Article

Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches

Social Science Computer Review
1-21

DOI: 10.1177/0894439320907027
journals.sagepub.com/home/ssc

Kohei Watanabe[1] **and Yuan Zhou**[2]

Abstract

There is a growing interest in quantitative analysis of large corpora among the international relations (IR) scholars, but many of them find it difficult to perform analysis consistently with existing theoretical frameworks using unsupervised machine learning models to further develop the field. To solve this problem, we created a set of techniques that utilize a semisupervised model that allows researchers to classify documents into predefined categories efficiently. We propose a dictionary making procedure to avoid inclusion of words that are likely to confuse the model and deteriorate the its classification performance classification accuracy using a new entropy-based diagnostic tool. In our experiments, we classify sentences of the United Nations General Assembly speeches into six predefined categories using the seeded Latent Dirichlet allocation and Newsmap, which were trained with a small "seed word dictionary" that we created following the procedure. The result shows that, while keyword dictionary can only classify 25% of sentences, Newsmap can classify over 60% of them accurately correctly and; its accuracy exceeds 70% when contextual information is taken into consideration by kernel smoothing of topic likelihoods. We argue that once seed word dictionaries are created by the international relations community, semisupervised models would become more useful than unsupervised models for theory-driven text analysis.

Keywords

text analysis, dictionary making, semisupervised learning, international relations, United Nations

Discourse analysis has been widely used in studies of international relations (IR) since the "Third Great Debate" (Banta, 2013; Holzscheiter, 2014; Lundborg & Vaughan-Williams, 2015; Milliken, 1999), in which the constructivist scholars challenged the dominant status of realism and liberalism. Aiming to understand the relationship between discourses and international politics, researchers have analyzed an extensive range of issues such as terrorism (Jackson, 2005), the Bosnian War

[1] University of Innsbruck, Innsbruck, Austria
[2] Kobe University, Kobe, Japan

Corresponding Author:
Kohei Watanabe, University of Innsbruck, Room 2.44, Universitaetsstrasse 15, 2nd Floor, Innsbrucki A-6020, Austria.
Email: kohei.watanabe@uibk.ac.at

(Hansen, 2006), and Japanese nationalism (Suzuki, 2015) in official speeches, government statements, newspaper articles, and academic works.

While discourse analysis is conducted qualitatively in those studies, an increasing number of IR scholars are embarking on quantitative analysis of large textual data such as debates at the United Nations (UN) General Assembly, employing natural language processing techniques (Baturo et al., 2017; Gurciullo & Mikhaylov, 2017; Schoenfeld et al., 2018). These speeches are a very useful source of information for IR scholars because representatives of the member countries express their foreign policy on various issues such as development, terrorism, and human rights in the forum. The speeches are approximately 15 min long and transcribed and translated into English. The content of the speeches enables IR scholars to infer and compare foreign policies of the UN member states over the years. Baturo et al. (2017) collected all the speech transcripts from 1970 to 2017 and made them available in the UN General Debate Corpus.

The corpus was analyzed to measure similarity of foreign policies between states by Baturo et al. (2017) and Gurciullo and Mikhaylov (2017) using quantitative text analysis techniques, but their classification or measurement schemes lacked consistency with earlier content analysis that was conducted manually (e.g., Brunn, 1999) due to the limitation of the machine learning models: Latent Dirichlet allocation (LDA) is a unsupervised model widely used in legislative or electoral studies to classify documents into topics, but it is nearly impossible for researchers to match topics identified by an unsupervised LDA and human coders; supervised models such as naive Bayes, Random Forest and support-vector machines (SVMs) allow researchers to define topics, but a large training set is often too expensive for resource-strapped researchers to create.

We aim to solve these problems by developing a set of techniques to identify predefined topics of documents using semisupervised machine learning models. In the following sections, we will point out the problems of existing methods for topic classification, explain our semisupervised document classification techniques, and demonstrate their effectiveness in a series of experiments. We will employ two semisupervised models, Newsmap (Watanabe, 2018b) and seeded LDA (Lu et al., 2011) both of which take seed words as weak supervision. The former is a variant of naive Bayes classifier trained on documents that contain seed words, while the latter is an LDA model fitted with word priors weighted by seed words. Therefore, the central piece of this article is the technique to find good seed words for topics.

In the experiments, we will classify sentences of the speeches in the General Debate Corpus into six topics ("greeting," "UN," "security," "human rights," "democracy," "development") that are considered important in the IR literature. We choose sentences as the unit of analysis not only because classification of smaller units enables fine-grained analysis, such as association between topics and countries or sentiments, but also because sentences have greater validity than paragraphs or quasi-sentences that contain multiple topics or lack objective boundaries (Däubler et al., 2012; Grimmer & Stewart, 2013). However, sentence classification is more challenging than classification of the larger text units due to the small number of words they contain.

The result of our experiments will show that researchers can perform topic classification of sentences by Newsmap with 60% accuracy, which is better than by the seeded LDA. The classification result is the best when the entropy-based diagnostic tool is used to identify words that are likely to make definition of topics less clear if included. Furthermore, over 70% of sentences can be classified accurately if the contextual information is exploited by kernel smoothing of topic likelihoods.

We believe that IR scholars can further develop the theory in the field by analyzing large corpora systematically and consistently with existing theoretical frameworks using semisupervised models. We recommend researchers to use these models because they achieve good balance between the *cost* and *control*, that is, they can control classification results without costly manual coding of documents. Once valid seed words for relevant topics are identified, they can repeat the

same analysis on different corpora using semisupervised models and compare the results for vigorous scientific debates.

