CHAPTER 25

Studying Ideational Change in Russian Politics with Topic Models and Word Embeddings

Andrey Indukaev

25.1   INTRODUCTION

Ideas are both a promising and challenging object for political and social science, especially in the case of Russian studies. The challenges and promises are of methodological and theoretical order. At the theoretical level, the study of the Russian political system quite often discards ideas because of the prevalent rent-seeking behavior of political and economic actors that make interests, not ideas, reign (Gel'man 2016b). However, ideas are of importance for politics and policy processes in any context (Carstensen and Schmidt 2016). Indeed, recent research shows that many aspects of Russian politics cannot be understood without taking into account the ideational dimension, making it a promising direction in the field of Russian studies (Wengle 2015; Dabrowska and Zweynert 2015). Pursuing this direction is challenging. First, ideas are hard to grasp and cannot be studied without a thorough and context-aware examination of meaning expression. Quite often, that implies using methods relying on a "close reading" of texts and requires that the texts where ideas are expressed are available. Within the context of Russian electoral authoritarianism, public expression of ideas in political arenas, through media and other channels, faces constraints that should be taken into consideration. Since the parliament is not a place of political debate, one does not have the data that could serve as a reference for capturing the legislature's ideological landscape, thus making it difficult to study the ideas in the Russian politics (Lowe et al. 2011; Slapin and

A. Indukaev (✉)
University of Helsinki, Helsinki, Finland
e-mail: andrey.indukaev@helsinki.fi

Proksch 2008). The political context and historical factors are also a source of the "public aphasia"—the lack of the register for the public discussion of the issues of common interest where ideas and ideological positions are expressed (Vakhtin and Firsov 2016).

The proliferation of digital communication, leaving a multitude of textual traces, implies that the study of ideas can significantly widen its scope. It is particularly relevant for Russia, where the Internet became a primary medium for oppositional politics and emerged as an arena for public debate, reducing the "public aphasia" and making the public expression less constrained (for more, see Chap. 2). In addition, a greater volume of textual data and new computational methods of textual analysis are becoming available. They promise a possibility to study the ideational dimension of politics without relying on big corpora of texts frequently and explicitly invoking political ideas, such as parliamentary debates or party manifestos. Indeed, ideational content can be captured even when it is sparsely distributed across a large volume of texts. Digital data and computational methods of text analysis give the opportunity to complete or scale up insights of qualitative analysis based on scarce sources of ideationally dense political texts.

Word embeddings (WE) and topic models (TM) refer to two groups of techniques of text processing and mining that are often used by researchers in social science and humanities (SSH) to study the ideational dimension of politics. This volume provides a discussion of both of them in detail (for more, see Chaps. 26, 23, and 24). An important feature of TM and WE, when used in SSH, is that there are no guidelines on how to apply them to research problems. Instead of treating them as "ready-to-use" methods, a sensible use of TM and WE in SSH implies nowadays developing a research design that takes into account the specificities of the research question and the data at hand. Thus, I find it important to complement contributions to this volume focusing on the overview of the methods with an illustration of their application. To do that, I will use WE and TM to study how ideas, influential in Russian politics, change over time. Particular attention will be accorded to explaining why and how each method was used given the peculiarities of the research question, of the Russian context, and of the data available. To allow the larger audience to apply the chapter's ideas in their own research, I will give preference to solutions that can be easily implemented in the R programming language (www.r-project.org) and indicate specific R packages used in each case.

The empirical focus of the chapter is on the ideas of innovation, technology, and economic development that played a key role in the modernization agenda of Dmitry Medvedev, on their evolution when the agenda was abandoned, and when some elements of it resurfaced in Russian politics thanks to Putin's fourth term's emphasis on digitalization as a key priority. More specifically, I explore the evolving relationship that the innovation, technology, and economic development maintain, in public discourse, with the political liberalization idea. Through its focus on digitalization, this chapter is connected to the first section

of the handbook, which studies digitalization as a sociopolitical phenomenon, and to the chapter of Anna Lowry on the digital economy.

The chapter will be organized as follows. In Sect. 25.2, I present the key ideational dimensions of Medvedev's modernization agenda and their evolution on the basis of a qualitative study. Then I formulate the research questions: (1) Did the ideas associated with modernization and its demise manifest themselves in the Russian media? and (2) How did the concept of digitalization embed itself in the existing set of ideas on technology and politics? Section 25.3 focuses on the overview of the methodology. In Sect. 25.3.2, I discuss topic modeling, and in Sect. 25.3.3, I discuss word embeddings and how it can be used to detect complex semantic relationships between words, revealing social and political representations. Finally, in Sect. 25.4, I apply TM and WE to the Russian media data by using specific approaches given the type of data at hand. I show that the modernization agenda influenced public discourse in Russia by promoting the idea that innovation, technology, and economic development are associated with political and social change, that this idea disappeared from the public discourse, and that the rise of the digitalization agenda did not bring it back.

25.2   IDEAS OF MODERNIZATION

As suggested above, qualitative analysis is essential for studying the ideational dimension of politics. Thus, any study of ideas using quantitative techniques should be accompanied by qualitative analysis or should build on such analysis done previously. In this chapter, I will apply TM and WE to study a case that I extensively studied in my doctoral dissertation, using a variety of qualitative techniques (Indukaev 2018). My focus will be on the ideas on the political role of innovation, technology, and economic development. These ideas have played an important role in Russia recently because of the political agenda of modernization that Dmitry Medvedev embraced during his presidency in 2008–2012. They were subject to a major transformation after the modernization program was abandoned. The transformation was a nontrivial one, making it an interesting object for the study. In the following section, I will describe the political context of the case study, outline the key features of the ideational change I am focusing on, and state the research questions and hypothesis.

25.2.1   Politics of Innovation, Technology, and Economic Development During Medvedev's Presidency and After

Medvedev's political platform positioned him as a more liberal and reform-minded president, without directly opposing him to Putin. Medvedev's political manifesto "Go, Russia!" presented economic and technological modernization of the country as a top priority, but also promised political liberalization. The latter promise relied, in part, on the planned change of the country's political system—including giving more power to the parliament and

making elections more inclusive. However, at the discursive level, the political change was subordinated to the imperative of economic modernization since, in Medvedev's reading of history, "democracy occurred on a mass scale, not earlier … than when the level of the technological development of the Western civilization made it possible to gain universal access to basic amenities: education, healthcare and information" (Medvedev 2009, n.p.). Technological and economic modernization is presented as a precondition to political modernization: "the technological development is a societal and political task of top priority because the scientific and technological progress is inextricably linked with the progress of political systems" (Ibid, n.p.). In Medvedev's political program, the idea of technological and economic development connoted the idea of political liberalization and social change, while the concept of modernization englobed both ideas.

