Text Mining from Party Manifestos to Support the Design of Online Voting Advice Applications

Buryakov, Daniil ; Hino, Airo ; Kovacs, Mate ; Serdült, Uwe

Abstract: Voting advice applications (VAA) allow potential voters to compare their own policy positions to political parties running for an election. One of the key design elements of a VAA are the policy statements representing the political space covered by political parties. VAA designers face the challenge of coming up with policy statements in a short time frame. Even with medium-sized corpora of texts such as party manifestos, the formulation and selection of policy statements serving as a stimulus in the VAA is a tedious and time-consuming task. In addition, there is the risk of human selection bias. This study proposes a system to aid VAA designers in policy statement selection and formulation. The system uses the BERT language model with semantic similarity calculation to mine party manifesto sentences that are relevant to already existing VAA statements. For the experiments, VAA statements stemming from the 2021 elections and party manifestos issued for the previous two Japanese elections were used. To expand the policy space, VAA statements from the 2019 European Parliament elections were added. Results show that the proposed system is able to analyze large amounts of text in a short time, and mines text that provides practical support for designing and improving VAAs.

DOI: https://doi.org/10.1109/besc57393.2022.9995398

2022 9th International Conference on Behavioural and Social Computing (BESC) | 979-8-3503-9814-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/BESC57393.2022.9995398

Text Mining from Party Manifestos to Support the Design of Online Voting Advice Applications

Daniil Buryakov
College of Information Science and Engineering
Ritsumeikan University
Kusatsu, Japan
ORCID : 0000-0002-9798-1794

Airo Hino
School of Political Science and Economics
Waseda University
Tokyo, Japan
ORCID : 0000-0001-5141-106X

Mate Kovacs
College of Information Science and Engineering
Ritsumeikan University
Kusatsu, Japan
ORCID : 0000-0001-5999-8061

Uwe Serdült
College of Information Science and Engineering
Ritsumeikan University
Kusatsu, Japan
Center for Democracy Studies Aarau (ZDA)
University of Zurich
Zurich, Switzerland
ORCID : 0000-0002-2383-3158

***Abstract*—Voting advice applications (VAA) allow potential voters to compare their own policy positions to political parties running for an election. One of the key design elements of a VAA are the policy statements representing the political space covered by political parties. VAA designers face the challenge of coming up with policy statements in a short time frame. Even with medium-sized corpora of texts such as party manifestos, the formulation and selection of policy statements serving as a stimulus in the VAA is a tedious and time-consuming task. In addition, there is the risk of human selection bias. This study proposes a system to aid VAA designers in policy statement selection and formulation. The system uses the BERT language model with semantic similarity calculation to mine party manifesto sentences that are relevant to already existing VAA statements. For the experiments, VAA statements stemming from the 2021 elections and party manifestos issued for the previous two Japanese elections were used. To expand the policy space, VAA statements from the 2019 European Parliament elections were added. Results show that the proposed system is able to analyze large amounts of text in a short time, and mines text that provides practical support for designing and improving VAAs.**

***Index Terms*—natural language processing, voting advice applications, machine learning, e-democracy**

I. Introduction

Voting advice applications (VAA) are online civic tech tools getting used primarily during the campaign phase of democratic elections. In some constituencies such educational tools reach many thousands of users. The main feature of a VAA allows potential voters to agree or disagree to a set of predefined policy statements, for example on a 5-point Likert scale. In a second step, the policy positions of a user are compared to the ones of all political parties running in that particular election, producing a score of overlap and eventually an ideological mapping in two-dimensional space (see Figures 1 and 2). One of the key challenges for VAA designers is to come up with those 20-30 policy statements. A second very time-consuming task is to find out the stance of all political parties regarding each of the policy statements. However, in this article, we are dealing with the former challenge, applied to a VAA in Japan [1]–[3].

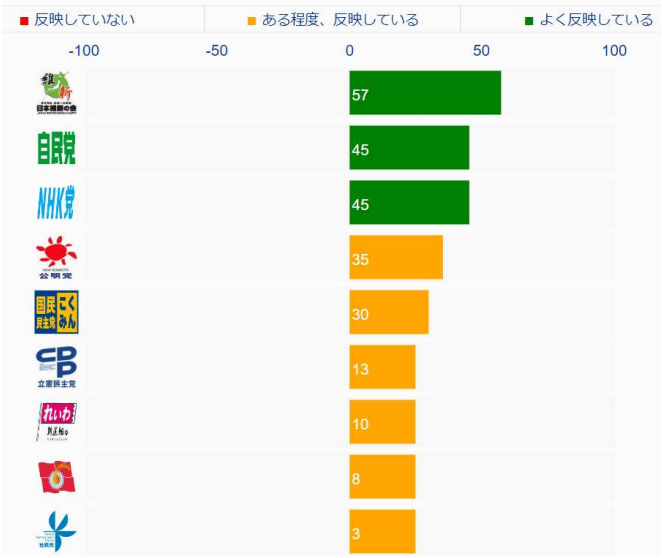

Fig. 1. User agreement with Japanese parties from FokusJapan [4].

