Engineering For Social Change
A. Cooper et al. (Eds.)

doi:10.3233/ATDE240849

Text-Mining of E-Participation Platforms: Applying Topic Modeling on Join and iVoting in Taiwan

Moritz SONTHEIMER[a,1], Jonas FAHLBUSCH[d], Tim KORJAKOW[e], Shuo-Yan CHOU[b, c]

[a] *Mechanical Engineering Department, National Taiwan University of Science and Technology*

[b] *Industrial Management Department, National Taiwan University of Science and Technology*

[c] *Intelligent Manufacturing Innovation Center, National Taiwan University of Science and Technology*

[d] *Chair for Work, Technology and Participation, Technical University of Berlin*

[e] *Computer Science Department, Technical University of Berlin*

Abstract E-participation platforms have emerged as digital tool to facilitate citizen engagement through online deliberation, voting and oversight processes. The digital add-on for *participatory democracy* (Carole Pateman) can be found in different countries around the world. In Asia, the rollout of two platforms in Taiwan, namely iVoting and Join, has captured the attention of Western media outlets. However, there is little literature on the exact content debated and agreed on these platforms up to this point. Utilising recent advancements in NLP, we explore the content of the proposals that were made. In our study, we combine new approaches of text mining with political analysis on Taiwan's e-participation platforms. The dataset, which includes 14,118 proposals from 2015 to 2022, has resulted in a distinct topic model being constructed for each platform. With the help of our method, we were able to cluster the proposals thematically and show which concerns were articulated and with how much approval. Based on a random sampling of 110 proposals, we were able to determine that our method assigns 81.82% of the proposals to the corresponding cluster. This can also significantly overcome language barriers, as we employed a translation pipeline within the text-mining process from Chinese into English. Our method is adaptable to e-participation platforms in various languages, providing decision-makers with a more comprehensive tool to understand citizens' needs and enabling the formulation of more informed and effective policies.

Keywords. Policy Informatics, Text Mining, LLMs, E-participation, Transdisciplinary Engineering

Introduction

Over the past ten years, Taiwan has witnessed significant innovations in public-private collaboration and digital co-creation. During the rise of the sunflower movement in the year of 2014, technical tools were deployed to make political processes more transparent,

[1] Corresponding author, Email: D10703815@mail.ntust.edu.tw.

e.g. live broadcasting of the occupation of Taiwan's parliament (Legislative Yuan) or the introduction of simplified visualisation of government spendings by g0v (a grassroot social movement community) instead of opaque Excel sheets by the state. One indication of this evolution, and a sign that the concerns of the movement should be translated into political actions, are the two e-participation platforms Join and iVoting. Though they represent complementary facets of Taiwan's digital deliberation initiatives, both were established to enhance informed decision-making and civic engagement. The primary difference between Join and iVoting lies in their scope, functionality and deployment. While iVoting focuses specifically on enhancing the electoral process through informed voting, Join provides a broader platform for civic engagement, allowing the creation, discussion, review and supervision of citizen proposals [1]. Join is utilised for engaging with governance issues at various levels, including national concerns, whereas i-Voting facilitates electronic voting on a range of decisions, especially local matters in Taipei.

In order to implement the idea of whole-citizen participation, the National Development Council promoted Join in February 2015. This platform empowers both Taiwanese citizens and foreigners residing in Taiwan (and are able to read and write Chinese) to engage in online discussions actively. It goes beyond simple e-petitioning by incorporating four key functionalities: initiating proposals, voting on others' proposals, commenting & discussing, and giving feedback on draft legislation. In contrast, iVoting or Intelligent Voting is Taiwan's first Voting Advice Application (VAA) aimed at enhancing democratic engagement through informed and intelligent voting. It was introduced during the 2012 Legislative Yuan elections. The platform allows users to become members, complete issue position diagnostics, and thus make informed voting decisions. From its launch in October 2011 to 2012, iVoting attracted 1,400 members and had over 40,000 visitors [2]. In recent years, however, there has been little activity on the site, which is also reflected in the decline in proposals. One factor might be the city government's use of the tool to present only pre-selected options for a controversial local redevelopment project in Taipei's Shezidao in 2016. This led to criticism regarding the lack of democratic choice and accountability, subsequently eroding public trust in the platform [3].

