XAI in Computational Linguistics:
Understanding Political Leanings in the Slovenian Parliament

Bojan Evkoski[*][†]**, Senja Pollak**[*]

[*]Jozef Stefan Institute
[†] Jozef Stefan International Postgraduate School
bojan.evkoski@ijs.si
senja.pollak@ijs.si

Abstract

The work covers the development and explainability of machine learning models for predicting political leanings through parliamentary transcriptions. We concentrate on the Slovenian parliament and the heated debate on the European migrant crisis, with transcriptions from 2014 to 2020. We develop both classical machine learning and transformer language models to predict the left- or right-leaning of parliamentarians based on their given speeches on the topic of migrants. With both types of models showing great predictive success, we continue with explaining their decisions. Using explainability techniques, we identify keywords and phrases that have the strongest influence in predicting political leanings on the topic, with left-leaning parliamentarians using concepts such as *people* and *unity* and speak about *refugees*, and right-leaning parliamentarians using concepts such as *nationality* and focus more on *illegal migrants*. This research is an example that understanding the reasoning behind predictions can not just be beneficial for AI engineers to improve their models, but it can also be helpful as a tool in the qualitative analysis steps in interdisciplinary research.

1. Introduction

Artificial intelligence, in particular machine learning, is extensively used for solving many real-world problems both in research (Jordan and Mitchell, 2015) and industry (Shinde and Shah, 2018). In recent years, with AI being more present and the introduction of stricter guidelines (Interpol, 2019) and regulations (Sartor and Lagioia, 2020), it became apparent that the sheer power of machine learning models for predicting tasks does not justify their widespread potentially inconsiderate usage. Today, understanding and explaining the previously considered black-box models is equally important as developing and training them. This process enables engineers to detect biases in the model, come up with ideas for improvement, and most importantly examine the security of the system. Explainable AI (or simply XAI) is gradually becoming a must both in industry and research.

Another aspect of AI interpretability is its effect of bridging the gap between heavy quantitative black-box research and qualitative research more common among humanities and social sciences scholars. As machine learning engineers get feedback from a model on why it makes a particular prediction, humanities and social sciences scholars would take these signals to further examinations that can lead to more qualitative explanations of why certain features play a significant role in a particular problem.

This work covers a case of interdisciplinary research between computational linguistics (or natural language processing) and political science where we explain AI models to gain insight from a political linguistics perspective. We focus on parliamentary debates, a salient research topic in both humanities and social science disciplines, such as sociology, political science, sociolinguistics, and history (Skubic and Fišer, 2022). Our goal is to bring more insight into the speeches of parliamentarians (MPs) of different political leanings. First, by training machine learning models to predict if a parliamentarian is left or right-leaning based on their speech. And then, by using explainability techniques on the models, extract data and derive knowledge on what actually differentiates left and right political speeches in the parliament. To make things more politically relevant and methodologically clear, we are focusing on the concerning topic of migrants and the European migration crisis (Barlai et al., 2017) where the left and right have evident divergent stances: left-leaning parties showing consistent support to immigrants from Asia and Africa throughout Europe, and right-leaning parties showing moderate to strong opposition against immigration to Europe and their country in particular (van der Brug and Harteveld, 2021). We apply our analysis to the Slovenian Parliament from 2014 to 2020. For data, we use the open-access ParlaMint dataset (Erjavec et al., 2022) which provides complete parliamentary transcriptions for 17 countries, including Slovenia.

The paper is organized as follows. In Section 2., we briefly cover the related work on using explainable AI in social sciences, as well as recent applications of computational linguistics for parliamentary debates. In Section 3. we provide a description of the ParlaMint SI dataset we are working with, as well as the preprocessing steps required before training the models. In Section 4. we dive into the training of the classical machine learning and the modern language models. Finally, we discuss the results in Section 5. and conclude in Section 6..

2. Related Work

By the end of the 2010s and the beginning of 2020s, explainable AI has become a topic that involved both computer and social scientists in discussions on how to interpret and use the knowledge drawn from model explanations (Miller, 2019). Social scientists urge the importance of qualitative investigation experts joining in XAI

projects (Johs et al., 2022). Yet, social qualitative investigation done by non-experts is still very common. Combining qualitative and quantitative approaches (mixed research) has a variety of hardships (Brannen, 2017), and researchers from both sides of the spectrum rarely have insight into the benefits of combining knowledge with the opposite side. With this work focusing on a specific social theme, we show an example of how XAI research can be the initial step for a more thorough qualitative approach (Molina-Azorin, 2016).