VAAs were introduced to Japanese elections in the early 21st century, initiated by academic researchers. Over the last 15 years and in particular for more recent elections, online newspapers, youth groups as well as academics launched their own respective VAA instances [5]. Typically nowadays, the Japanese public can chose between several VAAs on offer. For the organizers of a VAA the time period specified for election

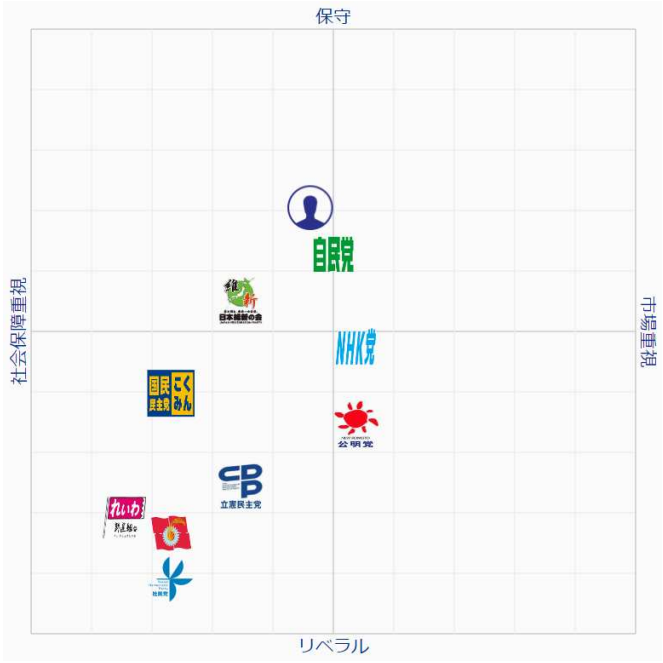

Fig. 2. Positioning of parties and the user in a two-dimensional space with a progressive-conservative and an economically liberal and more interventionist axis from FokusJapan [4].

campaigns is rather short. The Public Offices Election Law in Japan regulations stipulate that the official election campaign period for the Lower House election is for 12 days, and the one for the Upper House election is for 17 days. Normally, political parties make their manifestos publicly available close to the campaign period. VAA designers thus have little time to fully evaluate party positions before the start of the official campaign. It makes it even more problematic when Parliament, i.e. the Lower House, is dissolved, and a snap election is held. This is often the case. The likelihood of a snap election in Japan is quite high [6]. In the postwar history, it only happened twice that the mandate of the Lower House members was fulfilled. Since there is no dissolution of the Upper House, and elections are held every three years on a regular basis, it is slightly easier to plan a VAA in that case.

Several approaches have been taken so far to determine the supply-side of parties' or candidates' policy positions. The VAAs organized by a newspaper often field a survey to all the candidates and their responses are used as their positions. Responses to candidate surveys are sometimes used as a basis and the mean values from their responses serve as a proxy for party positions. Newspapers often allow users to explore matching results not only regarding political parties but also for single candidates. A more popular and common way to find out about party positions is to simply ask party headquarters to provide their take on all the policy statements covered in the VAA. Obviously, both of the approaches must rely on the responsiveness of every single political party, party headquarter or politician. Also, one has to assume that the given answers are honest and unbiased [7]. An alternative approach is to conduct a content analysis of party manifestos. This non-reactive way of deciding about the most relevant policy statements and the respective party positions is more transparent. Policy positions and party stances can be empirically rooted in the texts of the party manifestos as proof. However, this approach comes with a caveat as well. It simply takes a lot of time to go through all party manifestos.

The system proposed here is an attempt to reduce the burden of the VAA designers who have to come up with policy statements in a limited amount of time. Using a state-of-the-art language model and semantic similarity calculation methods, the system outputs party manifesto sentences corresponding to relevant VAA statements. Results demonstrate the practical usefulness of this approach for the VAA designers, assisting them in the tedious process of policy statement selection and formulation.

The remaining parts of the paper are organized as follows. Related work is presented in Section 2. Section 3 describes the proposed system in more detail. Section 4 introduces the Japanese case study with the data used and experiments made. In section 5 the results and a discussion thereof can be found. Finally, in Section 6, conclusions are made and directions for further work are outlined.

