THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Text Representations and Explainability for Political Science Applications

Denitsa Saynova

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2023

Text Representations and Explainability for Political Science Applications

Denitsa Saynova

Department of Computer Science and Engineering
Division of Data Science and AI
Chalmers University of Technology | University of Gothenburg
SE-412 96 Gothenburg, Sweden
Phone: +46 (0)31 772 1000

Printed by Chalmers Digitaltryck
Gothenburg, Sweden, 2023

To my family and friends

Text Representations and Explainability for Political Science Applications

Denitsa Saynova
Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

This work explores the utility of natural language processing approaches for the study of political behavior by examining two main aspects – representation and explainability. We investigate how current representation approaches capture politically relevant signals in a proportional representation system. In particular we test static word embeddings trained by transfer learning. We find that some signals in the embedding spaces can be validated from domain knowledge, however, there are multiple factors affecting the performance and stability of the results, such as pre-training and frequency of terms.

Due to the complexity of current NLP techniques interactions between the model and the political scientist are limited, which can impact the utility of such modeling. Therefore, we turn to explainability and develop a novel approach for explaining a text classifier. Our method extracts relevant features for a whole prediction class and can sort those by their relevance to the political domain.

Generally, we find current NLP methods are capable of capturing some politically relevant signals from text, but more work is needed to align the two fields. We conclude that the next step in this work should focus on investigating frameworks such as hybrid models and causality, which can improve both the representation capabilities and the interaction between model and social scientist.

Keywords: NLP, Political Science, Representation, Explainability

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] Annika Fredén, Moa Johansson, **Denitsa Saynova**, *Word embeddings on ideology and issues from Swedish parliamentarians' motions over time: A comparative approach.*
Submitted, under review.

[**Paper II**] **Denitsa Saynova**, Bastiaan Bruinsma, Moa Johansson, Richard Johansson, *Class Explanations: the Role of Domain-Specific Content and Stop Words. Proceedings of the 24rd Nordic Conference on Computational Linguistics (May 2023), 103–112.*

Other publications

Other publications by the author, not included in this thesis, are:

[**Paper a**] Brian Bonafilia, Bastiaan Bruinsma, **Denitsa Saynova**, Moa Johansson, *Sudden Semantic Shifts in Swedish NATO discourse. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), 184–193*, Toronto, Canada, July 2023.

Acknowledgments

I would like to thank my main supervisor Moa Johansson for her support and guidance. I would also like to thank my co-supervisor Richard Johansson for his help and feedback. I'm thankful to my project PI Annika Fredén, and my colleagues Bastiaan and Pasko for their endless enthusiasm for this work and fruitful discussions. Additionally, many thanks to my examiner Devdatt Dubhashi for his feedback and invaluable insight.

I want to thank my PhD colleagues at Chalmers and especially my office mates: Simon, Hanna and Deepthi for creating an engaging and positive working environment. I want to thank Lovisa for our discussions and collaboration. My thanks to the faculty and administrative staff of the DSAI division for their support and guidance and a special thank you to everyone in the Formal Methods unit for welcoming me and making the start of my PhD enjoyable and engaging.

Thank you to all my PhD colleagues from WASP HS, who have given me support and understanding and invaluable perspective on research.

Finally, I want to thank my friends - Dimitriya, Ivelina, David, and Igor and my family - my mother Vanya, my father Lyudmil, my sister Albena, and my husband Blagovest for their endless support and care.

This work is supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

Contents

Part I

Summary

CHAPTER 1

Introduction

The ubiquity of text data as well as advancements in natural language processing (NLP) and artificial intelligence (AI) methods for pattern discovery have sparked an interest in using that information for studying social and political behaviors. Through social media and large digitized libraries of parliamentary proceedings or newspaper articles, we can access information from millions of people or spanning several decades. The amount of data makes it unfeasible for the political scientist to manually process and apply traditional research methodologies, which brings the necessity to utilize automated modeling and NLP techniques.

There has been a growing body of work in recent years showing the advantages of utilizing NLP techniques for capturing, measuring and exemplifying political views and ideological positions from politically relevant text. This spans work from measuring polarization (Belcastro et al., 2020; Goet, 2019), to bias detection (Skubic et al., 2022), to tracking changes through time (Rodman, 2020).

There is, however, some misalignment in the goals and approach in the fields of NLP and in social science. NLP developments tend to be evaluated by collections of benchmark datasets that aim to assess a model's performance

along different linguistic tasks (Rajpurkar et al., 2016; Wang et al., 2018). That does not necessarily correspond to how such a model can be validated from a social science perspective, where there is an emphasis on the correlation between model results and predictions and real world events – for example, whether a model manages to capture a well-known temporal shift or event. Additionally, the approach to data collection may differ in their goal and focus. Machine learning tends to aim for vast sample sizes, commonly coming from multiple, diverse sources (e.g. books, Wikipedia articles, code, social media and more in the same training set). Social sciences generally are more concerned with the sampling of the data - its scope, the importance of a representative sample, the population that is studied, the possible biases, etc. Therefore it is crucial to evaluate how these modeling choices affect the suitability of NLP methods for studying political behavior and their strengths and weaknesses when applied to a social context. In this work, we focus on evaluating two aspects of the NLP pipeline for the discovery and measurement of political positioning from text – representation and supervised modeling.