Parliamentary discourse, although extensively researched in a qualitative setting (Ilie, 2015), is still underexplored in a quantitative manner. Naderi and Hirst (2015) use computational methods to analyze framing structures in speeches of the Canadian Parliament. Greene and Cross (2015) use dynamic topic modeling to explore the evolution of the political agenda in the European Parliament. Eskişar and Çöltekin (2022) analyze the emotion structure of speeches in the Turkish ParlaMint dataset.

Due to the European Migration crisis in 2015, researchers have been studying the political stances of politicians (Wallaschek, 2020), news media (Krotofil and Motak, 2018), and the effect of social media on the integration of migrants (Alencar, 2018). Computational techniques were also applied. For example, (Greussing and Boomgaarden, 2017) use techniques such as the bag-of-words and principal component analysis to understand media framing on the topic, while (Heidenreich et al., 2019) apply topic modeling to observe the dynamics of the migration narrative. In a more recent work by (Skubic et al., 2022), the authors create mention networks for multiple parliaments on the topic of migration and analyze the role of gender on speech influence.

3. Data

Our choice of parliamentary transcriptions is the richly annotated ParlaMint subset of the Slovenian Parliament. It covers all sessions of the National Assembly from August 2014 to July 2020, with more than 20M words. In order to prepare the data for training a political leaning model based on the transcriptions, we first applied a few preprocessing steps according to the metadata. First, we removed all guest and Chair speeches, leaving only regular MPs. Next, used the party name data for each parliamentarian to extract the main label (right or left) using Wikipedia's "political position" English metadata on the party. Far-right, right, and center-right are labeled as right; far-left, left and center-left are labeled as left. Finally, we applied speech selection according to the topic of our interest — migration. For this purpose, we prepared a set of keywords directly connected to the migration topic[1]. The speech selection is very straightforward. If any of our keywords appear in the transcription (observed as lemmas), we select the speech. Table 1 shows a general overview of the dataset magnitude, while Figure 1 shows a pie chart of speech share selections from the set of keywords. The

---

[1] The list of migration keywords was prepared in collaboration with social scientists Andreja Vezovnik and Veronika Bajt. It contains 95 lemmas with all their corresponding word forms prepared by Anka Supej.
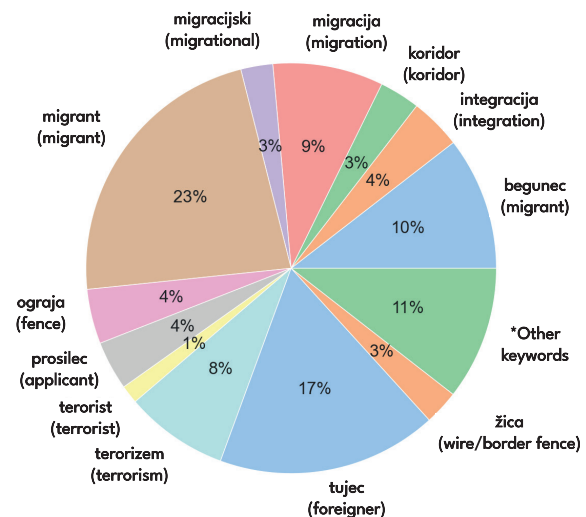

Figure 1: **Pie chart on the migration keywords.**

final dataset turned out to be quite balanced between the two classes, with 1519 of the speeches classified as "left", and 1455 classified as right.

4. Models and Explainability

We followed two approaches to **training machine learning models** for text classification. The first is a classical bag of words feature extraction combined with a Linear SVM classifier (Liu et al., 2010). The second approach is using the state-of-the-art Transformer language models (Vaswani et al., 2017). We applied several variations of both approaches, evaluated using 5-fold cross-validation, and compared the prediction accuracy. Finally, we selected the best model for each approach and trained it on the full dataset in order to apply explainability analysis.

For the classical approach, we trained a Linear SVM on three different bag-of-words types: unigram; unigram and bigram; unigram, bigram, and trigram. As the common practice, we used only lemmas which were already available in SI ParlaMint. We used a well-refined stopwords list for the Slovenian language to remove noise from the bag of words[2], but we also used a minimum frequency of 5 appearances in the whole dataset, and a maximum frequency of a word appearing in 35% of the speeches which proved effective in removing the procedural parts of speeches and improving the results. We optimized the model on each of the training folds in a nested 5-fold cross-validation and we evaluated the optimized model on the main testing folds, practically optimizing and evaluating five models of the same kind, reporting the average and its 95% confidence interval. Finally, using the hyperparameters of the best-performing model, we trained one final model which we used for the explainability experiments.