II. Related work

Over the past twenty years, the usage rate for VAAs has noticeably increased. The use of VAAs has proliferated in particular in democratic countries with multi-party systems. In such political settings it is more onerous for the potential voter to find out what the political parties stand for. [8] provides figures for the uptake of VAAs during the Swiss, Finish, Danish and German elections and found that up to 30% of voters have used a VAA. Taking into account the number of voters that might be influenced by VAAs [9], policy statements need to be carefully selected. They are key elements that affect the output of a VAA [10]. Also, statements have to meet some specific requirements: they should be relevant for the upcoming elections, capture the main issues of the electoral campaign and reveal the various dimensions of competition in the country's political system [11]. In the political science literature, for example, VAA statement quality is usually assessed post hoc by analyzing whether measurement scales of a higher order can be built. This process can also be built into a VAA semi-dynamically. In that way the cold start problem for a VAA, in the sense that one can not know whether a VAA statement produces the intended response from users or not, can be somewhat alleviated [12], [13]. In case the policy statements in a VAA fail to fulfill those requirements the whole VAA might be biased. One or several political parties could benefit from the fact that certain topics are covered or omitted in a VAA. Generally, statement selection is done manually by means of iterative discussions between VAA designers. In most cases this is done ad hoc based on expertise, or based on general population surveys in the best case. Corresponding party answers are coded by the VAA designers as well, sometimes supported by party representatives [14]. However, if done on the basis of party

manifestos, formulating twenty or more policy statements and exploring respective party positions require a considerable amount of time and effort.

With the development of natural language processing and machine learning methods, it becomes more common to tackle similar problems using modern technologies, at least partially. [15] applied sentiment analysis on Twitter data to update the most discussed political topics in a country and find out about their importance to candidates. A topic modeling approach was applied to petition data by [16] to discover citizen's policy suggestions. In a similar vain, text summarization techniques were used to harvest open government data related to politics from an e-petition platform such as JOIN [17], [18]. Topic modeling and language models can thus help to enhance democratic decision-making processes and analyze public opinion on politics-related issues. [19] developed an approach that intends to solve the problem of information overload caused by the large number of proposals initiated by the people as an input to a political system. The authors summarized comments related to proposals retrieved from the co-creation platform CONSUL. Based on the exploration of proposal topics they were able to group people according to their respective interests. [20] analyzed political comments related to two legislative proposals that were publicly accessible on the online portal of the Brazilian Chamber of Deputies. The authors demonstrated how topic modeling technologies may be used to find latent topics in comments and how this can enhance civic participation in the implementation phase of a governmental policy. In another study, topic modeling was used to analyze 12 million tweets from the "QAnon" platform related to US elections, which were uploaded between August and September 2020. According to the topic descriptions with word frequencies, a few issues were predominant among the tweets [21].

Using publicly available party manifesto data to help VAA designers in policy statement formulation, however, has not yet been investigated.

III. Proposed System

The proposed system has two inputs: party manifestos issued before the elections (Input 1) and VAA statements designed in previous years (Input 2). Figure 3 shows the general framework for the proposed approach. Before topic modeling is applied, the manifesto corpus is split into sentences (Sentence tokenization). After sentence-based topic modeling, each topic contains semantically similar sentences stemming from the party manifestos. For the Input 2, statements can originate from other regions of the world as well. Modern democracies these days are all dealing with very similar political issues. Dealing with health care, education, pension systems and tax issues is common for political systems whether in Japan or not. Opening up the policy space with statements from VAAs from outside Japan therefore helps to eventually uncover potentially relevant issues that were not present in the original VAA so far. Semantic similarity between topics and VAA statements is calculated by averaging the cosine similarity scores of sentence

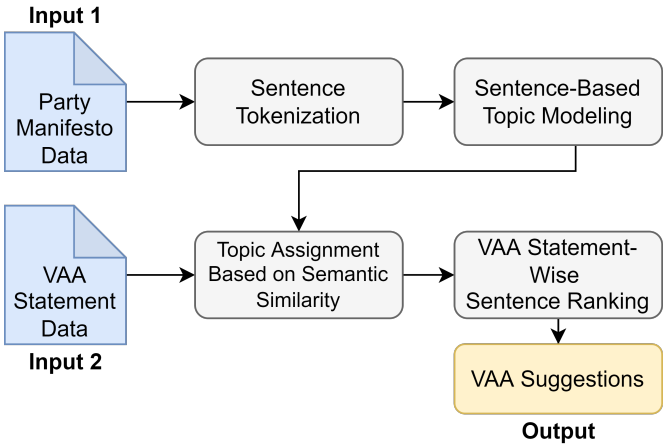

Fig. 3. Processing flow of the proposed system.

vectors of a given topic and a VAA statement vector. The vectors are created by a language model capable of producing contextualized word vectors. If the semantic similarity score between a VAA statement and a topic is above 0.5, the topic is assigned to a statement. From the pool of assigned topics, the sentences with the highest individual similarity scores are extracted and utilized as suggestions for the VAA designers. The methods and technologies used are further detailed below.