Understanding the topics that citizens discuss online and identifying the issues that garner the most support is crucial for policymakers and analysts [4]. Equally important is discerning which users frequently abandon topics, particularly in Taiwan's cyberspace that is subject to a high volume of false reports and external attacks [5]. As part of a participatory process with a *suggestional* nature, an optimal evaluation should encompass the citizens' requests for changes, along with the reasons for implementing these suggestions. Join proposals are encapsulated within both the concern / suggestion and the comments. However, this research focused exclusively on the proposals themselves, leaving the examination of citizens' comments for future study.

Although the term 'transdisciplinary' (TD) is inconsistently defined [7, 8], especially in engineering [7], our research emphasises the conceptual framework proposed by [8], which highlights that TD research should focus on real-world problems, foster collaboration (between scholars and non-academics), and involve evolving methodologies. With the deployment of our Natural Language Processing (NLP) method, we support people in politics and administration in gaining an improved overview of unfiltered and uncategorised online proposals. Existing research on Join sheds light on petition dynamics [1]. However, to understand thematic shifts and topical tendencies across a seven-year window, this research delves deeper into proposals and their concerns in the domain of policy informatics.

1. Leveraging Text-Mining Techniques and NLP in the Evolving Field of Policy Informatics

There is a recent rise in literature that tries to enhance traditional policy analysis methods by fusing them with computer science. The newly coined TD field, Policy Informatics, deals with the application of computer science methods and advancements within (conventional) policy analysis. Its transdisciplinary nature lies in the combination of concepts, methods, and processes from multiple fields to address policy and governance challenges through the use of computational tools and stakeholder collaboration [9] . In the field of text-mining techniques, this includes Named Entity Recognition (NER) to find important entities, Topic Modeling to uncover main themes, Sentiment Analysis to understand public opinion, Keyword Extraction to identify key terms and text summarization. In recent years and before the breakthroughs in NLP, significant progress has been published on these topics.

Bolivar [10] analysed how policymakers perceive the utility of social media platforms in facilitating civic engagement and participation in the context of public services. The findings indicate a recognition of social media's potential to enhance transparency, increase public participation, and foster a more interactive dialogue between the government and citizens. However, concerns are also raised about issues such as information overload, the quality of engagement, and the digital divide. Hagen et al. [11] provide a foundational review of techniques used to analyse text in the social sciences, with a focus on e-petition data through NLP. Their work showcases the potential of automated content analysis to extract emergent topics from vast quantities of textual data, offering a more nuanced understanding of public interests than conventional categorization schemes. By applying Latent Dirichlet Allocation (LDA) and its extensions, Hagen et al. theorise how topic modelling can uncover the latent thematic structures within e-petitions, highlighting the dynamic nature of public concerns and interests. Due to the rapid advances in the field of NLP, the use of LLM-based text mining for unstructured e-participation platforms is still a relatively rare approach in the field of policy informatics. We therefore propose an exemplary workflow to show how LLMs can be a useful tool in the discipline's toolbox.

LLMs, through pretraining and therefore contextualization on most of the human-created text, facilitate the extraction of meaningful insights from large-scale textual datasets, thereby augmenting the analytical depth achievable in policy research. This integration not only enables a much more nuanced exploration of policy implications, stakeholder sentiments, and regulatory outcomes, contributing to a more informed decision-making process, but also deals with problems such as changing language in e-petitions out-of-the-box [11]. The application of LLMs in policy analysis enables a paradigm shift towards leveraging computational intelligence for understanding implications and changes in policies in real time. It helps to enhance policy frameworks and promotes a proactive approach to predict and tackle emerging societal issues.

2. Case Study

In order to show how the latest advancements in LLMs can be used to facilitate Policy Informatics and enhance previously proposed methods, we revisit unstructured data from Taiwan's e-participation platforms Join and iVoting.

