Automated Emotion Annotation of Finnish Parliamentary Speeches Using GPT-4

Otto Tarkka, Jaakko Koljonen, Markus Korhonen,
Juuso Laine, Kristian Martiskainen, Kimmo Elo, Veronika Laippala
University of Turku
20014 University of Turku, Finland
ohitar@utu.fi

Abstract

Annotating datasets can often be prohibitively expensive and laborious. Emotion annotation specifically has been shown to be a difficult task in which even trained annotators rarely reach high agreement. With the introduction of ChatGPT, GPT-4 and other Large Language Models (LLMs), however, a new line of research has emerged that explores the possibilities of automated data annotation. In this paper, we apply GPT-4 to the task of annotating a dataset, which is subsequently used to train a BERT model for emotion analysis of Finnish parliamentary speeches. In our experiment, GPT-4 performs on par with trained annotators and the annotations it produces can be used to train a classifier that reaches micro F1 of 0.690. We compare this model to two other models that are trained on machine translated datasets and find that the model trained on GPT-4 annotated data outperforms them. Our paper offers new insight into the possibilities that LLMs have to offer for the analysis of parliamentary corpora.

Keywords: emotion analysis, parliamentary speeches, annotation, chatgpt

1. Introduction

Recent years have shown growing interest in the studies of sentiments and emotions in politics (see e.g., Fraccaroli et al. 2022; Orellana and Bisgin 2023). In the Finnish context, however, this is still an underdeveloped field of study. Koljonen et al. (2022) analyze emotion in post-WWII party manifestos, but analysis on modern plenary speeches in Finland is a largely unexplored territory. While sentiment analysis typically aims to categorise texts into two or three categories (positive, negative + neutral), emotion analysis aims at a more fine-grained classification, where texts are divided into emotion categories based on the emotion(s) they reflect. Sentiment and emotion classification have traditionally been done with dictionary based methods but they have given way to deep-learning approaches, which have shown greater classification accuracy (Widmann and Wich, 2023; Borst et al., 2023).

The downside of deep-learning is that it requires annotated training data. Data annotation is notoriously laborious, time consuming and often expensive. Crowd-sourcing platforms, such as Amazon's Mechanical Turk (MTurk) can be used to cut costs but there has been growing concern over both the quality of annotations and the ethical questions that using MTurk raises (e.g. Chmielewski and Kucker, 2020; Shmueli et al., 2021). When a ready-made dataset in the desired language is not available, and there are not sufficient resources to build a dataset from the ground up, there are few options available for researchers. One option is to machine translate an existing dataset to the desired language (Eskelinen et al., 2023). Very recently, a new option has emerged, which is leveraging ChatGPT or other similar *large language models* (LLMs) to do the previously laborious and costly annotation quickly and relatively cheaply.

This paper explores the possibilities of using GPT-4 to annotate a dataset that is used to train a BERT-based classifier for analysing emotion in Finnish parliamentary speeches. We create and evaluate an emotion annotated dataset and show that a BERT model trained on this data outperforms models trained on machine translated datasets. Our results show that GPT-4 is a promising tool for creating datasets for emotion analysis in parliamentary speeches. All training scripts and annotated data are available on GitHub.[1]

2. Background

GPT is a family of LLMs trained on massive natural language datasets that continue a given prompt with words that have the statistically best fit (Floridi and Chiriatti, 2020). ChatGPT has been further trained with conversational data to produce coherent responses to questions and to follow instructions. The version of ChatGPT that is most commonly used is also known as GPT-3.5. GPT-4 is an even larger and more capable multimodal model that performs well even in many academic and professional exams (OpenAI et al., 2023).

BERT is a language representation model which was first introduced in 2018 and outperformed state-of-the-art models in several *Natural Language Pro-*

[1] https://github.com/TurkuNLP/FinParl-emotion

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 70–76
May 20, 2024.

cessing (NLP) tasks (Devlin et al., 2018). It is still today the standard in many NLP tasks. There are often two stages in the BERT algorithm workflow: first, pre-training which uses masked language modelling and next sentence prediction, and second, fine-tuning (Rogers et al., 2020). We use FinBERT (Virtanen et al., 2019) as our base model, which we fine-tune with data annotated by GPT-4.

