Potential for Decision Aids based on Natural Language Processing

Maximilian Witte
University of Hamburg
Jasper Schwenzow
Universität Hamburg
Mark Heitmann
University of Hamburg
Martin Reisenbichler
Vienna University of Economics and Business
Matthias Assenmacher
Ludwig-Maximilians-Universität München

Cite as:
Witte Maximilian, Schwenzow Jasper, Heitmann Mark, Reisenbichler Martin, Assenmacher Matthias (2023), Potential for Decision Aids based on Natural Language Processing. *Proceedings of the European Marketing Academy*, 52nd, (114322)

Paper from the 52nd Annual EMAC Conference, Odense/Denmark, May 23-26, 2023

Potential for Decision Aids based on Natural Language Processing

Decision aids help consumers navigate the growing complexity of the choices they face. The emergence of large language models could spark a new generation of consumer decision aids based on vast amounts of unstructured text data. We evaluate the feasibility and social acceptance of such interactive decision aids in the context of political voting. In this research we (1) develop a new generation of interactive decision aids that enables human-like interaction to support individual voting behavior and (2) demonstrate the potential of such decision aids to adequately reflect complex political position and reach large parts of society. First empirical results indicate that AI-powered decision aids are particularly useful for specific sub-groups. We discuss practical challenges and solutions for building AI-powered decision aids based on free user input.

Keywords: NLP, interactive decision aids, voting advice applications

Track: Methods, Modelling & Marketing Analytics

1 Introduction

Every day, consumers make choices ranging from deciding which video to watch or which car to buy to which political party to vote for. These choices become more complex with greater product complexity, more extensive product assortments, or trade-offs between multiple highly valued goals. Confronting consumers with such complex decision problems can lead them to prolong their choice process or even avoid a decision (Iyengar & Lepper, 2000; Luce, 1998).

Marketing research has a long history of developing systems to aid human decision-making. Common examples include comparison matrices, which display product information in a compressed digestible format, and recommender systems, which reduce search space and complexity (Ansari, Essgaier, and Kohli 2000; Häubl & Trifts, 2000). Those decision aids proposed by marketing research traditionally leverage product and consumer data in structured, non-human-like formats, e.g., tables of technical features. Very recent advances in natural language processing (NLP) suggest there is potential in automated chat-like interfaces where users can discuss any topic with large scale language models and obtain surprisingly useful answers (e.g. https://chat.openai.com/).

We evaluate the potential of such approaches for decision-making problems. In order to evaluate if machine learning truly enables a new generation of decision aids two questions need to be answered: (1) Is it technically possible to build a reliable decision aid incorporating unstructured text input and (2) do consumers find such a decision aid system helpful in their process of decision making?

We train a large language model and provide answers to these questions. Specifically, we demonstrate the process of building an NLP-based decision aid for the example of political voting. We study this domain because it represents a highly complex decision, driven by many factors requiring in-depth knowledge. Other consumer decisions also have high degrees of complexity, e.g., financial services, automobiles, vacation packages. Political decisions likely represent an even more complex domain. Potential in this domain suggests potential in many others. Marketing research has also been increasingly interested in political decisions and preferences, and the polarization of political opinions (Cutright, Wu, Banfield, Kay, and Fitzsimmons, 2011; Jung, Garbarino, Briley and Wynhausen, 2017; Schoenmueller, Netzer and Stahl, 2022). We seek to add to this work by creating an interactive decision aid for political voting decisions and test its potential.

Related research has shown that early versions of non-interactive voting advice applications can influence people's voting behavior, especially in the short term (Mahéo, 2016). These applications typically consist of a pre-defined set of statements and corresponding opinions of the different parties (Garzia & Marschall, 2016). Our decision aid goes beyond this approach by allowing users to interactively ask about any topic in their own words that matter most to them personally. We utilize a language model to predict the respective level of agreement for multiple political parties. For instance, a consumer states, "We should ban cars from the city center," and the model provides the likelihood of agreement with this statement by all relevant political parties. To make these predictions plausible to users we provide best matching sources from party manifestos and tweets.

2 Method

To build a decision aid system capable of communicating with consumers via free-text input, we collect (1) annotated text data of political statements and (2) implement an algorithm that can learn the underlying distribution of opinions of political parties from natural language.

