A Weakly Supervised and Deep Learning Method for an Additive Topic Analysis of Large Corpora

Yair Fogel-Dror, Shaul R. Shenhav, Tamir Sheafer

CCR 3 (1): 29–59
DOI: 10.5117/CCR2021.1.002.FOGE

Abstract

The collaborative effort of theory-driven content analysis can benefit significantly from the use of topic analysis methods, which allow researchers to add more categories while developing or testing a theory. This additive approach enables the reuse of previous efforts of analysis or even the merging of separate research projects, thereby making these methods more accessible and increasing the discipline's ability to create and share content analysis capabilities. This paper proposes a weakly supervised topic analysis method that uses both a low-cost unsupervised method to compile a training set and supervised deep learning as an additive and accurate text classification method. We test the validity of the method, specifically its additivity, by comparing the results of the method after adding 200 categories to an initial number of 450. We show that the suggested method provides a foundation for a low-cost solution for large-scale topic analysis.

Keywords: topic analysis, deep learning, weak supervision, computational content analysis, natural language processing

"Political text as data" has emerged as an important trend in political science and communication studies in recent years. As the volume of and access to political texts continue to grow and computing resources become more available, we see an increasing need for research methods that focus on the systematic extraction of themes, topics, and concepts from large-scale news corpora (J. Grimmer & King, 2011; Justin Grimmer & Stewart, 2013; Quinn, Monroe, Colaresi, Crespin, & Radev, 2010). This paper builds on two

recent trends in this field, both of which aim to establish an accessible textual analysis method that can advance empirical research. The first is the use of topic models as an unsupervised topic analysis method to reduce costs by eliminating the need to manually code large amounts of text (Blei, 2012; Quinn et al., 2010). The second is the ability to incorporate external variables into the topic model and to explore and test theories by measuring relationships between topics and those variables (Lucas et al., 2015).

In this paper, we suggest that additivity—the ability to add topics to an existing model or even to merge two models—can further contribute to empirical research along these two lines. First, it makes this kind of research more accessible, as researchers can collaborate on projects and identify topics from different domains while reusing existing trained and labeled models. Second, it facilitates the testing of theoretical relationships between variables, as it allows for the addition of more topical variables to the theoretical model (e.g., testing whether a relation between variables holds while controlling for other variables). Last, by enabling the analysis of a different and possibly more general corpus, it expands the applicability of the empirical findings. Herein, we show how current methods are limited in these aspects and suggest that using weak supervision, in which the computer learns with "incomplete, inexact or inaccurate supervision" (Zhou, 2018, p. 44), can allow us to merge multiple topic models into a flexible and accessible method for topic analysis.

The outline of this paper is as follows: section 1 reviews current methods and their limitations; section 2 introduces our solution; section 3 describes the compilation of a training set using unsupervised learning; section 4 describes the supervised classifier; section 5 demonstrates and validates our solution's additivity; section 6 further validates our model; and section 7 presents our conclusion and highlights the advantages of our solution.

1 Current Methods of Large-Scale Content Analysis

As a computational content analysis method, topic modeling allows for large-scale analysis that allocates text to multiple categories with minimal human effort, which is mostly confined to the manual labeling of the categories that were extracted by the model. In this context, the computer looks for topics—distributions of words over a vocabulary—based mostly on the frequency and co-occurrence of words in an unsupervised approach without prior coding of text examples. For example, terms such as "game" and "football" are likely to appear more frequently in the topic "sport" compared with terms such as "politics" and "congress" (Blei, 2012). Topic models have

proven to be a powerful analytical tool that is highly suitable for large corpus analyses with multiple topics of interest (Blei, Ng, & Jordan, 2003; Justin Grimmer, 2010; Quinn et al., 2010). Recent developments have enhanced the ability to examine theoretical relations between external variables and corpus topics by incorporating covariant variables into a Structural Topic Model (STM). This model has further enhanced topic models' popularity among computational social science researchers (Roberts et al., 2014).

However, because topic models learn topics inductively instead of being given a list of predefined topics, they are sometimes difficult to use when testing a theory involving specific topical variables, which is the common scenario for theory-driven research (Collingwood & Wilkerson, 2012; Günther & Quandt, 2015; Guo, Vargo, Pan, Ding, & Ishwar, 2016; Roberts et al., 2014). In addition, the outcome can be affected by even small variations in processing steps or in the model's configuration. Therefore, achieving reliable, stable, and reproducible topic models is quite challenging. The problem is aggravated when the corpus is not fixed but continuously expanding, as is the case when collecting and analyzing political speeches, news, and social media during the course of a political campaign (Chuang et al., 2015; Denny & Spirling, 2018; Fokkens et al., 2013; Wilkerson & Casas, 2017). Topic models are also difficult to evaluate, leading to disagreements between researchers regarding the results of their analyses (Maier et al., 2018). All of this complexity compromises the ability of topic models to produce collaborative and replicable scientific results. Some of these limitations could be resolved if it were possible to add topics to an existing topic model. Unfortunately, there is no simple method for performing such an addition (Blei, 2012; Schwartz & Ungar, 2015).

These limitations may drive researchers to use topic models while exploring a corpus and building theory, and then use other, more appropriate methods to identify a given list of categories. One such method is dictionary analysis, in which a set of terms is searched for in the text to identify the corresponding predefined categories (Burscher, Vliegenthart, & De Vreese, 2015; Soroka, Young, & Balmas, 2015). Dictionaries are explicit, transparent, and additive. However, creating a valid dictionary is very costly, and adding categories to an existing dictionary may entail even higher costs, as all other categories should first be considered to prevent contradictions (Quinn et al., 2010). In addition, the accuracy of dictionary analysis may be compromised by the choice of terms, and in general, the method tends to suffer from low recall scores (Guggenheim, Jang, Bae, & Neuman, 2015; Guo et al., 2016). Recent methods have succeeded in reducing the subjective bias that may accompany the manual selection of words, which improves recall, but

these approaches further increase start-up costs for creating a dictionary (King, Lam, & Roberts, 2017).

Supervised learning, in which the computer learns the weight of each term and considers additional features, such as contextual information, usually results in more accurate classifications than dictionary analysis (Cambria & White, 2014; Justin Grimmer & Stewart, 2013). It also facilitates the creation of new categories by simply adding labeled text examples to the training set. As such, this method seems to be the best choice for a text classification designed to accurately identify predefined categories that also provides a more reliable, stable, and reproducible way to update the list of categories.

Despite its advantages, studies in the social sciences usually use supervised learning only to identify a small number of categories because of the high cost of identifying each one (Burscher, Odijk, Vliegenthart, de Rijke, & de Vreese, 2014; Quinn et al., 2010). In some cases, supervised learning is used merely as a filtering mechanism, and the actual in-depth analysis is performed manually (Nardulli, Althaus, & Hayes, 2015). Therefore, even though supervised learning seems to be a natural choice for theory-driven research, its high cost limits its use by social scientists, especially when the tested theory involves more than a few variables.

2 Weak Supervision as an Additive Alternative

To solve this problem, we suggest using a weakly supervised method, which reduces manual labor by splitting the training process into two phases. The methods involve first applying a low-cost labeling method to raw data, which minimizes human labor while creating a training set with labels that are useful despite being incomplete or not fully accurate. The training set is used to train a regular supervised or semi-supervised learning method to create a predictive deductive method (Hernández-González, Inza, & Lozano, 2016; Zhou, 2018). In this way, these methods can reduce the cost of human labor, thus leveraging very large training sets, while providing performance on par with fully supervised learning methods (Hoffmann, Zhang, Ling, Zettlemoyer, & Weld, 2011). Researchers have also demonstrated how weak and manual annotations can be combined to improve models' performances even further, thereby creating new paths for collaborative research initiatives (Deriu et al., 2017).