Representation in NLP (or the mapping from text to numerical values) relies on the idea of *distributional semantics* – one can use word co-occurrences to represent meaning. This results in text units (typically words) being represented as high-dimensional vectors in an embedding space. These embedding spaces have been shown to capture some word alignments such as gender, capital-country, and other relations (Mikolov, Sutskever, et al., 2013). Representation is also tightly connected to the concept of *transfer learning* – in order to leverage information from large textual corpora, one can pre-train models on "general" language and use those as starting point representations for adaptation to a down-stream task. In Paper I we investigate what politically relevant patterns can be captured in these embedding spaces. We train Word2Vec (Mikolov, Chen, et al., 2013) models on data from the Swedish parliament and compare the learned associations for a number of salient terms between the two main left (Social Democrats) and right (Moderates) parties. We find that some expected political signals are captured in the resulting embedding space, such as the left's association of "solidarity" with "welfare" and the right's with "security" and "stability". However, the lack of an exhaustive list of associations that should hold between words for each party makes the systematic investigation and evaluation of those representations difficult. We additionally find that there are multiple factors affecting the stability of the

results, including pre-training and word frequency.

Numerous advanced NLP architectures have shown great promise in learning complex tasks (Devlin et al., 2019; OpenAI, 2022; Touvron et al., 2023), which makes them a compelling possibility for modeling political behavior from data. This comes with its own set of difficulties however. First, current NLP methods tend to be based on complex, multilayered architectures, which make the task of interpreting the learned patterns very difficult (commonly referred to as the "black box problem"). Second, there is no guarantee that patterns that the model captures align with human intuition and domain knowledge, which can have many disadvantages. To this end, in Paper II, we investigate methods within explainable AI (XAI) and propose a novel approach for obtaining class-level explanations for a classifier, allowing a distinction between domain-specific content words and stop words in order to facilitate human-model interaction. A small scale examination suggests that when the model is more aligned with domain-specific words, it is also more accurate.

CHAPTER 2

Background

One way in which we can think about a wide range of natural language processing techniques is to divide the process into two components – representation of the input data (text) in a numerical format and learning a mapping from that representation to a particular output space (e.g. sentiment, categorization, etc.). In essence this results in a two-step mapping process – the first mapping text to a numerical vector (typically called *embedding*), the second mapping the numerical vector to an output variable. The way in which those mappings can be constructed can differ, but most commonly current NLP approaches rely on training neural networks. Therefore, in this chapter, we begin by discussing the theoretical framework behind current approaches for NLP representation in Section 2.1 and provide an overview of those in Section 2.2. We then turn our focus in Section 2.3 on the consequences of using neural networks as the predominant modeling approach in NLP by discussing connectionism and how that leads to the black-box problem. We conclude with an examination of the state and significance of human-model interaction and explainability in AI.

2.1 The Distributional Hypothesis

The Distributional Hypothesis, which is a cornerstone for most approaches to representation nowadays, can be summarized with Firth's well-known and highly-cited quote – *"you shall know a word by the company it keeps"* (Firth, 1957). The hypothesis is commonly attributed to both Harris (Harris, 1954) and Firth (Firth, 1957) and in essence states that we can represent words through the context in which they appear. This is practice means representing language through its statistical distribution – e.g. through collocations of words in a corpus. It is often claimed in NLP literature that what is captured with this approach is *meaning* (Bender & Koller, 2020), however it is worth discussing two caveats that relate to this issue.

Form and meaning. Coming from a structuralist view Harris proposes language can be *"described in terms of a distributional structure* [...] *and* [...] *this description is complete without intrusion of other features such as history or meaning"* – that is meaning is something that is derived from human experience and the structure of a language deviates in many respects from that external structure of meaning. He poses however that those two aspects – the external notion of meaning and the distributional structure of a language – are highly interconnected and one important aspect is that of difference. Stating that *"difference of meaning correlates with difference of distribution"* he provides a very powerful connection between meaning and form.

Context. Firth, on the other hand, comes from a more anthropological view of language and discusses at length what needs to be considered when we talk about "company" or – how do we define the relevant context from which we extract the distributional description. Whereas NLP approaches in this area tend to use the collocations of words within a specified window length in a corpus of text, Firth invokes the *context of situation* as a necessary parameter that needs to be accounted for. Firth provides a list of the three categories that comprise this idea of context: relevant features of participants, relevant objects, effect of the verbal action (Firth, 1957). He does not provide guidance for how to account for these in practice or how to determine relevance, but his ideas lay the groundwork for developments in the study of language in its social context.

Extensions to collocation

Recently there have been efforts to extend the purely textual co-occurrence-based notion of representation in NLP and introduce a broader view of the social and physical context of language (Bisk et al., 2020). We can broadly see two strategies that invoke different aspects of Firth's and Harris' work – *syntagmatic extension* and *comparative stratification* (Brunila & LaViolette, 2022).

Syntagmatic extension refers to extending representation beyond the simple textual collocation view by incorporating information from other contextual sources. In practice this can be anything from including representation of different abstraction levels in the text (for example adding document-level representation as in Le and Mikolov, 2014) to meta-data information regarding demographics of the speakers (Garimella et al., 2017) to including different modalities of data (Baroni, 2016).