Using ChatGPT for automating the annotation process is not a wholly original idea. In earlier research, ChatGPT has been used successfully in annotation tasks. For example, Gilardi et al. (2023) compare annotations between trained annotators and ChatGPT. They show that ChatGPT outperforms both crowd workers and trained annotators in a number of tasks with regard to inter annotator agreement. Malik et al. (2024), also, use ChatGPT to create annotated data to train a multi-label emotion classification model. Their model trained with data annotated by ChatGPT achieves satisfactory performance when using 8 emotion categories to classify emotions in tweets.

There is no one set of emotions that is universally used in emotion analysis, and, instead, papers in the field use a wide set of emotions. Bostan and Klinger (2018) compile and compare 14 datasets built for emotion classification using 12 different annotation schemes. Many papers use either Ekman's six basic emotions (Ekman, 1992) or Plutchik's wheel of emotions (Plutchik, 1982) as the basis of their set of emotions but often modify the taxonomy somewhat to suit the needs of the study. Others use a whole different set of emotions, such as the GoEmotions dataset, which employs a 28 category taxonomy (Demszky et al., 2020). The only pre-existing Finnish resources for emotion annotation that we are aware of are the XED corpus, which contains sentence-level multi-label emotion annotations for movie subtitles (Öhman et al., 2020) and the emotion lexicon SELF (Öhman, 2022).

3. Data

ParlamenttiSampo (Semantic Computing Research Group, 2021) contains the transcribed records of all plenary sessions of the Finnish Parliament (*Eduskunta*). To create our own dataset of emotion annotated plenary debates, we handpicked a number of plenary sessions discussing the reports of the Parliamentary Committees of the Finnish Parliament between the years 2017 and 2020. The 17 permanent Committees play an influential role in the decision-making in the Parliament. The Committees prepare e.g., legislative initiatives, government bills and reports for handling in plenary sessions. MPs are divided to the Committees proportionally in a way that reflects the strength of each party in the Parliament.

Each Committee works within their own field of expertise within the scope of a corresponding ministry. Thus, by choosing speeches from different committee reports, we assure that the speeches in our training data cover a variety of topics, terms and perspectives, which might evoke different emotional responses from MPs. This leads to a more representative dataset as parties tend to be more active in policy areas that are important to the party's key voter clientele (Bäck and Debus, 2016). We choose plenary debates from two different parliamentary terms to combat any bias caused by the changing dynamics between parties within parliamentary terms. Opposition politicians are inclined to have a greater incentive to persuade voters and reclaim their position as a credible alternative to become the governing party (Russell et al., 2017). In a competitive parliamentary system, opposition politicians tend to criticise government policies and, thus, their status of as an opposition MP is likely to affect their behaviour and rhetoric (Tuttnauer, 2018).

Our final data comes from 15 Committee reports consisting of 529 speeches, which were split into 6025 sentences using the Python NLP toolkit Trankit (Nguyen et al., 2021). We use the sentence as the unit of observation.

4. Methods

The steps we took in the creation of our dataset and model are as follows: First, we manually annotated a small set of sentences from parliament that act as the gold standard against which all evaluation is done. Then, we used GPT-4 to annotate the same set and evaluate its performance. Over multiple iterations and prompt engineering we reached results that are comparable to human performance. We then used GPT-4 to annotate a larger set of sentences using the same prompt. This data was then used as training data for a BERT model.

ID	Emotion	*N*	%
0	neutral	153	53
1	happiness/success	17	6
2	hopefulness/optimism/trust	33	11
3	love/praise	37	13
4	surprise (positive)	3	1
5	sadness/disappointment	3	1
6	fear/concern/mistrust	17	6
7	hate/disgust/derision	21	7
8	astonishment (negative)	6	2

Table 1: Emotion categories and gold standard labels.