The key projects associated with Medvedev's modernization agenda aiming at technological and economic development were associated with the ambition of political liberlaization. For example, the organizational design of the Skolkovo Innovation Center was influenced by the idea that the state should leave more space for the bottom-up initiative, making its mission focused less on concrete projects but more on the development of an "ecosystem" providing opportunities for unfettered innovative activities (Indukaev 2018). Rusnano, an institution aiming at nanotechnology development in Russia and created under Putin's patronage before Medvedev's election, associated itself with the political ambition of the modernization even more explicitly. Anatoly Chubais, the head of Rusnano, published on the organization's official website a short polemic text intended to defend the idea that modernizing the economy within the nondemocratic context is worth doing. One of his main arguments was that developing an innovative economy will bring to life a class of "scientific and technological intelligentsia," and that "true democracy will appear in the country only when there is a social class that really needs it" (Chubais 2009). Thus, Rusnano's investment in high-tech companies was framed as serving the cause of democratic transition.

Medvedev did not run for a second term and was not able to advance his political program. The ambitions of the political liberalization and social transformation that Medvedev's project included were discarded and have never completely regained their political standing. The situation is different with the ambitions of technological and economic modernization. They lost their priority status after Medvedev's departure. During Putin's third term, the projects inherited from the modernization era were not at the forefront of the political leadership's agenda, sidelined by the conflicts in the international arena and the conservative turn in the country-level politics. Many experts believed that the policy projects associated with modernization would be stopped, in particular Skolkovo (see, for example, Gel'man 2015). However, the project survived, and its budget was not significantly cut. Rusnano and other projects that aligned with the modernization agenda also remained active. Moreover, these organizations managed to align themselves with the import substitutions

agenda, *importozamêsenie* (for more, see Chap. 17), which was central to the field of the technological and economic development (Indukaev 2018). More importantly still, the Medvedev-era economic and technological development policy instruments regained political importance when Putin presented his fourth mandate as being centered around the ambitions of radical "digital transformation" (Rus. *cifrovizaciâ*) and of the "breakthrough" (Rus. *proryv*), the accelerated economic and technological development. Skolkovo, for example, immediately associated itself with Putin's agenda. Promoting technological and economic development, even though framed more in the digitalization than in the modernization terms, revived some elements of Medvedev's project.

25.2.2   *Research Questions*

The story I outlined above implies that in 2008–2012 innovation and technological and economic development were associated in Russian politics with the promise of political liberalization. This association was coined in the concept of modernization, which also served as a keyword (for more, see Chap. 17) of Medvedev's political program. When Putin replaced Medvedev as the head of state, the ideational configuration of modernization was discarded; innovation, technology, and economic development were not associated with the promise of liberalization any more.

In my previous research (Indukaev 2018), I detected the described change by qualitative analysis of political speeches and manifestos, policy documents, and institutional arrangements. Thus, the observed change concerns ideas expressed by top-level politicians and reflected in policy decisions. The first question I want to address in this chapter is whether the described ideational configuration and its change was reflected in the way innovation, technology, and economic development were discussed in the media.

The second research objective of this chapter is to extend the scope of my analysis to a new element, which started playing an important role in the political discourse on technology, innovation, and economic development, namely the idea of digitalization. At the top-level of the official discourse, I have not found any indices that Putin's promise of digitally enabled development was associated with the promise of political liberalization. Instead, digitalization is framed as prioritizing merely the quality of the citizens' life, and, not less important, the country's standing in the international arena. In contrast, digital technology was associated with liberalization during Medvedev's time, who suggested, "The growth of modern information technologies, something we will do our best to facilitate, gives us unprecedented opportunities for the realization of fundamental political freedoms, such as freedom of speech and assembly" (Medvedev 2009, n. p.). Moreover, the development of digital tools promising political empowerment and democratization was actively supported

by the state after 2012 and at the level of local politics (Chap. 3). One may suggest that digitalization is associated with political liberalization in public discourse, despite this association not being explicitly expressed by the political leadership. The second question of this chapter is whether this suggestion is valid.

25.3   Data and Methods

To extend the analysis based on writings and speeches by political leaders and policy documents to the ideas expressed by a wider audience, one needs appropriate data, such as Russian media data used in this chapter. Despite the limited freedom of speech, Russian media are not mere translators of the political leadership's perspective and can be used to assess how ideas spread within the Russian public. To analyze these data, I will use two families of computational textual analysis methods, topic modeling (TM) and word embeddings (WE). Apart from methodological reasons, described below, the choice is determined by the fact that topic modeling is among the most widely used among these techniques (Isoaho et al. 2019), and word embeddings could be expected to take the lead in the coming years.

25.3.1   Data

Integrum is the largest database of Russian media. It is a commercial product primarily aimed at business clients but is also used by researchers in their studies of Russian language and society (Chap. 17). In this chapter, I use this database. The research strategy is to assemble the corpus focused on technological and economic issues to detect how the political issues appear there. Thus, the query did not include the word *modernization*, since it has explicit political connotations. Instead, the query was made of terms related primarily to technological and economic development, innovation, and digitalization, but not to political change. The query looked as follows:

"иннов* OR роснан* OR сколков* OR венчур * OR нано* OR цифровиз* OR Электронная

Россия)/W2 OR (Цифровое Развитие)/W2",

where the symbols in Cyrillic represent stemmas of Russian words, "OR" is an operator, and "W2" is a context that is considered in the query.

When forming the query, I used wildcards to include all possible morphological forms of a word corresponding to a concept of interest (for the description of the search options, see Chap. 17). The promotion of innovative activities

was an important part of the modernization program, so any form and cognate word for *innovaciâ* (innovation) could be used in relation to this program. I use the stem with a wildcard "иннов*" (innov*) to capture all these forms. The stem "*венчур*" (venčur*) refers to venture capital, a specific form of investments in early-stage innovative firms, which was an important reference for the state's effort to promote innovation. Skolkovo was a flagship project of modernization, so I used "сколков*" (skolkov*) to get it mentioned. This part of the query returned a limited number of irrelevant documents because of the Russian word *skolok* (plural *skolki*) meaning, "pricked pattern," which should not influence the analysis because of the word's rarity. However, when building a query, a user should be aware that Russian words may have more frequent homonyms, which makes searching tricky.