Comparative stratification is based on what Harris calls "sublanguages" and Firth calls "restricted languages" – broadly speaking, the idea that we can split up language use according to the setting within which it is used and observe different distributional characteristics which are representative of the meanings within that setting. An example of how this can be applied in practice is studying diachronic semantic shifts – that is, splitting up the corpus in time periods and extracting separate representations for each period. These can then be compared to track shifts through time. The corpora can however be stratified along any dimension of interest, which, as we discuss later, can be a powerful tool for studying political behavior, where we can summarize a lot of domain knowledge into additional metadata along which we can produce meaningful splits of the data.

Implementation of the Distributional Hypothesis As discussed so far, the distributional hypothesis states that we can capture a representation that is linked to meaning by examining the context of a word. In practice this is done by translating text to a high-dimensional numeric space (called an embedding space), where a word is represented by its coordinates (a vector) in that space. The way we construct those spaces is based on the idea of collocations – words that appear in the same context (taken here to mean

similar words in their vicinity) are encoded closer to each other. We discuss the practical implementations of this idea in the next chapter.

2.2 Representation

There are several factors motivating the transformation of text to numbers in NLP applications. The first one is practical - machine learning methods are mathematical functions and as such require inputs to also have a mathematical form. Second, they can be used to impose a helpful structure on the input that makes the modeling step easier. There are many latent structures in a language that we may want to keep in its representation. Generally, we wish to represent texts that are similar in meaning or have some other semantic relationship as more similar embedding vectors.

There have been multiple approaches and paradigm shifts in text representation. The relevant work within NLP can be broadly split into *corpus statistics methods* and *neural network-based approaches*. Neural network approaches can themselves be split into static and contextual methods (see Patil et al., 2023 for an extensive survey).

Corpus statistics methods are a group of methods for representing texts as fixed-length vectors. This allows the use of linear algebra for calculating distances and therefore similarities of texts. The classical approach here is to represent the corpus as some form of document-term matrix (see Table 2.1), where the rows correspond to documents in the data, the columns represent the terms in the vocabulary and the values can be binary (signifying presence/absence), counts, frequencies or other weighted values (e.g. TF-IDF). As such, this representation does not keep a reference to the word order and is commonly referred to as a Bag of Words (BoW) approach.

These methods are interpretable in the sense that we can identify what each dimension of the document or term vector represents. However, the sparsity of these representations (most values of the matrix are zero) means the approach is not scalable to large datasets and training models is more time-consuming. As mentioned, these document representations also do not capture word order or synonyms.

To overcome these issues dimensionality reductions techniques have been proposed to transform these representations to a low-dimensional space. An

	I	cats	football	...	like	playing	with
doc 1	1	0	1	...	1	1	0
doc 2	1	1	0	...	1	1	1
doc 3	0	1	1	...	0	1	0

Table 2.1: Document term representation example for three documents
Doc 1: I like playing football
Doc 2: I like playing with cats
Doc 3: Cats playing football

established method is latent semantic analysis (LSA) (Deerwester et al., 1990) that uses singular value decomposition (SVD) to cluster together words that appear in similar contexts into "topic" dimensions and represents text in terms of those latent dimensions rather than counts for each word in the vocabulary. It is based on the idea of distributional semantics that words that appear in the same context share some similarity and can therefore be clustered together. This allows to deal with both synonyms and to an extent – polysemy. It does still carry some of the weaknesses of the BoW approach on which it is based and is harder to interpret.

Neural network vector representations are the predominant approach in recent years. Similarly to LSA discussed above, this family of methods is based on the common idea of capturing word meaning through collocation. These embeddings are typically trained through language modeling (LM) tasks or translation tasks. Language modeling is the task of learning the joint probability of strings of text. This is usually done by masking out some part of the string and teaching the model to reconstruct that.

Word2Vec (Mikolov, Chen, et al., 2013) is one of the first proposed models in this family. It is a neural network with a single hidden layer. It can be trained to predict either context (surrounding words) from a single word (skip-gram negative sampling – SGNS) or, alternatively, to predict a word given the surrounding context (continuous bag of words – CBOW). For both of these, we need to define how big the context is (i.e. how many words before and after we need to consider). A larger context makes the training computationally more expensive and a typical value is around 5. For a corpus with vocabulary size n and embedding vector size m the data is processed in the following steps for the CBOW setting: The input context is represented as the summed