A. BERT Topic Modeling

Topic modeling is an unsupervised machine learning technique used to find hidden topics from a collection of documents. One of the most popular topic modeling approaches called LDA (Latent Dirichlet Allocation) [22] uses the BOW (Bag-of-Words) model to process documents, and relies on a probabilistic approach to cluster them into topics. BOW, however, ignores word context and ordering [23], and LDA suffers from the sparsity problem [24]. Thus, it is not optimal for clustering short texts such as sentences, and works better with longer documents.

BERT (Bidirectional Encoder Representations from Transformers) [25] is one of the state-of-the-art language models widely used for various natural language processing tasks. It is a context-aware language model that can generate word embeddings, and also sentence embeddings by averaging the values across token embeddings. In natural language processing, tokens represent a single piece of information, be it in the form of words, a punctuation or numbers. An embedding is a numerical vector representation of a natural language word. The distance between vectors can be used to measure semantic similarity between words. In contrast to conventional embedding methods that produce static vectors like Word2Vec [26], BERT generates sentence-specific dynamic vectors, which help to better incorporate word context into the vectors. Therefore, in this study, the context-aware BERT topic modeling technique [25] is utilized to cluster the sentences from party manifestos. Figure 4 depicts the processing steps of BERT topic modeling.

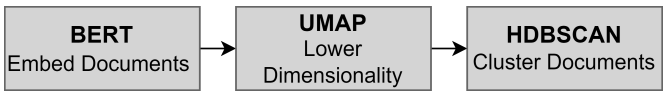

Fig. 4. The processing steps of the BERT topic modeling algorithm.

Clustering with BERT topic modeling works as follows. After generating sentence embeddings with BERT, dimensionality reduction is applied to create lower-dimensional sentence embeddings of the data using UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction). Next, semantically similar sentence embeddings are clustered with HDBSCAN (Hierarchical density-based spatial clustering of approximation with noise) [27]. It calculates the probability of a document being a part of clusters.

B. Similarity Calculation

The semantic similarity between a topic and a VAA statement is measured by averaging the sum of all similarity scores of sentences from a given topic and a VAA statement. The similarity score of the sentence and the VAA statement is calculated by measuring the cosine of the angle between their vectors in the interval of $[0, 1]$. Thus, the similarity score $CosSim$ of two vectors A and B is calculated by:

$$CosSim(A, B) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\mathbf{B}\|}$$

where the numerator is a dot product of the vectors, and the denominator is the product of the two vectors' length. The closer the cosine angle to 1, the more semantically similar the contents are. A score of 0 suggests that the sentences are unrelated, whereas a score above 0.5 shows that they are more similar than dissimilar.

For the proposed system, the vectors are produced using the sentence-BERT model that demonstrated higher computational efficiency and performance gain over the original BERT model [28]. The sentence-BERT model is re-trained for similarity calculation tasks and produces sentence vectors which can be used to compare sentences.

IV. Case Study

The current paper uses FokusJapan, a VAA in Japan, as a case study. However, the proposed system is applicable in any other country where party manifesto data is openly available [29], and where there is an offer of VAAs. The method proposed in the study demonstrated its usefulness and indeed helped the process of formulating the policy statements for the Japanese VAA called FokusJapan [4], designed for the House of Councillors elections held on 10 July 2022.

A. Data

The party manifestos retrieved for the case study originate from the official party websites updated before the 2021 Japanese Upper House elections and 2022 House of Councillors elections. The main function of a manifesto is to inform voters about policies a party supports and intends to enact if elected. It is a rich source of information regarding the political and ideological landscape, allowing to depict the divergence of political attitudes across parties. The party manifesto data stems from the following political parties: *Liberal Democratic Party, Komeito, Constitutional Democratic Party of Japan, Democratic Party for the People, Japanese Communist Party, Japan Innovation Party (as known as Ishin), Social Democratic Party of Japan, Reiwa Shinsengumi and NHK Party*.