Nanotechnology promotion and the designated organization Rusnano were major projects of technological development and were also associated with modernization. Again, the "*нано*" (nano*) part of the query returned a lot of irrelevant documents because of many words beginning with "нанос" ("*nanos*"), in particular the verb *nanosit'*, the meaning of which is "to inflict." That included, for example, a significant amount of crime news. The corresponding documents were removed during the corpus preparation. The query "цифровиз*" ("*cifroviz*") aims at various forms of the word *cifrovizaciâ* (digitalization), a distinctive term that Putin introduced into political language as a Russian equivalent of digitalization. The query also included the names of two major policy instruments in the field of digitalization, *Èlektronnaâ Rossiâ* (Electronic Russia) and *Cifrovoe razvitie* (Digital Development) programs.

The list of media included 12 sources from the category "Central press," 2 from "Central news agencies," 39 from "Central internet media," 13 from "Central TV and radio," and 20 regional newspapers, news agencies, and internet media as well as the websites of the Russian government and the presidency. The list composition aimed at a coverage of a variety of sources, including pro-government and more oppositional ones, and also media specialized in technology or economics, regional media, in particular from the regions actively engaged in development projects, such as Tomsk, Novosibirsk (Indukaev 2019), and Tatarstan. The time period covered spans from October 1, 2007, until January 1, 2019, starting about half a year before Medvedev's inauguration. The query produced 320,000 documents, among which a random selection of 160,000 was used to work with, because of computational limitations of the used setup. The corpus was preprocessed: all characters were transformed to lowercase, and punctuation and number were stripped. Using the collocation functionality of text2vec R package (CRAN.R-project.org/package=text2vec), the most common multi-word expressions, such as *tehnologičeskoe razvitie* (technological development) were transformed into tokens, such as *tehnologičeskoe_razvitie*. The stopwords were removed using "stopwords-iso" list from R stopwords package (CRAN.R-project.org/package=stopwords). The resulting corpus had 45,295,399 tokens.

25.3.2   *Topic Modeling*

The fact that modernization became a slogan for President Medvedev's term sparked active research in the field of Russian studies focused on a vast range of subjects connected to the topic of modernization (Gel'man 2016a; Mustajoki and Lehtisaari 2017). The only use of methods of quantitative textual analysis that I am aware of is the study of "the attitudes of the people towards modernization" in Russia. It was done through exploring the media publication available in the Integrum database (Chap. 17; Laine and Mustajoki 2017). The authors showed how economic, educational, and political preconditions of modernization were debated. To do so, the authors focused on the uses of the word *modernizaciâ* (modernization) that explicitly refer to country modernization. Applying an iterative search procedure to a dataset containing about 10,000 occurrences of the word, the authors extracted 100 passages where the necessary preconditions to the modernization of the country were discussed.

In this chapter, I analyze the political concept of modernization in the context of a larger set of ideas on the political role of innovation, technology, and economic development. This leads me to use the corpus that covers quite a large spectrum of discussions of innovation, technology, and the corresponding government's activities and to look there for evidence regarding the research questions focused on political ideas. To do that, I will approach the corpus in a way that gives the opportunity to explore the totality of its ideational content but also to focus on particular ideas and concepts. Topic modeling is a great method to start this exploration.

To put it simply, topic modeling is based on the assumption that the documents in a given corpus are generated as a mixture of a determined number of topics—technically, bags of words grouped together based on their tendency to co-occur in the corpus (for more, see Chap. 24). Many variations and extensions of the method are available (for more, see Chap. 23); however, the basic intuition stays the same. Initially, topic modeling was developed as a tool for the retrieval of information that can summarize the thematic content of a large collection of documents. Yet, the key issue that researchers in social sciences and humanities encounter when using topic modeling is that the there is no universal rule for interpreting the output of the topic model—the "topic" that emerges as output—as well as no universal way to integrate TM into the research design and to adapt it to a specific research question (Isoaho et al. 2019). In what follows, I discuss how to use the method to answer research questions related to the study of ideas.

In many studies using topic modeling, the thematic content of a corpus is predetermined, and the method is used instead to detect various ideological perspectives on a given topic. In these approaches, a topic or a set of topics, detected by the model, are interpreted as being associated with a specific perspective on the thematic content. Typically, scholars analyze these perspectives using the concept of "issue dimension" (Nowlin 2016) or, more commonly,

that of a frame (see e.g. DiMaggio et al. 2013; Fligstein et al. 2017; Ylä-Anttila et al. 2018). However, quite often, using an issue-specific corpus does not guarantee that the topic model will output topics corresponding to the frames. In the known examples of the use of topic modeling in frame analysis by Fligstein et al. (2017) and DiMaggio et al. (2013), both interpret some sets of topics among the output of the model as corresponding to the frames, while not attributing other topics to any frame. Indeed, the topic model outputs, even within an issue-specific corpus, cannot be automatically seen as frames in most cases (see Isoaho et al. 2019). The association between a topic produced by a topic model and a frame, an "issue dimension," or any other comparable analytical category is a matter of interpretation which does not rely exclusively on the topic model output but invokes other quantitative or qualitative methods and a theoretical perspective on the issue.

Another use of TMs for studying ideas is to treat TM output more as topics in the literal sense—basically, a coherent theme appearing in the corpus—and not to interpret them as ideational perspectives. When other methods are used to reveal these perspectives, topic modeling can be used to offset the influence of thematic content of analyzed texts on the ideational perspective (Jelveh et al. 2018; Lauderdale and Clark 2014). Other approaches suggest modification of the Topic Modeling algorithm in a way that assumes that word choice in texts is determined both by the ideological perspective and by the topic in the mainstream understanding of a term—the theme of a text (Magnusson et al. 2018; Ahmed and Xing 2010).

In the next section, I apply TM to summarize the thematic content of the corpus. Then, I focus on the topics that are of interest for the study of ideas on innovation, technology, and economic development. I will not use the family of approaches described in the previous paragraph. However, the insight that there are a variety of possible relationships between a topic detected by TM and an ideological perspective—from equivalence to independence—will be key to understanding the limitations of TM-based analysis of ideas. To overcome these limitations, I will use another family of techniques based on word embeddings.

