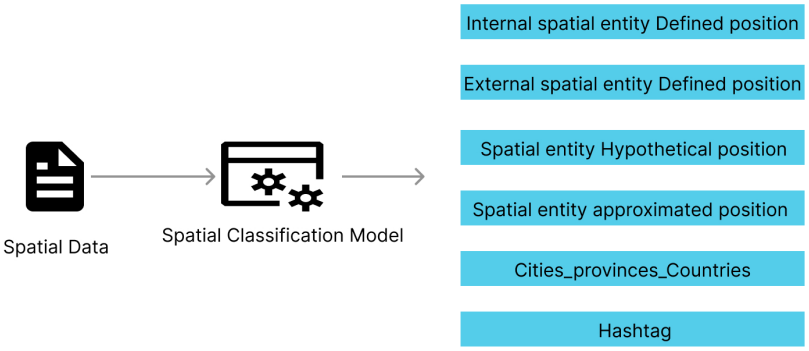

Fig. 4 – Illustrative diagram of the first spatial classification model

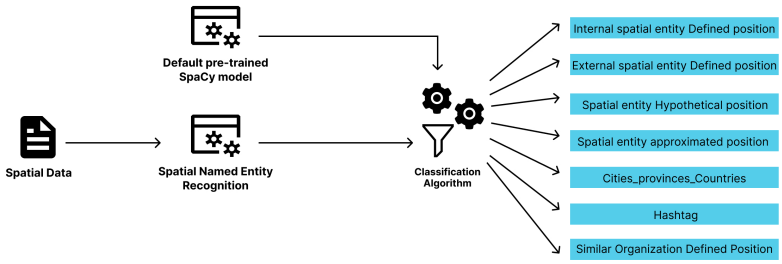

Fig. 5 – Illustrative diagram of the second spatial classification solution

temporal patterns from the annotated corpus, and a Temporal Entity Classification model that automatically classifies the temporal entities detected by the Temporal NER model. Figure 6 shows the temporal entity that this model is designed to detect in the text. Figure 7 shows the temporal patterns that this model is designed to classify.

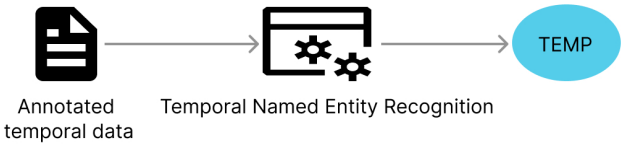

Fig. 6 – Illustrative diagram of the Temporal NER model

- Design of Solution 2 of the Temporal Model: This solution uses the pre-trained SpaCy model to detect the date from the (DATE) entity after translation, then applies the NER model trained with collected temporal data to identify deictic expressions and temporal hashtags. The extracted information is processed through a classification algorithm based on logical classification rules. Figure 8 illustrates the second solution as well as the temporal patterns that this solution is designed to classify.

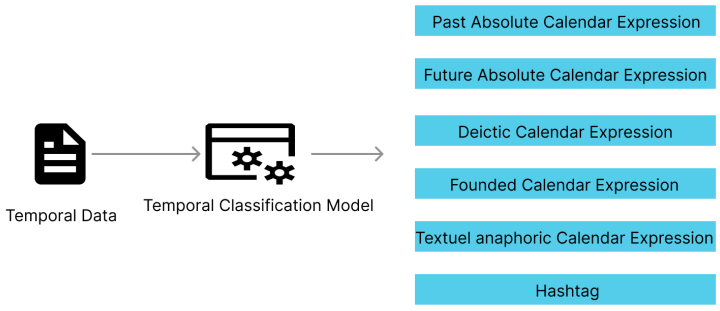

Fig. 7 – Illustrative diagram of the temporal classification model

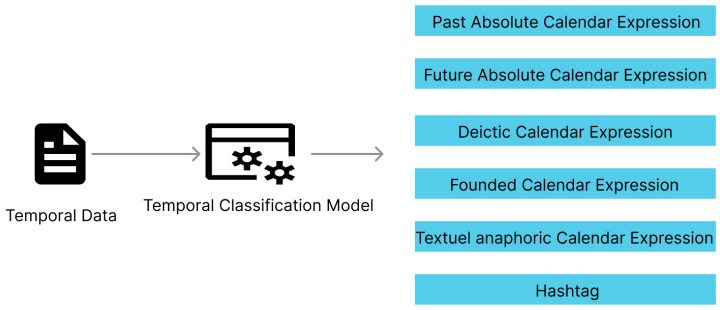

Fig. 8 – Illustrative diagram of the temporal classification model

5. Discussion and Conclusion

This study explores the integration of AI models for semantic, spatial, and temporal analysis in citizen participation data on digital platforms . Automating data analysis through AI enhances the ability to identify nuanced patterns, supporting scalable and efficient decision-making processes tailored to citizen needs (Marzouki et al., 2022a). Semantic analysis captures stakeholders' experiences and concerns, spatial analysis further contributes by mapping out the geographic dimensions of citizen inputs, thereby aligning community needs with spatial strategies in urban planning (Lafrance et al., 2019), and temporal analysis tracks changes over time (Andrienko et al., 2010), fostering proactive governance (Marzouki et al., 2022b).