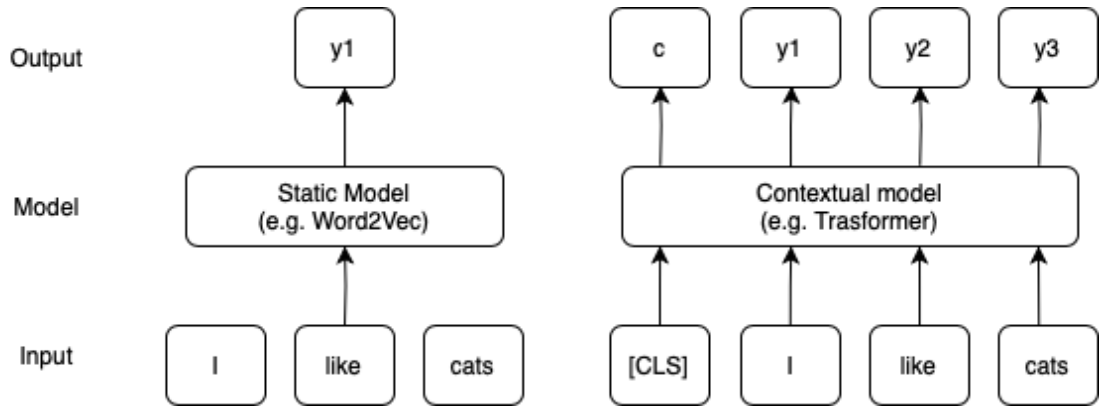

Figure 2.1: Conceptual representation of the input-output relationship in static and contextual encoding models. Contextual embeddings can also produce embeddings at different levels at the same time, represented here by the embedding for the [CLS] token, which contains a representation for the full text.

one-hot encoded individual words – that is – a vector x of size $n \times 1$ with ones at the indexes of the words in the context and zeroes everywhere else. We then have the following calculation through the two layers of the network (represented by the $n \times m$ embeddings matrix W and the $m \times n$ output layer matrix Z):

$$\begin{aligned} x^T W &= h \\ hZ &= a \\ y &= softmax(a^T) \end{aligned} \tag{2.1}$$

We pass the result through a softmax function to obtain a probability for each index in the vocabulary in the $n \times 1$ vector y. The network is trained with the logarithmic loss to maximize the probability for the true label word. For the SGNS setting the flow is similar but the input is instead the single middle word and the output is the context words.

Once the network has been trained, we can use the internal weights of the hidden layer as vector representations for the terms in the vocabulary. That is, the ith row of the W matrix is the $1 \times m$ embedding for the ith word in the vocabulary. This produces *static* representations, that is – a one-to-one mapping from a word token to a numerical vector. To represent text in this setting we can then use an aggregation of those word embeddings.

Finally, the attention mechanism (Bahdanau et al., 2015) and the transformer architecture (Vaswani et al., 2017) have introduced *contextual embeddings*. These are still vector representations learned by a neural network, however, in contrast to static methods, which produce a single embedding for a token, contextual embeddings provide an embedding for a token depending on the full text sequence in which it appears (see Figure 2.1). This allows for highly flexible representations and even more attentiveness to the context.

Transfer learning is one of the main applications of representation learning. The central idea is to condense "knowledge" from one task and use it to solve another in order to reduce the data size requirements for learning the second task. In the context of NLP we can use general text (that tends to be easier to access in big quantities) in order to extract general linguistic patterns and features which can then be used as a representation for learning the mapping for a second task with a much smaller dataset. This is a two-step process consisting of *pre-training* and *adaptation*. In the first step a model is trained on large amounts of data to learn a general representation of the inputs typically through language modeling or machine translation tasks. During the adaptation phase the learned patterns are used to train a model for a new task – e.g. classification, named entity recognition, entailment, etc.

Adaptation can be done in two ways – *feature extraction* or *fine-tuning*. As exemplified in Figure 2.2 in feature extraction the pre-trained model is used by freezing the weights, passing the new data through the network and using the final layer's output as feature vectors that are then fed into a down-stream model trained on the new task and data. In fine-tuning, the weights of the pre-trained model are further changed by additionally training the full network on the new data, however, some layers may still be frozen or frozen during only the initial stages of the fine-tuning process. The choice of if and when to freeze the pre-trained model's weights is dependent on the type of representations that are learned by the layers in the network and the tasks for the model. Empirical results (Peters et al., 2019) show that when the pre-training and adaptation tasks are similar, fine-tuning performs better, whereas, when the tasks are substantially different, feature extraction might be more effective.

Comparative stratification in practice As discussed in Section 2.1 one way to go beyond simple text collocation representations is to stratify the data

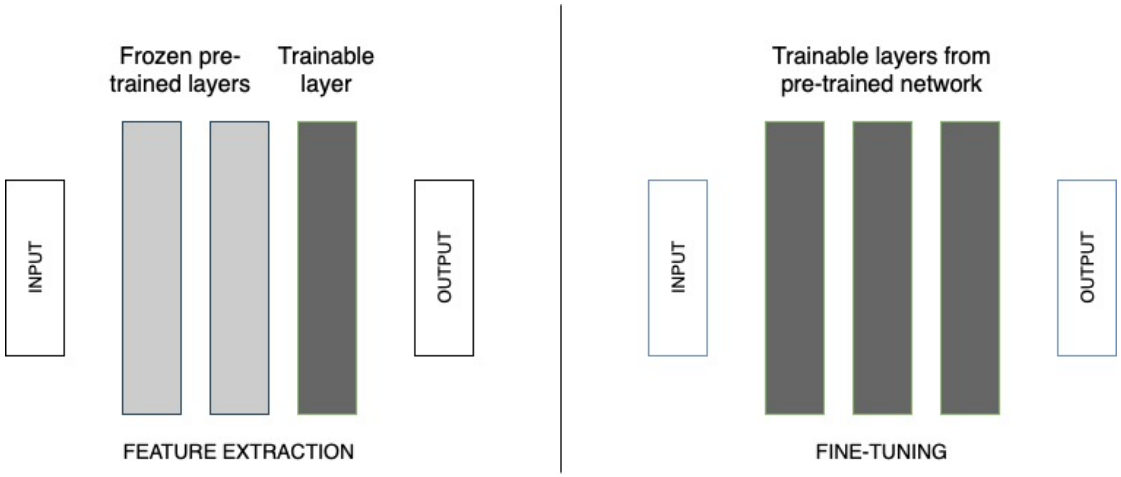

Figure 2.2: Representation of the two approaches to transfer learning.