2.1. Data Collection from Join and iVoting

Our dataset includes 14,118 proposals from Join and 454 from iVoting. The proposals were submitted between 2015 and 2022. The information captured for each proposal encompasses title, content, submission date, proposer's username, background, rationale, and vote count. Personal identifiers such as gender, age, and real names were optional for submission and are not present in our dataset. As mentioned, the iVoting platform has relatively few proposals and even some - exemplified by the Shezidao case - that are government-initiated. This presents challenges for direct comparison. Consequently, the subsequent figures focus on displaying the quantity of proposals on the Join platform throughout the analysed period, see Figure 1.

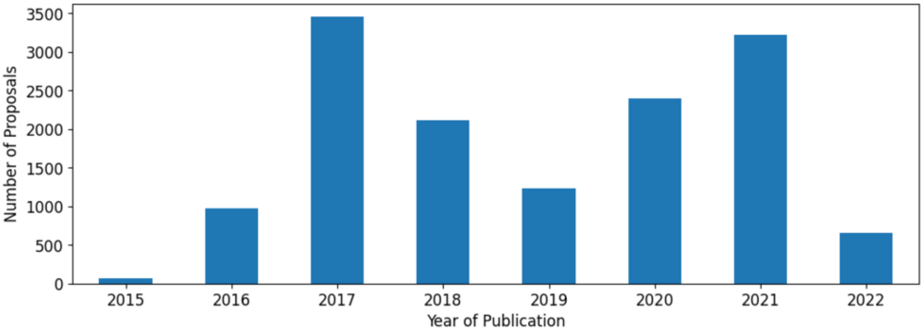

Figure 1. Number of publications on Join from 2015 to 2022.

On Join, before opening a proposal for endorsement through public voting, the public administration first assesses its coherence and suitability. Proposals with offensive content will be sorted out. The approved ones remain available on the platform for a duration of 60 days to gather supporting votes and further comments from the public. After this deliberative period, if at least 5,000 votes in support have been accumulated, then the relevant government ministry responsible for the issue area of the proposal must take action. Consequently, the responsible authority has to contact the proposer initiatively to understand his/her demands. It is obligatory to organise a meeting for the discussion of policy formulation, ensuring that the minutes of this meeting are made publicly available. Subsequently, the responsible authority is mandated to issue a well-founded response to the proposal within a two-month period. Based on this official procedure, we decided to develop text modelling approaches to uncover main themes of the proposals in order to create categories with the help of NLP.

2.2. Topic Modeling

In our specific setup, we use BERTopic's [12] modular approach and address the challenge of analysing Chinese texts by incorporating a preliminary step of translation into the BERTopic pipeline, leveraging the Google Translate API to convert these texts into English. This approach enables the application of LLMs as embedding models in the BERTopic [12] pipeline to datasets containing Chinese language, neglecting the original vocabulary of the embedding model, significantly expanding the method's applicability and enabling insights from a broader range of sources. While translation introduces potential bias and variance due to cultural and linguistic nuances that may not be perfectly captured, we mitigate these effects through several strategies. First, we evaluate the introduced bias by comparing the topic modelling retrieved from translated texts with a topic modelling of the raw Chinese texts where the extracted topics are translated. Then, the texts are fed to a pipeline that consists of a state-of-the-art embedding model [13], UMAP [14] for dimensionality reduction, HDBSCAN [15] for clustering in the latent space and a representation generation process consisting of a concatenation of c-TF-IDF and KeyBERT to generate concise descriptions of topics. To evaluate the model, the top-3 cluster associations of each proposal are then evaluated against the human assessment of the cluster association. In case the human assessment and the top-3 cluster associations coincide, the model has correctly classified the proposal. In the case that the human proposal does not coincide with the top-3 cluster associations, the model has not correctly classified the proposal and the classification is evaluated as false. If through the ambiguous nature of the proposals, the proposal could be assigned to either the top-3 clusters or the human assessed cluster, then the classification is evaluated as not uniquely assignable. Finally, the accuracy can be calculated by dividing the total number of correctly classified proposals by the number of total evaluated proposals.

Table 1. Thematic topics and number of related proposals on Join and iVoting.