Opposed to the common understanding of machine learning models, a Linear SVM classifier is not a black box that cannot be explained in its calculations. It creates a hyperplane that uses support vectors to maximize the distance between the two predicted classes (in our case "left"

---

[2] https://github.com/stopwords-iso/stopwords-sl

Dataset	#Parties	#Speakers (Left/Right)	#Speeches (Left/Right)	AWS	MSS
SI ParlaMint	12	353 (114/48)	75122 (70.6%/20.4%)	1244.1	30.0
SI ParlaMint (MPs)	12	166 (114/48)	31151 (55.9%/40.9%)	1210.5	125.5
SI Parlamint (MPs on migration)	12	153 (105/48)	2974 (51.1%/48.9%)	854.7	12.0

Table 1: **General statistics of the Slovenian ParlaMint dataset and two of its subsets.** AWS - Average words per speech. MSS - Median speeches per speaker.

and "right"). The weights that represent the Linear SVM equation hyperplane are the vector coordinates that are orthogonal to the hyperplane. So, their direction indicates the predicted class. The absolute size of the coefficients in relation to each other can then be used to determine feature importance for the data separation task. For our task, the minimum (most negative) weight coefficients correspond to words in the bag-of-words representation that led the model to classify a speech as a "leftist", while the maximum (most positive) weight coefficients correspond to words that led it to classify a speech as "right". Using this explainability logic, we derived an understanding of what makes the difference between a leftist and a rightist ideology's thought and word choice.

For language models, we used two types: the multilingual BERT and the SloBERTa. The multilingual BERT (Devlin et al., 2018) was trained in 104 languages and is one of the first successful language models which serves as an appropriate baseline for fine-tuning the model to any kind of textual task. SloBERTa (Ulčar and Robnik-Šikonja, 2021) uses the BERT architecture and it is monolingual, meaning it is trained purely on Slovenian data, making it more suitable for tasks where the model is applied specifically to Slovenian text, as is ours. Both pretrained models were fine-tuned on the raw SI ParlaMint transcriptions using 30 epochs. The Learning rate of the BERT model was 5e-6, for the SloBERTa was 6e-6. As in the SVM case, we optimized the number of epochs and the learning rate using 5-fold nested cross-validation. For preprocessing, we used the standard respective tokenizers of BERT and SloBERTa. As both models are limited to 512 tokens, we applied an automatic truncation, using the first 512 words in each speech for training and predicting. Both models are limited to a maximum of 512 tokens, and speeches above that limit were truncated.

For **explaining the deep learning models,** we used a technique introduced Lundberg and Lee (2017) called by SHAP. It is a powerful technique that uses classic equations from cooperative game theory to compute explanations of model predictions. Shapley values are feature importance values derived when a model is trained on all feature subset combinations, mathematically calculating the importance of each. Calculating Shapley values in this manner is computationally expensive, so what SHAP manages to do is derive these values by sampling approximations. Reading the SHAP results is very similar to reading the explainability analysis of the Linear SVM: more negative values refer to words leading the model to classify a speech as "leftist" and more positive values refer to the "rightist" words. There are two ways of presenting the token importance using Shapley values: its total Shapley in the entire dataset or its maximum/minimum recorded value. The issue with the first technique is that it priori-

Model	Accuracy
Random Baseline	0.511
Tf-Idf + Lin. SVM (1-grams)	0.866 ± 0.01
Tf-Idf + Lin. SVM (1, 2-grams)	0.903 ± 0.01
Tf-Idf + Lin. SVM (1, 2, 3-grams)	**0.913 ± 0.02**
Multilingual BERT	0.819 ± 0.01
SloBERTa	**0.877 ± 0.03**

Table 2: **Classification scores on predicting political leanings based on speech transcriptions.**

tizes common words such as stopwords and should be generally avoided. We use the second technique, as it leads to more sensible results. A third variation of normalizing the Shapley sum according to the number of occurrences is also an option, yet we leave out this experimentation and plan it as our further work.

5. Results and Discussion

Table 2 shows the results of classifying speeches on the topic of migration as "leftist" or "rightist", depending on the political affiliation of the speaker in the Slovenian Parliament. The classical Tf-idf approach using unigrams, bigrams and trigrams showed the best results, with around 91% accuracy. From the language models, the SloBERTa outperformed BERT yet with only 87% accuracy. One reason behind the worse performance of the language models could be its inability to operate on very long sequences, as more than 60% of the speeches in our dataset contain more than 512 tokens (maximum sequence for BERT).