2.1 Methods for Low-cost Labeling with Predefined Categories

Our approach applies unsupervised learning to a large volume of news articles to compile a training set that is then used to train a separate supervised classifier. It is true that other low-cost methods can be used as alternatives for human coding, such as crowdsourcing (Dehghani, Zamani, Severyn, Kamps, & Croft, 2017; Rudkowsky et al., 2018). However, our method provides better use of the available resources, so that projects with more funding can use crowdsourcing to create a training set, while those with more constrained funding can take advantage of access to experts to verify and interpret the outcomes of the unsupervised learning method. We believe one of the reasons for the popularity of topic models in the computational social sciences, and specifically in communication studies, is that many social scientists have more access to experts than they do to funding. Additionally, the specifics of a particular research project can make crowdsourcing less attractive. In a pilot study we performed with a group of six undergraduate coders, it took approximately three months of manual labor to compile a dataset of 10,000 labeled sentences with reasonable inter-coder reliability for less than twenty categories. In a case such as ours, which was likely to entail a larger number of categories, the scale of the coding labor required made crowdsourcing infeasible.

Another weakly supervised learning alternative asks experts to define simple rules, or labeling functions, that are applied for labeling a large number of texts and, by doing so, create the training set (Ratner et al., 2020). This approach is a good fit for some studies, especially when there are available human experts who wish to identify a predefined list of categories, with a clear understanding of the characteristics of each category. However, in many communications studies, researchers might prefer to let the categories inductively emerge from the text, and therefore use topic modeling as a starting point. As a result, and for its simplicity, topic modeling, which is by far more common than crowdsourcing, offers a more accessible starting point. Also, asking experts to define rules might be costly and complex, depending on the potential number and nature of categories. The higher this number is, the higher the cost would be. Defining such rules to identify a large number of thematic categories, for example, might result in an extremely complex task.

2.2 Our Approach: Low-cost Labeling with Inductively Defined Categories

We, therefore, used unsupervised learning as the first step of our weakly-supervised topic analysis method. More concretely, we used topic models

as the unsupervised method. Our unique contribution here is the transformation of the output of the topic models into a labeled training set (as described in section 3.3), which renders the use of a weakly supervised solution practical. Using first topic modeling and then supervised learning allowed us to enrich topic models with additivity, namely the ability to add categories to the training set and train the supervised classifier to identify existing and new topics. We could thus gain from one of the beneficial characteristics of supervised learning. Adding categories to a training set is not cost-free, as adding any new category should consider existing ones so as not to duplicate categories in the same training set. Nevertheless, this process is usually simple enough, as there is no need to repeat the human labor already invested in the previous version, that is, with no need to retrain and relabel the original topic models (see the demonstration in section 5).

2.3 Method Characteristics and the Advantages of Additivity

Our approach would be especially beneficial in the following three scenarios. Firstly, when a researcher is interested in identifying a large number of categories, e.g., when the researcher's interest is broad. In this case, our approach can serve as a low-cost alternative for supervised learning. Unlike supervised learning, in our approach, the researcher does not control the categories that emerge by the topic model. However, the ability to train additional topic models until the desired categories are identified plays an important part. Our method allows the researcher to include all topic models (old and new) in the same training set without the need to retrain and relabel models. This may compensate for the inability to manually define the list of categories.

A second scenario where we expect our method to be beneficial is when a researcher initiates a study with an inductive exploration of a corpus, for example, news articles regarding US politics in an election year. After this exploration, the researcher may consolidate a theory based on this exploration and then wishes to test this theory in a broader context, such as a non-election year or local newspapers covering local elections. In this case, our approach enriches the unsupervised learning method with the virtues of supervised learning – replicability, validity, and, as detailed above, additivity.

Lastly, researchers are sometimes interested in collaborating between separate research projects. This is true for internal collaboration between different projects within the same lab and for external collaborations between researchers from different labs who are willing to put their resources to better use and expand their theoretical premises. For example, say a

Table 1: Comparison of Computational Content Analysis Methods

	Characteristic	*Lexicon*	*Supervised learning*	*Topic modeling*	*Crowdsource-based weak supervision*	*Labeling-functions-based weak supervision*	*Topic-modeling-based weak supervision*
Research Goals	Categories definition	Deductive	Deductive	Inductive	Deductive	Deductive	Inductive
	Replication	Easy	Easy	Hard	Easy	Easy	Easy
	Validation	Medium	Easy	Hard	Easy	Medium	Medium
	Addition	Medium	Easy	Hard	Easy	Medium**	Easy
	Preferred research type	Confirmatory	Confirmatory	Exploratory	Confirmatory	Confirmatory	Both
Resources	Human resource	Experts	Experts	Experts	Non-experts	Experts	Experts
	Corpus size (for training)	*	Small	Medium	Medium	Large	Large
	Cost per category	High	High	Low	Medium	Medium	Low
	Cost per size of training set	*	High	Low	High	Low	Low

Note: The table presents an evaluation of the goals each method supports, and the resources it incurs, and it is partially based on (Quinn et al., 2010; Roberts et al., 2014).

** The lexicon approach does not include a training set.*

*** Adding categories to labeling-functions-base weak supervision might be hard, as this approach resembles the creation of a lexicon, yet this aspect depends on the type of categories.*

researcher has two projects with shared characteristics, such as the political debate in the US regarding two separate issues – gun control and taxes. Training a separate model for each project is a reasonable starting point. Then, it might also be beneficial to combine the two models into a larger one, to support even broader research interests. At the least, it would serve as a starting point for a model that already includes some categories of US politics, public spending, and civil rights, for other research projects.

Such collaboration is clearly possible when using supervised learning. However, it is not really feasible for unsupervised and inductive studies, and therefore we are not aware of such collaborations. Our approach is capable of overcoming some of the major barriers to such collaborations. Starting from a commonly used method such as topic modeling, with the ability to merge multiple topic models into the same method, makes this collaboration more practical and accessible. Table 1 summarizes the characteristics of the main approaches mentioned above to simplify the comparison. It can also assist in the selection of the appropriate approach given concrete research goals and resources.

In all, our method should enable a text classification of a large number of categories that are inductively defined, with minimal cost, and at the same time support theory testing with mechanisms for replication, validation, and the addition of topics. In other words, it enables the adding of topics to an existing model without compromising the models' ability to identify existing categories, and without the need to define the categories and manually label text examples from scratch. We provide evidence of the reduction of costs and the additivity offered by our method in section 5. In section 6, we show how the validation of such a model could be easier compared to a topic model with a large number of categories.

3 Training Set Compilation

Our solution is composed of two main phases: first, compiling a training set, and then, training a supervised learning classifier. The training set compilation phase consists of the following steps: (1) collect a corpus of texts (news articles in our case) that belong to a single general subject (e.g., crime, sports); (2) train a topic model at the article level for each corpus; (3) convert topics from the article level to the sentence level; (4) create clusters of sentences based on topic association scores; (5) manually label the clusters; and (6) add the labeled sentences to the training set (see Figure 1 for a schematic overview of the process). In the following, we describe the

process in detail, while illustrating it with our example case of a large-scale topic analysis of news articles.

3.1 Collecting Articles for Single-Subject Corpora

We envision a common scenario in which a researcher collects multiple corpora, each relevant to a single general subject (an area of interest) that the researcher wants to divide into more specific categories (e.g., breaking down politics into subcategories of elections, policy, and political campaigns). Also, we found it technically preferable to train a topic model on a collection of news articles relevant to a single general subject because such a corpus makes it easier to identify and label topics. To demonstrate the training set compilation method, we collected articles from the LexisNexis archive from January 1995 to March 2017, starting with a list of approximately 700 news sources (see the Supplementary Materials). For each general subject, we identified the names of substantially similar newspaper sections (e.g., economy, markets, and finance) based on the collaborative judgment of three experts. We then collected all of the articles found in these sections, without any filtering.