The VAA data in the form of policy statements stemming from 7 Japanese and 2 European VAAs was manually retrieved and utilized to help VAA designers in policy statement formulation. The statements were issued for 2021 Japanese Upper House elections and the 2019 European Parliament election, respectively. The policy statements were extracted from VAAs originating from the following sources: *Mainichi Shimbun, Asahi Shimbun, JX PRESS, Shimotsuke Shinbun, Japan Choice, FokusJapan, EUandi* [30] and *EUvox* [31]. In the end, there were 135 and 50 statements covering the most important policy areas in Japan and the EU, respectively. Table I presents an example of statements from VAAs utilized for the case study. While most VAAs use policy statements to trigger a response of agreement or disagreement from the users, others formulate a question to ask for agreement or disagreement, or for choices of categorical answers.

TABLE I
A few Examples of VAA Policy Statements

Voting Advice Application	Policy Statement or Question
Mainichi Shimbun	Same-sex marriage should be recognized
Asahi Shimbun	Japan's defense capabilities should be strengthened
JX PRESS	Japan should stop the selective surname system
Shimotsuke Shinbun	Do you agree with the government's measures against corona?
Japan Choice	What do you think about the consumption tax?
FokusJapan	The minimum wage should be raised to 1500 yen or more
EUandi	Euthanasia should be legalized
EUvox	Women should be free to decide on matters of abortion

Note that the language used for policy statement formulation used in VAAs should be straightforward and easily understood by all strata of the population. Double negation, acronyms and foreign words should be avoided in order not to confuse the users. Since most users these days consult a VAA via a smartphone, the necessity for brief sentences becomes even more important and only accentuates the complexity of the VAA designers' work.

B. Experiments

Experiments were conducted on a workstation equipped with an NVIDIA Geforce RTX 2080 GPU, an Intel i9-9920X CPU, and 128 GB RAM. Although using a GPU is not a requirement for reproducing the experiments of this study, word and sentence embedding computation becomes significantly faster compared to using only the CPU.

Party manifestos from the nine Japanese political parties were split into sentences before performing BERT topic modeling. The whole manifesto corpus contains 13,408 sentences. The average number of characters per sentence is 71.8. The sentence-BERT model was used to produce sentence embeddings. It was utilized because the original BERT model, which had not been trained for similarity calculation, typically underperforms in this area [32]. A sentence-BERT model pretrained on Japanese text was used to calculate the semantic similarities of VAA statements and party manifesto sentences. For calculating the similarity scores between VAA statements designed for the 2019 European Parliament elections and sentences from Japanese party manifestos, the multilingual version of sentence-BERT model was applied [33]. The model was pre-trained on a large text collection, including Wikipedia, news data, etc. in Japanese.

Multiple topic models were built with an increasing number of topics k. Since a VAA usually consists of 20-30 statements, k was set to 30 initially. Ultimately, this choice is arbitrary. In total, five models were made, from $k = 30$ to $k = 50$, with incremental steps of 5. Based on the authors' manual assessment, the model with 35 topics was selected to be used in the experiments. Based on human inspection and comparing to the other models, when k was set to 35, the model was able to produce more consistent yet diverse topics. Since topic labels would not affect the final output of the system, topic labeling was not conducted. On average, a topic contained 305 sentences. Approximately 20% of the sentences of the whole corpus (2,725 sentences) were not assigned to any topic and considered as outliers. In other words, if a certain subject is only represented by one or a few sentences, it will not form a topic. Keeping only the topically relevant part of the corpus narrowed down the amount of data used in later steps significantly.

To choose candidate topics for extracting the semantically most relevant sentences for a given VAA statement, the similarity between each topic and a VAA statement is measured. This was done by calculating the average cosine similarity of manifesto topic sentences and VAA statements. Considering the numerical range of the cosine similarity measure, the similarity threshold for topic assignment was set to 0.5. Since in this case study, the statements are originating from multiple VAAs, statements discussing similar issues were merged. The whole process resulted in 32 statements to which topics were assigned. In the next step, the semantic similarity between each sentence of the assigned topics and the corresponding VAA statements was calculated, and the sentences were ranked based on their similarity scores. The top five sentences were used as suggestions for the VAA designers. The decision to show only the top five sentences was purely pragmatic. In our view, five sentences as suggestions for a VAA policy statement are the amount of text that can be handled and still allows for a broad enough choice of options to a VAA designer. In total, 160 sentences from the manifestos were used. Since it is essential for VAA designers to understand the degree of inter-party polarization, party names indicating the manifesto source of the sentences are shown in the output as well.