4.1. Data Annotation

251 sentences from three plenary debates were manually annotated. Initially, we planned to use Ekman's six basic emotions to categorise the sentences but testing showed that the annotators struggled to assign sentences to these categories consistently. Hence, we chose to create our own set of emotions based on the emotions that we observed in the data. After further test rounds and discussion, we decided on a final set of 8 emotions + neutral (see Table 1). The final evaluation data, which we refer to as the gold standard, was annotated by four expert annotators (ann1-ann4). The annotators were native speakers of Finnish and all were familiar with the practices and typical rhetoric of the Finnish Parliament. The emotion label of a sentence was chosen by majority vote. If a sentence did not have a single winning label, all winning labels were accepted as a possible labels, which is why the numbers in Table 1 add up to more than 251. 31 sentences in the gold standard have more than one label, 27 of which have two and four have four labels.

The emotion categories in Table 1 were chosen because annotation tests showed that they reflected the data well and to create an annotation scheme where the emotions are balanced in terms of sentiment: four emotions express positive and another four negative sentiment. This is to prevent the formation of catch-all categories, which might oversimplify and distort the analysis. To ease annotation and make the emotions more clearly defined, we decided to refine the emotion categories by specifying their different manifestations: for example, love in the context of parliamentary speech can also be understood as praise or admiration. A challenge that emerged was to distinguish between true emotion and rhetorical strategy. In other words, what should be classified as an emotion, instead of a mere performance? We followed a definition commonly employed by psychologists in viewing emotion as a subcategory of *affect*, wherein affect is embodied and unconscious, while emotions are more structured and patterned expressions of affects, anchored in language.

4.2. Prompting and Model Training

Interacting with LLMs requires prompt engineering, which refers to formulating and manipulating the model input in such a way that desired results are achieved. We tried multiple different prompts and compared the results to our gold standard before settling on the final version. We found that writing the prompt in English, even though our data is in Finnish, improved the results. This effect is likely explained by the fact that GPT-4 performs worse on low resource languages, such as Finnish, compared to its performance on English (OpenAI et al., 2023). We also tried including the preceding and following sentences for each example as context but found that this only confused the model and led to worse results. We noted that keeping the instructions short and concise led to higher inter-annotator agreement than including detailed explanations for each class. Finally, using GPT-4 gave better results than the standard GPT-3.5, which is why this is the model we decided to use despite its higher cost. To save some cost, the re-occurring formulaic greeting *Arvoisa puhemies!* ('Honoured chairman!') and its variations were automatically given the "neutral" label. In total, the cost of annotating our data using the OpenAI API was around $60.

annotators	κ	F1 micro	F1 macro
ann1-ann2	0.406	0.602	0.379
ann1-ann3	0.476	0.685	0.429
ann1-ann4	0.145	0.590	0.355
ann2-ann3	0.553	0.713	0.529
ann2-ann4	0.624	0.729	0.590
ann3-ann4	0.518	0.673	0.481
average	0.499	0.655	0.416
gold-GPT-4	0.554	0.725	0.495

Table 2: Inter-annotator agreement between different human annotators and between gold standard and GPT-4.

GPT-4 was then used to annotate a dataset of 6025 sentences. The specific version of the model used is *gpt-4-0125-preview*, which is the most recent version of the model at the time of writing. The 251 sentences used for annotation evaluation were kept separate and the remaining 5774 sentences were split into train and validation sets with a 90-10 split. These data were used to train a BERT model. The model was evaluated against the gold standard annotations. We used a grid search to optimize the hyperparameters of the training stage. We used a learning rate of 3.16e-05, batch size of 32 and a label smoothing factor of 0.1.

As a baseline for our model, we trained two other BERT models using two machine translated datasets. Machine translating datasets has been shown to be a resource efficient way to create datasets that can produce better results than using multilingual models (Eskelinen et al., 2023). We produced the Finnish translations using DeepL[2]. The first baseline model is trained on the Many Emotions (ME) dataset. ME combines emotion annotated sentences from three separate datasets: Daily Dialog (Li et al., 2017), GoEmotions (Demszky et al., 2020) and Emotion (Saravia et al., 2018). These datasets source from transcriptions of casual conversations, Reddit posts and Twitter mes-

[2] https://www.deepl.com/translator

sages, respectively. The second baseline model is trained on the HunEmPoli dataset, which contains emotion annotated sentences from the Hungarian parliament (Üveges and Ring, 2023). We test the performance of these baseline models against the gold standard. Since the labels in the datasets differ slightly from our labels, we harmonise the labels before comparison by removing sentences and combining labels where necessary.