3.2 Training a Topic Model at the Article Level for Each Corpus

Before training each topic model, we performed standard preprocessing on each corpus: cleaning; lemmatization; and the removal of punctuation, stop words, common and rare terms, and short texts (for a thorough explanation of these steps, see, for example, Jacobi, van Atteveldt, & Welbers 2016). We then estimated the number of topics based on the size of the corpus (generally 25 to 100 topics). Finally, we trained several Latent Dirichlet Allocation (LDA) topic models (Blei et al., 2003) until the human coders were satisfied with the results at the labeling step (as described in section 3.5).[1]

3.3 Converting Topics from the Article-Level to the Sentence Level

Mixed-membership topic models such as LDA or STM have a useful advantage—their fit for analyzing news articles since those articles are more likely to contain multiple topics compared to other texts. However, this feature also creates a challenge, as labeling and validating such topic models by reading entire articles is difficult when a researcher cannot exactly identify the section of the article that expresses a specific topic (Maier et al., 2018).

For example, in our demonstration, we trained a topic model on a corpus with "crime" as the general subject. When we conducted an article-level analysis of the distribution of topics for a given article entitled "Police: Man

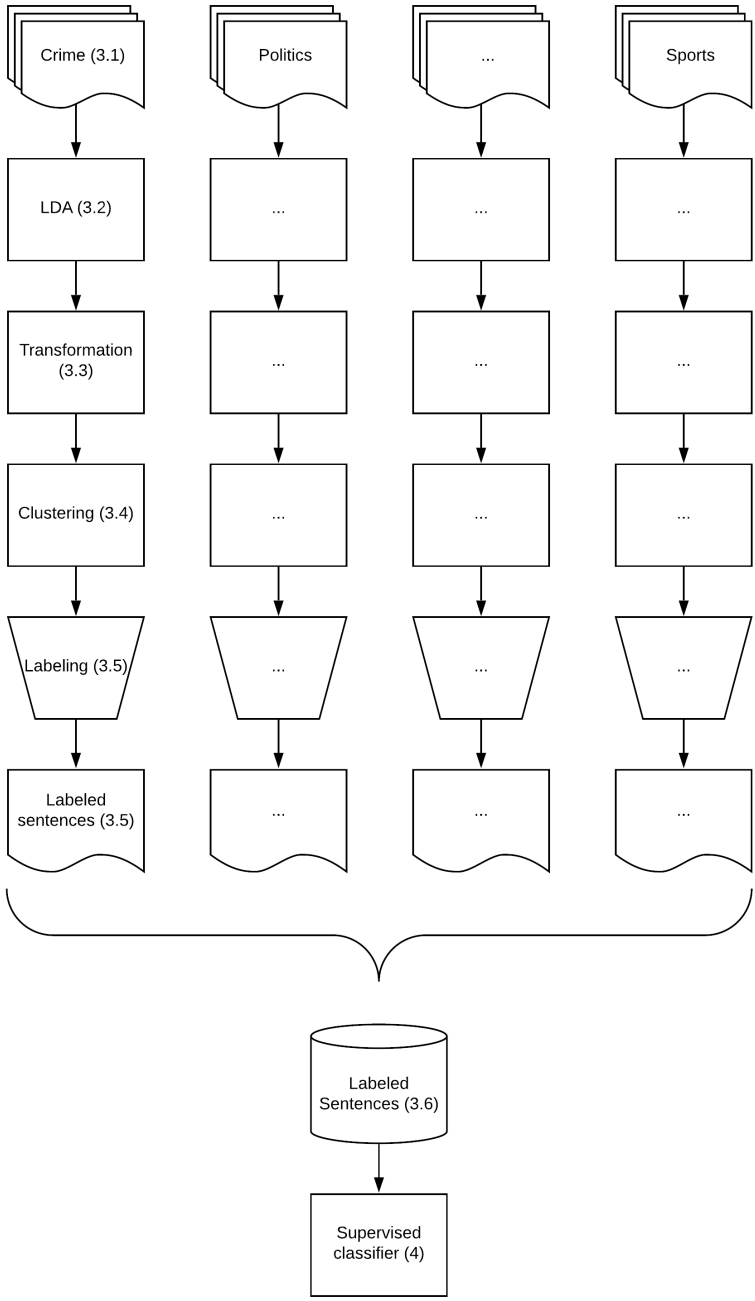

Figure 1: Schematic Representation of Process

Note: Process overview (section numbers in parentheses).

arrested in Waterloo police chase sold heroin, crack cocaine," the two topics with the highest percentages were topic #5 (22.5%) and topic #32 (18.6%). Because the percentages were quite similar, it was difficult to conclusively identify a single main topic for the entire article or clearly identify which section of the article discussed each topic.

Compared to articles, sentences tend to be more focused and hence associated with fewer topics. This makes them much easier to label manually and to use in training a supervised algorithm (Leetaru & Schrodt, 2013). However, we must consider that two sentences with similar content can have different meanings, depending on the context of the article, among other parameters. We thus began our analysis at the article level and then moved to the sentence level before we labeled topics. This allowed us to use the rich contextual information at the article level to train the topic model before moving to the sentence level.

Next, we calculated a "topic association score," representing the level of association between sentence *s* and topic *k*. For each topic, the topic association score considers both the broader context of the distribution of topics at the article level and the specific content of the distribution of each sentence's words over the vocabulary.

Formally, the topic model results in a distribution of topics (Θ_d) for each document *d*, a probability of topic *k* occurring in document *d* ($\vartheta_{k,d}$), and a probability of word *w* occurring in topic *k* ($\varphi_{k,w}$). For each sentence *s*, we calculated a topic association score ($TA_{k,s}$) using equation (1). For each topic *k* in the distribution of topics in the document (Θ_d), we multiplied the proportion of topic *k* in document *d* ($\vartheta_{k,d}$) by the sum of the values of the corresponding *phi* for each word *w* in the sentence ($\varphi_{k,w}$):

$$(1)\quad \mathrm{TA_{k,s}} = \theta_{\mathrm{k,d}} * \sum_{\mathrm{w\ in\ s}} \varphi_{\mathrm{k,w}}$$

This results in better differentiation between topics at the sentence level because, instead of a single distribution of topics constant throughout the entire article, each sentence receives different topic association scores based on its specific content (see the follow-up example in section 3.5).

3.4 Clustering Sentences Based on Topic Association Scores

The goal of the training set compilation phase is to replace the manual labeling of individual sentences, which is an extremely labor-intensive task. We achieved this goal by creating clusters—automatically created groups of sentences—that could be labeled collectively. To this end, we shifted our focus from sentences to clusters, with each cluster corresponding to a topic

in the topic model. Instead of taking into account the topic association scores assigned to specific sentences, we were interested in collecting sentences with the highest scores for each topic. We first computed the topic association scores for sentences from the entire corpus. Then, for each topic, we extracted all sentences with a standardized topic association score above two (that is, the top 5% from all sentences), which we used as a minimum threshold for creating sentence clusters. These groupings were then reviewed by the human experts.

3.5 Labeling the Clusters Manually

Human experts played three roles during the training set compilation phase. First, they judged whether the topic model resulted in "good enough" clusters in terms of clarity and coherence. If not, we reconfigured and retrained the topic model. Once the clustering was considered to be good enough (usually within the first or second attempt), the human experts inferred a label for each cluster by manually reviewing a random sample of sentences. To ensure that the sentences were read in context, we provided the experts with the entire article and title for each sentence. To ensure that the sentence clusters were coherent, we asked the experts to establish an exact threshold for each cluster. This unusual use of human coding was done by arranging the collected sentences in five groups based on their standardized topic association score (2–2.5, 2.5–3, 3–3.5, 3.5–4, and above 4). We then asked the experts to indicate the exact threshold for each topic that would provide a coherent cluster. We required the label and threshold to correspond to each other since a more general label might lead to choosing a lower threshold, which would include more sentences. For example, consider a case where sentences with the highest topic association score for a given topic are all related to US-Russia relations, while sentences for that same topic but with lower association scores are also related to US-Mexico relations. The human experts were responsible for deciding whether to choose a higher threshold and a narrower label, such as the topic "US-Russia relations," or to choose a lower threshold and a broader label, such as "US foreign affairs." This process sometimes required several discussions and iterations until the experts agreed on both the label and threshold.

