PBL5 Final Report AY 2025 – Digital Governance Systems Laboratory

Name: Lin Chung-Hsi Student ID: 26002304790 Date: July 22nd, 2025

Summary and discussion of 4 initial papers

Paper 1

[1] N. Goodman, H. A. Hayes, and S. Dunbar, "Absentee online voters in the Northwest Territories: Attitudes and impacts on participation," in *Proc. Int. Joint Conf. Electron. Voting*, Cham, Switzerland: Springer Nature, Sep. 2024, pp. 89–106.

(*subject/main idea*)

The paper tried to analyze the e-voting population within Northwest Territory, Canada. In the NWT, only absentee voters who applied for e-voting are allowed to vote through this mode, which is the main focus of the study. To understand these absentee voters, the paper draws comparisons to municipal elections in Ontario Province, a more urban region. Here are the key findings of this paper.

1. The high internet availability and digital literacy for NWT absentee voters, better than the e-voters in Ontario.
2. Accessibility is the main rational for NWT voters to choose e-voting, Convenience for Ontario
3. Normally, 10% wouldn't have voted without this vote mode, however, 43% shared the same sentiment in NWT.
4. A logistic regression is performed to analyze the dependency of different factors support this sentiment.
5. For NWT voters, the two main factors are: They are uncommitted voters; They are living outside of Yellowknife (The largest city in NWT, so they live in more rural area relative even to NWT residents). These two factors incite the sentiment stated above.
6. For Ontario voters, uncommitted voting records shares the same influence, but the second largest factor is age. There is less chance a voter wouldn't have voted with increasing age. The author cites potential research for this finding.

(importance/usefulness/relevance/contributions)

This paper highlights the implementation of e-voting for more rural area. As most countries have their population living in wide spectrums of rurality, it is important to implement e-voting in the entire spectrum to uphold equality during elections. Understand how the rurality of NWT impacts the voting population's experience and sentiment towards the new mode is very helpful.

(criticisms)

Even though the paper pointed out unintuitve observation and drew interesting results, the data used in this paper seems a little problematic. As the e-voters from NWT is limited to absentee voters, the survey data is not a good match for comparison. Thus, all the interesting conclusions are all tied to unique situation of these absentee voters, which is helpful to understand these absentee voters, but not very insightful for other demographics.

Paper 2

[2] Y. W. Chiu, H. C. Huang, C. J. Lee, and H. P. Hsieh, “PEPO: Petition executing processing optimizer based on natural language processing,” in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Jul. 2023, pp. 3150–3154.

(*subject/main idea*)

This paper proposes a system to solve the underlying problems of a petition hotline in Taiwan that allows citizens to submit petition to the government to act on issues the citizen wishes to solve. The underlying problem is how the personnel responsible for this hotline, liaisons, are facing overwhelming amounts of petitions and manually classifying and forwarding the cases to corresponding department is often inaccuracy and time consuming, resulting in low satisfaction from the citizen.

The proposed system, PEPO, aims achieve three things.

1. Utilizing natural language processing with the content of the petition to give accurate department classification. PEPO would suggest the three most relevant departments to the liaison.
2. Classify the importance of the cases to either general, risky, or emotional. The purpose is to quickly let the liaison understand the severity of the case.
3. Generate a template text for the liaison to quickly modify and respond to citizens.

(importance/usefulness/relevance/contributions)

The achievement that should be highlighted PEPO’s accuracy on the first goal, classifying the case to relevant departments. Manually, the accuracy for department annotation is only around 60%, where as PEPO achieved 80% within the 37328 data points. This suggest that real world implementation of PEPO should be considered for its boost in accuracy and effectiveness in reducing wasted work hours. This kind of development is a great step towards digital government.

(criticisms)

For response generation, there wasn’t any evaluation metrics on showing the effectiveness of this function, or whether this goal is achieved. Aside from quantitative results, they could also include qualitative feedback from the liaisons, which could identify problems or usefulness of this function. Going along this lines they could also install surveys at the end of the service to evaluate the effectiveness on improving government institutes’ servicability and quality by referencing citizen’s experience.

From what I can find with the professor, the PEPO system is not implemented to any public sources I can find in Mandarin. This is not a criticism, but I would want to see how feasible and effective this system in real world conditions since I am a Taiwanese.

Paper 3

[3] N. Kamoen, T. McCartan, and C. Liebrecht, “Conversational agent voting advice applications: A comparison between a structured, semi-structured, and non-structured chatbot design for communicating with voters about political issues,” in *Proc. Int. Workshop Chatbot Res. Des.*, Cham, Switzerland: Springer Int. Publishing, Nov. 2021, pp. 160–175.