5. Results

We use Cohen's Kappa (κ) and F1 metrics to evaluate *inter-annotator agreement* (IAA). The numbers in Table 2 attest to the difficulty of the annotation task: despite many test rounds and discussion, IAA remained modest. When discussing the annotation results, we noticed that in many cases there is no single correct label for a given sentence and, instead, different interpretations are equally valid. The subjectivity of emotion annotation and subsequent low IAA has been noted before in the literature (Öhman, 2020).

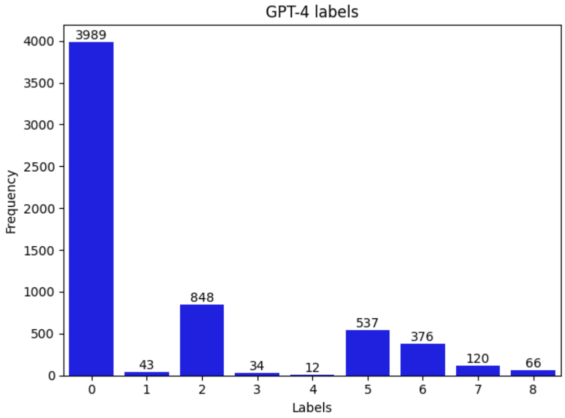

Figure 1: Distribution of labels in final GPT-4 annotated dataset.

IAA between human annotators and GPT-4 was calculated by comparing GPT-4 annotations to the gold standard. For sentences with multiple labels in the gold standard, any of the possible labels are counted as correct since GPT-4 agrees with at least one human annotator. As the numbers in Table 2 show, GPT-4 reaches human level accuracy in the task.

The model trained on GPT-4 annotated data, which we here call the GPT-4 model, reaches a macro F1 of 0.411 and a micro F1 of 0.690 meaning that the model performs well overall but struggles with some classes. This is understandable considering the distribution of labels in the datasets shown in Figure 1. The plot in Figure 2 shows that the model tends to over-predict class 0 (neutral) and seems to combine most sentences with positive sentiment in class 2 as is the case with the GPT-4 annotations, too. This suggests that the results could improve via further prompt engineering, although positive classes were also difficult for human annotators to distinguish. The comparison between models in Table 3 shows that the baseline models perform much worse. Surprisingly, even the in-domain HunEmPoli dataset does not seem to fit our data well. This might be because of differing annotation schemes and instructions, or due to cultural differences between the two parliaments. The ME model only predicts the emotions *neutral* and *joy* in our evaluation set, suggesting that casual conversation and internet discourse are too distinct from parliamentary discourse to be used as our training data. These results support the use of GPT-4 as a resource efficient method of creating training data.

GPT-4 model	**ME model**	**HunEmPoli model**
GPT-4 annotated parliamentary speeches	machine translated Many Emotions	machine translated parliamentary speeches
c. 6,000 sentences	c. 550,000 sentences	c. 19,000 sentences
9 labels	7 labels	6 labels
micro **0.690** macro **0.411**	micro 0.574 macro 0.138	micro 0.261 macro 0.182

Table 3: Model comparison

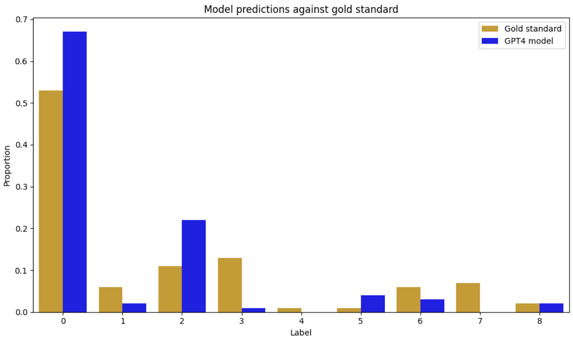

Figure 2: Model predictions vs gold standard.