The models showed that predicting parliamentary political leaning from speech transcriptions is possible. Since statistical models can differentiate "leftist" and "rightist" speeches, we investigated if their explanation through feature importance makes sense from a political linguistics perspective. Figures 2 and 3 show the results of the feature importance analysis. Here, we can notice some obvious similarities and differences between the classical and the language model approach. The first difference is that the SVM feature importance can catch not only tokens (words) but also bigrams and trigrams, which can be useful for a better linguistic context. Since it uses lowercase lemmas, it mitigates duplicates in different word forms, yet it can potentially lead to ambiguities. The most important difference is that the SVM feature importances are the general values for the whole dataset, while the Shapley values are detected maximums for a particular comment, which does not mean that these words have the same negative (leftist) or positive (rightist) effect for all samples.

Both model interpretations show that right-wing parliamentarians' motif is to mention the country name (Slovenia) and its forms. The SVM interpretation also reveals that the rightists emphasize their party names. Regarding migration, their motif is to focus on the illegal aspect of the migration, with the SVM models catching the
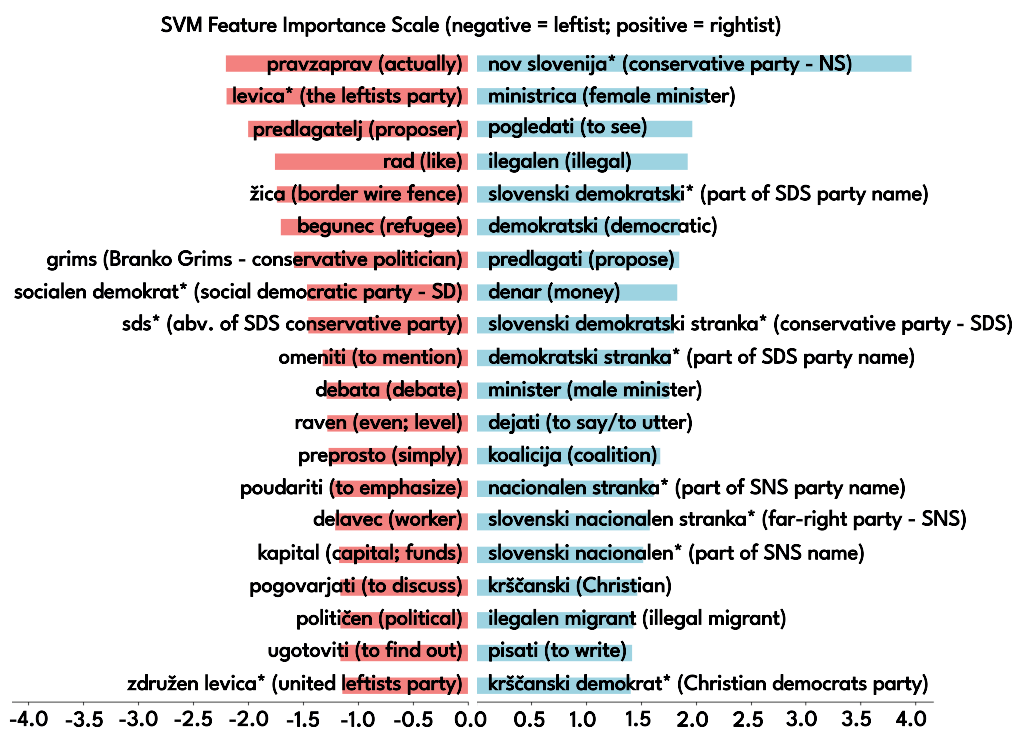
58

SVM Feature Importance Scale (negative = leftist; positive = rightist)

pravzaprav (actually)	nov slovenija* (conservative party - NS)
levica* (the leftists party)	ministrica (female minister)
predlagatelj (proposer)	pogledati (to see)
rad (like)	ilegalen (illegal)
žica (border wire fence)	slovenski demokratski* (part of SDS party name)
begunec (refugee)	demokratski (democratic)
grims (Branko Grims - conservative politician)	predlagati (propose)
socialen demokrat* (social democratic party - SD)	denar (money)
sds* (abv. of SDS conservative party)	slovenski demokratski stranka* (conservative party - SDS)
omeniti (to mention)	demokratski stranka* (part of SDS party name)
debata (debate)	minister (male minister)
raven (even; level)	dejati (to say/to utter)
preprosto (simply)	koalicija (coalition)
poudariti (to emphasize)	nacionalen stranka* (part of SNS party name)
delavec (worker)	slovenski nacionalen stranka* (far-right party - SNS)
kapital (capital; funds)	slovenski nacionalen* (part of SNS name)
pogovarjati (to discuss)	krščanski (Christian)
političen (political)	ilegalen migrant (illegal migrant)
ugotoviti (to find out)	pisati (to write)
združen levica* (united leftists party)	krščanski demokrat* (Christian democrats party)