Join		**iVoting**	
Cluster	**Count**	**Cluster**	**Count**
Government Policies and Regulations	4126	Governance and Sustainability in Taiwan	247
Traffic Safety Regulations	3219	Traffic Management	81
Education Policies in Taiwan	1747	Taipei Public Transport Infrastructure	26
Labor Rights and Regulations in Taiwan	890	Education Technology and Reform	25
Government Response to COVID-19 Epidemic	727	Public Servant Conduct and Accountability	24
Energy Transition and Sustainable Development	672	Waste Management and Recycling in Taiwan	18
Housing Market Regulation	500	Taiwan's Epidemic Prevention Efforts	15
Criminal Justice Reform	498	Labour and Social Welfare in Taiwan	10
Political Status of Taiwan and its Relations to China	476	Inappropriate Content	8
Electoral Reform	369		
Animal Welfare and Protection Policies	281		
Tobacco Control and Smoking Regulations	240		
Gender and Military Service	187		
Media Regulation and False Information Online	187		

3. Results

3.1. Clustering of datasets from Join & iVoting platform

The method introduced was applied to cluster the text-data from Join and iVoting. The cluster topics, ranked in descending order by the total number of proposals, are shown in Table 1. The biggest clusters in both datasets represent the noise cluster of the HDBSCAN-method, which can be neglected. The data indicates that the largest clusters address traffic and educational topics in both datasets. Although far fewer proposals were made on iVoting, the topics are roughly identical to those on Join. However, a distinct local character is also evident; for example, Taipei's public transport infrastructure and waste management were specifically addressed. The semantic map of the clustering with traffic concerns as the most discussed topic can be seen in Figure 2. The smallest clusters in the Join platform are concerned with the Military service and the regulation of media. The smallest clusters in iVoting are concerned with labour laws. The inappropriate content cluster for the iVoting dataset has been due to proposals that have not been

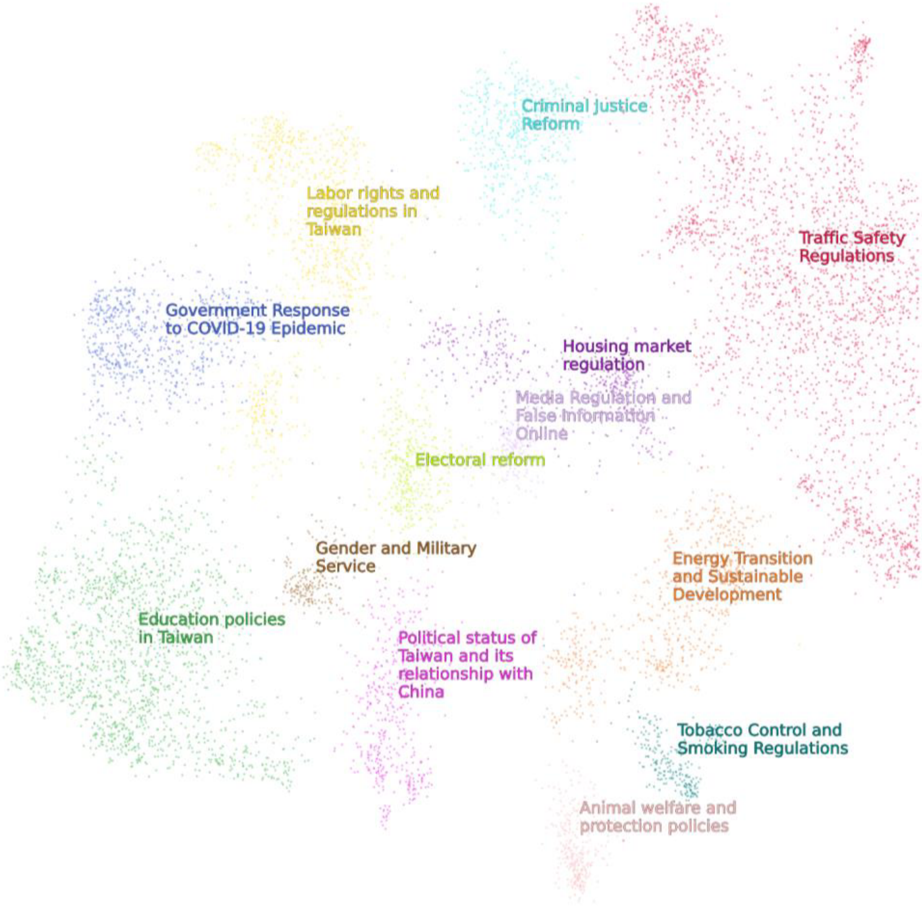

Figure 2. A Semantic map of all Join proposals and their associated clusters.