2.1 Data collection and augmentation

Training a machine learning model on text classification in a specific domain typically requires domain-specific annotated data containing text statements and corresponding labels, e.g., statement: "There should be a lower income tax", label party A: "Yes". We assemble a data set from five resources during the modeling process. An overview of the data set assembly process is shown in Figure 1.

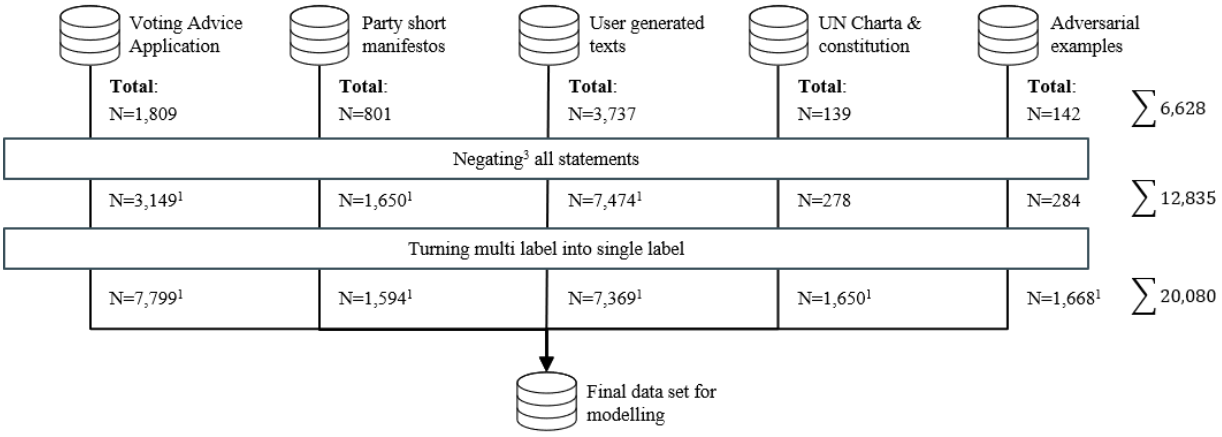

Figure 1: Overview of the modeling process. (1) Differences to N x 2 due to deletion of duplicates.

We train a model for the German political landscape because traditional non-interactive voting advice applications provide a particularly extensive list of 1,809 political statements between 2002 and 2021. The resource contains statements and six corresponding labels of the six most popular political parties in Germany, namely (1) conservatives (CDU/CSU), (2) social democrats (SPD), (3) greens (Bündnis 90/Die Grüne), (4) liberals (FDP), (5) socialists (Die Linke) and (6) nationalists (Alternative für Deutschland). We extended this data with manually extracted statements from the 2021 party manifestos (N=801), which we attributed only to the corresponding parties, meaning these statements only have one label. Additionally, we collected user-generated statements by asking participants for political statements and stating which parties they think would support those statements the most and the least. After an independent manual quality assessment, this added another 3,737 statements with two corresponding labels each to our data set.

Every statement in the data set at that point is politically controversial, meaning it would have both agreement and disagreement labels. In decision problems in general and political voting in particular, users can also be interested in topics that are actually non-controversial between party alternatives (e.g., “Slavery should not be allowed.”). Recent work has established the targeted enrichment of data sets as a strategy to increase model robustness and address related issues (Bakhtin et al., 2022; Gupta et al., 2021). In our case, we add 91 statements from the German constitution stating basic human rights, 48 statements from the United Nations Human Rights Charta, as well as 142 manually created adversarial examples. We attribute equal labels for all parties to these statements. Another potential problem is that a single word can flip the meaning of a sentence and subsequently its labels (e.g., “Let’s tax the wealthy” vs. “Let’s not tax the wealthy”). We automatically generate such negated sentences and assign labels according which increases the final data set to 12,835 labeled statements (see Figure 1).

2.2 Modelling

Next, we train different text classification models to predict the political opinions of the six major parties based on political statements. In recent years transformer-based architectures have emerged as the most capable for language-related tasks (Vaswani et al., 2017). By taking pre-trained models and fine-tuning them on a desired new task (transfer learning), highly accurate results can be achieved even with significantly smaller resources compared to the initial training (Hartmann, Heitmann, Siebert, and Schamp, 2022).