We now turn back to the example of the news article presented in section 3.3, describing a drug dealer in Waterloo who caused a car accident while fleeing from police. The original topic model we trained operated at the article level with two main topics. Yet, the goal of the training set compilation phase was to identify texts that were more strongly associated with each topic. When we shifted our focus from the article level to the sentence

level, the picture became clear. Sentences involving drivers and vehicles received higher scores on the first topic (#5), while sentences involving drugs received higher scores on the second topic (#32—see Table 2). After reviewing a sample (N=~100) of sentences with high topic association scores for each topic collected from the entire corpus, the human experts assigned the label "Crime—Drivers & Vehicles" to topic #5 and "Drugs" to topic #32.

The manually inferred label was then propagated to all sentences within each cluster. Therefore, unlike traditional methods of manual labeling done with supervised learning, the human experts only reviewed a small fraction of each group of sentences, but the label they inferred was assigned to a much larger group of similar sentences.

Table 2: Labeling Categories by Reviewing Sentences

Sentence Text	*Topic Association Scores*	
	Topic #5	*Topic #32*
"He allegedly refused to stop, and intentionally crashed into an unmarked Sheriff's vehicle, causing damage and a hand injury to a deputy."	**1.64**	0.03
"McCullough caused damage to the field with the vehicle, and became stuck in mud."	**1.12**	0.01
"Seneca County Sheriff's deputies announced additional charges Thursday for a Rochester man allegedly connected to selling illicit drugs in the area."	0.34	**1.88**
"McCullough was charged with two counts of third-degree criminal sale of a controlled substance, two felony counts of third-degree criminal possession of a controlled substance, two counts of sale of an imitation controlled substance."	0.12	**2.08**

Note: Example of labeling categories using topic association scores for sentences.

At this point, we separated the training set compilation phase from the training of the supervised learning classifier by imposing two rules. First, we focused our attention on a sentence's binary association with each category, rather than its actual topic association score. This decision made it easier for us to compile a training set of labeled sentences. Second, some words, such as stop words, were removed during preprocessing and were therefore not given a *phi* value by the topic model and did not contribute to their sentence's topic association score. However, the clustering and manual label inference were performed using the original sentences, including all words, so that the supervised classification method would be able to take them all into account.

3.6 Adding the Labeled Sentences to the Training Set

We aggregated the labeled sentences into a single training set. In cases where the label of a topic learned by one topic model overlapped with the

label from another topic model, we merged both groups into a single group of sentences with one label.

The purpose of the entire process is to train a supervised classifier, and therefore validation should focus on the supervised classifier, while the training set is assumed to contain noise. Nevertheless, in a previous study, we evaluated the correctness of the training set compilation phase by manually reviewing labeled sentences. This evaluation validated the clustering method and the labels given to clusters, and, as a by-product, helped to train our human coders and to fine-tune the process. We, therefore, recommend researchers who are interested in applying the process to conduct this evaluation, which we describe in more detail in Online Appendix 1.

4 Designing and Training a Deep Learning Sentence Classifier

We now turn to the second phase of our weakly supervised method, in which we used the compiled training set to train a supervised deep learning classifier. A deep learning model usually outperforms classical learning models, as it can learn how to efficiently represent raw data using its hidden layers (dos Santos & Gatti, 2014; Lai, Xu, Liu, & Zhao, 2015). Unfortunately, deep learning models also usually depend on large amounts of data, sometimes millions of labeled examples (LeCunn, Bengio, & Hinton, 2015). This is likely one of the most significant barriers to using deep learning in the computational social sciences, especially when the goal is to identify a large number of categories. Yet it is also where we gain the most benefit from the low-cost, unsupervised compilation of the labeled training set. Our design of the supervised classifier may not be optimal (many other designs can be used as alternative methods of supervised learning), but it provides a valid demonstration of a sufficient method. In the interest of concision, we provide only a brief description of the classifier. (For detailed explanations, see Online Appendix 2. We recommend that researchers who are new to the field of deep learning review this appendix before reading the next section).

4.1 Preprocessing Sentences

Because deep learning models automatically learn how to represent raw data, the preprocessing of text input varies somewhat from classical machine learning techniques. Instead of removing stop words or symbols from the text (Lai et al., 2015), we used only the Stanford CoreNLP

tokenization tool (Manning et al., 2014) and converted the tokens to lower case. Finally, we removed sentences with fewer than five tokens, assuming they did not contain enough information regarding the relevant category.

4.2 Model Architecture

In our architecture, sentences are represented by a fixed-length vector. To allow the model to analyze the complete sentence, we chose a length of 100 words (including punctuation marks), which covers more than 99% of the cases (based on a sample of 10 million sentences). Shorter sentences are padded with zeros at the beginning, which the model ignores. The model's input layer then embeds the words of these fixed-length sentences into a vector representation, based on GloVe pre-trained vectors (Pennington, Socher, & Manning, 2014).[2]

We added a long short-term memory (LSTM) layer to allow the model to learn from sequential information (word order) and multiword patterns (Bengio, Courville, & Vincent, 2012; Hochreiter & Schmidhuber, 1997; Lai et al., 2015). The LSTM layer was configured to contain 100 memory units so that an entire sentence could be stored in memory simultaneously. To reduce the risk of overfitting the training set, we added a dropout regularization method, configured with a rate of 20% for the input and the recurrent features of the LSTM layer (Gal & Ghahramani, 2016; Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014).

We experimented with different architectures to classify sentences based solely on the text but did not achieve a reasonable level of accuracy. This finding was consistent with our understanding of discourse, whereby the same sentence may have different meanings in different contexts. As a simple solution, we added the article's title as a contextual input to the model and duplicated the embedding and LSTM.

We concatenated the output of the two LSTM layers into a 200-length vector. The vector was fed into a fully connected network with a number of modules equivalent to the number of categories plus 30, with a "ReLU" activation function (Krizhevsky, Sutskever, & Hinton, 2012; Nair & Hinton, 2010). This layer was connected to the output layer with the same number of modules as the number of categories.

Even though a sentence is usually more focused than an entire article, it still may refer to more than one category, especially when the categories are not mutually exclusive. In fact, sentences in the political domain commonly contain multiple topics (consider, for example, sentences from political debate on public health spending).

We, therefore, designed the model's output layer to predict a multilabel classification, such that multiple categories may be predicted for each sentence. To this end, the layer minimized a weighted binary cross-entropy loss with a sigmoid activation function (Kurata, Xiang, & Zhou, 2016; Nam, Kim, Loza Mencía, Gurevych, & Fürnkranz, 2014). This loss function creates a multilabel classification by separately providing the probability for each category to be true. All categories with a greater than 50% probability of being true were marked as identified. In the end, we used Adam optimizer to minimize the loss function (Kingma & Ba, 2014).

4.3 Training the Sentence Classifier

Once the choice of layers and the individual layer sizes (number of modules) were set, we tuned the hyperparameters. To reduce the risk of overfitting, which can occur during the selection of the best hyperparameters, we split the sentence-level data into three sets: training, validation, and test. We trained the model with different hyperparameters using the training set, chose the best configuration based on the accuracy calculated on the validation set, and tested the accuracy of the final model using the test set (Justin Grimmer & Stewart, 2013). We also halted training before any degradation of performance on the validation set, after two epochs (training-cycles) (Srivastava et al., 2014).

5 Adding Topics Iteratively

One of the advantages of supervised learning is the ability to add more categories to the training set by adding text examples labeled with new categories. Typically, a researcher will simply add manually labeled text examples for each new category. Alternatively, the researcher can conduct additional iterations using our method: collect an additional corpus with a general subject, train a topic model, convert its outputs to clusters of sentences, infer a label for each cluster, and add the sentences with the new labels to the training set.