-4.0 -3.5 -3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0

Figure 2: **SVM feature importance.** The left (red) side contains phrases that have the highest significance for the model predicting a speech as "leftist". The same explanation is for the right (blue) side for the "rightist" speeches. Phrases marked with ⋆ refer to party names.

SloBERTa Shapley Values (negative = leftist; positive = rightist)

-25.40	-------	(presiding) predsedujoči	Novi (New)	-------	+30.00
-25.25	-------	(we work) delamo	Nove (New plu.)	-------	+29.03
-24.27	-------	(Thank you) Hvala	Slovenske (Slovenian)	-------	+27.99
-24.07	-------	(United) Združeni	Slovenska (Slovenian)	-------	+24.93
-21.75	-------	(jokes) vici	Slovenski (Slovenian)	-------	+24.92
-20.89	-------	(part-of-word token) kevi	premalo (too little)	-------	+22.95
-18.71	-------	(peak) vrhunec	Slovenije (Slovenia)	-------	+21.72
-17.89	-------	(chosen f.) izbrala	časi (the times)	-------	+20.09
-17.63	-------	(touched) dotaknil	gospe (ladies)	-------	+18.58
-17.53	-------	(discuss) razpravi	kolega (colleague)	-------	+17.48
-17.41	-------	(bet) stavili	Kr (part-of-word token)	-------	+16.65
-17.15	-------	(you go out) izhodiš	Spoštovani (Dear)	-------	+16.62
-16.85	-------	(task) naloga	naj (most)	-------	+16.05
-16.74	-------	(consistent) skladna	Sloveniji (Slovenija)	-------	+14.93
-16.38	-------	(a word) besedo	poglejte (look)	-------	+14.65
-15.55	-------	(or) ali	v (in)	-------	+14.54
-15.22	-------	(not) ni	žal (unfortunately)	-------	+14.13
-14.97	-------	(jokes) vice	vaše (yours)	-------	+13.99

Figure 3: **Maximum and minimum Shapley values for the SloBERTa model.** Left/right (red/blue) side shows the words that had the highest signal in predicting a "leftist/rightist" speech.

phrase "illegal migrant" as something that almost exclusively right-wing politicians use. The leftists tend to use the word "begunec" (angl. refugee) instead of "migrant", as the meaning behind the former refers to someone who is fleeing. They tend to use words such as "united" and "debate" and they often times refer to their party opponents by the party abbreviations instead of the full name. One interesting case is that the SVM manages to recognize the leftists emphasizing the word "grims". Grims is the surname of the right-wing politician Branko Grims who is strongly against immigrants in Slovenia and Europe and was the most active politician in the parliament on the topic.

Although much more can be drawn from the feature importance, this is where our analysis stops. The model

interpretability through the list of words that had the most meaningful impact on the classification opens up a great possibility for further qualitative work. Researchers studying political linguistics could observe the usages of these lists and find patterns in the connotation of the words used. They could confirm or debunk ideological concepts used by the different sides of the political spectrum. Or, they could analyze if these words and phrases actually contain the primary message of a sentence or if their role is more on the stylistic side of political speeches.

6. Conclusion

Explaining the decision-making of machine learning models (known as XAI) can be a great tool in interdisciplinary research. When the goal is not just to classify, but to understand patterns that appear across classification groups, XAI can complement qualitative research.

In this work, we used XAI to bridge the gap between computational and political linguistics. We developed classical machine learning as well as deep learning language models that can classify parliamentary speeches as "leftist" and "rightist" for the topic of migration. We applied our approach to the Slovenian parliament using the ParlaMint dataset with data from 2014 to 2020. With both approaches showing great predictive success, we applied methods (such as calculating the Shapley values) that can explain the decisions of our models and show which words and phrases differentiate "leftist" from "rightist" speeches and vice versa. While left-leaning parliamentarians use concepts such as "unity" and "debate", the right-leaning parliamentarians put more emphasis on the national symbols (mentioning Slovenia) and their party names.