As our data set consists of German sentences, early explorative tests revealed that we need a model pre-trained on a large German text corpus to achieve accurate results. We select recent and user-approved German pre-trained language models from the Huggingface library for fine-tuning and benchmarking: BERT and Electra (Chan, Schweter, and Möller, 2020; Clark, Luong, Le, and Manning, 2020; Devlin, Chang, Lee, and Toutanova, 2018). We then fine-tune both models on our assembled data. Hyperparameter tuning is automized by employing random search for 40 epochs per model.

Two approaches can be used to make political inferences for an interactive decision aid: (1) a naïve multi-label approach and (2) a condensed single-label approach. To accomplish the second, we multiply the statements and add the parties from the labels as pre-fixes to the beginning of each statement. This approach offers two advantages over the first: It solves any missing label problem, and it increases regularization. An example of the reformulated task is shown in Figure 2.

Statement	**Label_1:** Conservatives	**Label_2:** Social Democrats
"Increase minimum wage."	0	1

Statement	**Label**
"Conservatives: Increase minimum wage."	0
"Social Democrats: Increase minimum wage."	1

Figure 2: Example of the multi-label task and single-label task reformulation.

We use hold-out validation to evaluate the technical performance of the machine-learning models. We choose to sample a random test set of 1,200 statements with equal weighting of the (1) voting advice application, (2) party manifesto data, and (3) user-generated statements. Notably, we use the UN Charta, the German constitution, and the adversarial statements only as additional training data, as those statements helped us to create more robust models.

Next, we fine-tune the models using random-search for hyperparameter optimization for approximately 120 hours using one NVIDIA Tesla K80 GPU. We fine-tune one Electra and two BERT models, one using the described single-label approach and one using the naïve multi-label (benchmark) approach. For the latter, we only use the voting advice application data set as training data because of missing labels in the other data sets. The combination of

values for the hyperparameters (1) batch size, (2) learning rate, and (3) number of epochs trained that results in the highest performance per model are displayed in Table 1.

Model	Batch size	Learning rate	Number of epochs
BERT (multi-label)	32	1.67e-5	8
BERT (single-label)	32	1.95e-5	11
Electra	32	2.21e-5	13

Table 1: Optimal hyperparameter settings for fine-tuned language models.

3 Results

To understand the potential of interactive decision aids, we evaluate the fine-tuned model in two ways. First, we assess the accuracy of the model based on unseen hold-out data to verify predictions. Second, we test user acceptance of the best-performing model and collect real-world feedback from a representative population sample.

3.1 Technical evaluation

We evaluate the technical performance of the three proposed models calculating three popular metrics accuracy, precision, and recall for hold-out test data set. These models are (1) a BERT multi-label model fine-tuned on the voting advice application data set, (2) a BERT single-label model fine-tuned on all five data sets, and (3) our final model, the single-label German Electra model fine-tuned on all five data sets (see section 2.2). The results of all three models are displayed in Table 2.

BERT multi-label				**BERT single-label**				**Electra**			
	Acc. in %	Prec. in %	Rec. in %		Acc. in %	Prec. in %.	Rec. in %		Acc. in %	Prec. in %	Rec. in %
VAA	77.2	83.1	71.7	VAA	86.0	85.9	85.9	VAA	88.2	88.2	88.2
PMO	52.0	48.9	48.4	PMO	79.8	79.8	79.8	PMO	85.5	85.4	85.6
UGC	69.5	77.4	60.7	UGC	90.5	90.5	90.5	UGC	92.0	92.0	92.1
Total	66.2	69.5	60.7	Total	85.4	85.4	85.4	Total	88.6	88.6	88.6

Table 2: Technical evaluation of fine-tuned multi-label BERT (baseline), BERT, and Electra model. Data sets: (1) Voting Advice Application (VAA), Party manifestos (PMO), User-generated content (UGC).

Compared to the accuracy of 66.2% of our initial multi-label BERT model, both single-label models fine-tuned on the extended data set, achieve a substantial increase in accuracy of 19.2% (BERT single-label accuracy of 85.4%) and 22.4% (Electra 88.6%) respectively. All models achieve the highest accuracy in classifying user-generated statements. With accuracy values of 90.5% (multi-label BERT-based) and 92% (based on Electra) the language model can achieve a remarkable level of correct inferences suggesting its usefulness as a decision aid. This is promising given relatively little text input was provided. Other marketing applications can leverage much more information from social media, professional reviews, traditional media, retailers or vendors. It may do even better to understand which alternatives match best to which types of user wants, needs and desires.