(*subject/main idea*)

This paper set out to distinguish effects between different types of the new rising use of Conversational Agent Voting Advice Applications, or CAVAA. The paper identified three type of agent implementation: structured, non-structured, and semi-structured. These CAVAAs are implemented during the Dutch nation election period. The three can be understood like this:
(Structured): There are buttons on the user interface allowing users to request for additional information. One button provides information on policy specific terms, the other for information that helps understanding the semantics of the policy. Lastly, a third button would let the user continue.
(Non-structured): Instead of buttons, there is a text box that allows the user to inquire a chatbot for more information. The chatbot is trained for provide clarification to related terminologies and respond with the current state of affairs in mind
(Semi-structured): A combination of both structured and non-structured, with features from both.

The author explored several aspects for evaluation and formulated hypothesis on them.

1. Perceived Political Knowledge
2. Voting Intensions
3. Perceived Ease of Use
   Hypothesis 1: structured / semi-structured > non-structured
4. Perceived Usefulness
   Hypothesis 2: semi-structured > structured / non-structured
5. Perceived Playfulness
   Hypothesis 3: structured > semi-structured / non-structured

(importance/usefulness/relevance/contributions)

Through the collected data, the author concluded the following.

H1 is supported. Non-structured is indeed inferior in perceived ease of use than semi-structured and structured
H2 is disproved, CAVAA versions makes no significant difference to Usefulness.
H3 partially matches the result. Structured does appear more playful than non-structured, but no significance and be found between semi-structured with the other two versions.

(criticisms)

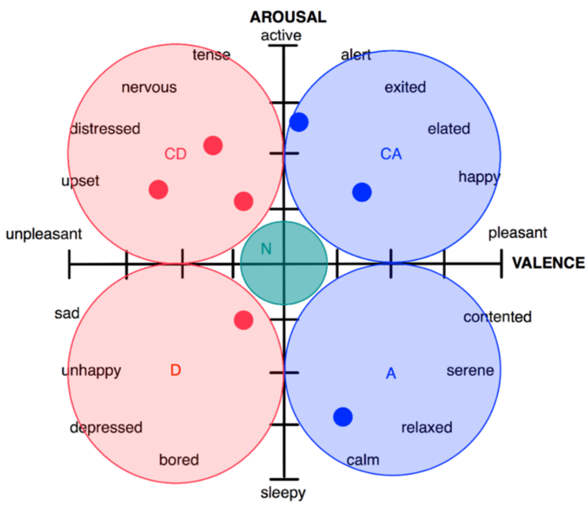
There is a noticeable impact by the novelty of trending AI chatbots on the general public’s perception on their experience interacting with one. The author did not consider this novelty factor influencing perceived playfulness. Novelty, intuitively, can also vary by the participant’s age and level of education. These two are important factors frequently discussed in the study, which makes it more important to take in consideration of the novelty effect. The study should have made further discussions.

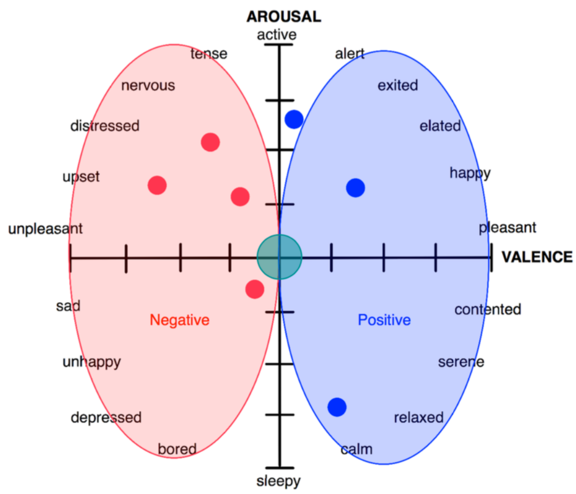
Paper 4

[4] L. Terán, U. Kakenova, and E. Portmann, “Analyzing and integrating dynamic profiles on voting advice applications,” in *Proc. 10th Int. Conf. Theory Pract. Electron. Governance*, Mar. 2017, pp. 62–69.

(*subject/main idea*)

This paper set out to explore implementation of a dynamic profile used in VAAs. With politician twitter accounts, the studies filtered and categorized the tweets and performed sentiment analysis to these tweets. Then, a profile vector is constructed with the resulting emotional matrix. As a result, a table indexed by political topics is labelled by the predicted sentiment of the politician on the Likert scale (CA A N D CD). Notably, this paper proposed the distribution of the Likert scale on Russell’s emotion space. Shown in the figures below, the paper express agreement-disagreement as positive and negative valence values and the intensity of intension (agree to completely agree) with the magnitude of the arousal values.