25.3.3 Word Embeddings: Semantic Change and Interpretable Dimensions

Word embeddings is a family of techniques that represent words as numerical vectors in a way that the relative positions of vectors in the embedding space reflect the relations of semantic proximity of corresponding words (for more, see Chap. 26). To put it simply, the semantic proximity of two words corresponds to the geometric proximity of two vectors that represent the words. The term "word embeddings" is often used to refer both to the vectors representing the words and to the techniques used to obtain the vectors.

The capacity to represent semantic proximity as a geometric one opens avenues for many advanced approaches for studying the ideational dimension in

large corpora. One of these approaches comes from the studies of how the meanings of words change in time: diachronic lexical semantics change. Distributional semantics is one of the advanced computational approaches to semantic change in linguistics. Since the introduction of neural word embeddings, the methods of distributional semantics have manifested significant progress (for a comprehensive review, see Tahmasebi et al. 2018; Kutuzov et al. 2018; and also Tang 2018).

Distributional semantics using WE analyzes semantic shifts following the logic that a relative position of words vectors in multidimensional embedding space is a reflection of the meaning shift. The techniques used may vary. In most cases, researchers use diachronic corpora: for example, a bigger corpus sliced into a set of subcorpora corresponding to consecutive time periods. Then, word embeddings are created for each subcorpus. The vectors representing a word of interest and corresponding to different time periods in the subcorpora may have a different position relative to other words' vectors. That could imply a change of semantics of the word of interest. One of the most used techniques is to focus on the change of semantic "neighbors" of a word—the words whose vectors are the closest to the vector representing the word of interest. For example, a word is expected to have changed its meaning if there was a significant change of the list of top ten words most semantically similar to it. For example, in the word embedding space based on the corpus of English texts dating to the 1850s, the world "broadcast" had words like "seed," "sow," and "scatter" as its nearest neighbors, but in the embeddings based on a 1990s corpus, it neighbored "bbc," "radio," and "television." That suggests that the old meaning "throwing seeds" was replaced by the new one, "disseminating information" (Hamilton et al. 2018, 2).

The methods of distributional semantics can be used to analyze ideational change, even though they were not designed for this purpose. Within the study of semantics, the change of word meaning can be explained by "sociocultural" causes (Kutuzov et al. 2018, sec. 2), which opens an avenue for research that interprets semantic change not as a language's internal affair, but as an indicator of an ideological transformation in the society. Also, the methods of distributional semantics can be used to analyze synchronic variation instead of diachronic change. For example, Azarbonyad et al. (2017) used word embeddings-based metrics of semantic similarity to contrast the viewpoints of Labor and Conservative parties on democracy.

The malleability of words and concepts, the fact that their meaning can vary in time and across different social and political contexts of use, is an essential feature of political language. In the case of the studies of Russian politics, this malleability is of great importance. The change of political language is not primarily associated with public debates on political arenas but is related to opaque political processes that are not always intelligible. Moreover, compared to democratic systems, abrupt political change, and, correspondingly, changes in political discourse are not a feature of Russian politics. At the surface, the political system manifests continuity, and its political discourse is subject to a

gradual change. That makes this change less obvious. Using the methods of distributional semantics, I will show how the concept of modernization, central to Medvedev's political program, gradually changed its meaning while staying an important element of the political discourse on technology, innovation, and economic development.

When analyzing the ideational change through looking at how concepts central to the political discourse change their meaning, a question arises of how to include new concepts in the analysis. Indeed, the ideational configuration can evolve not only through semantic drifts of its key elements. One of the possible paths to ideational change is the rise of new ideas and new concepts. In my inquiry on political ideas on innovation, technology, and economic development, one can easily detect such a new element—the concept of digitalization. In this case, analyzing how new politically important concepts are different compared to old ones becomes an important problem, and word embeddings provide an opportunity to do it.

An important feature of word embedding is that any two words can be characterized not only by the distance between them—that means the *length* of the difference vector obtained by subtracting the vector representing word A from the vector representing word B. In addition to it, the *direction* of the difference is informative, as it can reveal fine-grained aspects of the semantic relationship between two words. For example, the vectors for the words "queen" and "king" can have a relatively small distance and be neighbors in the embedding space built on a sufficient volume of data—a trivial result, since both words designate a monarch. However, if one looks at the direction of the difference between two-word vectors in an embedding space, one can make an interesting observation. The difference between the vector "king" and the vector "queen" will be almost the same as the difference between the vector "man" and the vector "woman" (Mikolov et al. 2013). Thus, one can conclude that it is possible to determine in the embedding space a vector whose direction summarizes the semantic difference between male and female, or in other words, a "gender" dimension. This logic can be extended to other forms of semantic relationship, for example those opposing "rich" and "poor," or the "affluence" dimension. This approach is thoroughly presented by Kozlowski et al. (2019) in a recent article. The authors calculated word embeddings on Google Ngram's corpus with the help of standard techniques but used the resulting vectors in a way that made it possible for them to extract what they call "cultural dimensions." The technique assembles antonym pairs for a dimension, such as "poor"-"rich" for the "affluence" dimension, and then calculates the difference vector for each pair and the average difference vector. Thus, any word in the corpus can be located as being more or less related to the affluence. Authors show, for example, how certain activities are located on an "affluence" dimension, tennis, for example, being more related to affluence than boxing. The method was proved to capture cultural representations existing in society and revealed through other means, such as surveys or experimental studies. Comparable approaches are being actively developed, such as one proposed by

Bodell et al. (2019), who modified a word embeddings algorithm in a way that the resulting embedding space dimensions are interpretable.

The approaches like the one developed by Kozlowski et al. (2019) convincingly show that one can construct, in an embedding space, the vectors that capture the semantic relationships corresponding to cultural representations within a society. Such approaches do not have to focus exclusively on culture but can be applied to the study of political ideas and representations. For example, Rheault and Cochrane (2019) use a modified version of the word2vec model, which, based on a parliamentary debate corpus, creates an embedding space that, after applying the dimensionality reduction, produces a vector that represents the opposition between the right and the left ideological perspectives.

One of the questions of this chapter is how a new conceptual element—namely, digitalization—fits into the existing ideational landscape. To answer it, I will rely on the approaches described above by constructing vectors that correspond to key dimensions of this ideational landscape.

25.4   Results

In this chapter, I analyze the Russian media to see how its language reflected the events in which the association between political liberalization and innovation, technology, and economic development was brought into political discourse by Medvedev, but vanished after his departure. Also, I am going to look in the media, for evidence that the digitalization agenda revived in the public discourse the political liberalization promise of the modernization agenda.