along a salient dimension. This allows to both improve embeddings for downstream tasks and to study the semantic shifts along that dimension. In political and social science this allows us to account for a multitude of relevant contexts. One commonly studied dimensions is time, with multiple examples of tracking diachronic semantic shifts (Hamilton et al., 2016; Rodman, 2020; Tahmasebi, 2018), but work in this area also spans political leanings (Goet, 2019; Spinde et al., 2021), demographics (Hovy, 2015) and social network groups (Yang & Eisenstein, 2017) among others.

When considering NLP for political science application, this wide use of embedding representations makes it crucial to examine their capability to capture politically relevant information as well as to investigate how different design choices affect the quality of the representations.

2.3 Connectionism and the black box problem

The predominant paradigm in AI and NLP is connectionism or the idea that complex cognitive (or any information processing) systems can be described by connected layers of simple processing units exemplified in practice by the (artificial) neural network. Even though there is fruitful research in trying to combine these with other ideas such as symbolic AI, the connectionism view has been dominating both research and applications, which can be seen in the

rapid development of deep neural networks.

Neural networks are universal function approximators (Hornik et al., 1989), meaning they can be used to represent any function and empirical results show they can be trained to solve complex natural language tasks (Cohen et al., 2022; Devlin et al., 2019; OpenAI, 2022). However, this has pushed research somewhat into optimizing for performance at the expense of other requirements for NLP, for example, cost, robustness, statistical power, and handling of social bias (Bowman & Dahl, 2021; Ethayarajh & Jurafsky, 2020). One important issue stemming from the ever-growing complexity is *explainability* often referred to as the "black box problem" – that is, not being able to understand the internal process by which a model makes its prediction and instead having access only to a mapping from inputs to outputs.

There is no consensus on a single concise definition of explainability within AI, however, it generally refers to a human's ability to understand the model's decision process. This can range from understanding the input features' contributions, to counterfactual understanding of the minimal changes required to change the model's prediction, to localizing where in the model layers a particular calculation/processing occurs, to many other aspects of understanding. For an exhaustive review of explainability aspects refer to Nauta et al., 2022, for a review of current methods refer to Danilevsky et al., 2020; Madsen et al., 2022. Our inability to understand how models make their decisions can lead to errors (Alcorn et al., 2019; Finlayson et al., 2019; Obermeyer et al., 2019; Su et al., 2019), unintended biases (Buolamwini & Gebru, 2018; Hovy & Prabhumoye, 2021) and ultimately mistrust in the system. This can be a barrier to adopting high-performing black-box prediction models for studying political behaviors, since it can be an obstacle to validating the model.

Local and Global explanations

There are many facets to defining what an explanation is. One important dimension along which we can place the methods in this area is local vs global. Local methods are concerned with explaining single instances of model prediction. That is, the question they aim to answer is of the form *Why is this text predicted as being class X?*. Global methods aim at explaining some general behavior of the model – for example – answering questions such as *What does the model see as class X's important features?*. In XAI local methods tend to be more researched and developed (Nauta et al., 2022). Local methods are

predominantly based on creating salience maps – a correspondence between input feature and its contribution towards the model prediction. Among the most widely used approaches in this category are LIME (Ribeiro et al., 2016), Shapely values (Shapley, 1952), and gradient-based approaches (Baehrens et al., 2010; Sundararajan et al., 2017). Out of these, Shapely values-based approaches, even though more accurate representations of the black box model (Ethayarajh & Jurafsky, 2021), tend to be computationally expensive (for example the attention flow implementation for transformer architectures (Abnar & Zuidema, 2020)). Gradient-based approaches and LIME are easier to compute, with LIME having the advantage of being model-agnostic making it more flexible and versatile.

LIME is a widely used post-hoc, model-agnostic local explainability method. This makes it versatile and relatively easy to use. Additionally it is computationally more efficient compared to other methods. The model approximates a black-box model locally in several steps: First, for a particular input instance, we sample "around" the data point by augmenting it. The way this augmentation is done depends on the type of data – e.g. for continuous data we can add small random values along each dimension. We discuss implementations for text further down. Second, the new data points are passed through the black-box model for inference to get their respective labels, essentially creating a new dataset locally around the input. Third, an explainable model such as regularized logistic regression or a decision tree is fit to this data (typically weighted by the distance between the new datapoints and the original input). Finally, we can use the explanations from this new model (e.g. the weights of the regression model) as attribution scores of the features for the original input instance that we aim to explain.