(importance/usefulness/relevance/contributions)

Creating and updating a VAA profile requires experts and is time consuming and potentially outdated. The purpose of these studies is to help the generation and updating of VAA profiles. The creation process of these profile sometimes requires data collection, surveys that not all politicians responding to. The method explored expands the sources of these data and works without any interaction with politicians. Twitter in particular is widely used in Japan, where the implementation of this method is suitable.

(criticisms)

Both of these papers lack in evaluation, it would present more value when the generated dynamic profile is compared to existing VAA profile. Without comparisons, it is hard to determine the effectiveness and accuracy of the model. However, with proper evaluation, it is possible that this model is ready to be implemented.

Presentations

I made the following 3 presentations during this semester in this class:

- 1st presentation (Week 06, May 16th, 2025)
  - (Dynamic VAA Profile Generation Using Sentiment Analysis with Twitter Data)
  - In this presentation, I proposed the idea to replicate a methodology in a journal in Japanese settings. I also laid out the three main tasks in the project: data scraping from Twitter, topic extraction with some clustering methods, and sentiment analysis that should output similar result format as the journal described.
- 2nd presentation (Week 10, June 11th, 2025)
  - (Dynamic VAA Profile Generation Using Topic Extraction and Sentiment Analysis on Politician Tweets)
  - In this presentation, I focused on the three major parts of the methodology. I detailed the problems I faced in collecting the data and what preprocesses I deemed appropriate to apply after seeing the raw data. I proposed two clustering methods after reading into other journals and laid out the advantages and drawback of both. I also presented the sentiment dataset that I found for training. However, I discovered the issue of mismatching emotion spaces between the dataset and the journal. The journal proposed the sentiment classification with Russell's Arousal Valence space, whilst the dataset is annotated with Plutchik's Eight Emotions.
- 3rd presentation (Week 14, July 11th, 2025)
  - (Dynamic VAA Profile Generation Using Topic Extraction and Sentiment Analysis on Politician Tweets)
  - In this presentation, I mainly focused on the reasons behind the poor results. I wasn't able go in details due to the short time, unfortunately. The reasons can be deduced to three fundamental issues: low quality and imbalanced data, difficulty of precisely clustering tweets to strictly specific topics, and that there is no prior research in transforming between emotion spaces.

Problem statement

The topic for my project involves Voting Advice Applications, VAAs. In multiple academic journals, there were comments about the inherent inefficiency and cost associated with creating, curating, and updating VAAs. Most VAAs consist of political questions where the user would answer typically in the form of a Likert scale. The application would rank the user's answer to politicians or parties that has similar answers. Simply put, VAA are matching political sentiments between the voters and the candidates. However, the process of gathering answers(sentiments) from the candidates is costly. It takes experts with surveys from candidates to generate these set of sentiments. There are also problems where politicians are reluctant to respond to the survey or changing/updating their opinions during the election period. Also, the VAA questions are susceptible to change from external factors/events such as Covid. The project aims to generate sentiments through an automated approach, referencing from a journal that would be discussed later. The journal described the methodology of collecting social media posts of these politicians, and ideally, extracting relevant information representing their sentiments through their words online. In Japan, as politicians are vocal on a monopoly platform, X (formerly Twitter), this project should be able to utilize similar methodology described in the journal.

Summary and discussion of 4 additional papers

[5] L. Terán and J. Mancera, "Dynamic profiles using sentiment analysis and Twitter data for voting advice applications," *Gov. Inf. Q.*, vol. 36, no. 3, pp. 520–535, 2019.

[6] F. Ravenda, S. A. Bahrainian, A. Raballo, A. Mira, and F. Crestani, "A self-supervised seed-driven approach to topic modelling and clustering," J. Intell. Inf. Syst., pp. 1–21, 2024.

[7] S. Basu, Y. Yu, and R. Zimmermann, "Fuzzy clustering of lecture videos based on topic modeling," in Proc. 14th Int. Workshop Content-Based Multimedia Indexing (CBMI), Bucharest, Romania, Jun. 2016, pp. 1–6.

[8] A. Bhuvaneswari and M. Kumudha, "Topic modeling based clustering of disaster tweets using BERTopic," in Proc. 2024 MIT Art, Design and Technology School of Computing Int. Conf. (MITADTSoCiCon), Apr. 2024, pp. 1–6.