First, it is important to get an idea of the corpus thematic composition, to understand whether it can be used to answer the research questions. The keywords used in the query, in particular "innov*," match with words that have a multiplicity of meanings. For example, the word "innovacija" (an innovation) is often used to refer to new features of products. As a consequence, the corpus is composed of many documents unrelated to the research question. In general, one does not know precisely what is being discussed in the corpus. In this situation, topic modeling is an appropriate method to start with, as it can reveal the composition of the corpus.

The corpus was analyzed using the text2vec library for R (Selivanov and Wang 2018). This library has an advantage of being developed with computational efficiency in mind. Topic modeling is implemented there using WarpLDA algorithm, which is significantly more efficient than other algorithms for Latent Dirichlet Allocation (Chen et al. 2016). The disadvantage of this implementation is that it does not take into account topic correlation and does not allow the inclusion of covariates, such as date, in the topic modeling process, which is possible to do with the much slower Structural Topic Models (STM) version of TM (Roberts et al. 2019). However, as this chapter uses a large corpus necessary to calculate word embeddings, a more efficient but less sophisticated algorithm was preferred.

Given the size of the corpus, the number of topics to be calculated had to be set high. I first ran a model with 50 topics, a part of which were interpreted as relevant to the chapter's research questions. Then, to check the stability of these relevant topics, I ran models with 45 and 55 topics, respectively, and saw the same topics reappear. This technique was used to validate the model, showing that the results are robust enough to resist minor changes in the model parameter—number of topics.

Analyzing TM output showed that the corpus includes many documents that discuss themes irrelevant to the research questions. For example, many topics focus on specific products and services, such as mobile phones or cloud services; others correspond to themes dominating the Russian media space, such as Ukrainian politics or war in Syria. As mentioned, many topics were interpreted as being relevant, such as the one focused on nanotechnology. However, one topic stands up as being central to my analysis.

The topic that is the most prevalent in the corpus is the one that is clearly associated with the modernization agenda. I analyzed the 50 most representative words for this topic (using the tex2vec function get_top_words, setting lambda to 0.3). The list includes various forms of the words and expressions *gosudarstvo* (state), *èkonomika* (economy), *modernizaciâ* (modernization), *čelovečeskij_kapital* (human capital), *srednij_klass* (middle class), *strana* (country), *reforma* (reform), *otstavanie* (retardation), *proizvoditel'nost'_truda* (labor productivity), *konkurencii* (genitive for competition), *preobrazovanij* (genitive for trasformations), *peremeny* (changes), *strukturnyh_reform* (genitive for structural reforms), *obŝestva* (genitive for society), and *razvityh_stran* (genitive for developed countries). These words are characteristic for the topic and suggest that it is associated with Medvedev's idea of modernization. First, it appeared in a corpus that was built without using *modernizaciâ* (modernization) in the query but focusing on documents mentioning innovation and policy tools in the fields of innovation, technology and economic development. It suggests that the debate of modernization is associated with the debate on innovation and technological and economic development, as it was in Medvedev's program. Second, the topic combines words referring to economic development with words referring to social and political change and reforms. Third, terms like "developed countries" and "retardation" suggest the importance of the rhetoric where the country's modernization is seen as "catching up" with the most developed countries. Last, the words referring to the state are frequent in the topic, suggesting that the modernization is considered at the state level. All these dimensions of modernization are present in Medvedev's manifesto "Go, Russia" (Medvedev 2009). A close reading of the top ten documents where the topic is the most prevalent confirms my interpretation. All the documents debate the ideas that are present in Medvedev's program.

As I described in Sect. 25.3.2., TM is often used to analyze political ideas by associating a topic with a certain ideational perspective. The "Modernization" topic that I described can be associated with a specific perspective on the relationship between political and social change and economic development,
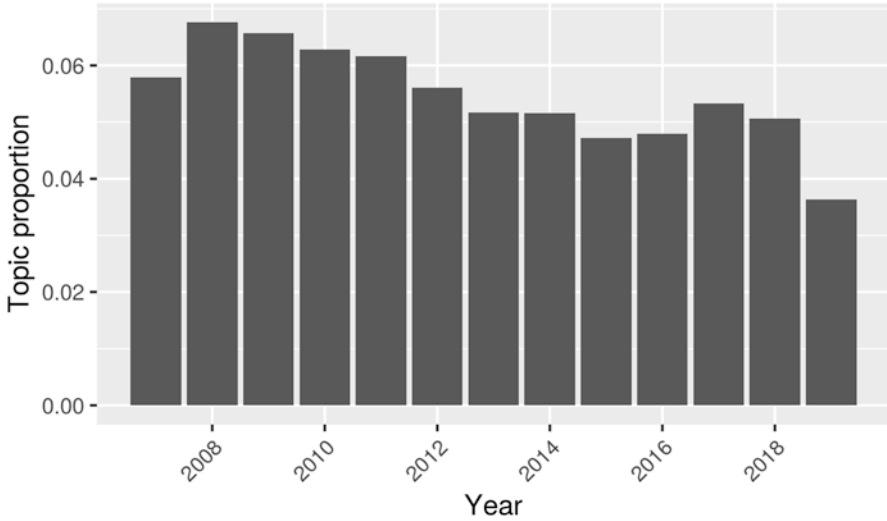
Fig. 25.1 "Modernization" topic prevalence dynamic

technology, and innovation. If one accepts the idea that this topic is an indicator of a certain ideological position, one can attempt to assess the ideational change looking at topic dynamics. The topic prevalence in time corresponds, in part, with what could be expected based on the case description in Sect. 25.2. The topic peaked in 2008 and declined gradually after that (Fig. 25.1). However, after reaching a minimum in 2015, the topic started rising again, with a second peak in 2017, the year Putin started promoting his agenda of digitalization. That could suggest that the revival of technological development as a central element of the political leadership agenda revived the connotation between technological and sociopolitical change. However, using the methods of distributional semantics described in Sect. 25.3.3, I will show that this interpretation does not hold.

Modernization, despite its clear association with Medvedev's political program, is a concept that has a rich and a malleable meaning, and actors can use it in ways that can highlight various dimensions of the meaning and even attempt to redefine it. I will show that there was a change of meaning which erased the Medvedev era's ideological association between modernization and political reform, as suggested by the qualitative analysis in Sect. 25.2.

To analyze meaning change, I used a technique based on word embeddings. To calculate word embeddings, I used an implementation of the GloVe algorithm (Pennington et al. 2014) provided by text2vec package in R.