There are a number of design choices for this approach. A default implementation we follow later in this work [1] has the following hyper-parameters: The interpretable model that is fit to approximate the black-box model is a ridge regression. This is fit on 5000 sampled data points. New data points are sampled by removing a random number of words (between 1 and length of example) from the example 5000 times. We follow the default BoW approach where a word is removed with all its occurrences (if multiple exist in the example). An alternative is to consider position as well, allowing for a word

[1] https://lime-ml.readthedocs.io/en/latest/

to have different contributions scores at different positions in the sentence. This may have benefits in terms of model-alignment (Section 2.3) when the black-box model has position encoding, but poses more difficult choices if we want to aggregate the individual results.

SP-LIME is an approach proposed in the original LIME paper for aggregating instance level LIME explanations in a way which can present a more general view of the patterns learned by the model, thus providing a global explanation. The method provides *"a set of representative* [text] *instances"* (Ribeiro et al., 2016) as explanations of model behavior. This is done by searching for representative instances in a set of datapoints in the following way: We obtain LIME explanations for each instance in the set. We then calculate the score for feature j as $I_j = \sqrt{\sum_{i=1}^{N} W_{ij}}$ where N is the number of explained data points and W is the explanation matrix containing the local importance of the features. We also define a coverage function for a set of examples V as:

$$c(V, W, I) = \sum_{j=1}^{d'} \mathbb{1}_{[\exists i \in V : W_{ij} > 0]} I_j \tag{2.2}$$

The intuition behind this definition is to capture the most diverse and at the same time highest scoring (in terms of their explanation features) text examples. That is, if we have two texts with very similar features as explanations, we might not gain more information from seeing both.

Finally, we also define a budget B of how many texts we want to present as explanation of the model. And based on this, the task translates into optimizing the following function:

$$Pick(W, I) = \underset{V, |V| \leq B}{\operatorname{argmax}}\ c(V, W, I) \tag{2.3}$$

Since this is NP-hard, the problem is solved by a greedy search algorithm that adds the instance i that results in the highest marginal coverage gain, defined as $c(V \cup \{i\}, W, I) - c(V, W, I)$. As we show later in our work, this can often lead to high-scoring examples containing very frequent words (i.e. stop words) as explanations, due to their high coverage.

Model and human perspective in XAI

Another important aspect of explainability that needs to be considered is the distinction between model and human explanations. Both in terms of how to summarize the model's behavior in a way a human can track and in terms of how a model and a human will process information and make inferences.

Technically speaking, neural networks are explainable in the sense that they are deterministic at inference time, so we can track through the calculations and see the causal connections between the activation of all neurons in all layers. The reason we need explainability methods is that these sequences are prohibitively large and complex for a human to process. Therefore, explainability is some abstraction from the original model. This gives rise to an obvious trade-off – explanations true to the models vs explanations that are understandable by a human. These aspects are commonly referred to as *functionally-grounded* and *human-grounded* explanations (Doshi-Velez & Kim, 2017; Madsen et al., 2022).

What we need to further consider is the different capabilities in pattern recognition. A machine learning model has advanced statistical capabilities, allowing it to detect weak signals from large amounts of data that would be beyond a human's computational capabilities. However, they are also restricted by the data they have been trained on, by definition having no access to external information, whereas, a human brings their extensive world knowledge to any task they solve, which implies the reliance on and use of a much richer information and representation of the world. In other words, models are better at detecting small distributional differences, whereas humans have domain and other types of external knowledge, which may lead them to "focus" on different pieces of information when making inferences. For example, a person could classify a text as right-leaning because they identify the message as calling for the lowering of taxes, while a model attaches importance to words such as "the", "and", etc., since those could have slightly different distributions between parties. While both can be valid patterns from a model perspective (and lead to better predictive accuracy) only the first type of features can be aligned with political background knowledge.

CHAPTER 3

Summary of Included Papers

3.1 Paper I

In Paper I we investigate the utility of word embedding methods for capturing politically relevant signals in the Swedish parliament. Being a proportional representation system, the Swedish parliament is characterized by the parties' need to collaborate in coalitions to reach a majority (Bäck & Bergman, 2016), which may lead to less polarized views and thus smaller differences in the text distributions. Therefore patterns that can indicate agreement and disagreement of meaning and word use are of particular interest. We base our approach on previous works that successfully use embeddings as a tool to track semantic shifts through time (Rodman, 2020) and measure similarity of political parties (Goet, 2019). We additionally examine the effects of several design choices on the results and the stability of the embeddings.

Methodology. We base our work on a corpus of Swedish parliamentary motions that contain early signals of political direction from individual party members and are therefore of particular interest. To simplify the task of detecting differences in language use we focus on the two main parties repre-

senting the right-left political spectrum – Moderates and Social Democrats. We additionally explore the temporal shifts in language use by stratifying the data into two time spans - 1988-2009 and 2010-2020, which mark the periods before and after the entrance of the radical right-wing Sweden Democrats into parliament. We examine 10 terms covering topic indicated as important for voters (Fredén & Sikström, 2021) and compare how those are embedded differently between parties and between time periods.