[5] is the subsequent paper to [4], written by the same author. They tackled the topic more completely and holistically—for example, by analyzing the algorithms that match users to candidates, analyzing VAA designs through comparisons, and more. Although the paper didn't provide deeper discussion on clustering and the reasoning behind the legitimacy of the Likert scale distribution, it offered more depth in its refined methodology for sentiment analysis.

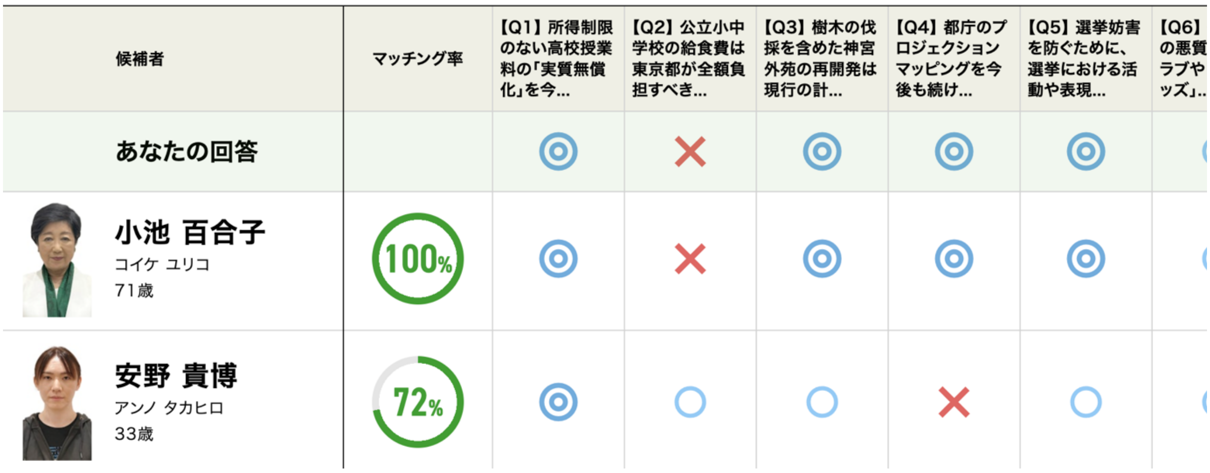
[6] is a paper performing seed-driven self-supervised topic modeling. The paper designed and explored the architecture, and compared it with conventional clustering methods such as K-means and other neural topic modeling models using five benchmark datasets. The meaningful insight of this paper is that they concluded this approach performs better than conventional methods by producing more semantically meaningful clusters. Needless to say, this laid the foundation for my decision to approach clustering using the methodology presented in this paper. However, due to the recency of this paper (2024), there are few resources on semi-supervised clustering for the type of corpus I am working with.

[7] is a paper performing multilabel classification with fuzzy clustering on lecture video caption texts. The idea is that a lecture video can be useful for multiple subjects but is difficult to label manually. The implementation uses the word embedding of Wikipedia text for each subject as seeds to perform C-means clustering. This methodology is a great approach, but since my task doesn't involve multilabel classification, I could not fully utilize its advantages. However, this paper established classification based on extracting keywords from lecture transcripts, which is analogous to my topic (VAA questions are long, with a detail paragraph). This revealed that extracting keywords from long topic texts is important for my project.