The data used is the same as described in the corresponding section, but with one major adjustment, which is due to my choice of research design appropriate for detecting the change in word meaning and use. Dubossarsky et al. (2019) recently demonstrated that the Temporal Referencing technique

has significant advantages over other approaches of detecting genuine semantic change. The idea of the method is, first, to focus on a limited set of words whose change is going to be studied. Then, the corpus is not sliced into sub-corpora corresponding to different time intervals; instead, the word embeddings are calculated for the entire corpus. However, the corpus is modified: the words of interest are replaced by "time-specific tokens." It simply means that, for example, if one wants to study how the meaning of the word "modernization" changes from year to year, one replaces the word "modernization" in the documents dated by 2007 by "modernization_2007" and does the same for every other year. The rest of the words, whose semantic change is not analyzed, stay intact. In this chapter, I use the described method to trace the semantic change of two words: *modernizaciâ* (modernization in singular) and *innovacii* (innovations in plural). To keep the research design simple, I worked with two periods January 1, 2007–January 1, 2012 (label "_before"), and January 2, 2012–January 1, 2019 (label "_after"). This change was made because by the end of 2011, it was clear that Medvedev would not keep the presidency, and the promise of political reform was not to be fulfilled.

To detect the change in the meaning of *modernizaciâ* "modernization," I compare the list of semantic neighbors of *modernizaciâ* before January 1, 2012, to semantic neighbors of *modernizaciâ* after 2012. In addition, I analyze how the list of neighbors changed: which words became less semantically close to the word of interest and which words got closer. When one looks at the "neighbors" of *modernizaciâ*, one sees quite a radical change. In Table 25.1 are provided the top 30 words closest to *modernizaciâ* before and after January 1, 2012.[1] Modernization after 2012 does not have a semantic proximity to democracy (*demokratiâ*), fight with corruption (*bor'ba_s_korrupciej*), reforms (*reformy*), and politics (*politika*), being associated mostly with terms related to technological advances, efficiency (*povyšenie èffektivnosti*), and retooling (*tehničeskoe perevooruženie*). The meaning of the concept changed, and the specific association between the modernization and the promise of political liberalization evaporated after Medvedev's departure. This result refutes the idea based on the topic modeling analysis of the "modernization" topic, which suggested that Medvedev's modernization discourse resurrected around 2017: instead, the very concept of modernization changed its meaning.

The fact that the concept of modernization lost its association with political change does not completely rule out that the other key concepts referring to technological and economic development do not manifest it. Digitalization became the most important concept in the technological and economic development projects by the government after 2016. Revealing the association of this concept with the idea of political liberalization in public discourse is a good way to assess the scope of the ideational change that happened after 2012. Would it be possible that the digitalization project took the role of a technology development project bearing also a promise of a political change? To explore it, I used the approach following the insight that vectors in the embedding space can capture "dimensions of cultural meaning" and ideological

Table 25.1 Top-30 semantic neighbors of the word *modernizaciâ* (modernization) before and after January 1, 2012. Neighbors appearing only in one list are highlighted

Before January 1, 2012	After January 1, 2012
nevozmožna	rekonstrukciâ
modernizacii	modernizaciû
modernizaciû	masštabnaâ
reforma	optimizaciâ
innovacionnoe_razvitie	zamena
bor'ba_korrupciej	innovacionnoe_razvitie
korennaâ	proizvodstvennyh_moŝnostej
nužna	provedena
politika	tehničeskoe_perevooruženie
nikakaâ	perestrojka
neobhodima	povyšenie_èffektivnosti
demokratiâ	modernizacii
strukturnaâ	vozmožna
perestrojka	kompleksnaâ
revolûciâ	èkspluataciâ
innovacii_before	obnovlenie
vozmožna	dal'nejšaâ
innovacionnaâ_èkonomika	korennaâ
èkonomičeskaâ	infrastruktury
strategiâ	nevozmožna
političeskaâ	realizaciâ
ser'eznaâ	kardinal'naâ
reformy	privatizaciâ
demokratizaciâ	reforma
nastoâŝaâ	strukturnaâ
dolžna	bama
razvitie	tehnologičeskaâ
diversifikaciâ	transformaciâ

dimensions (Kozlowski et al. 2019, 905). To operationalize this insight, I followed the approach by van Lange and Futselaar (2019), which is less robust than the one proposed by Kozlowski et al. (2019), but requires less data and less preparation. The authors suggest that to create a vector that captures the distance of any word to a given perspective, it is enough to detect the words that indicate the perspective, then to create an aggregate vector that is an average of vectors of each of these words. The proximity of any word to a given perspective is then measured as a cosine distance between the words' vector and the aggregated vector representing the perspective. Van Lange and Futselaar's strategy to construct the aggregated vector is to focus on a concept that epitomizes an ideological perspective and then to find all the words that refer to a concept and do not have multiple meanings.

My analysis was limited by constructing two vectors, corresponding to two ideational perspectives on technology: innovation and economic development. The first perspective—"Political liberalization"—frames these phenomena as associated with social and political change. The second one, however, is focused on development, efficiency, competitiveness—the issues that economy and

technology face and that are seen as apolitical. I labeled this perspective "Economy and technology." To construct the vectors for each perspective, I first compiled, based on qualitative knowledge of the case, the list of words that are markers of a position. Next, among these words, I selected those that are frequent in the corpus (more than 150 occurrences) because the embedding vector's quality is sensitive to word frequency. Next, I looked at the top neighbors of each word and selected, based on a qualitative analysis, those that are good markers of the ideational perspective, again selecting only frequent words. Finally, I excluded words with multiple meanings. As a result, the "Political liberalization' vector consists of *reformy, demokratiâ, demokratii, liberal'noj, liberalizaciâ, liberalizacii, svobody, prav_svobod,*[2] *strukturnye_reformy.* The "Economy and technology" vector consists of *diversifikaciâ, diversifikacii, diversificirovat', importozamêsenie, konkurentosposobnost', povysenie_konkurentosposobnosti, èkonomiceskoe_razvitie.*