In our discussion, "keywords" refers to words that are used to obtain frequency counts of words for document classification in the classic dictionary-based approach, while "seed words" refers to words used to train semisupervised document classifiers (Newsmap and seeded LDA). These words are both in a dictionary, but we distinguish between these two types of lexica by calling them "keyword dictionary" or "seed word dictionary."

For the following examples and experiments, we preprocessed the texts in the following manner to construct a document-feature matrix using an R package *quanteda* v.1.3.14 (Benoit et al., 2018). We divided speeches into sentences, segmented these sentences into tokens (words), and removed numeric tokens and punctuation marks based on the rules and character classes defined by International Component for Unicode (ICU). After compounding sequences of tokens for multiword expressions in the dictionaries (e.g., "human rights" and "climate change") into single tokens, we removed grammatical words and infrequent words (less than 10 times in the corpus).[1]

Problems of Existing Methods

Dictionary Analysis

Dictionaries are sets of keywords in predefined categories corresponding to certain concepts, and often used in quantitative analysis as a robust theory-driven method. For example, the Lexicoder Topic Dictionary (Albugh et al., 2013) contains 1,387 keywords under 28 topics (e.g., macroeconomics, civil rights, health care, agriculture) based on the Comparative Agenda Project's coding scheme. Since keywords for categories are manually selected, dictionary analysis allows researchers to assess not only the "criterion validity" but also the "content validity" of the measurement (cf Adcock & Collier, 2001; Sartori & Pasini, 2007).

However, dictionaries sometimes produce false results due to their strong context dependency (Grimmer & Stewart, 2013). Therefore, if researchers wish to study documents in fields or languages that are new to quantitative text analysis, they have to create a dictionary from scratch, but it is not easy to collect hundreds of words manually (King et al., 2017). The number of keywords required for a dictionary depends on the concepts to capture and documents to analyze, but it tends to be very large when they are complex and sparse.

Even if dictionaries are available, researchers can only produce simple frequency counts in dictionary analysis. They usually obtain frequency counts of words in documents using a dictionary and divide them by their lengths (i.e., total number of words) to gauge the salience of the concepts, which is only relative to other categories. Therefore, keyword-based classification does not offer theoretically grounded threshold for the reliability of classification results based on probabilities such the likelihood ratio.

Unsupervised Topic Models

A variety of unsupervised topic models such as the LDA (Blei et al., 2003), the correlated topic model (CTM; Blei & Lafferty, 2007), and the Bayesian hierarchical topic model (Grimmer, 2010) have been used to identify topics in documents. Roberts et al. (2016) developed a structural topic model (STM), which allows researchers to discover the relationship between topics and documents' attributes (e.g., country and year) by incorporating them into topic identification. These "topic models" assume that documents are mixture of multiple topics and identify user-provided number of topics as clusters of words solely based on their co-occurrences.

However, unsupervised topic models almost always produce topics that are inconsistent with the theoretical framework. To highlight this problem, we fitted an unsupervised, or unseeded, LDA

Table 1. Topics and Words in Unseeded (Unsupervised) LDA.

Topic	LDA topic	Words
Greeting	6	people, world, time, year, like, Africa, mr, African, role, need
	8	new, republic, session, many, among, national, us, sustainable, mr, parties
	16	states, international, support, cooperation, today, regional, resources, president, order, mr
UN	3	united nations, international, world, development, people, support, efforts, community, social, security council
	15	united nations, international, must, social, make, challenges, respect, order, time, nations
	20	world, must, us, years, general assembly, united nations, like, end, need, system
Security	7	world, new, security, well, role, us, global, made, progress, developing
	9	also, peace, country, one, global, president, process, countries, government, reform
	12	government, security, development, countries, like, one, support, president, organization, stability
	18	government, countries, security, years, developing, need, work, end, must, time
Human rights	2	people, world, political, states, efforts, state, region, human rights, end, war
	4	development, human rights, today, process, country, challenges, now, towards, rights, agenda
	19	time, organization, human, must, human rights, situation, efforts, president, system, effective
Democracy	10	also, can, Africa, political, role, people, secretary-general, particular, just, every
	13	political, economic, global, general assembly, African, within, sustainable, weapons, conference, goals
Development	5	global, order, work, social, Africa, situation, security council, challenges, way, goals
	11	states, development, new, efforts, democracy, way, regard, also, ensure, assembly
	17	can, global, united nations, developing, now, year, among, regional, problems, nuclear

Note. Words are weighted by FREX by their exclusivity and LDA topics are manually mapped to substantive topics. LDA = latent Dirichlet allocation; UN = United Nations.

model to discover 20 topics in the UN General Assembly speeches in the post–Cold War period (Table 1).[2] To make topics more interpretable, we applied the frequency and exclusivity (FREX) method that down weights common words (Roberts et al., 2014) and manually mapped 18 LDA topics to the six substantive topics, excluding two that are very difficult to interpret (aka "garbage topics"). We are fully aware that such post hoc interpretation and selection of topics risks error and distortions, but this is a common practice among the users of unsupervised topic models. We also evaluated the result of classification by the model using a manually coded data set (explained in the Experiment section) and found its performance was very poor: The F1 scores ranged between .01 ("democracy") and .28 ("development"), making the overall score to be F1 = .13.[3]

Semisupervised Topic Classification Techniques

Computer scientists have developed various types of semisupervised techniques, which exploit not only labeled documents but also unlabeled documents in training models (Chapelle et al., 2010). For example, Blum and Mitchell's (1998) "co-training" technique extracts features from a small number of labeled documents and uses these features to expand the training set for a better classification result. In their experiment, they iteratively expanded the training set by including the most reliable classification results from a naive Bayes classifier and achieved significantly lower prediction errors than a model without the expansion. Zelikovitz and Hirsh (2000) have proposed a technique to

expand training sets by nearest neighbor classification of unlabeled documents and demonstrated that this technique is the most effective when labeled documents are short and scarce. More recently, external data such as Wikipedia have been used as a source of universal knowledge to improve classification results (Banerjee et al., 2007; Schönhofen, 2009). In the same vain, Phan et al. (2008) also trained topic models (LDA and Latent Semantic Analysis) on external data to accurately classify short documents using a supervised model.