conformed with the regulations of the platform. In the data set of Join, proposals with inappropriate content were not listed but sorted out beforehand.

3.2. Topics over time on both platforms

Figure 3 shows the clustered proposals over time. Each of the clusters is binned into 2-months intervals. The resulting plot of each topic's frequency can be compared to the events and can reflect the popularity of specific topics on the e-participation platforms. There are three peak phases that were fuelled by media and public debates. It can be seen here that the platform has always been a vehicle for discussing these debates online and making suggestions. In February 2017, the Legislative Yuan declared that proposed regulations on trade, investment, and intellectual property rights undergo a 60-day comment period via the Join platform for public and business feedback [16]. As a result of this instruction, a growing number of online proposals were submitted, as can be seen in the table. The submissions reached a peak in autumn 2017, when ride-hailing regulation (Uber vs. Taxi), infrastructure projects and public transit fares were critically reflected in different proposals on Join. In the same year, a 16-year-old student proved that the platform could have some use when her proposal for a "nationwide progressive ban on the use of disposable utensils" received more than 5,000 votes in a short time. As a result, the Environmental Protection Administration restricted government bodies, schools, and department stores from providing disposable plastic straws. We assume that such success stories, which were also spread on social media, had a certain signal effect that prompted further proposals to be submitted on various topics. While many different topics were posted at the end of 2017, two topics that are not related to each other stand out in the period from 2020 to 2022 and thus in the context of the pandemic: "COVID-19 Medical Response Measures" and "Reforming Education"

With the beginning of COVID-19 in 2020, citizens (despite spending more time in front of displays than usual) started to use Join to make proposals on social distancing measures, mask distribution methods, and digital health solutions. During that time and before the “new curriculum guidelines of 12-Year Basic Education” was amended in February 2021, concerns and general shortcomings in education were controversially debated on Join. These topics highlight an overall demand for making education more inclusive, engaging, and fair. Leaving these two topics aside, it is noticeable that the subject that caused the most concern over the entire period was the area of traffic management and regulation. This is also reflected in Figure 3.

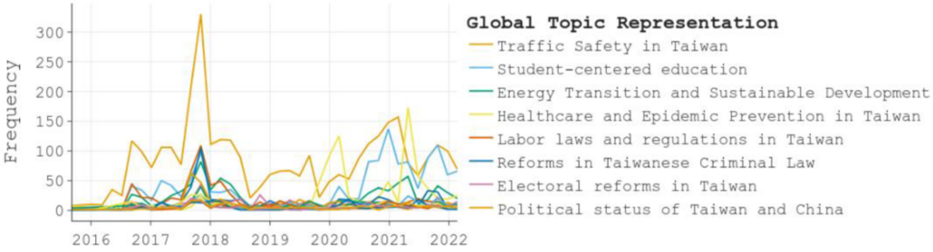

Figure 3. Topics over time on Join.

3.3. Evaluation of the Topic Clustering Approach

In order to establish whether our topic modelling and classification approach would agree with a human domain expert, we performed a qualitative analysis of the performance of our approach. A domain expert was presented with 110 randomly sampled proposals and the predicted topics from the topic model and was tasked to decide if the topic is correctly assigned to the proposal given the list of all extracted topics. The topic model approach can classify topics with an accuracy of 81.82 %. It misclassified topics in 10.91% of the cases. In 7.27%, the proposals can not be assigned unambiguously to one of the clusters even under careful consideration of the evaluators.