I calculated the distance to the two aggregated vectors for the vectors representing words that are central to the research question, including the two vectors for *modernizaciâ* before and after January 1, 2012, and the same for *innovacii.* Fig. 25.2 shows that *modernizaciâ* before January 1, 2012, was closer to the "Political liberalization," but during the period after January 1,
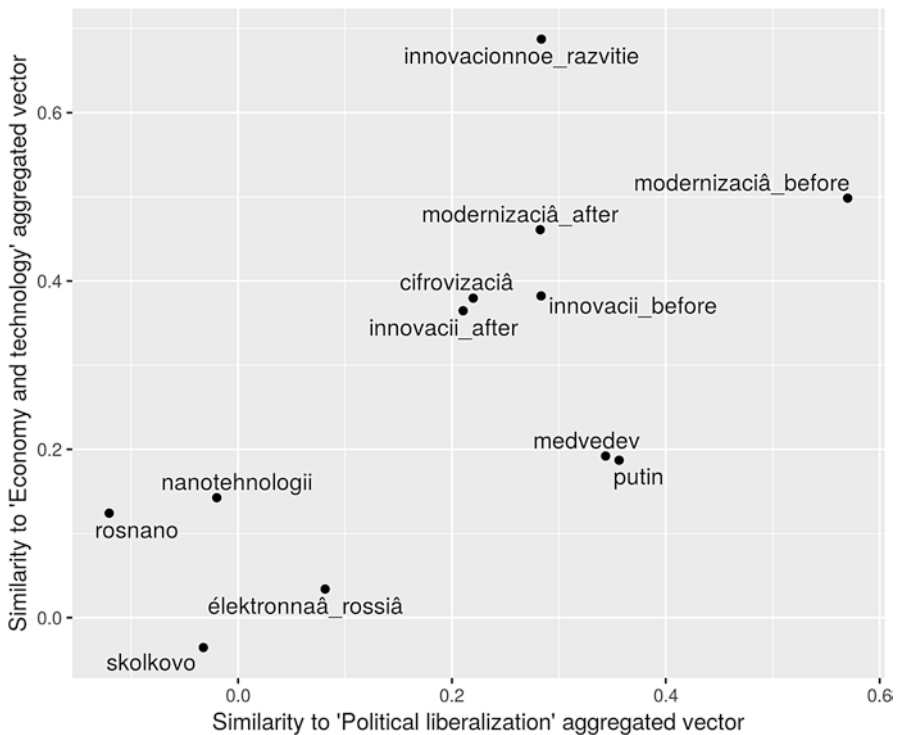

Fig. 25.2 Projection of keywords on a two-dimensional space

2012, it joined other terms, such as *innovacii*, becoming less associated with the idea of political change. The graph also gives an answer to our second research question. The position of *cifrovizaciâ* (digitalization) vis-à-vis the "Economy and technology" and "Political liberalization" vectors is almost identical to that of *modernizaciâ* after January 1, 2012, and that of *innovacii*. That suggests that the digitalization program, in contrast to Medvedev's modernization, is not associated with the question of political liberalization at the level of public discourse.

25.5   Conclusion

In this chapter, I used computational methods of textual analysis to study a recent case of ideational change in Russian politics. Based on my prior qualitative analysis of policy documents and political communication of political leadership, I outlined the main contours of this change. When Dmitry Medvedev was president, innovation and technological and economic development were associated with the promise of political liberalisation and played key role in the modernization agenda endorsed by the new president. The modernization agenda was abandoned after the end of Medvedev's mandate, but technology and innovation regained political importance when Putin chose digitalization as a priority project. However, the political liberalization was not associated with this new promise of technological and economic development.

I showed that the described story of ideational change could be observed at the level of Russian media discussing innovation, high technology, and policy projects of technological development. A good illustration of this case is the semantic change of the concept of modernization. This key concept of Medvedev's agenda had a close connotation to political and social change but became an apolitical term referring to mere economic and technological development. Moreover, the analysis corroborated the hypothesis that the digitalization, as a new political concept, does not have connotations to political and social change.

From the methodological point of view, this chapter serves as an example of the application of two popular methods of text mining—word embeddings and topic models—to the study of ideational change. An interesting result is that the analysis revealed how topic modeling can provide misleading results and how the methods of distributional semantics and multidimensional ideological mapping can help to avoid an erroneous interpretation. More precisely, believing that a topic can indicate the presence of a certain ideological position across time may lead to errors. As I showed, the topic centered on modernization, while being coherent and well present in the corpus, cannot be seen as a proxy for Medvedev era's ideational perspective on the political role of technology and economic development. This insight seems to be of great relevance for Russian politics, where ideational change takes place with not much public debate and can be overlooked by a researcher.

This chapter also provides a successful exploration of the possibilities that word embeddings provide for the study of the ideational dimension of politics. The capacity of WE to capture complex semantic relationships gives an opportunity to construct multidimensional "ideational spaces" in which words can be located according to their proximity to two or more ideational perspectives. I believe that this promising and actively developing branch of text analysis will be of great use for Russian studies, given the lack of simple ways to identify ideational oppositions that structure Russian public life.

Notes

1. The authentic context is *bor'ba s korrupciej*, but the stopword "*s*" ("with") is deleted.
2. The authentic context is *prav i svobod*, but the stopword "*i*" ("and") is deleted.

References

Ahmed, Amr, and Eric P. Xing. 2010. Staying Informed: Supervised and Semi-Supervised Multi-View Topical Analysis of Ideological Perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1140–1150. EMNLP '10. Stroudsburg, PA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1870658.1870769.

Azarbonyad, Hosein, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words Are Malleable: Computing Semantic Shifts in Political and Media Discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1509–1518. CIKM '17. New York, NY: ACM. https://doi.org/10.1145/3132847.3132878.

Bodell, Miriam Hurtado, Martin Arvidsson, and Måns Magnusson. 2019. Interpretable Word Embeddings via Informative Priors. *ArXiv:1909.01459 [Cs, Stat]*, September 10.

Carstensen, Martin B., and Vivien A. Schmidt. 2016. Power Through, Over and in Ideas: Conceptualizing Ideational Power in Discursive Institutionalism. *Journal of European Public Policy* 23 (3): 318–337. https://doi.org/10.1080/1350176 3.2015.1115534.

Chen, Jianfei, Kaiwei Li, Jun Zhu, and Wenguang Chen. 2016. WarpLDA: A Cache Efficient O(1) Algorithm for Latent Dirichlet Allocation. *ArXiv:1510.08628 [Cs, Stat]*, March. http://arxiv.org/abs/1510.08628.

Chubais, Anatolij. 2009. O Modernizacii i Liberalizacii [On Modernization and Liberalization]. 28 December. http://www.rusnano.com/about/press-centre/first-person/76380.