While the above models are aimed at increasing the efficiency of supervised learning, seeded models are aimed at improving the interpretability of results by using prior lexical knowledge. Lu et al. (2011) developed a technique to weight prior distribution of topics over words to detect sentences that mention specific issues. In their semisupervised LDA model,[4] pseudocounts are added to user-defined topic seed words before an LDA model is fitted. Jagarlamudi et al. (2012) modeled topics as a mixture of user-defined topics and other topics by adding extra parameters. Watanabe (2018b) developed Newsmap, a semisupervised model for geographical classification of documents, by combining dictionary analysis and the naive Bayes classifier. The model was also used for semisupervised topic classification (Watanabe, 2018a).

Naive Bayes models can classify sentences into topics accurately because each sentence usually has only one topic. However, creation of a seed word dictionary for topic classification is more difficult than for geographical classification since there are many more potential seed words for topics than for countries. Seed word selection is a two-step process in which analysts collate candidate words that potentially help to define the categories and choose only words that have little ambiguity as seed words. Usually, short texts require a greater number of seed words than long texts because there is a smaller chance that seed words occur in shorter texts than in longer texts. However, seed words that match unrelated texts significantly deteriorate the classifier's performance because false positive cases confound the true association between topics and words in the machine learning process (Jagarlamudi et al., 2012). We call words that match intended documents *true seed words* and those that match unintended documents *false seed words*. Therefore, the goal in dictionary making for semisupervised models is maximizing the number of true seed words while minimizing the number of false seed words.

Semisupervised Classification

Newsmap is a semisupervised learning technique originally created to classify short news summaries according to their geographical focus (Watanabe, 2018b). Unlike full-supervised models, Newsmap does not require a manually coded training set but a seed word dictionary. First, the model searches the entire corpus for seed words in the dictionary in order to assign labels to documents; second, the labels are used to estimate the association between the labels and features. This dictionary-based learning is advantageous because (1) training of new models does not require any manual coding, (2) searching a corpus for dictionary words is not computationally intensive, and (3) dictionaries can be ported to different projects without or little modification.

However, Newsmap requires users to define all the relevant topics in a seed dictionary because it estimates features' association with a topic by comparing between their frequencies in documents with the label and all other labels, ignoring documents without labels. This exhaustiveness of topics is important not only to produce the best classification results but also to avoid arbitrary post hoc selection of topics.

Seed Word Selection

A common approach to evaluating the contribution of seed words to classification accuracy is using a manually labeled "gold standard" (King et al., 2017), but it might risk overfitting the seed word

dictionary to the sample and making it less portable to other projects. Use of multiple samples for evaluation reduces this risk but manual coding of large amounts of the document is expensive. Therefore, we devised an heuristic approach that does not require manually coded documents in seed word selection. In this approach, analysts use *coverage* and *entropy* as diagnostic statistics to judge whether a seed word would positively or negatively contribute to classifiers' performance. Since these statistics are computed without manually coded documents, their diagnosis is only probabilistic, but we will demonstrate that they offer analysts good guidance in dictionary making in our experiments.

Seed Word Coverage

Seed word coverage indicates the proportion of the documents in which seed words occur. While keyword coverage refers to the proportion of documents that can be classified regardless of its accuracy in simple dictionary-based topic classification, seed word coverage indicates the proportion of documents that can be used to train a model in semisupervised learning. Therefore, a higher seed word coverage is more desirable, but sentence-level classification requires many seed words to achieve this due to the small number of words in each unit.

Average Feature Entropy (AFE)

AFE statistic builds on the entropy function that is commonly used to measure uniformity of distribution produced by discrete variables in information processing:

$$\mathrm{AFE} = \frac{1}{m}\sum_{i}^{m} \mathrm{H}(F_i + 1),$$

$$F_i = [f_1, f_2, f_3, \ldots, f_n],$$

$$\mathrm{H}(F_i) = -\sum_{j}^{n} P(f_j)\log_2 P(f_j),$$

where m is the total number of features, H is the entropy function, and F_i is a vector of frequencies of features co-occurring with seed words for n topics. Feature frequencies are smoothed by adding one before computing entropy.

By taking frequency of co-occurrences between seed words and all other words, AFE predicts randomness of labels given to documents that tend to deteriorate the classifiers' performance: AFE becomes high when features co-occur equally frequently with seed words for many topics, but it becomes low when they co-occur only with seed words for few topics (AFE = 0 when features co-occur only with seed word for one topic). To compute this statistic, we first apply a seed word dictionary to the corpus and aggregate feature frequencies by topic labels, then compute entropy for each feature.[5]

Contextual Smoothing

A trained Newsmap model can recognize a very large number of features thanks to the semisupervised learning, but some of the sentences lack topic indicators at all. In such a case, however, we can classify those sentences by considering the topics of surrounding sentences as we did in our manual coding. Contextual smoothing does not require us to modify Newsmap's algorithm because we can simply postprocess the likelihood of topics predicted by the model: We train the classifier and predict topics in the same way as described above and smooth the likelihood using a kernel smoother to reclassify sentences into topics with the highest scores.