4. Discussion

Overlapping of topics can lead to ambiguity in assigning the topics to a distinct cluster, which is the result of hard-clustering. In comparison, soft-clustering can assign a sample to a number of clusters. However, this is much harder to evaluate for humans, since humans can not easily assign a probability to a given proposal. It would be highly subjective. Even further, it was a challenge in this research to come up with appropriate clusters for the text corpus. A first evaluation of the text corpus by a social scientist has led to the following topics: Transport, Law & Justice, Education, Finance, Environment & Climate, Energy, Social, Military & National Security. When assigning the topics to some of these clusters, there is a definite overlap for these categories. Especially critical is the Law & Justice category, since all proposals should be realised in a law eventually. This also explains that one of the clusters from the topic model is called "Government Policies and Regulations". All proposals that do not match to one of the other clusters according to their probability will get assigned to this cluster. This cluster is made up of a very heterogeneous mixture of proposals, since all proposals in the dataset should follow the goal of becoming government policies and regulations eventually. Therefore, it is of high interest to keep this cluster as small as possible. However, in our experiments, this cluster still contains 28.57% of all proposals for the Join Dataset. It still needs to be evaluated, how the size of this cluster could be reduced.

To evaluate our results on the text corpus, a cross reference with historic events has been conducted. The timestamps of the proposals and their occurrence can be evaluated with the real occurrence of events in the society. For example, the Covid-19 event clearly has had an effect on the proposals on both platforms. The proposals related to Covid-19 have spiked at the end of 2019 and in the beginning of 2021. Both spikes can be related to specific events. The former marks the outbreak of the pandemic, the latter the significant increase of local cases in Taiwan during the year of 2021.

The fact that the original text corpus is collected in Mandarin-Chinese language and the evaluation was mainly conducted in English is of importance for this study as well. The embedding with the BAAI/bge-large-en-v1.5 model into a mathematical vector space made it possible to decouple language from proposals. Even though the translations associated with each vector encountered a loss of meaning, LLMs seem to be able to cope with this loss of translation given the high accuracy on the test set. We even suggest that they can understand machine-translated language in a good manner and translate it to a more human-friendly manner with the help of the representation tuning module as part of BERTopic.

Overall, our study reveals that LLMs have the potential to significantly enhance the field of Policy Informatics by providing advanced tools for data analysis, natural language processing, and simulation of policy impacts and interpretation. Our approach can help to guide political decision makers in a data-driven fashion. The data derived can clearly show where participants on e-participation platforms see a necessity for new government regulations. With these engineering tools in the right hands and a strong interdisciplinary collaboration, social change within a society can be accelerated. However, it also needs to be mentioned that the participants of e-participation platforms fall under the phenomenon of the self-selection bias. It needs to be considered that the data might not be representative with the rest of the society. So these proposals might only give a direction, which then must be confirmed with the whole democratic corpus.

5. Conclusion

This study investigates how recent advances in natural language processing (NLP), specifically large language models (LLMs), can enhance the capabilities of Policy Informatics as a subdomain of TD Engineering. We conduct a case study applying LLMs to analyse text corpora from two Taiwanese e-participation platforms: Join and iVoting. Utilising LLMs, we extract and evaluate the major topics discussed on these platforms over time. We then qualitatively assess the accuracy of this topic modelling approach using a manually annotated test set. The findings demonstrate the potential for incorporating such NLP/LLM methods into urban social movements, where data-driven approaches can support initiatives for positive social change. Modern democratic participation can be facilitated through evidence-based analysis of public discourse on e-participation platforms. This study paves the way for leveraging cutting-edge NLP to gain insights from civic technology platforms and better understand the perspectives and priorities of the citizenry and therefore solving ill-defined, society-relevant problems with the given transdisciplinary framework.

In conclusion, our study demonstrates that transdisciplinary cooperation between informatics and policy (data) can enhance our understanding of online-expressed views. However, several areas require further research. Firstly, a deeper analysis of the sentiments in the proposals is essential to understand user motivations and concerns. Additionally, investigating the demographics of platform users, particularly identifying if they are predominantly well-educated, urban individuals with digital skills, is crucial for assessing representativeness. Furthermore, examining user comments is vital to uncover community dynamics and feedback mechanisms. This analysis could reveal trends and patterns not evident from proposals alone. Addressing these areas will enhance our understanding of the platforms, their users, and the implications for digital civic engagement.