5.1 Illustrating Additivity

An interesting possibility is to decompose one of the existing categories into more specific ones. For example, we trained a version of our supervised classifier using sixteen different corpora, as described in section 3. This version (henceforth referred to as Version 15) was trained to identify 450 categories and enabled research projects focused on the economy and politics.

One of the categories was guns and gun control in the United States, which we extracted from corpora collected using the names of newspaper sections, such as politics and crime. For the sake of this illustration and to allow the collaboration with a separate project in our lab that focuses on the issue of gun control and the use of guns in the US, we wanted to decompose this category into more focused categories. We used the trained supervised classifier to identify news articles in the category of guns and gun control in the US (e.g., all articles in which this category was identified by more than 10% at the article level). By doing so, we created a new corpus that was focused on gun control in the US and used it to train an additional topic model. The ability to identify more nuanced categories can enable further empirical research on this topic.

Another example was a research project focused on the news coverage of political campaigns in the US. For this project, and also to demonstrate the addition of more nuanced contextually relevant categories, we have repeated this process and decomposed two additional categories, "elections & primary campaigns" and "conflicts" (a general category consists of various kinds of conflict), which we considered likely to be relevant to these two projects. We have also collected another corpus from the opinion sections of various newspapers to add more diverse perspectives on potentially relevant political issues. After training a topic model for each corpus and running the rest of the training set compilation method, we labeled the resulting groups of sentences.

At this point, we had to ensure the coherency of the training set, which is always a challenge when adding categories. Consider a group of sentences that was extracted in this newer version. The human coders were asked to label this group while considering all labels that were already included in the training set in the previous version. If the human coders inferred this group of sentences with a label that was similar enough to a category already included in the training set, they needed to judge whether to merge both groups of sentences by assigning the new group with the existing label. By doing so, the model would have additional text examples for the same category, which should translate into better performances. However, assigning the group of sentences with a new and different label would end up in the model identifying two categories that are similar or even practically identical, instead of one. In this case, one can expect a decrease in performances, as the same category would now be split between two topics.

Another concern is that differentiating one category over another is not straightforward. In fact, it is often a theoretically driven question, which should be adapted to specific research questions. For example, it is up to

the researcher to decide whether to separate between the categories of trials and verdicts or between similar types of political scandals or campaigns from different periods.

These concerns are not unique to our approach and may accompany the compilation of any training set with a large number of categories. However, we applied some measures to reduce this risk. First, we made sure that at least one expert was involved in, or at least supervised, all labeling efforts. Second, we leveraged the sample of sentences used for the labeling (see section 3.5) to help coders decide whether two groups of sentences belonging to the same category. Third, we kept track of the model's performances and looked for categories that showed a decrease in performance between the two versions. Last, the researchers who led the various studies were involved in the labeling effort and provided some theoretical guidance regarding the desired boundaries between close categories.

After labeling the groups of sentences created with the training of the four topic models, we added the resulting labeled sentences to our training set.

Table 3: Collected Corpora

General subject ("Context")	*Articles*	*Topics*	*Labeled Sentences*
Economy	11,002,527	75	17,066,574
Education	281,716	50	710,898
Elections & Primary Campaigns*	300,205	50	1,144,320
Energy & Natural Resources	100,435	50	291,640
Guns & Gun Control in the US*	25,707	25	94,396
Health	381,093	50	2,494,971
Immigration	13,767	25	28,384
International	4,433,328	75	14,647,669
Legal, Crimes, & Police	949,554	50	16,639,412
Mideast & The Arab World	107,031	75	1,608,840
National Elections & Political Conflicts*	2,129,710	100	8,975,413
National Security	190,136	50	917,377
Opinion	5,000,000	100	1,222,735
Politics	953,437	50	24,121,219
Science	113,954	50	551,606
Sport	5,000,000	75	1,761,938
Technology	200,303	75	3,956,669
Tourism	688,952	50	5,195,551
Transportation & Vehicles	305,683	50	1,369,054
Weather	147,198	50	582,371

Note: Collected corpora, each with a general subject, were used to train LDA models with the corresponding number of topics.

** The general subject was a topic identified by a previous version of the model. Other general subjects were defined using newspapers' section names.*

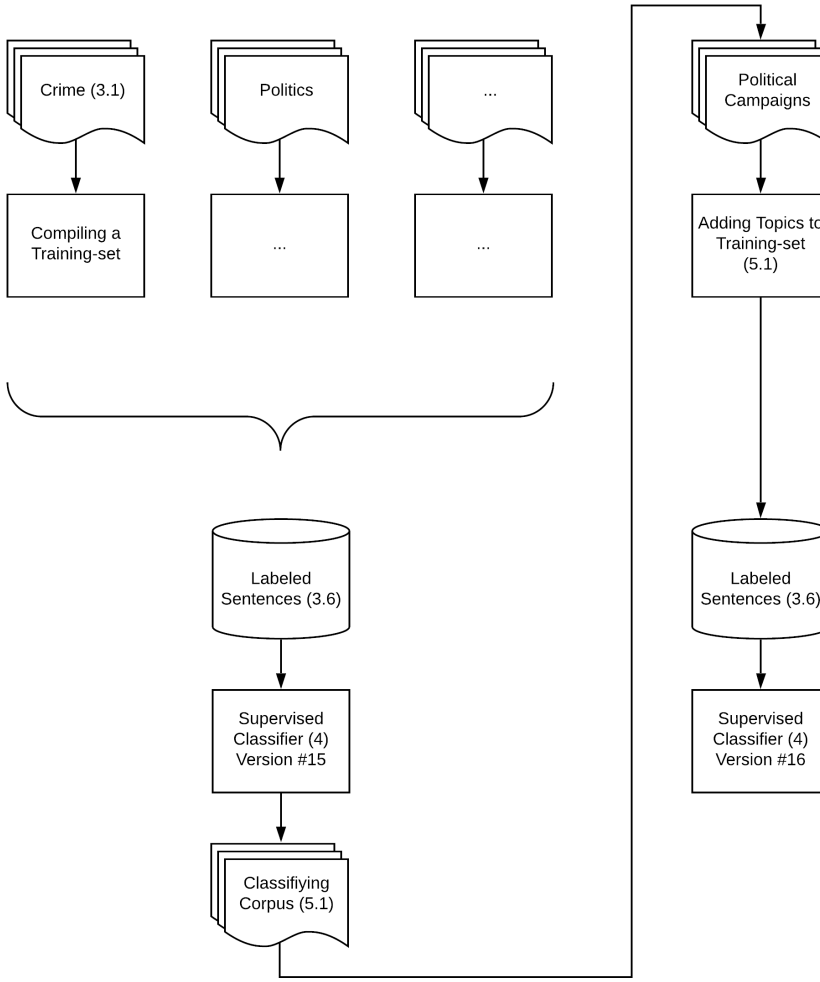

Figure 2: Adding Topics to the Model by Decomposing Existing Categories

Note: Overview of the process of adding topics to the model (section numbers in parentheses).

Combined with the illustration of the original training set used for Version 15, we analyzed 20 corpora containing approximately 30 million articles, resulting in a training set containing about 100 million sentences labeled with a total of 651 topics (see Table 3). We used this training set to train a new supervised classifier (referred to as Version 16). The entire process is illustrated in Figure 2.

The training set compilation phase, including the additional 201 categories (from a training set of 100 million sentences), required approximately

400 work hours by human experts who were tasked with inferring a label and setting a threshold for each cluster. To compare it with standard labeling of a topic model, these numbers are translated to an average of 20 work hours per topic model, each with about 50 to 100 topics, including the assessment of the topic model's quality, labeling of sentence groups, and setting up a threshold for each. We believe this amount of labor is not higher, at the very least, than the cost of using topic modeling on a single corpus.