Political inferences are also more complex than other consumers decision domains because the parties themselves are often not clear in their positions. This is reflected in the lower accuracies for the short-party manifestos. Within their manifestos, parties typically extend their otherwise punctuated statements, making them softer and more heterogeneous to achieve higher consent. However, even for this text group an accuracy well above 80% appears adequate for an interactive advice application. Based on its overall highest performance, we use the fine-tuned Electra model for social evaluation.

3.2 User acceptance

To evaluate the user acceptance of the interactive decision aid, we set up an online survey asking 505 panelists representative of the German population to try out a simple website taking text statements and providing the likely support of the six main parties. Additionally, we included the option to retrieve tweets from members of a selected party and text snippets from long-form party manifestos related to the given statements to clarify why certain results appear and provide further context. Afterwards, the subjects were asked about their experience using the application. Considering the simplicity of the website and beta version of the application, overall acceptance is high. Specifically, almost half (47.7%) of participants consider the application helpful, 56.4% have fun using it, and 62.4% rating the application as interesting, suggesting overall openness to such services. Additionally, mean expressed trust in the decision aid (3.8) is higher than in politicians (3.0), private media (3.5) and social media (2.8) and comparable only to public media (4.0), rated on a scale from 1 – 7 (very high distrust – very high trust). Obviously, there is also a significant number of users who do not see a benefit in its current form. On the other hand, it would represent serious

consequences for the formation of political opinions if half the population could be reached when such applications become freely available. While our survey does not control for social exchange, news coverage, or other types of social dynamics that will also influence societal acceptance it does indicate substantial potential.

In terms of individual users, the application is accepted differently by voters of different parties. People who state they would vote for the Greens have the highest probability of finding the new voting advice application useful (62.2%), followed by liberals (54.5%), social democrats (51.3%), and socialists (51.3%) (Figure 3). Conservative, right-leaning voters find it less helpful on average with the nationalists scoring the lowest (36.0%), indicating that it could be comparatively more difficult to engage these voters using a new voting advice application with free text input. In addition, differences in experienced helpfulness of the application among different groups of voters are moderated by the people's artificial intelligence (AI) affinity. When splitting the voter groups along their expressed attitude towards AI, it becomes evident that people with a positive attitude also tend to find the application more and similarly helpful between different parties. These results indicate that the emergence of AI-driven decision aids may reach any specific sub-groups of the general population. However, the extent depends on the share of people with a positive attitude towards AI (Figure 3).

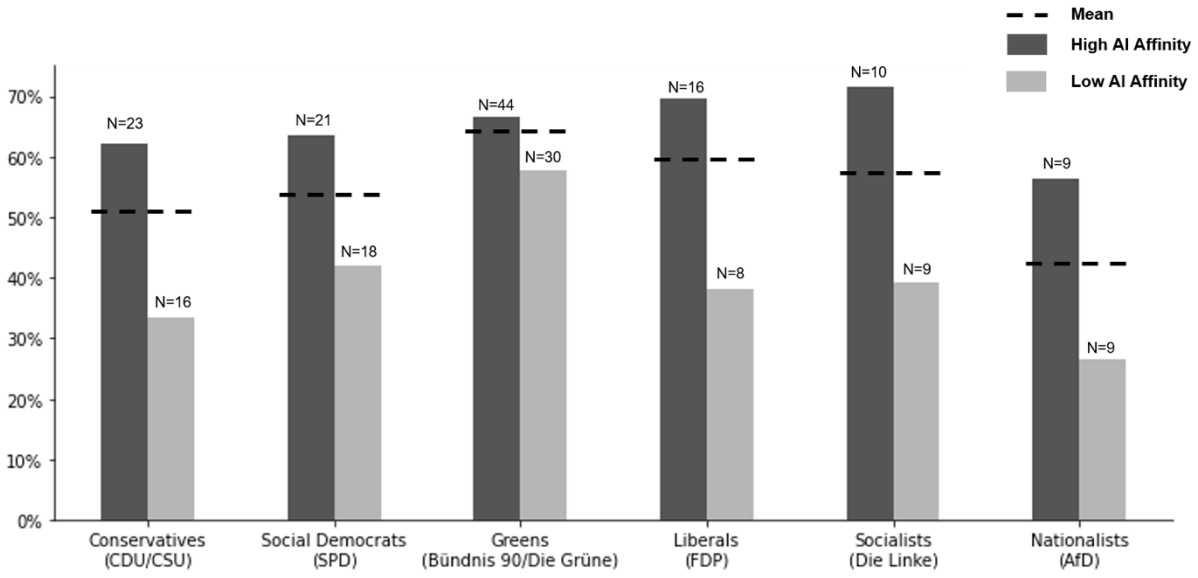

Figure 3: Percentage of subjects finding the application helpful depending on voting choice and affinity to artificial intelligence.