We leave an opening for further work in multiple directions. One is the improvement of the interpretability methods for models that work with text. Different feature importance techniques could be applied and a comparative study could help us understand the strengths and weaknesses of the methods. Another direction is exploring the ways the model explanations can be used in interdisciplinary research. How to use the words that give the most value to models' predictions in a qualitative continuation of the work. Or, how to understand and communicate model flaws with experts from the social sciences in order to improve them.

References

Alencar, Amanda, 2018. Refugee integration and social media: A local and experiential perspective. *Information, Communication & Society*, 21(11):1588–1603.

Barlai, Melani, Birte Fähnrich, Christina Griessler, and Markus Rhomberg, 2017. *The migrant crisis: European perspectives and national discourses*, volume 13. LIT Verlag Münster.

Brannen, Julia, 2017. Combining qualitative and quantitative approaches: an overview. *Mixing methods: Qualitative and quantitative research*:3–37.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Erjavec, Tomaž, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaur Barkarson, Steinór Steingrímsson, et al., 2022. The parlamint corpora of parliamentary proceedings. *Language resources and evaluation*:1–34.

Eskişar, Gül M Kurtoğlu and Çağrı Çöltekin, 2022. Emotions running high? a synopsis of the state of turkish politics through the parlamint corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*.

Greene, Derek and James P Cross, 2015. Unveiling the political agenda of the european parliament plenary: A topical analysis. In *Proceedings of the ACM web science conference*.

Greussing, Esther and Hajo G Boomgaarden, 2017. Shifting the refugee narrative? an automated frame analysis of europe's 2015 refugee crisis. *Journal of ethnic and migration studies*, 43(11):1749–1774.

Heidenreich, Tobias, Fabienne Lind, Jakob-Moritz Eberl, and Hajo G Boomgaarden, 2019. Media framing dynamics of the 'european refugee crisis': A comparative topic modelling approach. *Journal of Refugee Studies*, 32(Special_Issue_1):i172–i182.

Ilie, Cornelia, 2015. Parliamentary discourse. *The International Encyclopedia of language and social interaction*:1–15.

Interpol, Unicri, 2019. Artificial intelligence and robotics for law enforcement. *Interpol/Unicri, Lyon/Turin*.

Johs, Adam J, Denise E Agosto, and Rosina O Weber, 2022. Explainable artificial intelligence and social science: Further insights for qualitative investigation. *Applied AI Letters*, 3(1):e64.

Jordan, Michael I and Tom M Mitchell, 2015. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

Krotofil, Joanna and Dominika Motak, 2018. Between traditionalism, fundamentalism, and populism: a criti-

cal discourse analysis of the media coverage of the migration crisis in poland. In *Religion in the European refugee crisis*. Springer, pages 61–85.

Liu, Zhijie, Xueqiang Lv, Kun Liu, and Shuicai Shi, 2010. Study on svm compared with the other text classification methods. In *2010 Second international workshop on education technology and computer science*, volume 1. IEEE.

Lundberg, Scott M and Su-In Lee, 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pages 4765–4774.

Miller, Tim, 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Molina-Azorin, José F, 2016. Mixed methods research: An opportunity to improve our studies and our research skills.

Naderi, Nona and Graeme Hirst, 2015. Argumentation mining in parliamentary discourse. In *Principles and practice of multi-agent systems*. Springer, pages 16–25.

Sartor, Giovanni and Francesca Lagioia, 2020. The impact of the general data protection regulation (gdpr) on artificial intelligence. *European Parliamentary Research Service*.

Shinde, Pramila P and Seema Shah, 2018. A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*. IEEE.

Skubic, Jure, Jan Angermeier, Alexandra Bruncrona, Bojan Evkoski, and Larissa Leiminger, 2022. Networks of power: Gender analysis in selected european parliaments. In *2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS)*.

Skubic, Jure and Darja Fišer, 2022. Parliamentary discourse research in sociology: Literature review. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*.

Ulčar, Matej and Marko Robnik-Šikonja, 2021. Sloberta: Slovene monolingual large pretrained masked language model.

van der Brug, Wouter and Eelco Harteveld, 2021. The conditional effects of the refugee crisis on immigration attitudes and nationalism. *European Union Politics*, 22(2):227–247.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wallaschek, Stefan, 2020. Contested solidarity in the euro crisis and europe's migration crisis: A discourse network analysis. *Journal of European Public Policy*, 27(7):1034–1053.