Table 2. Speeches Chosen for the Experiment.

Year	Country	Greeting	UN	Security	Human	Dem.	Dev.	Total
1991	Australia	6	67	46	14	0	18	145
1992	Libya	11	18	93	12	1	29	153
1993	Ukraine	6	36	79	0	2	33	150
1994	Sweden	4	26	72	11	1	16	126
1995	Japan	2	28	43	2	0	34	107
1996	Brazil	8	44	22	2	2	28	98
1997	Zimbabwe	7	21	22	0	13	20	76
1998	Mexico	9	27	30	3	1	24	85
1999	Chad	11	8	61	2	1	34	106
2000	United States	7	42	26	14	10	14	106
2001	Germany	4	7	80	18	2	11	118
2002	Russia	0	15	60	2	0	10	87
2003	Afghanistan	7	0	49	1	3	23	76
2004	Myanmar	5	9	42	16	30	12	109
2005	Tunisia	4	2	14	0	0	12	28
2006	Colombia	2	1	29	1	12	19	62
2007	United Kingdom	6	11	23	1	3	54	92
2008	Equatorial Guinea	5	3	7	1	1	29	41
2009	Turkmenistan	4	19	32	1	0	16	68
2010	Palau	4	7	23	1	0	49	80
2011	China	5	0	46	0	0	85	131
2012	Norway	8	15	28	18	5	1	67
2013	Slovenia	4	4	42	4	0	1	51
2014	Tuvalu	9	17	11	1	0	54	83
2015	France	3	16	34	23	2	36	111
2016	Saudi Arabia	2	2	35	8	0	10	55
2017	Estonia	0	13	36	15	1	31	96

Note. Topics are identified by manual coding of sentences. UN = United Nations; Dem. = democracy; Dev. = development.

Experiments

The data for our experiments are a subset of the General Debate Corpus (Baturo et al., 2017). After excluding speeches during the Cold War to ensure the consistency of topics, we sampled one speech every year from a different country (Table 2). In choosing sample countries, we kept the balance in their international influence, geographical location, and levels of industrialization. We segmented these speeches into sentences and manually classified them into five topics ("UN," "security," "human rights," "democracy," "development"). These topics are based on the previous studies on the UN (Smith, 2006; Zanotti, 2005), but we added "greeting" as a category for speakers' opening remarks, although they are unlikely to be of IR scholars' interest.

Manual Coding

We employ manually classified sentences as the gold standard against which we test our diagnostic tools and analytical techniques. However, sentence-level classification is not an easy task even for human coders because sentences often (1) contain more than one indicator or (2) lack a clear indication of topics. If a sentence contains multiple topics, we classify them into the most salient one; if sentences mention the UN's role in issues such as human rights protection or security assurance, we classify them into those substantive topics. If a sentence does not contain clear topic indicators, we resort to topics in surrounding sentences to classify it.[5] One of the authors, who has

Table 3. Seed Words Used in the Experiments.

Topic	Knowledge Based	Frequency Based
Greeting	greet*, thank*, congratulat*, sir, express*	great*, mr, wish*, hop*, contribut*, anniversar*, welcom*
UN	united nations, international court*	security council, general assembly, organization*, reform*, secretary-general, resolution*, permanent member*, charter*, session*, conference*
Security	secur*, kill*, attack*, dispute*, victim*	peac*, terror*, weapon*, nuclear*, conflict*, war*, disarmament*, threat*, cris*, solution*, settlement*, force*, destruction*, militar*, violence*, arm*, fight*
Human rights	human rights, violat*, race*, dignit*, protect*, citizen*, educat*	humanitarian, child*, women, refugee*, communit*, people, respect*, responsib*, food*, health*
Democracy	democra*, autocra*, dictator*, vote*, represent*, elect*, leader*	president*, party, institution*, government*, law*, republic*, free*, leadership*, legal*
Development	develop*, market*, investment*	econom*, climate change, assistance*, sustain*, povert*, trade*, grow*, social*, environment*, prosperit*, progress*, financ*, cooperation*
Total # of words	29	66

Note. Knowledge-based seed words define categories, while frequency-based seed words increase coverage. UN = United Nations.

much experience in analyzing the corpus from a more substantive point of view, performed the manual coding.[6]

Seed Word Selection

Seed word selection has a critical importance in seeded models. However, humans can judge the relevance of words in a list but cannot easily create such a list themselves (King et al., 2017). Therefore, we constructed two sets of seed words based on knowledge or frequency. *Knowledge-based seed words* are smaller sets of words selected based on researchers' background knowledge in the field, while *frequency-based seed words* are larger sets of words chosen from the most frequent words in the corpus. We consider knowledge-based seed words superior to frequency-based seed words because knowledge-based seed words offer operational definitions of the concepts and have greater external validity that ensures portability across corpora.

For our experiment, we read the debate transcripts carefully and consulted glossaries and indices of relevant books to select knowledge-based seed words; then we obtained a list of the 300 most frequent words in the entire corpus and manually classified the words into relevant topics as frequency-based seed words; if the knowledge-based and frequency-based sets had common elements, we removed them from the former, except the words that define the topics (i.e., names of the topics), to highlight the impact of frequent seed words in the experiment (Table 3). In other words, we aim to identify words that are both frequent and defining in our experiments.