V. Results and Discussion

The proposed system produced five sentences with the highest similarity scores above 0.5 for each selected VAA statement. The obtained results can be an additional source of information for the VAA designers, in addition to the manually explored conventional data sources, such as party candidate websites, surveys, etc. Table II shows an example of suggestions produced by the system. All the suggestions produced by the system are labeled by party names indicating the source manifestos and can be found at the following link for further consultation: https://github.com/DBurya/FokusJapanVAASuggestions. The statement in Table II refers to the political discussion in Japan regarding the reform to make tuition fees for higher education free. This topic has received much attention in recent years and is part of the ongoing education costs issue, which are pointed out as being 'high' compared to other developed countries. While the Liberal Democratic Party (LDP) and Komeito plan to reduce the burden of tuition fees for private school students and increase the number of interest-free scholarships, many opposition parties are calling for a tuition-free system. For example, Japan Innovation Party, Democratic Party for the People, and Reiwa Shinsengumi want to make tuition fees for higher education completely free and allow scholarships that will not need to be paid back. The suggestions produced by the proposed system stay in line with the aforementioned political

TABLE II
A translated example of system suggestions for a VAA statement

VAA statement
Higher education should be completely free of charge.
System suggestion 1
[Japanese Communist Party] We are aiming for free education, halving tuition, abolishing entrance fees, and eliminating school lunch fees - halving tuition at universities and vocational schools, and making them free in the future.
System suggestion 2
[Japan Innovation Party] Zero educational burden for the next generation of children, all education such as early childhood education, high school, university, etc. so that we can receive an equal quality education regardless of the financial situation of the family.
System suggestion 3
[Reiwa Shinsengumi] We will create a society where people can go to graduate school for free without owing debt if students are willing to learn.
System suggestion 4
[Komeito] Expand the scholarship program and tuition reduction/exemption (new system to support students) to middle-income families, including families with multiple children and students of science, engineering, and agriculture, who especially need to reduce their burden, so that anyone can enter university if they wish, regardless of their family's financial situation.
System suggestion 5
[Democratic Party for the People] Reduce tuition fees for higher education, including universities and graduate schools, and extend non-repayment scholarships to middle-income families.

landscape, and LDP is not presented, which might indicate the system's ability to provide results related to current political reality. This shows the usefulness of the proposed system for the VAA designers since it can provide them with the results reflecting the political climate of a country.

To test the proposed system's output in a real-word scenario, the produced results were sent to the FokusJapan group of VAA designers as a support for policy statement formulation on the occasion of the House of Councillors elections held on July 10, 2022. For the 2022 elections, the VAA FokusJapan comprised twenty policy statements, among of which nine were formulated based on the results of the proposed system (see Table III). Some of these statements were identical to the statements used in the previous elections, which means that the proposed system can not only suggest to implement some new statements, but also use the same statements with the source manifesto provided. The rest of the other eleven statements were selected based on the manual assessment by the VAA designers based on the party manifestos. Experiments demonstrated that the obtained results can ease the burden for VAA designers in terms of time and effort. Policy statements provided by the system can be used as an additional source of information. Since the system output includes the opinions of political parties towards certain issues, it also helps with statement elimination. If there is no disagreement on a statement among the parties, meaning that it does not differentiate between the political parties, the statement can be excluded from the VAA. Furthermore, based on the similarity score, the system allows to quantify how critical a certain issue/topic is for a certain party. Only manual assessment of this topical importance would be a time-consuming and laborious task.

TABLE III
SELECTED STATEMENTS FOR THE FOKUSJAPAN VAA ORIGINATING FROM THE PROPOSED SYSTEM

(Japanese VAA) [Original] Statement 1
新型コロナに関して、感染抑止よりも経済活動を優先すべきである。
[English] Statement 1
Economic activity should be prioritized over infection control of coronavirus.
(Japanese VAA) [Original] Statement 2
温室効果ガスを削減するために炭素税を導入すべきである。
[English] Statement 2
A carbon tax should be introduced to reduce greenhouse gases.
(Japanese VAA) [Original] Statement 3
原子力発電所は早めに廃炉すべきである。
[English] Statement 3
Nuclear power plants should be decommissioned in advance.
(Japanese VAA) [Original] Statement 4
消費税を5%以下に下げるべきである。
[English] Statement 4
The consumption tax should be reduced to 5% or less.
(Japanese VAA) [Original] Statement 5
富裕層に対する課税を強化すべきである。
[English] Statement 5
Taxation on the rich should be strengthened.
(Japanese VAA) [Original] Statement 6
最低賃金を1500円以上に引き上げるべきである。
[English] Statement 6
The minimum wage should be raised to 1500 yen or more.
(European/Japanese VAAs) [Original] Statement 7
年金の支給額を徐々に減らすべきである。
[English] Statement 7
The amount of pension payment should be gradually reduced.
(Japanese VAA) [Original] Statement 8
高等教育を完全無償化すべきである。
[English] Statement 8
Higher education should be completely free.
(Japanese VAA) [Original] Statement 9
被選挙権年齢を引き下げるべきである。
[English] Statement 9
The eligibility age to be elected should be lowered.