4 Conclusion and Outlook

Decision aids have become popular tools to enhance consumer decision-making. Recent advances in natural language processing could enable a new generation of these tools across domains that allow consumers to ask questions in their own words. We set out to evaluate the (1) technical possibility and (2) social acceptance of such decision aids in the particularly challenging domain of voting decisions. Our results indicate that fine-tuning a text classification model to predict political statements robustly is feasible and a considerable fraction of a representative panel considers using it, demonstrating that AI-driven interactive decision aids can be helpful to a large percentage of the general population in critical decisions. Providing such tools could help marketers and governments reach specific population sub-groups. However, user acceptance may depend predominantly on the user's attitude towards AI. A natural concern, especially given the high trust in the system, is a potential bias in the advice provided by an algorithm. Our service was presented as scientific research to survey participants, we do not know how participants would respond to a commercial solution.

There are obvious limitations to this research. We built the voting advice application based on readily available large-scale language models and trained them with very limited resources. This approach already results in remarkably high levels of accuracy that appear sufficient for a significant share of the population. However, high accuracy levels can mask dramatic individual errors that would require more extensive user testing to fully explore. This may result in even higher user consent than what we have obtained. Although other consumer decision-making domains such as insurance, entertainment, fashion, etc. can involve also highly sophisticated preferences and deliberations, political decisions are particularly complex. Among other things, party members themselves can contradict each other. We suspect extensions to other domains are possible but further training and testing would be needed to assess potential across domains.

References

Ansari, A., Essegaier, S., & Kohli, R. (2000). Internet Recommendation Systems. Journal of Marketing Research, 37, 363–375.

Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried D., Goff A., Gray J., Hu H., Jacob A. P., Komeili M., Konath K., Kwon M., Lerer A., Lewis M., Miller A. H., Mitts S., Renduchintala A., Roller S., Rowe D., Shi W., Spisak J., Wei A., Wu D., Zhang H. & Zijlstra M. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*.

Chan, B., Schweter, S., & Möller, T. (2020). German's Next Language Model. *International Committee on Computational Linguistics: Proceedings of the 28th International Conference on Computational Linguistics*, 6788–6796.

Clark, K., Luong, M.-T., Brain, G., Brain, Q. V. L. G., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv*

Cutright, K. M., Wu, E. C., Banfield, J. C., Kay, A. C., & Fitzsimons, G. J. (2011). When your world must be defended: Choosing products to justify the system. *Journal of Consumer Research*, *38*, 62–77.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186.

Garzia, D., & Marschall, S. (2016). Research on Voting Advice Applications: State of the Art and Future Directions. *Policy & Internet*, *8*, 376–390.

Gupta, P., Tsvetkov, Y., & Bigham, J. P. (2021). Synthesizing Adversarial Negative Responses for Robust Response Ranking and Evaluation. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3867–3883.

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2022). More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing*.

Häubl, G., & Trifts, V. (2000). Consumer Decision Making in Online Shopping Environments: The Effects of Interactive Decision Aids., 19, 4–21.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*, 995-1006.

Jung, K., Garbarino, E., Briley, D. A., & Wynhausen, J. (2017). Blue and Red Voices: Effects of Political Ideology on Consumers' Complaining and Disputing Behavior. *Journal of Consumer Research*, *44*, 477–499.

Luce, M. F. (1998). Choosing to Avoid: Coping with Negatively Emotion-Laden Consumer Decisions. *Journal of Consumer Research*, *24*, 409–433.

Mahéo, V. A. (2016). The Impact of Voting Advice Applications on Electoral Preferences: A Field Experiment in the 2014 Quebec Election. *Policy & Internet*, 8, 391–411.

Schoenmueller, V., Netzer, O., & Stahl, F. (2022). Polarized America: From Political Partisanship to Preference Partisanship.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *2017-December*, 5999–6009.