After stemming the collected seed words to create glob patterns, we trained a Newsmap model and tested its performance by the F1 score, which is one of the standard measures in computer science. Overall, knowledge-based seed words (F1 = .53) achieved slightly better than frequency-based seed words (F1 = .52); a larger set combining these two performed the best (F1 = .57). Classification accuracy is highest in "development" (F1 = .65) and lowest in "democracy" (F1 = .36) and "human rights" (F1 = .37) in knowledge-based seed words, but "greeting"

Table 4. Classification Results by Newsmap.

Topic	Knowledge Based	Frequency Based	All
Greeting	.495	.235	.343
UN	.500	.481	.585
Security	.563	.638	.627
Human rights	.370	.255	.359
Democracy	.362	.276	.349
Development	.654	.642	.678
Overall	.535	.520	.570

Note. Inclusion of frequency-based seed words leads to lower F1 scores in "Greeting," "Human Rights," and "Democracy."

Table 5. Topics and Words in Seeded (Semisupervised) LDA.

Topic	Words
Greeting	also, general assembly, session, work, like, year, president, take, government, years
UN	international, united nations, community, states, support, security council, organization, role, member, reform
Security	people, many, terrorism, war, end, situation, conflicts, security, however, still
Human rights	world, must, can, us, new, human rights, one, today, challenges, human
Democracy	peace, cooperation, political, process, country, security, region, state, efforts, Africa
Development	development, countries, economic, global, social, developing, sustainable, poverty, resources, trade

Note. Many of them are seed words but relevant topic words are also identified. LDA = latent Dirichlet allocation.

Table 6. Classification Results of Seeded (Semisupervised) LDA.

Topic	Knowledge Based	Frequency Based	All
Greeting	.324	.332	.315
UN	.310	.421	.433
Security	.388	.464	.472
Human rights	.190	.158	.187
Democracy	.092	.117	.147
Development	.539	.567	.560
Overall	.348	.400	.407

Note. Inclusion of all the frequency-based seed words leads to lower F1 scores in "Greeting" and "Human Rights." LDA = latent Dirichlet allocation.

(F1 = .23), "human rights" (F1 = .25), and "democracy" (F1 = .27) are even lower in frequency-based seed words (Table 4).

We also fitted a seeded LDA with the knowledge-based seed words for comparison and obtained a more interpretable model than the unseeded LDA without FREX weighting (Table 5).[7] Although there are words that are strongly associated with unexpected topics (e.g., "peace" and "security" in "democracy"), many of the words in our frequency-based seed words are correctly identified as topic words (e.g., "sustainable" and "poverty" in "development"). However, the seeded LDA achieved only F1 = .40 overall even when all the seed words are used (Table 6), although it is a significant improvement from the unseeded LDA (F1 = .13). Contrary to Newsmap, the seeded LDA performed better with the frequency-based set than the knowledge-based set and combining these two sets changed its classification performance only very little. The poor

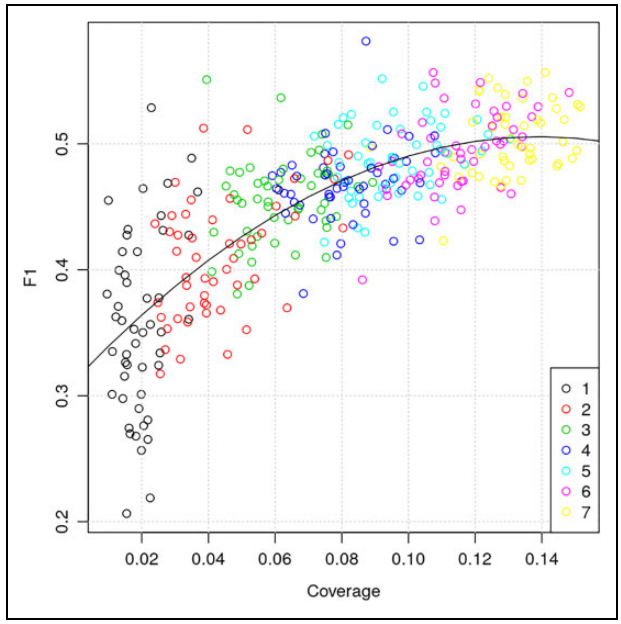

Figure 1. Correlation between seed word coverage and the classifier's performance (F1). *Note.* F1 is strongly correlated with seed word coverage. The colors of the circles indicate the number of seed words for each topic.

performance of the seeded LDA also indicates that it is very difficult to classify sentences accurately using existing tools.

Experiment 1: Seed Word Coverage

We randomly drew 1– 7 seed words from each topic and trained a Newsmap model to understand the relationship between the seed word coverage and the classification performance. Figure 1 shows that seed word coverage is strongly correlated ($r = .74$) with the F1 scores: They score .4 when the coverage is around 4%, but they increase to .5 when the coverage is 10%. Furthermore, the coverage is roughly the function of the number of seed words: The coverage is less than 5% when each topic has one to two seed words, but it increases to more than 10% when each has six to seven seed words. This also suggests that the optimal size of seed dictionary is approximately six to seven words for each topic in this task.

The strong positive correlation between the number of seed words and classification performance suggests that the most effective strategy in classification by Newsmap is adding as many seed words as possible to a dictionary. However, the improvement of performance slows down as the size of the dictionary grows because newly added words are often false seed words that confuses the learning process as discussed in Experiment 3.