VI. CONCLUSIONS

In the current study, a system was proposed to support the VAA designers in the statement formulation process by extracting VAA statement-wise suggestions from a party manifesto data corpus. A case study was conducted utilizing policy statements from both Japanese and European VAAs as well as party manifesto data issued on the occasion of previous Japanese elections. The system output was sent to a group of VAA designers (FokusJapan) and has shown practical benefits for the task of selecting and formulating policy statements. The case study demonstrates how analyzing a large amount of textual data with the help of machine learning techniques can contribute to the further development of online civic tech tools such as VAAs. In addition to the efficiency gain, the use of the proposed system also helps to build a more solid empirical foundation for a critical part of a VAA, namely the selection and formulation of policy statements.

However, while the proposed system was tested in a real-world scenario and had demonstrated convincing results in practical use, there are limitations of the study that should be pointed out. During the use of BERT topic modeling, 20% of sentences were not assigned to any topic and considered as outliers. However, they might hold information which can be used to formulate new policy statements not yet covered in the VAA. Furthermore, in collaboration with VAA designers, additional tests with different VAAs are needed to further test the functionality of the system.

In future work, the proposed system can be expanded to process additional data sources such as comments from social media, online news or e-petitions. Such an expansion of the data corpus would allow to avoid a one-sided approach when statements are created. In addition to documents issued by political organizations, text data originating from public political debate would enter the system as input as well. Also, the obtained results should be better justified by using further test metrics.

REFERENCES

[1] J. Wheatley, "Cleavage structures and dimensions of ideology in english politics: evidence from voting advice application data," *Policy & Internet*, vol. 8, no. 4, pp. 457–477, 2016.

[2] D. Garzia and S. Marschall, "Research on voting advice applications: state of the art and future directions," *Policy & internet*, vol. 8, no. 4, pp. 376–390, 2016.

[3] D.-C. Liao, W. B. Chiou, J. Jang, and S. H. Cheng, "Strengthening voter competence through voting advice applications: an experimental study of the ivoter in taiwan," *Asian Education and Development Studies*, vol. 11, no. 2, pp. 213–223, 2022. [Online]. Available: https://doi.org/10.1108/AEDS-03-2020-0043

[4] "FokusJapan: Japanese voting advice application," http://www.fokusjapan.com/, 2022, accessed: 2022-07-07.

[5] H. Tsutsumi, T. Uekami, K. Inamasu, H. I. Levy, J. Song, and Y. Shinada, "The impact of voting advice applications on voters' behavior and political interest," in *CeDEM Asia 2018: Proceedings of the International Conference for E-Democracy and Open Government; Japan 2018*. Edition Donau-Universität Krems, 2018, p. 123.

[6] A. Hino and H. Ogawa, "Japan: Political development and data for 2017," *European Journal of Political Research Political Data Yearbook*, vol. 57, no. 1, pp. 162–175, 2018.

[7] V.-J. Ilmarinen, V. Isotalo, J.-E. Lönnqvist, and Åsa von Schoultz, "Do politicians' answers to voting advice applications reflect their sincere beliefs? comparing publicly and confidentially stated ideological positions in a candidate-centred electoral context," *Electoral Studies*, vol. 79, p. 102504, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0261379422000622

[8] M. Germann and K. Gemenis, "Getting out the vote with voting advice applications," *Political Communication*, vol. 36, no. 1, pp. 149–170, 2019. [Online]. Available: https://doi.org/10.1080/10584609.2018.1526237

[9] S. Munzert and S. Ramirez-Ruiz, "Meta-analysis of the effects of voting advice applications," *Political Communication*, vol. 38, no. 6, pp. 691–706, 2021.

[10] J. Lefevere and S. Walgrave, "A perfect match? the impact of statement selection on voting advice applications' ability to match voters and parties," *Electoral Studies*, vol. 36, pp. 252–262, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0261379414000420

[11] D. Garzia and S. Marschall, "Voting advice applications," 03 2019. [Online]. Available: 10.1093/acrefore/9780190228637.013.620

[12] M. Germann, F. Mendez, J. Wheatley, and U. Serdült, "Spatial maps in voting advice applications: The case for dynamic scale validation," *Acta Politica*, vol. 50, no. 2, pp. 214–238, 2015.