Another alternative is supervised learning. Before evaluating the human effort demanded by our method, we consider that there was some overlap between categories extracted from different topic models. In these cases, a single label took twice the labeling effort. Yet, even when counting only the final labels, each category demanded less than two hours on average. Compared to what we have seen in our pilot study mentioned above, the same total amount of work of directly labeling sentences resulted in a dataset of only 10,000 labeled sentences for less than 20 categories. Putting aside these smaller number of categories and size of the training set, and the additional cognitive effort that it would take to label hundreds of categories, even with only 20 categories, each demanded more than 50 work hours on average. Considering the cost per size of the training set we compiled in this paper, even when ignoring the much higher number of categories, our method demanded 2.4 minutes to complete the labeling of 10,000 sentences, compared to the 400 hours of manual labeling.

5.2 Testing Additivity

To test the additivity of our model, we performed reliability tests at the sentence and article levels. In both tests, we compared the classifications made by the two versions of the model: Version 15, with 450 categories, and Version 16, with 651 categories.

For the first test, we compared the classifications made by the two versions on a reserved test set of 5.99 million sentences, sampled from the compiled training set. As we do not have a gold standard based on human coding that would have allowed an external verification of accuracy, we treated the two versions as two coders and tested their inter-coder reliability. We did not expect complete agreement, since any addition of categories to the model could affect the model's classification of other related categories. Our expectation was that the agreement between the two versions would be sufficiently high to indicate the stability of the model despite the addition of new categories.

The comparison of the versions shows high levels of agreement for most categories. The weighted average of Krippendorf's α was .79 (see the

detailed inter-coder agreement scores per category in the Supplementary Materials).

For the second reliability test, we analyzed a corpus of 1.8 million news articles from *The New York Times* published between January 1995 and July 2017 to compare the results of the two versions at the article level (we expect this to be the more common use case).[3] To do so, we aggregated the classifications of sentences into categories determined by the classifier and applied them to the article level. Categories were assigned a percentage at the article level based on the number of sentences in which it appeared (see Online Appendix 3 for the details about this aggregation process).

We compared the classifications made by the two versions at the article level in two ways. We first measured the Krippendorf's α and found that the alpha's weighted average was again high (α=.76). We also measured the correlation between the results using Pearson's r, which showed a strong correlation (weighted average r=.81).

6 Validation

Supervised methods offer a direct method for evaluating model performances by comparing the results of the classification method with a test set reserved before the training phase. We, therefore, evaluated our model using this test set, which resulted in accuracy measures for every category and on average. Then, as our solution is weakly supervised, we added more validations that are more common in unsupervised learning.

6.1 Direct Assessment of Model Performances

We began the validation process using the held-out test set of about 6 million labeled sentences, in which most (95.1%) were originally labeled with a single expected category during our training set compilation phase. After classifying the test set with our trained classifier, we identified multiple categories per sentence in most cases (80.2%), although this number was usually small (M=2.45, STD=1.13). To evaluate performance, we counted every classification as a true positive if one of the identified categories was true according to the test set.

The model achieved satisfactory levels of accuracy. Given the nature of the test set and the fact that new topics were added without updating previous existing examples in the training set, we only have information regarding one expected label for each sentence in the test set in most cases. We do not know, for example, if the sentence is also relevant to categories

that were added to the dataset later (as they did not exist at the time the sentence was added to the test set). We, therefore, do not have information regarding false positives (the model falsely identifying a category when it should not have). We have information only regarding false negatives (the model failing to identify a category when it should have) and true positives (the model succeeding in identifying an expected category).

Following this step, we calculated recall scores per category (the number of true positives divided by the sum of true positives and false negatives). As we did not have information about false positives, we did not calculate the equivalent precision scores (the number of true positives divided by the sum of true positives and false positives) (see the Supplementary Materials). We also have the overall number of true positive cases (where at least one of the identified categories was the expected one) and the overall number of assumed false positives (where none of the identified categories was correct, so we assume the sentence was falsely identified). We, therefore, calculated the average precision ($Precision_{mean}$=75.6%) and the weighted averaged recall ($Recall_{mean}$=75.7%).

These results are consistent with acceptable levels of accuracy despite the high resolution of the unit of analysis (i.e., sentence) and the large number of identified categories (N=651) (Justin Grimmer & Stewart, 2013). This finding suggests that our model provides a valid method for conducting a weakly supervised analysis of a large number of categories.

6.2 Semantic Validation

Usually, weakly supervised learning models are evaluated by comparing their results with a benchmark dataset containing similar categories. As the discipline currently does not possess such a dataset, and our label definitions may differ from those of other researchers, we followed some of the validation steps used when validating topic models (e.g., Barberá et al. 2018). We provide two datasets that may reassure researchers that the model's assumptions and predictions match its theoretical premises. First, we provide the test set used for the additivity test (section 5.2) in the Supplemental Materials. Each row in the test set contains the tokens of the title of the article and the sentence analyzed in the training set compilation phase, the expected label attached to the sentence during this phase, and the labels predicted by the two versions of the model.

Second, we provide a sample dataset of news articles analyzed during the additivity test. To create this dataset, we collected a sample of articles (up to 100) where more than 10% of the article was applicable to the category (this dataset contains only article titles, publication dates, and

LexisNexis identifiers to allow for replication without violating copyrights). Although this is a relatively low threshold (in some cases, only two or three sentences were classified into the category), it is usually sufficient to get a sense of the article's main topics, which can then be validated using its title. Also, to enable a more in-depth examination of these results, we provide similar results for this dataset at the sentence level to show the exact classifications made by our model.

6.3 Predictive Validity

The process of assessing the model's predictive validity is less time consuming than the process of semantic validation and tests whether the model is well correlated with external events for selected categories. Such an approach to validity (Quinn et al., 2010) requires a relatively consensus-driven and clear timeline of events to compare with to measure the precision and the recall of the model, i.e., to ensure the predicted spikes in the category's timeline are related to relevant events and that the model did not miss any major events. Here, we illustrate the predictive validity of four categories.

To perform this test, we analyzed a corpus of news articles from *The New York Times*. We aggregated the resulting classifications by averaging

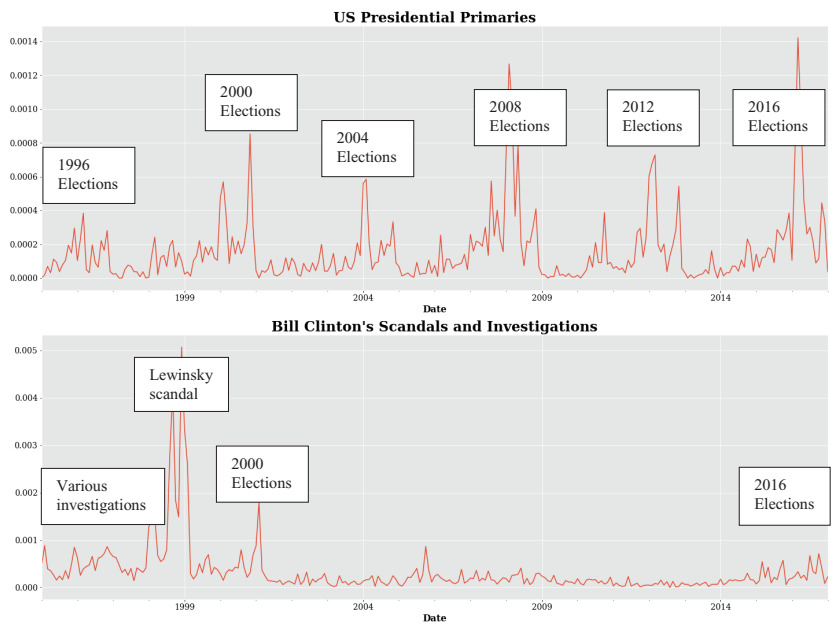

Figure 3: Predictive Validity Over Time

Note: The Y-axes represent media attention to a category per month.