Experiment 2: Average Feature Entropy

We incrementally expanded the dictionary by randomly drawing a seed word from the candidate set to reveal the relationship between AFE and the F1 scores (Figure 2). Starting only with the knowledge-based seed words as the most reliable benchmark, we sampled the frequency-based seed words, added them one by one ("weapon*," "econom*," "war*," etc.) to the dictionary and

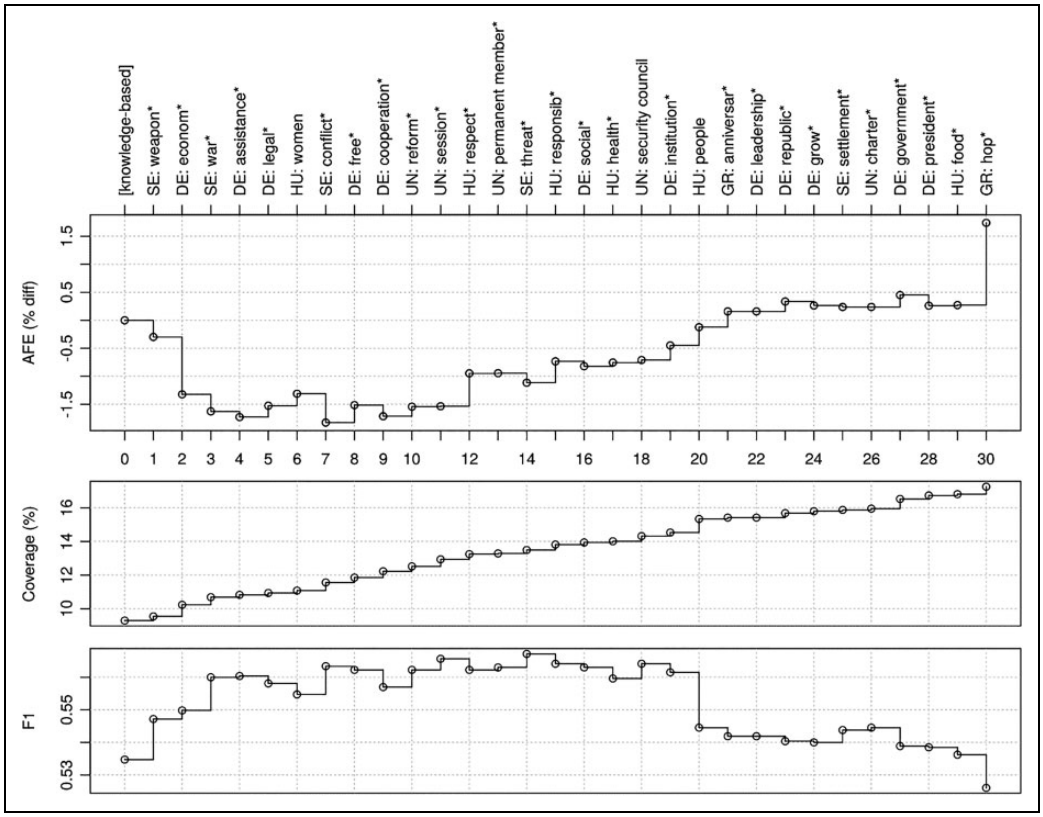

Figure 2. Changes in the seed word coverage, entropy (average feature entropy), and classification accuracy (F1) by frequency-based seed words. *Note*. Seed word coverage increases monotonously as more seed words are included, but F1 decreases sharply when false seed words ("people" and "op*") are added.

computed the AFE and the F1 scores. The classifier's performance gradually increased as we added more seed words, but F1 fell sharply coinciding with a rise in AFE when we included "people" (20th) and "hop*" (30th). Changes in F1 and AFE in opposite directions can be also found in "legal*" (5th), "women" (6th), "respect*" (12th), "responsib*" (15th), "institution*" (19th), "anniversary*" (21th), and "government*" (27th).

We repeated the same procedure 100 times to understand the relationship between AFE and F1. Figure 3 shows that changes in the AFE and F1 are negatively correlated ($r = -.49$, $p < .001$), namely, it is twice more likely that F1 decreases when a new seed word increases AFE. This result indicates that changes in AFE can be used to identify risky seed words that lead to lower classification performance if included.

Experiment 3: Selection Criteria

The results of Experiment 1 and 2 suggest that it is possible to reduce the risk of adding seed words that are likely to deteriorate the classifier's performance using the AFE statistic. In order to test this possibility, we computed changes in AFE when the frequency-based seed words are added to a dictionary only with the knowledge-based seed words.

The result showed that 7 words in "greeting," 10 words in "human rights," and 6 words in "democracy" are risky seed words (Table 7). The risky seed words are concentrated in the three

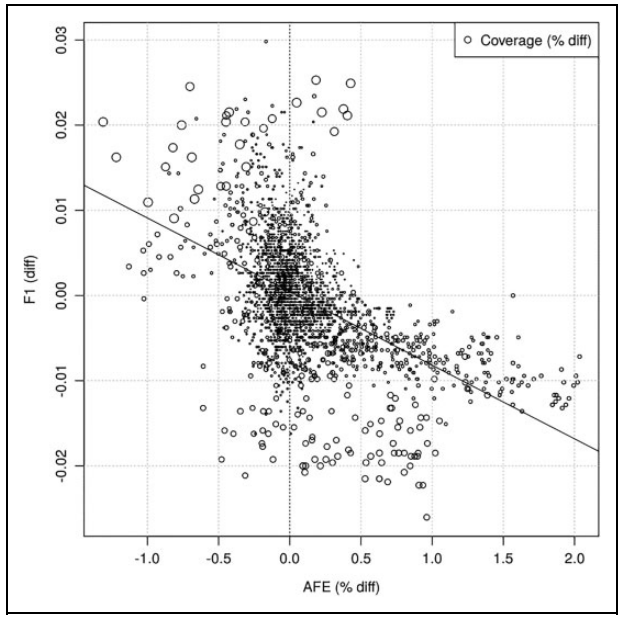

Figure 3. Correlation between entropy (average feature entropy [AFE]) and classification accuracy (F1) changes ($r = -.49$). *Note*. Increase in AFE often leads to lower F1, but large increase in seed word coverage often leads to higher F1.