[13] M. Germann and F. Mendez, "Dynamic scale validation reloaded," *Quality & Quantity*, vol. 50, no. 3, pp. 981–1007, 2016.

[14] K. Gemenis, "An iterative expert survey approach for estimating parties' policy positions," *Quality & Quantity: International Journal of Methodology*, vol. 49, no. 6, pp. 2291–2306, 2015.

[15] L. Terán and J. Mancera, "Dynamic profiles using sentiment analysis and Twitter data for voting advice applications," *Government Information Quarterly*, vol. 36, no. 3, pp. 520–535, 2019.

[16] L. Hagen, "Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?" *Information Processing & Management*, vol. 54, no. 6, pp. 1292–1307, 2018.

[17] 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)*, 2021, pp. 199–204.

[18] D. Buryakov, M. Kovacs, V. Kryssanov, and U. Serdült, "Using Open Government Data to Facilitate the Design of Voting Advice Applications, EGOV2022: EGOV-CeDEM-ePart 2022 Conference, Linköping, Sweden, September 6-8, 2022."

[19] M. Arana-Catania, F.-A. V. Lier, R. Procter, N. Tkachenko, Y. He, A. Zubiaga, and M. Liakata, "Citizen participation and machine learning for a better democracy," *Digit. Gov.: Res. Pract.*, vol. 2, no. 3, jul 2021. [Online]. Available: https://doi.org/10.1145/3452118

[20] N. F. F. d. Silva, M. C. R. Silva, F. S. F. Pereira, J. P. M. Tarrega, J. V. P. Beinotti, M. Fonseca, F. E. d. Andrade, and A. C. P. d. L. F. de Carvalho, "Evaluating Topic Models in Portuguese Political Comments About Bills from Brazil's Chamber of Deputies," in *Intelligent Systems*, A. Britto and K. Valdivia Delgado, Eds. Springer International Publishing, 2021, pp. 104–120.

[21] A. Anwar, H. Ilyas, U. Yaqub, and S. Zaman, "Analyzing QAnon on Twitter in Context of US Elections 2020: Analysis of User Messages and Profiles Using VADER and BERT Topic Modeling," in *DG.O2021: The 22nd Annual International Conference on Digital Government Research*, ser. DG.O'21. ACM, 2021, p. 82–88.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 993–1022, mar 2003.

[23] R. Silveira, C. G. Fernandes, J. A. M. Neto, V. Furtado, and J. E. P. Filho, "Topic modelling of legal documents via legal-bert," in *RELATED 2021, Relations in the LegalDomain Workshop, in conjunction with ICAIL*. Online: CEUR-WS.org, 2021, pp. 64–72. [Online]. Available: http://ceur-ws.org/Vol-2896/

[24] H.-Y. Lu, Y. Zhang, and Y. Du, "Senu-ptm: a novel phrase-based topic model for short-text topic discovery by exploiting word embeddings," *Data Technologies and Applications*, vol. 55, no. 5, pp. 643–660, 2021. [Online]. Available: https://doi.org/10.1108/DTA-02-2021-0039

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *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)*. Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: 10.18653/v1/N19-1423

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[27] L. McInnes and J. Healy, "Accelerated hierarchical density based clustering," *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, vol. 10, pp. 33–42, Nov 2017. [Online]. Available: 10.1109/icdmw.2017.12

[28] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *CoRR*, pp. 3982–3992, 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[29] P. Lehmann and M. Zobel, "Positions and saliency of immigration in party manifestos: A novel dataset using crowd coding," *European Journal of Political Research*, vol. 57, no. 4, pp. 1056–1083, 2018. [Online]. Available: https://ejpr.onlinelibrary.wiley.com/doi/abs/10.1111/1475-6765.12266

[30] A. Reiljan, F. F. da Silva, L. Cicchi, D. Garzia, and A. H. Trechsel, "Longitudinal dataset of political issue-positions of 411 parties across 28 European countries (2009–2019) from voting advice applications EU profiler and euandi," *Data in Brief*, vol. 31, pp. 1–9, 2020.

[31] K. Gemenis, F. Mendez, and J. Wheatley, "Helping citizens to locate political parties in the policy space: a dataset for the 2014 elections to the european parliament: social and behavioural sciences," *Research Data Journal for the Humanities and Social Sciences*, vol. 4, no. 1, pp. 13–26, 2019.

[32] B. Li, H. Zhou, J. He, M. Wang, Y. Yang, and L. L. 0005, "On the Sentence Embeddings from Pre-trained Language Models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Association for Computational Linguistics, 2020, pp. 9119–9130. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.733/

[33] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084