The integration of AI-driven methodologies represents a significant advance in the transformation of digital citizen participation platforms. This approach addresses data scalability and complexity, improving decision-making processes in urban governance. By bridging the gap between citizen data and actionable insights, it fosters more inclusive and data-informed policy-making.

Our research aims at automating the analysis of large volumes of textual data, in order to provide actionable insights for city decision-makers. Future work will include collecting real citizen participation data, testing the proposed models, and developing a participatory platform integrated with decision support mechanisms. Further refinement of these AI models will enhance the effectiveness of citizen participation platforms, including the analysis of other data types, like images and videos, to provide a comprehensive understanding of citizens' living contexts.

References

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., & Tominski, C. (2010). Space, time and visual analytics. *International Journal of Geographical Information Science*, *24*(10), 1577–1600.

Anis, A. (2022, March). Pytorch lstm: The definitive guide. https://cnvrg.io/pytorch-lstm

Bondielli, A., Passaro, L., & Lenci, A. (2018). Corenlp-it: A ud pipeline for italian based on stanford corenlp. *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, 57–61.

Bonsón, E., Royo, S., & Ratkai, M. (2015). Citizens' engagement on local governments' facebook sites. an empirical analysis: The impact of different media and content types in western europe. *Gov. Inf. Q.*, *32*, 52–62.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Burgess-Allen, J., & Owen-Smith, V. (2010). Using mind mapping techniques for rapid qualitative data analysis in public participation processes. *Health Expectations*, *13*(4), 406–415. DOI: https://doi.org/10.1111/j.1369-7625.2010.00594.x.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., & Derczynski, L. (2014). *Developing language processing components with gate version 8*. University of Sheffield Department of Computer Science.

Dasagrandhi. (2021). Understanding named entity recognition pre-trained models. https://blog.vsoftconsulting.com/blog/understanding-named-entity-recognition-pre-trained-models#:~:text=Stanford%5C%20NER

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Elsherif, H. M., Alomari, K. M., AlHamad, A. Q. M., & Shaalan, K. (2019). Arabic rule-based named entity recognition system using gate. *MLDM (1)*, 1–15.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.

Lafrance, F., Daniel, S., & Dragićević, S. (2019). Multidimensional web gis approach for citizen participation in urban evolution. *ISPRS International Journal of Geo-Information*, *8*(6), 253.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marrero, M., et al. (2013). Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards Interfaces*, *35*(5), 482–489.

Marzouki, A., Mellouli, S., & Daniel, S. (2022a). The identification of stakeholders' living contexts in stakeholder participation data: A semantic, spatial, and temporal analysis. *Land*, *11*(6), 798.

Marzouki, A., Mellouli, S., & Daniel, S. (2022b). Understanding issues with stakeholders participation processes: A conceptual model of spps' dimensions of issues. *Government Information Quarterly*, *39*(2), 101668.

Meersman, R. (1997). Introduction: An essay on the role and evolution of data(base) semantics. In *Database applications semantics* (pp. 1–7).

Mohanan, M., & Samuel, P. (2016). Open nlp based refinement of software requirements. *International Journal of Computer Information Systems and Industrial Management Applications*, 293–300.

Mohit. (2014). Named entity recognition. In *Natural language processing of semitic languages* (pp. 221–245).

Partalidou, E., Spyromitros-Xioufis, E., Doropoulos, S., Vologiannidis, S., & Diamantaras, K. (2019). Design and implementation of an open source greek pos tagger and entity recognizer using spacy. *IEEE/WIC/ACM International Conference on Web Intelligence*, 337–341. DOI: https://doi.org/10.1145/3350546.3352543.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1–67.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson.

Sanford, C., & Rose, J. (2007). Characterizing eparticipation. *International Journal Of Information Management*, *27*(6), 406–421. DOI: https://doi.org/10.1016/j.ijinfomgt.2007.08.002.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sayadi, K. (2017). *Classification du texte numérique et numérisé. approche fondée sur les algorithmes d'apprentissage automatique* [Doctoral dissertation, Université Pierre et Marie Curie-Paris VI].

Schmidhuber, J., & Hochreiter, S. (1997). Long short-term memory. *Neural Comput*, *9*(8), 1735–1780.

Siva Rama Rao, A. V., Vamsi, P. V. V., Rashmika, N., Hemanth, K., & Aditya Kumar, K. (2022). Named entity recognition using stanford classes and nltk. *Proceedings of Second International Conference on Sustainable Expert Systems: ICSES 2021*, 583–597.

Widiastuti, N. I. (2019). Convolution neural network for text mining and natural language processing. *IOP Conference Series: Materials Science and Engineering*, *662*(5), 052010.

Zhang, M., Wang, K., Zhang, C., Chen, H., Liu, H., Yue, Y., & Qi, X. (2011). Using the radial basis function network model to assess rocky desertification in northwest guangxi, china. *Environmental Earth Sciences*, *62*, 69–76.