Table 7. Risky Seed Words Identified by Average Feature Entropy.

Topic	Seed Words
Greeting	great*, mr, wish*, hop*, contribut*, anniversar*, welcom*
Human rights	humanitarian, child*, women, refugee*, communit*, people, respect*, responsib*, food*, health*
Democracy	party, institution*, law*, republic*, free*, legal*

Note. These words are likely to confuse learning process if included in the seed word dictionary.

topics because AFE correctly predicted that these frequency-based seed words deteriorate the classifier's performance (Table 4). If the frequency-based seed words are added to the dictionary excluding the risky seed words, its performance improved to F1 = .61, which is significantly higher than only with the knowledge-based seed words (F1 = .53) or with all the seed words (F1 = .57). Conversely, if only the risky seed words are added to the dictionary, the classifier's performance decreased significantly (F1 = .44), showing that risky seed words identified by AFE are false seed words. Changes in the classifier's performance by risky words are shown in Figure 4.

Experiment 4: Contextual Smoothing

We constructed a dictionary excluding the risky seed words and trained a Newsmap model to achieve the best performance of the classifier, but the model still cannot classify sentences that lack topic indicators at all. Human coders can resort to surrounding topics to classify such sentences but Newsmap models cannot. To mimic manual classification, we predicted topics of individual sentences and smoothed the topic likelihood ratios by Daniell kernel with window size $m = 3$ to reclassify them. With this window size, we take topics of three preceding and succeeding sentences into consideration to identify the topic of the current sentence. Figure 5 shows that the smoothed

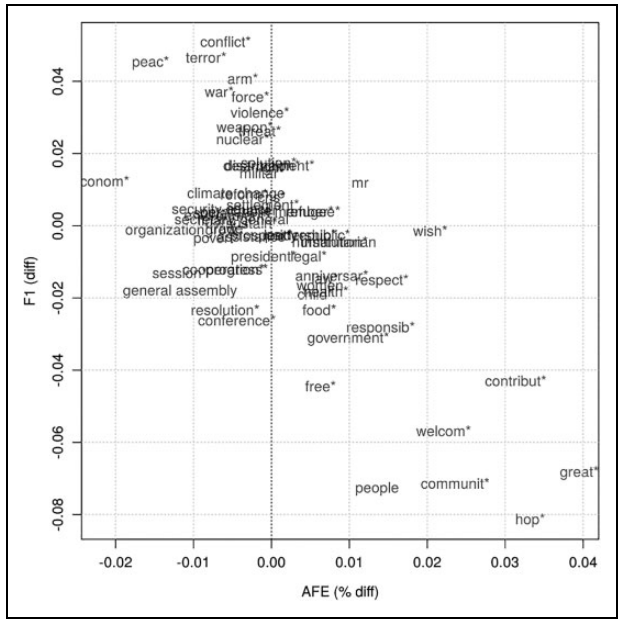

Figure 4. Changes in F1 by knowledge-based seed words. *Note*. Words in the right-hand side of the dotted lines are considered risky seed words.

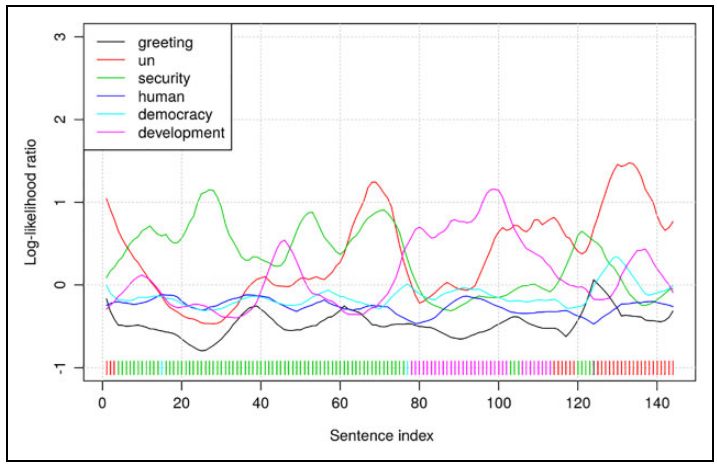

Figure 5. Smoothed topic likelihood ratios in speech by Ukraine in 1993. *Note*. Contextual smoothing captures actual transition of topics. The speech appears shorter because smoothing omits first and last sentences.

likelihood ratios can capture the actual transition of topics in a speech made by the representative of Ukraine in 1993.

The contextual smoothing improved the result of classification from F1 = .61 to .72 overall. Table 8 and Figure 6 show that smoothing had a significant positive effect in most of the cases, especially in Ukraine (+.22), Mexico (+.24), Russia (+.26), Sweden (+.22), and France (+.21). Although the classification accuracy of the United States (F1 = .56) and France (F1 = .61) is still low in absolute terms because of the low F1 scores in "human rights" and "emocracy" that speakers of the two countries tend to emphasize, the result clearly shows the effectiveness of

Table 8. Classification Results by Countries.