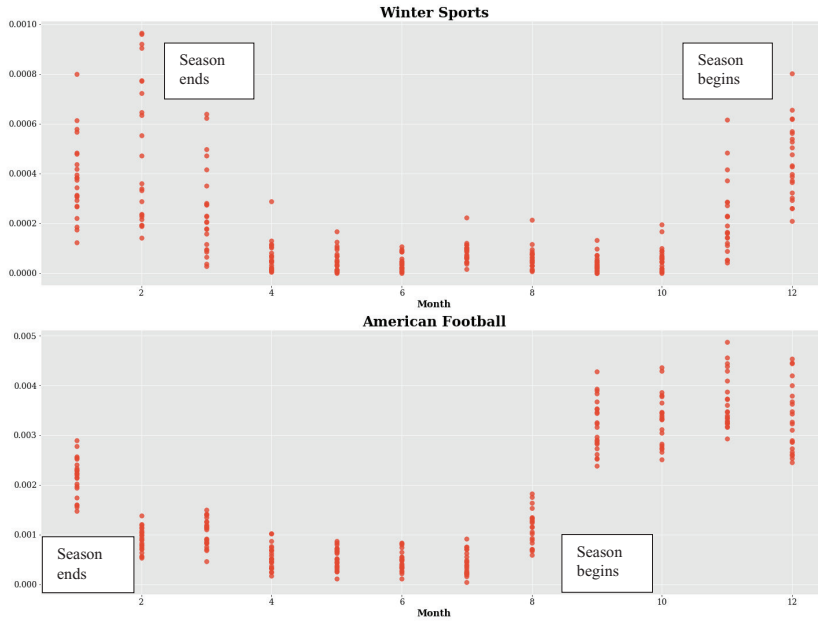

Figure 4: Predictive Validity by Seasonality

Note: The Y-axes represent media attention to a category per month. The X-axes represent the same month in every year.

their scores from the article to the daily level, and from the daily to the monthly level. This resulted in measurement for monthly media attention per category.

Figure 3 shows two categories representing specific events with a relatively easy-to-define timeline. The upper chart shows the presidential primary elections in the United States, which occur every four years. As expected, we can see a repeating lower attention / higher attention sequence: when an incumbent president is running for office, his or her victory in the primary election is almost certain; therefore, it attracts less attention.

The lower chart shows a category representing scandals and investigations related to President Bill Clinton. The main spike clearly indicates the Lewinsky scandal, with additional smaller spikes during the 2000 and the 2016 elections (when Senator Hillary Clinton ran for office). The small spikes before 1998 called for a closer examination. We filtered all pre-1998 news articles that our method labeled with this category and reviewed their titles. This analysis showed that these news articles did, in fact, deal with various investigations relevant to President Clinton (see the full list in the Supplementary Materials).

Another type of predictive validity is illustrated in Figure 4. The figure shows the monthly media attention paid to two seasonal categories, for which we can expect to find an annually repeating pattern. To show this cycle, we collapsed the 23 years of data into one calendar year, in which each data point represents a single year-month of data. We used sports categories as exemplars of expected periodical patterns—the categories of United States winter sports and American football—under the assumption that these categories will correspond to the seasonal calendar. The upper chart shows the category of winter sports, which are much higher during the winter months in the United States than the rest of the year. The lower chart shows the category of American football. This category also follows the expected periodic cycle, representing the beginning of the season in September and its end with the Super Bowl in late January or early February.

7 Conclusion

Labeled datasets are the most important component for advancing the process of automatic meaning-making. However, researchers struggle to gain access to labeled texts. In this paper, we offer a very effective and efficient method for generating labeled texts and show how researchers can use it for large-scale text analyses. The method proposed in this paper benefits from advances made in topic modeling to develop a low-cost method of topic analysis that meets the needs of theory-driven research: a collaborative, reusable, and additive method.

Throughout the training process, we used three types of topic analysis methods, each of which defines topics slightly differently. We first utilized topic models, which define topics as distributions over the vocabulary. We then converted the outputs of the topic models to topic association scores and created clusters of sentences, each representing a category. Last, we labeled these clusters and aggregated them into a training set, which we used to train a weakly supervised classifier that calculates the weights of features to predict each category based on the entire training set.

We do not claim that a topic originally identified by the topic model is identical to its corresponding cluster of sentences; the results of the supervised classifier might be identical to the results of the topic model but are not necessarily so. Although this might be considered as a potential pitfall of combining unsupervised and supervised learning, we find that this process is nevertheless well suited to our goals. Specifically, the combination of unsupervised and supervised methods allowed us to inductively and

efficiently learn how categories are represented in the news, to add more categories, or to further divide categories, without the need to retrain and relabel the topic model. Our method enables researchers to first explore a corpus using a topic model, where categories are learned and emerge from the text and not given in advance. Then, the researchers may choose to embed their topic model in a topic analysis method with a larger list of supported categories that are known at this point. As such, this approach enables the researchers to test a theory with a specific set of categories, quite similar to supervised learning, but with dramatically reduced costs. We also believe this should allow for collaboration between different research projects and help researchers test more complex theories by incorporating increasing numbers of categories and variables into their theoretical models.

Combining unsupervised and supervised methods entails some caveats. For example, the supervised method might create the illusion that validation through comparison with a test set would be relatively simple. However, this test set was automatically created, and therefore should be treated with care, and the method should be further validated through other means. A second aspect we described and should be performed with care is collaboration. Our illustration is of a collaboration between different researchers within the same lab, which allowed the same human expert to supervise the labeling of the entire training set. This step might not be possible when the collaboration is done between two separate labs. In this case, keeping track of the model's performances might be even more crucial. Reviewing measures for the semantic similarity between categories might also be useful for locating cases of incoherent labeling, but we leave this direction for future research.

A more technical aspect that calls for future research is the preprocessing step. Preprocessing choices should be appropriate to the method and may differ between the two phases of our proposed solution. For example, when training a topic model, it is very common to remove stop words, while training a deep learning classifier does not necessarily include this step. Therefore, removing stop words before training the topic models may lead to a failure to identify a potential difference between related topics (such as two perspectives on the same topic, or two different styles, such as discussing the same topic with different levels of confidence, or from a personal or a collective perspective). In our illustration, we implemented the common preprocessing method used for topic models to show how common methods for training such models could be used. Nevertheless, we believe these aspects worth further investigation and experimentation in future research.

The suggested method is composed of multiple steps, some of which require specific choices of algorithms and configurations. Ours is not the only possible combination, and other clustering methods may be used in place of the one we developed. Our focus in this paper was not on creating a better topic model or even a context-aware clustering method, but rather on showing how such a combination of methods might be used to create a low-cost and additive method for a large-scale topic analysis with a high degree of resolution and a large number of categories. However, we do believe that the use of LDA as the starting point for our solution makes it much more relevant and accessible to researchers.

Compared with our pilot study, in which we manually labeled sentences, the advantages of our proposed approach are very clear. We were able to label over 30 times more categories and 5,000 times more sentences with the same amount of human labor. We achieved this goal by leveraging context both in the compilation of the training set and in the weakly supervised classifier architecture (i.e., by incorporating the title). In addition, the low cost of compilation allowed us to create a very large dataset of labeled sentences, which makes it possible for us to use deep learning as the classification method. Last, our use of multilabel classification at the sentence level also contributed to a more accurate and realistic sentence classification. Given the demonstrated capability of the model to incorporate additional topics and to refine the training set, we believe this approach could be of great use to the discipline.

Disclosure statement

We are reporting that we have a patent (U.S. Patent No. US9772996) that corresponds to some aspects of the research reported in the enclosed paper. Regardless, we will grant a royalty-free license for the academic use of the method described herein.

Acknowledgments

This research was funded by the Israeli Ministry of Science, Technology and Space; The Leonard Davis Institute for International Relations; and Yissum Research Development Company of the Hebrew University of Jerusalem.