Based on previous work aimed at tracking overall word use change, we use static embeddings by training Word2Vec models. Due to limited data, we employ a transfer learning approach, leveraging "general language" learned as a starting point for adapting to the different strata of data. Since our task of learning embeddings is the same as the pre-training task, during adaptation we opt for fine-tuning rather than feature extraction. We compare two pre-trained models – an external model trained on general Swedish text available from the Nordic Language Processing Library (NLPL) word embedding repository[1] and a model pre-trained on other Swedish parliamentary data.

Fine-tuning is done from both of these pre-trained models on each strata (party and time period combination). We then extract the 20 closest words in the embedding space to the term of interest based on cosine similarity. To estimate the stability of the resulting embeddings we perform a bootstrapping of the data by training 10 versions of the model for each strata and calculating the mean and standard deviation distances between vectors.

We evaluate the results by manual investigation of the top 20 closest words to the terms of interests. We additionally investigate the stability of results by looking at the number of words that are more than one standard deviation above the score of the twentieth word in the list (indicating that those appear in the list more reliably and are less likely to appear due to the stochastic nature of the model or small variations in the data).

Results. From manual investigation of the results, we find that some word associations correlate with expected party views. For example, in the crime dimension we see an association with tax crimes for the Social Democrats and with gang crime and assault for the Moderates. Additionally, in the solidarity dimension we observe association with peace and welfare for the Social Democrats and security and stability for the Moderates. These associations

[1] http://vectors.nlpl.eu/repository/20/69.zip

correspond well to the general tendencies of left-wing parties to center on social and economic welfare, whereas the right-wing focus more on security. When looking at the stability of the embeddings, we see the most salient difference is between terms (rather than between pre-training approaches or parties). Similarly to previous works (Borah et al., 2021; Wendlandt et al., 2018), we find a roughly logarithmic relationship between stability and frequency. However due to the large variation in this relationship, we suggest there might be other factors contributing to this effect more connected to the types of words – for example value laden versus policy terms.

Contributions. Denitsa Saynova contributed to the design of the study, training the models, and aggregating and summarizing the results as well as the writing of the paper. Annika Fredén and Moa Johansson contributed to the design of the study, interpretation of the results, writing of the paper and supervision of the project.

3.2 Paper II

In Paper II we explore the state of current XAI methods and their utility for the social sciences. We identify the need for developing suitable class-level explanations and propose a novel approach that provides ranked feature lists for a binary text classifier that separate domain specific content words from stop words.

Methodology. We propose a four step algorithm for producing class-level explanations: First, we run an instance explainability method on a selected set of datapoints (in our application we use LIME). Second, we aggregate the instance-level explanation features into their respective lists for the two classes. Third, we propose two scoring approaches – one based on frequency normalization and one on principal component analysis of embeddings – to rank the feature lists along a dimension from domain specific to stop words. Finally, we propose the use of "keywords in context" (KWIC) to exemplify the texts in which those features appear and allow the examination of the validity of those patterns. We test this approach on a black box model (in our application a BERT classifier) trained to predict party from text. The corpus we use for the case-study contains debates transcripts from the Swedish

Riksdag.

Results. Both our scoring functions result in domain specific words at the top of the lists and stop words at the bottom, with the normalization approach resulting in mainly function words at the bottom of the lists. We further see that the top words refer to taxes and employment which reflects the studied texts and the left/right dimension in Sweden. Through deeper analysis with KWIC of the term "labor market policy" (identified as important for Social Democrats) we show how these features can be validated with domain knowledge. Finally, based on a small sample of datapoints, we find that the model performs better for texts that have predominantly domain specific content word features as explanations.

Contributions. Denitsa Saynova contributed to the design of the study and the proposed novel XAI method, implementation of the methods as well as the writing of the paper. Bastiaan Bruinsma contributed to the design of the study and XAI method, providing the political science context and framing of the studied materials and wring of the paper, Moa Johansson and Richard Johansson contributed to the design of the study and XAI method as well as writing of the paper and provided supervision for the project.

CHAPTER 4

Concluding Remarks and Future Work

In this work we investigate how current NLP approaches can be used for the study of political behavior.

In **Paper I** we investigate the utility of representation learning for capturing alignment and disagreement between parties and across time. We find that several choices in the model design have an effect on the types of patterns we can discover. We additionally comment on the stability of the results.

In **Paper II** we focus on the human-model interaction aspect and develop a novel XAI approach for class-level explanations for a text classifier. This four step method is post-hoc, model agnostic and adaptable and allows us to identify features the model associates with a particular class as well as sort them by relevance to a political scientist (i.e. from politically-charged, domain words to stop words).

Future Work

There are several weaknesses to the purely connectionist and collocation-based approach to text modeling that we have discussed. To address these, in future work we will investigate how bringing in other perspectives and frameworks

can help improve both the performance and the understanding of the model. In particular, we will focus on two aspects: First, we wish to bring our explainability work into the context of causality. This can be beneficial both as it is a more natural framework for human understanding and because it can align our methodologies with the model's inner workings. Second, we wish to extend the representation framework we consider and explore other ways of representing text at different levels - for example by looking at hybrid models that combine the distributional semantics view with more structured representation of knowledge (e.g. databases or knowledge graphs).

References

Abnar, S., & Zuidema, W. (2020). Quantifying Attention Flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4190–4197). ACL.

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., & Nguyen, A. (2019). Strike (With) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4840–4849.