6. Conclusion

In this paper, we have shown that GPT-4 can be used to create an emotion analysis dataset that can then be used to train an emotion classifier. The work presented in this paper is still ongoing as we continue refining the annotation, prompting and training procedures in the near future. This emerging methodology shows much promise as it makes the previously expensive and time consuming process of manual annotation much faster and cheaper. Machine translating existing datasets can still be a useful method for obtaining training data but, depending on the task, domain and availability of datasets, using an AI assistant such as GPT-4 might be a viable option. In the future, as the technology matures and costs are reduced, their use in data annotation could become commonplace, although they do raise their own set of challenges that must be overcome (see Ziems et al. 2024).

Many open questions still remain and there is much research being done is this emerging field. One open question is the viability of using AI assistants for other annotation tasks, as there is no guarantee that quality annotation is possible for all tasks and datasets. In fact, Heseltine and von Hohenberg (2024) show that GPT-4 annotations vary between tasks and languages. Additionally, much more is still to be learned about optimal prompting strategies. For example, Hu and Collier (2024) measure the effect of introducing persona variables, such as gender, political orientation and level of education in the prompt. We encourage other researchers in the field to continue experimenting with similar methods to advance resource efficient data annotation.

7. Acknowledgements

This research was funded by the Research Council of Finland [grant number 353569]. The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources and professor Filip Ginter for his sage advice and support in the undertaking of this research.

8. Bibliographical References

Hanna Bäck and Marc Debus. 2016. *Political Parties, Parliaments and Legislative Speechmaking*. Palgrave Macmillan UK.

Janos Borst, Jannis Klähn, and Manuel Burghardt. 2023. Death of the dictionary? – The rise of zero-shot sentiment classification. In *Computational Humanities Research Conference (CHR 2023)*.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Oswald Campesato. 2021. *Natural Language Processing Fundamentals for Developers*. Mercury Learning & Information.

Michael Chmielewski and Sarah C. Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv preprint.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Anni Eskelinen, Laura Silvala, Filip Ginter, Sampo Pyysalo, and Veronika Laippala. 2023. Toxicity detection in Finnish using machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 685–697, Tórshavn, Faroe Islands. University of Tartu Library.

Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4).

Nicolò Fraccaroli, Alessandro Giovannini, Jean-François Jamet, and Eric Persson. 2022. Ideology and monetary policy. the role of political parties' stances in the European Central Bank's parliamentary hearings. *European Journal of Political Economy*, 74.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1).

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. ArXiv preprint.

Juha Koljonen, Emily Öhman, Pertti Ahonen, and Mikko Mattila. 2022. Strategic sentiments and emotions in post-second world war party manifestos in Finland. *Journal of Computational Social Science*, 5.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*.

Usman Malik, Simon Bernard, Alexandre Pauchet, Clément Chatelain, Romain Picot-Clémente, and Jérôme Cortinovis. 2024. Pseudo-labeling with large language models for multi-label emotion classification of French tweets. *IEEE Access*, 12:15902–15916.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Emily Öhman. 2020. Emotion annotation: Rethinking emotion categorization. In *Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries (DHN 2020)*.

Emily Öhman. 2022. SELF & FEIL: Emotion lexicons for Finnish. In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries conference*.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *The 28th International Conference on Computational Linguistics (COLING 2020)*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston

Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 technical report.

Salomon Orellana and Halil Bisgin. 2023. Using natural language processing to analyze political party manifestos from New Zealand. *Information*, 14(3).

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Meg Russell, Daniel Gover, Kristina Wollter, and Meghan Benton. 2017. Actors, motivations and outcomes in the legislative process: Policy influence at Westminster. *Government and Opposition*, 52(1):1–27.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Or Tuttnauer. 2018. If you can beat them, confront them: Party-level analysis of opposition behavior in European national parliaments. *European Union Politics*, 19(2):278–298.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. ArXiv preprint.

Tobias Widmann and Maximilian Wich. 2023. Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. *Political Analysis*, 31(4):626–641.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? ArXiv preprint.

István Üveges and Orsolya Ring. 2023. HunEmBERT: A fine-tuned BERT-model for classifying sentiment and emotion in political communication. *IEEE Access*, 11:60267–60278.

9. Language Resource References

Semantic Computing Research Group. 2021. *Parlamenttisampo*. PID https://parlamenttisampo.fi/fi/.