Supplemental Materials

The Supplemental Materials could be found at:
https://doi.org/10.17605/OSF.IO/DV8GX

1 In theory, other topic models could be used. We chose LDA as it is currently popular and requires fewer theory-specific assumptions (such as the involvement of a covariate variable in STM).
2 See nlp.stanford.edu/projects/glove/
3 See the analyzed dataset and code in the Supplemental Materials.

References

Barberá, P., Casas, A., Nagler, J., Egan, P., Bonneau, R., Jost, J. T., & Tucker, J. A. (2018). Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review*, 1–19. https://doi.org/10.1017/S0003055419000352

Bengio, Y., Courville, A., & Vincent, P. (2012). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. Retrieved from https://ieeexplore.ieee.org/abstract/document/6472238

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. Retrieved from http://www.jmlr.org/papers/v3/blei03a.html

Burscher, B., Odijk, D., Vliegenthart, R., de Rijke, M., & de Vreese, C. H. (2014). Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis. *Communication Methods and Measures*, *8*(3), 190–206. Retrieved from https://doi.org/10.1080/19312458.2014.937527

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015). Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts? *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 122–131. https://doi.org/10.1177/0002716215569441

Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Computational Intelligence Magazine*, *9*(2), 48–57. Retrieved from https://doi.org/10.1109/MCI.2014.2307227

Chuang, J., Roberts, M. E., Stewart, B. M., Weiss, R., Tingley, D., Grimmer, J., & Heer, J. (2015). TopicCheck: Interactive Alignment for Assessing Topic Model Stability. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Retrieved from http://www.aclweb.org/anthology/N15-1018

Collingwood, L., & Wilkerson, J. (2012). Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods. *The Journal of Information Technology and Politics*, *9*(3), 298–318. Retrieved from https://doi.org/10.1080/19331681.2012.669191

Dehghani, M., Zamani, H., Severyn, A., Kamps, J., & Croft, W. B. (2017). Neural Ranking Models with Weak Supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM* (pp. 65–74). https://doi.org/10.1145/3077136.3080832

Denny, M. J., & Spirling, A. (2018). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. *Political Analysis*, *26*(2), 168–189. Retrieved from https://doi.org/10.1017/pan.2017.44

Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Müller, S., Cieliebak, M., … Jaggi, M. (2017). Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification. In *Proceedings of the 26th international conference on world wide web. International World Wide Web Conferences Steering Committee* (pp. 1045–1052). Retrieved from http://arxiv.org/abs/1703.02504

dos Santos, C. N., & Gatti, M. (2014). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *COLING* (pp. 69–78). Retrieved from http://www.aclweb.org/anthology/C14-1008

Fokkens, A., van Erp, M., Postma, M., Pedersen, T., Vossen, P., & Freire, N. (2013). Offspring from Reproduction Problems: What Replication Failure Teaches Us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 1691–1701). Retrieved from http://www.aclweb.org/anthology/P13-1166

Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*.

Grimmer, J., & King, G. (2011). General Purpose Computer-Assisted Clustering and Conceptualization. *Proceedings of the National Academy of Sciences*, *108*(7), 2643–2650. Retrieved from https://doi.org/10.1073/pnas.1018067108

Grimmer, Justin. (2010). A Bayesian Hierarchical Topic Model for Political Texts : Measuring Expressed Agendas in Senate Press Releases. *Political Analysis*, *18*(1), 1–35. Retrieved from https://doi.org/10.1093/pan/mpp034

Grimmer, Justin, & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Guggenheim, L., Jang, S. M., Bae, S. Y., & Neuman, W. R. (2015). The Dynamics of Issue Frame Competition in Traditional and Social Media. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 207–224. Retrieved from https://doi.org/10.1177%2F0002716215570549

Günther, E., & Quandt, T. (2015). Word Counts and Topic Models. *Digital Journalism*, *4*(1), 75–88. Retrieved from https://doi.org/10.1080/21670811.2015.1093270

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling. *Journalism & Mass Communication Quarterly*, *93*(2), 332–359. Retrieved from https://doi.org/10.1177%2F1077699016639231

Hernández-González, J., Inza, I., & Lozano, J. A. (2016). Weak supervision and other non-standard classification problems: A taxonomy. *Pattern Recognition Letters*, *69*, 49–55. https://doi.org/10.1016/j.patrec.2015.10.008

Hochreiter, S., & Schmidhuber, J. J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1–32. Retrieved from https://doi.org/10.1162/neco.1997.9.8.1735

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., & Weld, D. S. (2011). Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 541–555). Association for Computational Linguistics.

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling. *Digital Journalism*, *4*(1), 89–106. https://doi.org/10.1080/21670811.2015.1093271

King, G., Lam, P., & Roberts, M. E. (2017). Computer-Assisted Keyword and Document Set Discovery from Unstructured Text. *American Journal of Political Science*, *61*(4), 971–988. https://doi.org/10.1111/ajps.12291

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *ArXiv*, 1–15. Retrieved from http://arxiv.org/abs/1412.6980

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet large scale visual recognition challenge. *Advances in Neural Information Processing Systems*.

Kurata, G., Xiang, B., & Zhou, B. (2016). Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 521–526). https://doi.org/10.18653/v1/n16-1063

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent Convolutional Neural Networks for Text Classification. In *Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 2267–2273). Retrieved from http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/download/9745/9552

LeCunn, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, *521*(7533), 436–444. Retrieved from https://doi.org/10.1038/nature14539

Leetaru, K., & Schrodt, P. a. (2013). GDELT: Global Data on Events, Location and Tone, 1979-2012. In *2013 Annual Meeting of the International Studies Association* (Vol. 2). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.686.6605&rep=rep1&type=pdf

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, *23*(2), 1–24. https://doi.org/10.1093/pan/mpu019

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., … Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, *12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Baltimore, Maryland, USA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P14-5010

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning* (*ICML-10*). https://doi.org/10.1123/jab.2016-0355

Nam, J., Kim, J., Loza Mencía, E., Gurevych, I., & Fürnkranz, J. (2014). Large-scale multi-label text classification – Revisiting neural networks. In *Joint european conference on machine learning and knowledge discovery in databases*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-44851-9_28

Nardulli, P. F., Althaus, S. L., & Hayes, M. (2015). A Progressive Supervised-Learning Approach to Generating Rich Civil Strife Data. *Sociological Methodology*, *45*(1), 148–183. https://doi.org/10.1177/0081175015581378

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543. Retrieved from http://www.aclweb.org/anthology/D14-1162

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science*, *54*(1), 209–228. Retrieved from https://doi.org/10.1111/j.1540-5907.2009.00427.x

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2020). Snorkel: rapid training data creation with weak supervision. *VLDB Journal*, *29*(2–3), 709–730. https://doi.org/10.1007/s00778-019-00552-1

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., … Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, *12*(2–3), 140–157. https://doi.org/10.1080/19312458.2018.1455817

Schwartz, H. A., & Ungar, L. H. (2015). Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 78–94. Retrieved from https://doi.org/10.1177%2F0002716215569197

Soroka, S., Young, L., & Balmas, M. (2015). Bad News or Mad News? Sentiment scoring of Negativity, Fear, and Anger in News Content. *The ANNALS of the American Academy of Political and Social Science*, *659*(May), 108–121. https://doi.org/10.1177/0002716215569217

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* (*JMLR*), *15*, 1929–1958. Retrieved from http://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf

Wilkerson, J. D., & Casas, A. (2017). Large-scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Sciene*, *20*(May 2017), 1–18. Retrieved from https://doi.org/10.1146/annurev-polisci-052615-025542

Zhou, Z. H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, *5*(1), 44–53. https://doi.org/10.1093/nsr/nwx106

About the Authors

Yair Fogel-Dror, Shaul R. Shenhav, Tamir Sheafer
The Hebrew University, Jerusalem, Israel