Dabrowska, Ewa, and Joachim Zweynert. 2015. Economic Ideas and Institutional Change: The Case of the Russian Stabilisation Fund. *New Political Economy* 20 (4): 518–544. https://doi.org/10.1080/13563467.2014.923828.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding. *Poetics, Topic Models and*

the Cultural Sciences 41 (6): 570–606. https://doi.org/10.1016/j.poetic.2013.08.004.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. *ArXiv:1906.01688 [Cs]*, June. http://arxiv.org/abs/1906.01688.

Fligstein, Neil, Jonah Stuart Brundage, and Michael Schultz. 2017. Seeing Like the Fed: Culture, Cognition, and Framing in the Failure to Anticipate the Financial Crisis of 2008. *American Sociological Review* 82 (5): 879–909. https://doi.org/10.1177/0003122417728240.

Gel'man, Vladimir. 2015. *Modernizacija, instituty i «poročnyj krug» postsovetskogo neopatrimonializma [Modernization, Institutions and the "Vicious Circle" of Post-Soviet Neo-Patrimonialism]* . Saint Petersburg: European University at St.Petersburg Press. https://tuhat.helsinki.fi/portal/files/46596892/M_41_15_final.pdf.

———. 2016a. *Authoritarian Modernization in Russia: Ideas, Institutions, and Policies.* London: Routledge. http://ebookcentral.proquest.com/lib/helsinki-ebooks/detail.action?docID=4644407.

———. 2016b. The Vicious Circle of Post-Soviet Neopatrimonialism in Russia. *Post-Soviet Affairs* 32 (5): 455–473. https://doi.org/10.1080/1060586X.2015.1071014.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2018. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *ArXiv:1605.09096 [Cs]*, October. http://arxiv.org/abs/1605.09096.

Indukaev, Andrey. 2018. *Revaloriser La Recherche En Russie Moderne (2000–2018). De La Politique de l'innovation à l'entrepreneuriat Scientifique?* Thesis, Paris Saclay. http://www.theses.fr/2018SACLN047.

———. 2019. Siberian Software Developers. In *From Russia with Code: Programming Migrations in Post-Soviet Times*, ed. Mario Biagioli and Vincent Lepinay, 195–212. Durham, London: Duke University Press.

Isoaho, Karoliina, Daria Gritsenko, and Eetu Mäkelä. 2019. Topic Modeling and Text Analysis for Qualitative Policy Research. *Policy Studies Journal*. https://doi.org/10.1111/psj.12343.

Jelveh, Zubin, Bruce Kogut, and Suresh Naidu. 2018. *Political Language in Economics. SSRN Scholarly Paper ID 2535453*, Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=2535453.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review* 84 (5): 905–949. https://doi.org/10.1177/0003122419877135.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic Word Embeddings and Semantic Shifts: A Survey. *ArXiv:1806.03537 [Cs]*, June. http://arxiv.org/abs/1806.03537.

Laine, Veera, and Arto S. Mustajoki. 2017. Preconditions of Russian Modernisation: A Media Analysis. In *Philosophical and Cultural Interpretations of Russian Modernisation*, ed. Arto S. Mustajoki and Katja Lehtisaari, 175–190. London; New York: Routledge, Taylor & Francis Group.

Lauderdale, Benjamin E., and Tom S. Clark. 2014. Scaling Politically Meaningful Dimensions Using Texts and Votes. *American Journal of Political Science* 58 (3): 754–771. https://doi.org/10.1111/ajps.12085.

Lowe, Will, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. Scaling Policy Preferences from Coded Political Texts. *Legislative Studies Quarterly* 36 (1): 123–155. https://doi.org/10.1111/j.1939-9162.2010.00006.x.

Magnusson, Måns, Richard Öhrvall, Katarina Barrling, and David Mimno. 2018. Voices from the Far Right: A Text Analysis of Swedish Parliamentary Debates. *SocArXiv*, April. https://doi.org/10.31235/osf.io/jdsqc.

Medvedev, Dmitri. 2009. «Rossiâ, Vperëd!» ["Go, Russia!"]. Gazeta.Ru, September 10. http://www.gazeta.ru/comments/2009/09/10_a_3258568.shtml.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*, January. http://arxiv.org/abs/1301.3781.

Mustajoki, Arto S., and Katja Lehtisaari, eds. 2017. *Philosophical and Cultural Interpretations of Russian Modernisation*. London; New York: Routledge, Taylor & Francis Group.

Nowlin, Matthew C. 2016. Modeling Issue Definitions Using Quantitative Text Analysis. *Policy Studies Journal* 44 (3): 309–331. https://doi.org/10.1111/psj.12110.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543, Association for Computational Linguistics, Doha, Qatar. https://doi.org/10.3115/v1/D14-1162.

Rheault, Ludovic, and Christopher Cochrane. 2019. Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. *Political Analysis*: 1–22. https://doi.org/10.1017/pan.2019.26.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. Stm: Package for Structural Topic Models. *Journal of Statistical Software* 91 (2). https://doi.org/10.18637/jss.v091.i02.

Selivanov, Dmitriy, and Qing Wang. 2018. *Text2vec: Modern Text Mining Framework for R* (version 0.5.1). https://CRAN.R-project.org/package=text2vec.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science* 52 (3): 705–722. https://doi.org/10.1111/j.1540-5907.2008.00338.x.

Tahmasebi, Nina, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Diachronic Conceptual Change. *ArXiv:1811.06278 [Cs]*, November. http://arxiv.org/abs/1811.06278.

Tang, Xuri. 2018. A State-of-the-Art of Semantic Change Computation. *Natural Language Engineering* 24 (5): 649–676. https://doi.org/10.1017/S1351324918000220.

Vakhtin, Nikolai, and Boris Firsov, eds. 2016. *Public Debate in Russia: Matters of (Dis)Order. Russian Language and Society Series*. Edinburgh: Edinburgh Edinburgh University Press.

van Lange, Milan, and Ralf Futselaar. 2019. Debating Evil: Using Word Embeddings to Analyse Parliamentary Debates on War Criminals in the Netherlands. *Contributions to Contemporary History* 59 (1). http://ojs.inz.si/pnz/article/view/322.

Wengle, Susanne A. 2015. *Post-Soviet Power: State-Led Development and Russia's Marketization*. New York: Cambridge University Press.

Ylä-Anttila, Tuukka, Veikko Eranti, and Anna Kukkonen. 2018. Topic Modeling as a Method for Frame Analysis: Data Mining the Climate Change Debate in India and the USA. *SocArXiv*, March. https://doi.org/10.31235/osf.io/dgc38.