References

[1] H.-Y. Huang, M. Kovacs, V. Kryssanov, and U. Serdült, Towards a Model of Online Petition Signing Dynamics on the Join Platform in Taiwan., In: *2021 Eighth International Conference on eDemocracy & eGovernment (ICEDEG)*. pp. 199–204 (2021).

[2] D. Liao and B. Chen, Strengthening Democracy: Development of the iVoter Website in Taiwan, In: D. Liao, B. Chen, and M.J. Jensen, Eds. *Political Behavior and Technology: Voting Advice Applications in East Asia*. pp. 67–89. *Palgrave Macmillan US*, New York (2016).

[3] A. Su, *Shezidao as a Limit Case for Democracy in Taiwan? Perspectives from Design with Jeffrey Hou*, https://newbloommag.net/2023/04/17/shezidao-dispute/.

[4] M. Liebeck, K. Esau, and S. Conrad, Text Mining für Online-Partizipationsverfahren: Die Notwendigkeit einer maschinell unterstützten Auswertung, *HMD Praxis der Wirtschaftsinformatik*. 2017, vol. 54, pp. 544-562.

[5] A. Rauchfleisch, T.-H. Tseng, J.-J. Kao, and Y.-T. Liu, Taiwan's Public Discourse About Disinformation: The Role of Journalism, Academia, and Politics, *Journalism Practice,* 2023. vol. 17, no. 10, pp. 2197–2217.

[6] G. Tress, B. Tress, and G. Fry, Clarifying Integrative Research Concepts in Landscape Ecology, *Landscape Ecology,* 2005, vol. 20, no. 4, pp. 479–493.

[7] S. Lattanzio, A. Nassehi, G. Parry, and L.B. Newnes, Concepts of transdisciplinary engineering: a transdisciplinary landscape, *International Journal of Agile Systems and Management, 2021,*. vol. 14, no. 2, p. 292-312.

[8] F. Wickson, A.L. Carew, and A.W. Russell, Transdisciplinary research: characteristics, quandaries and quality, *Futures,* 2006, vol. 38, no. 9, pp. 1046–1059.

[9] S.S. Dawes and M. Janssen, Policy informatics: addressing complex problems with rich data, computational tools, and stakeholder engagement, *Proceedings of the 14th Annual International Conference on Digital Government Research*. pp. 251–253, 2013.

[10] M.P. Rodríguez Bolívar, Policy Makers' Perceptions About Social Media Platforms for Civic Engagement in Public Services. An Empirical Research in Spain, In: J.R. Gil-Garcia, T.A. Pardo, and L.F. Luna-Reyes, Eds. *Policy Analytics, Modelling, and Informatics: Innovative Tools for Solving Complex Social Problems*. Springer International Publishing, Cham, 2018, pp. 267–288.

[11] L. Hagen, T.M. Harrison, and C.L. Dumas, Data Analytics for Policy Informatics: The Case of E-Petitioning, In: J.R. Gil-Garcia, T.A. Pardo, and L.F. Luna-Reyes, Eds. *Policy Analytics, Modelling, and Informatics: Innovative Tools for Solving Complex Social Problems*. pp. 205–224. Springer International Publishing, Cham, 2018.

[12] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, http://arxiv.org/abs/2203.05794, 2022.

[13] S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff, C-Pack: Packaged Resources To Advance General Chinese Embedding, http://arxiv.org/abs/2309.07597, 2023.

[14] L. McInnes, J. Healy, and J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, http://arxiv.org/abs/1802.03426, 2020.

[15] C. Malzer and M. Baum, A Hybrid Approach To Hierarchical Density-based Cluster Selection., In: *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. pp. 223–228 (2020).

[16] American Chamber of Commerce in Taiwan, *Participate on Join.gov.tw to Make Taiwan Better*, https://amcham.com.tw/2017/02/participate-join-gov -tw-make-taiwan-better/, (2017).