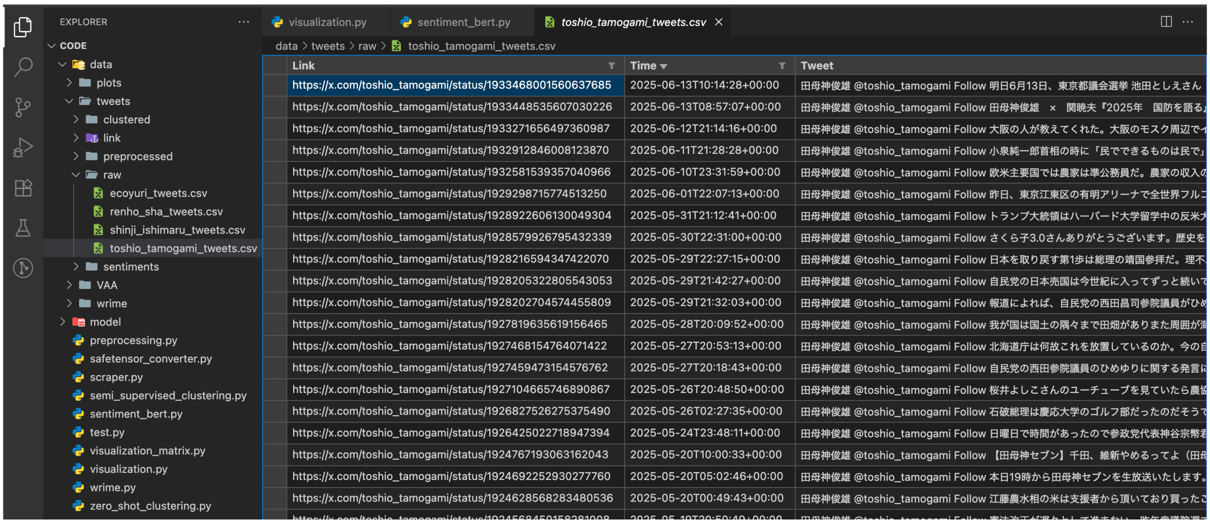
[8] is a paper with a similar application to my project, where the goal is to extract information from Twitter during disasters. Topic modeling is the main focus of this paper, detailing the process of extracting tweets containing similar information pointing to the same event or location. Their methodology utilizes BERTopic to model topics unsupervised. They compared it with conventional methods and demonstrated that it outperformed them with the noisy social media texts. This paper provided valuable insights that, combined with [6], guided my decision to implement semi-supervised topic modeling with BERTopic.

Data acquisition

I selected the 2024 Tokyo Governor election as the target election. I selected the top 4 voted candidates as data source. First, I scraped the VAA questions from "Tokyo MX with 投票マッチング", which has the VAA questions regarding this election. More importantly, this VAA publicly displays the sentiment of all candidates in the form of Likert scale. These sentiment data was also scraped from the site. Here is a screen shot to the VAA result page:



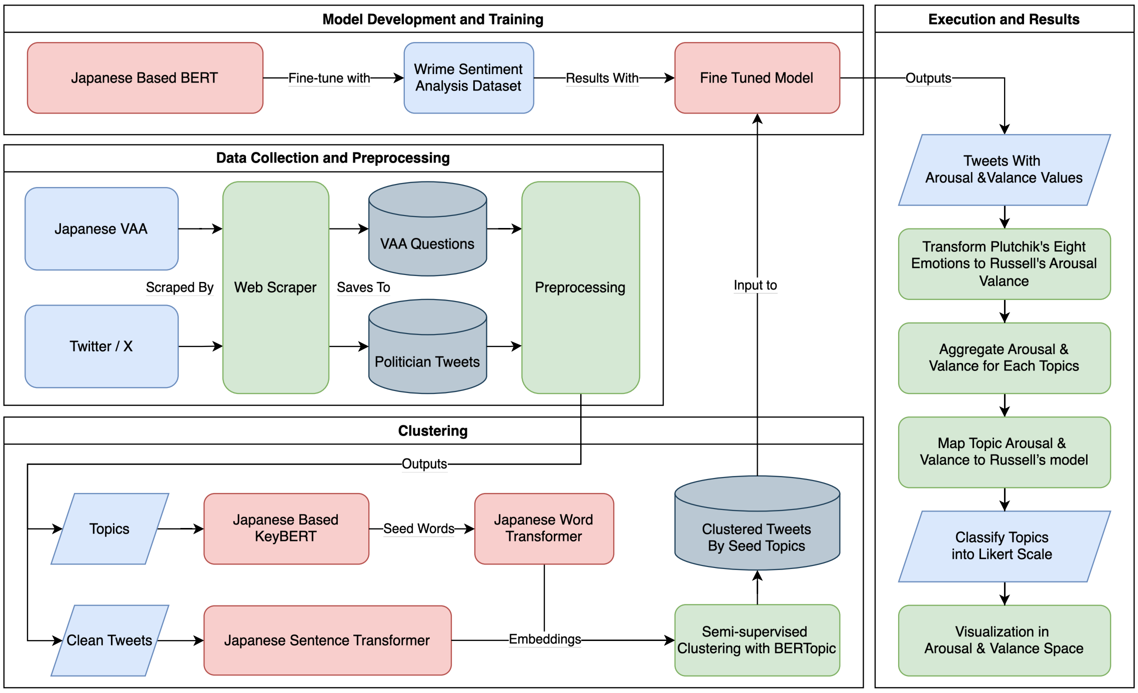
Secondly, the text data I need for the project are tweets from the accounts of four Japanese politicians (Koike Yuriko, Ishimaru Shinji, Renhou, Tamogami Toshio). I collected them by coding a program that utilizes a browser automation library to interact with each candidate's Twitter profile page. To avoid being banned from bot detection—a problem I encountered several times during the project—the program pauses for several seconds between each action. One by one, it copies the HTML section containing each tweet's text and saves it to a CSV file. The process took around 20 hours total, with stops and checkpointing, to finish scraping. Here is a screen shot of the csv file for Tamogami Toshio's tweets:



The CSV file contains the following columns:

- "link": Link to the tweet, containing the username of each candidate.
- "time": Time of posting the tweet.
- "tweet": Text content of the tweet.