Bäck, H., & Bergman, T. (2016). The Parties in Government Formation. In J. Pierre (Ed.), *The Oxford Handbook of Swedish Politics* (pp. 206–226). Oxford University Press.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K.-R. (2010). How to Explain Individual Classification Decisions. *J. Mach. Learn. Res.*, *11*, 1803–1831.

Bahdanau, D., Cho, K., & Bengio, Y. "Neural machine translation by jointly learning to align and translate" [3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015]. English (US). In: 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015. 2015, January.

Baroni, M. (2016). Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, *10*(1), 3–13.

Belcastro, L., Cantini, R., Marozzo, F., Talia, D., & Trunfio, P. (2020). Learning Political Polarization on Social Media Using Neural Networks. *IEEE Access*, *8*, 47177–47187.

Bender, E. M., & Koller, A. (2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). "Experience Grounds Language". *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8718–8735.

Borah, A., Barman, M. P., & Awekar, A. (2021). Are Word Embedding Methods Stable and Should We Care About It? *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, 45–55.

Bowman, S. R., & Dahl, G. (2021). "What Will it Take to Fix Benchmarking in Natural Language Understanding?" *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4843–4855.

Brunila, M., & LaViolette, J. (2022). "What company do words keep? Revisiting the distributional semantics of J.R. Firth & Zellig Harris". *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4403–4417.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st conference on fairness, accountability and transparency* (pp. 77–91). PMLR.

Cohen, A. D., Roberts, A., Molina, A., Butryna, A., Jin, A., Kulshreshtha, A., Hutchinson, B., Zevenbergen, B., Aguera-Arcas, B. H., Chang, C.-c., Cui, C., Du, C., Adiwardana, D. D. F., Chen, D., Lepikhin, D. (, Chi, E. H., Hoffman-John, E., Cheng, H.-T., Lee, H., ... Chen, Z. (2022). LaMDA: Language Models for Dialog Applications. In *Arxiv*.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A Survey of the State of Explainable AI for Natural Language Processing. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 447–459.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *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)*, 4171–4186.

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning.

Ethayarajh, K., & Jurafsky, D. (2020). "Utility is in the Eye of the User: A Critique of NLP Leaderboards". *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4846–4853.

Ethayarajh, K., & Jurafsky, D. (2021). Attention Flows are Shapley Value Explanations. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 49–54.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, *363*(6433), 1287–1289.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 10–32.

Fredén, A., & Sikström, S. (2021). Voters' Sympathies and Antipathies Studied by Quantitative Text Analysis: Evidence from a two-wave panel experiment in Sweden during covid-19. *Annual Midwest Political Science Association Conference.*

Garimella, A., Banea, C., & Mihalcea, R. (2017). "Demographic-aware word associations". *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2285–2295.

Goet, N. D. (2019). Measuring Polarization with Text Analysis: Evidence from the UK House of Commons, 1811–2015. *Political Analysis*, *27*(4), 518–539.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of*

the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1489–1501.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359–366.

Hovy, D. (2015). "Demographic Factors Improve Classification Performance". *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 752–762.

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8), e12432.

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, II–1188–II–1196.

Madsen, A., Reddy, S., & Chandar, S. (2022). Post-Hoc Interpretability for Neural NLP: A Survey. *ACM Computing Surveys*, *55*(8), 1–42.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013*.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2022). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *CoRR*, *abs/2201.08164*.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

OpenAI. (2022). Introducing ChatGPT. Retrieved August 18, 2023, from https://openai.com/blog/chatgpt

Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A Survey of Text Representation and Embedding Techniques in NLP. *IEEE Access*, *11*, 36120–36146.

Peters, M. E., Ruder, S., & Smith, N. A. (2019). "To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks". *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 7–14.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). "SQuAD: 100,000+ Questions for Machine Comprehension of Text". *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101.

Rodman, E. (2020). A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. *Political Analysis*, *28*(1), 87–111.

Shapley, L. S. (1952). *A Value for N-Person Games*. RAND Corporation.

Skubic, J., Angermeier, J., Bruncrona, A., Evkoski, B., & Leiminger, L. (2022). Networks of power: Gender analysis in selected european parliaments. *2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS)*.

Spinde, T., Rudnitckaia, L., Hamborg, F., & Gipp, B. (2021). "Identification of Biased Terms in News Articles by Comparison of Outlet-Specific Word Embeddings". In K. Toeppe, H. Yan, & S. K. W. Chu (Eds.), *Diversity, divergence, dialogue* (pp. 215–224). Springer International Publishing.

Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, *23*(5), 828–841.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International conference on machine learning*, 3319–3328.

Tahmasebi, N. (2018). A Study on Word2Vec on a Historical Swedish Newspaper Corpus. *Proceedings of the Digital Humanities in the Nordic*

Countries 3rd Conference, DHN 2018, Helsinki, Finland, March 7-9, 2018., 25–37.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355.

Wendlandt, L., Kummerfeld, J. K., & Mihalcea, R. (2018). Factors Influencing the Surprising Instability of Word Embeddings. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2092–2102.

Yang, Y., & Eisenstein, J. (2017). "Overcoming Language Variation in Sentiment Analysis with Social Attention". *Transactions of the Association for Computational Linguistics*, *5*, 295–307.