Approach and methods used

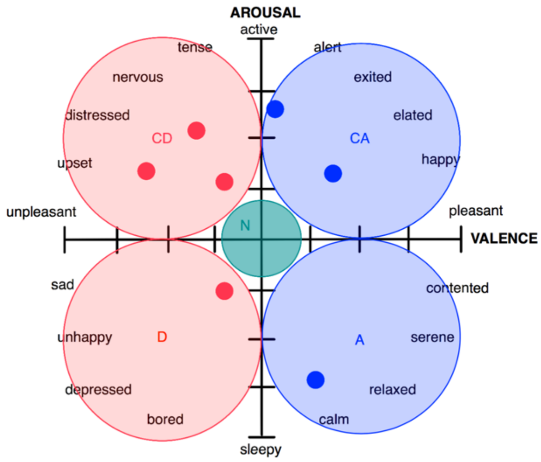


The first part of handling the data is preprocessing. I identified several artifacts that were irrelevant to the text. For example, there were HTML tags, username, special characters, time stamps, and many more that would tamper the performance of both clustering and sentiment analysis. These artifacts were removed as preprocessing before clustering and sentiment analysis.

Clustering was the part where I struggled the more upon, where the accuracy of the methods I read were low and incapable of differentiating detailed context. I tested with two clustering methods, zero-shot clustering and semi-supervised clustering. Zero-shot is a method where the cluster topics are predefined, in this case, is simply the VAA questions. First, both the tweet and the questions are put through a Japanese sentence transformer that should entail the semantic in high dimensional vectors, embeddings. Then, I treat the embedding of predefined topics as centers of the clusters and assign each tweet to the closest cluster center based on cosine similarity between their embeddings. Threshold value for the cosine similarity would be set to filter out irrelevant tweets. As the difference between relevant and irrelevant tweets is arbitrary, it is tricky to determine the threshold value. The sentence transformer used here is from *"sonoisa/sentence-bert-base-ja-mean-tokens-v2"* published on Hugging Face.

The other clustering method is semi-supervised clustering, where the model considers predefined topics as highly probable topics during modeling. The resulting set of topics wouldn't guarantee the predefined topic to exist, but the boost in their presence can be adjusted by tuning parameters. As a result, this topic modeling method would cluster irrelevant tweets to topics that I can simply discard, avoiding the arbitrary threshold problem that zero-shot presents. The libraries that realize this method are BERTopic and KeyBERT. After obtaining the output set of topics, I can apply the process in zero-shot to cluster the tweets. The filtering process is to simply skip tweets that is not clustered to any of the predefined topics.

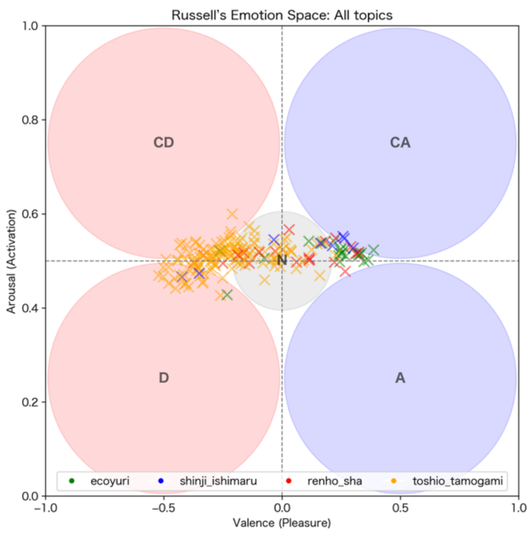
After labeling each tweet to a VAA question, I need to perform sentiment analysis to these tweets to approximate the candidate's sentiment to these topics. I acquired a sentiment dataset called WRIME which has 40 thousand Japanese sentences annotated with Plutchik's eight emotions. I used a pretrained Japanese-based Bert ("cl-tohoku/bert-base-japanese-v3"), and finetuned it with a fully connected output layer that would fit to the dataset. This model would take Japanese text as input and output an array of numerical values matching Plutchik's model. However, this format doesn't match the emotion space the reference literature uses, therefore, a transformation process needs to occur. Though not supported by literature, I converted the eight emotions to eight vectors on Russell's arousal Valence. The transformation would be a dot product between the array and matrix. The result would be a two-dimensional vector that has Arousal and Valence as elements.

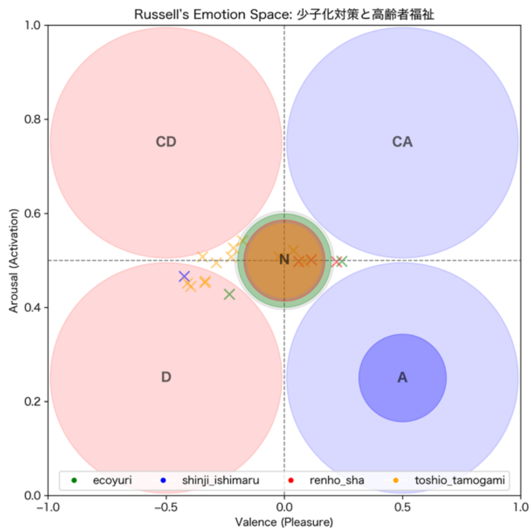


Plutchik's eight emotions, $x = [1.49 \quad 0.0. \quad 2.21. \quad 0.92. \quad 0.26. \quad 0.18. \quad 0.14. \quad 2.05]$

Transformation matrix, $A = \begin{bmatrix} 0.6 & 0.6 \\ 0.6 & 0.2 \\ -0.7 & 0.6 \\ 0.0 & 0.8 \\ -0.8 & 0.1 \\ -0.6 & 0.4 \\ -0.8 & 0.7 \\ 0.3 & 0.6 \end{bmatrix}$ Russell's Arousal Valence, $Ax = [0.51 \quad 0.33]$

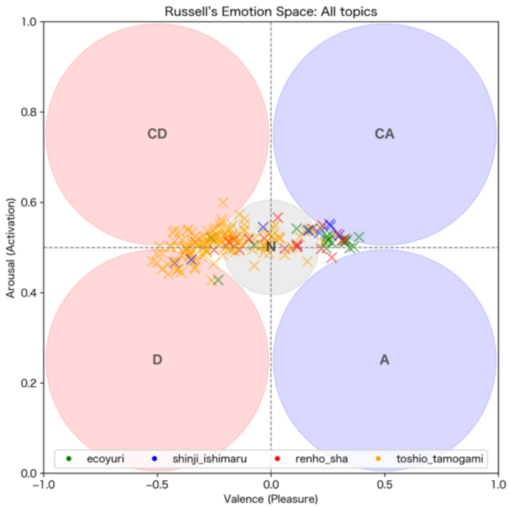
Visualization simply involves plotting the values we obtain for each tweet on the Russell's emotion space. Each tweet is represented by a marker and color coded by the politician. I also replicated the Likert scale distribution that was presented in the reference literature. For each topic/question, there would be a plot of the tweets that is clustered to the topic and true labels for the politician's sentiment scraped from the VAA. True sentiments are represented as larger points that resides in the corresponding region of the Likert scale.

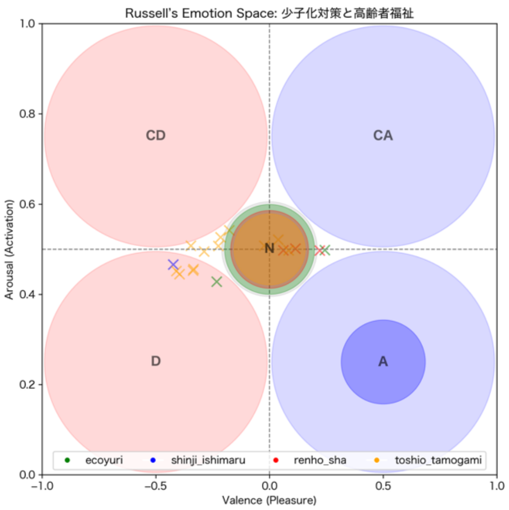




Results

Using the figure from the previous section, here on the right, the result are tweets represented by markers positions according to the arousal valence values of the tweet's text evaluated by the sentiment model. The arousal values of the data points are concentrated toward 0.5, and the valence values are also low variated, never crossing magnitude of 0.5. In general, these data points don't variate away from the Neutral region much and almost all predictions mismatch the true label. There are issues scatted across the three major parts of the methodology. Let's discuss what and why each part contributed to this outcome.





First, the data are poor in both quality and quantity. In total, there are only 2111 raw tweets, after preprocessing, only 841 remains. Moreover, most of the tweets are irrelevant, which are filtered by the clustering methods, leaving only 192 tweets. The reason why most tweets get filtered out is because of the sheer number of irrelevant tweets compared to the valuable ones. Although politicians have become more active and vocal on social media platforms, the word they produces are also varies due to the environment of social media platforms. There were a lot of tweets that were only for greetings and general information, such as "good morning", "I attended a conference", "I'll be speaking here today". Most tweets don't represent their opinion directly or not at all. That leads to a lot of the topic having less to no data points, since there are not enough tweets of these politician speak about all of these topics. In particular, PFAS was part of a VAA question, but none of the tweet contains the keyword PFAS. This highlights the inherent lack of data by collecting data from social media.

Secondly, the quality of clustering for both methods are very limited. Though both can successfully filter out completely irrelevant tweets, others with somewhat high similarity values are challenging. During the presentation, I described the "Different Tax" and "Partial Answer" problem I observed from the clusters. The "Different Tax" problem is when the politician opinionates about a certain policy, however, the policy mentioned in the VAA question is the same type but a different one. This scenario has high semantic similarity but is a mislabel. The "Partial Answer" problem is when the VAA question entails two or more topics, and the tweet only resembles its argument regarding one of the topics mentioned. For example, a VAA question inquires the priority between birth rate and senior health care. Unintendedly, tweets regarding only either would be considered as high similarity. This is to express that when similarity between the embedding reaches some threshold, the accuracy of classification diminishes and returns when similarity overcomes the semantic ambiguity.

Lastly, there needs to be further discussion about whether sentiment analysis through emotion spaces is effective to approach political sentiments. The figure on the right is the result table of the reference literature. The accuracy of their result is also extremely low, often giving contradictory results against the experts, CA opposing CD. Though the paper avoided discussing limitation in this regard, but politically related statements are often long and has consistent tones. Meaning that politician would naturally express their opinion in a neutral and stable arousal level, which is also what we observed in the results. Potentially, it is not an appropriate way to evaluate the intensity (arousal) of their opinion in Russell's emotion space. With this said, I have to acknowledge that my result went through a transformation between emotion spaces, which isn't completely representative for comparison with the reference literature.

Table 1. Comparison of Answers: Experts vs. Twitter

	@BarackObama		@SenJohnMcCain		@HillaryClinton		@MittRomney	
Question #	Twitter	Expert	Twitter	Expert	Twitter	Expert	Twitter	Expert
1	CA	CA	CA	CD	CA	A	CA	D
2	**No answer**	CD	CA	CA	CA	D	CA	A
3	**No answer**	CA	CA	CD	**No answer**	A	**No answer**	CA
4	CA	CD	**No answer**	CD	**No answer**	CD	CA	D
5	CA	D	**No answer**	CD	**No answer**	D	**No answer**	D
6	CA	CD	CA	N	CA	CD	**No answer**	CA
7	CA	CD	CA	CA	CA	A	CA	CA
8	CA	CD	CA	CA	CA	CD	**No answer**	CA
9	CA	CA	**No answer**	CD	**No answer**	CA	**No answer**	CD
10	CA	A	CA	CD	CA	A	**No answer**	D
11	**No answer**	A	CA	CD	**No answer**	A	**No answer**	CD
12	**No answer**	CD	CA	A	CA	CD	**No answer**	A
13	**No answer**	A	**No answer**	A	**No answer**	A	**No answer**	D
14	CA	A	CA	A	CA	CA	CA	CD
15	CA	CA	**No answer**	D	**No answer**	CA	**No answer**	D
16	**No answer**	D	CA	D	CA	CD	CA	D
17	CA	D	**No answer**	D	**No answer**	D	**No answer**	D
18	**No answer**	CD	CA	A	CA	D	**No answer**	CD
19	D	A	D	CD	D	N	D	CD
20	**No answer**	CD	CA	CA	CA	CD	CA	A
21	**No answer**	CA	CA	CD	**No answer**	CA	CA	CD

Note: Completely agree (CA), Agree (A), Neutral (N), Disagree (D), Completely disagree (CD)

To conclude, the result of this project is rather underwhelming in the sense of accuracy. However, the process explored the limitation and expandability of research in this direction. Particularly the emotion space transformation is a notable part of this project that need further research to consolidate the validity of my project's approach.