Country	Code	Year	Raw (F1)	Smooth (F1)	Change
Australia	AUS	1991	.662	.755	.093
Libya	LBY	1992	.561	.678	.117
Ukraine	UKR	1993	.545	.764	.219
Sweden	SWE	1994	.638	.856	.217
Japan	JPN	1995	.750	.794	.044
Brazil	BRA	1996	.670	.830	.160
Zimbabwe	ZWE	1997	.704	.648	−.057
Mexico	MEX	1998	.638	.878	.240
Chad	TCD	1999	.581	.733	.152
United States	USA	2000	.531	.564	.033
Germany	DEU	2001	.664	.773	.109
Russia	RUS	2002	.575	.840	.265
Afghanistan	AFG	2003	.566	.620	.053
Myanmar	MMR	2004	.500	.569	.069
Tunisia	TUN	2005	.750	.600	−.150
Colombia	COL	2006	.562	.808	.245
United Kingdom	GBR	2007	.643	.721	.078
Equatorial Guinea	GNQ	2008	.783	.794	.012
Turkmenistan	TKM	2009	.601	.500	−.101
Palau	PLW	2010	.536	.561	.025
China	CHN	2011	.757	.839	.081
Norway	NOR	2012	.560	.635	.075
Slovenia	SVN	2013	.545	.744	.199
Tuvalu	TUV	2014	.663	.788	.124
France	FRA	2015	.395	.608	.213
Saudi Arabia	SAU	2016	.825	.844	.020
Estonia	EST	2017	.604	.659	.054

Note. We find improvement in the F1 scores all but three countries (Tunisia, Turkmenistan, and Zimbabwe).

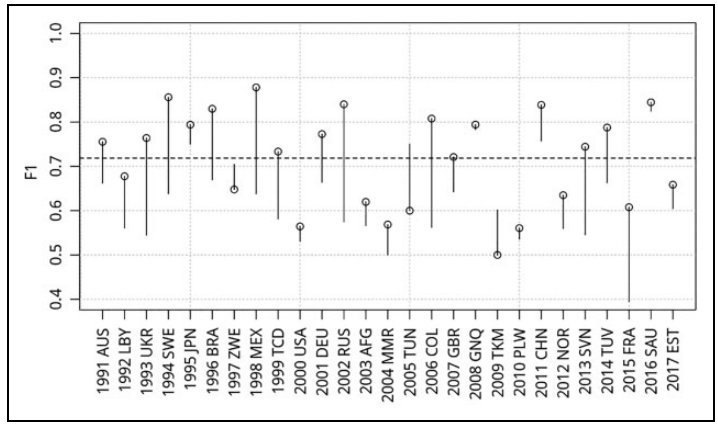

Figure 6. Classification results of individual speeches. *Note*. Circles are F1 scores after contextual smoothing, and lengths of lines indicate the magnitude of changes from raw F1 scores.

contextual smoothing in sentence classification. In Tunisia (−.15), Turkmenistan (−.10), and Zimbabwe (−.05), however, contextual smoothing had negative effect on the classification result. This is primarily because their speeches are relatively short in the number of sentences, and the

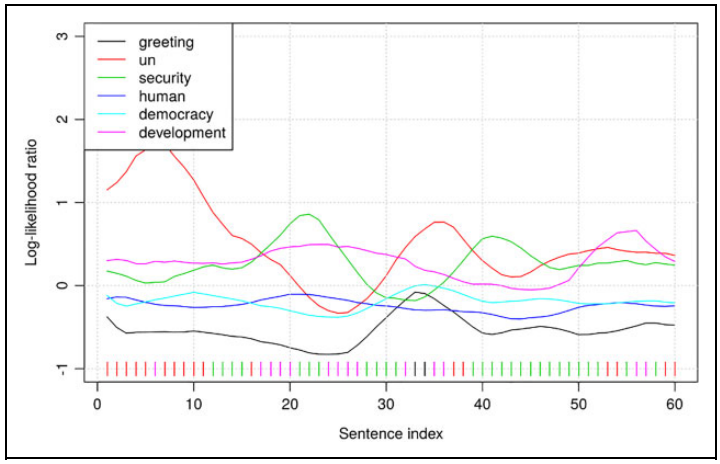

Figure 7. Smoothed topic likelihood ratios in speech by Turkmenistan in 2009. *Note.* Change of topics are too fast for contextual smoothing. The speech appears shorter because smoothing omits first and last sentences.

window size was too large. For example, Figure 7 shows that smoothing does not keep up with the rapid change in the topics in the speech by the representative of Turkmenistan in 2009, which was only in 68-sentence long.

Proposed Dictionary-Making Procedure

Our experiments suggest that the best strategy in dictionary making for semisupervised models is increasing the number of seed words to maximize the seed word coverage while avoiding false seed words that damage classifiers' performance. Researchers can easily increase the number of seed words if they collect frequent words from corpora but some of them can be false seed words. Therefore, we propose the following procedure as the best practice in seed word dictionary making employing the AFE statistic.

Identify Categories

Researchers identify category of document classification based on literature in the field. Categories should be identified based on the theoretical framework of the study or adopted from earlier quantitative or qualitative text analysis studies.

Operationalize Categories

Researchers define categories by a set of words that are collected from either the corpus or other sources (i.e., thesauri, glossaries, and indices of books). Selection of words should be solely based on the researchers' knowledge of relevant discourses ignoring their frequencies in the corpus (knowledge-based selection). Researchers should always consider whether the operationalization is valid for the same categories in other corpora.

Improve Seed Word Coverage

Researchers collect a set of words that are frequent in the corpus to increase the number of documents that can be used to train the semisupervised machine learning model (frequency-based

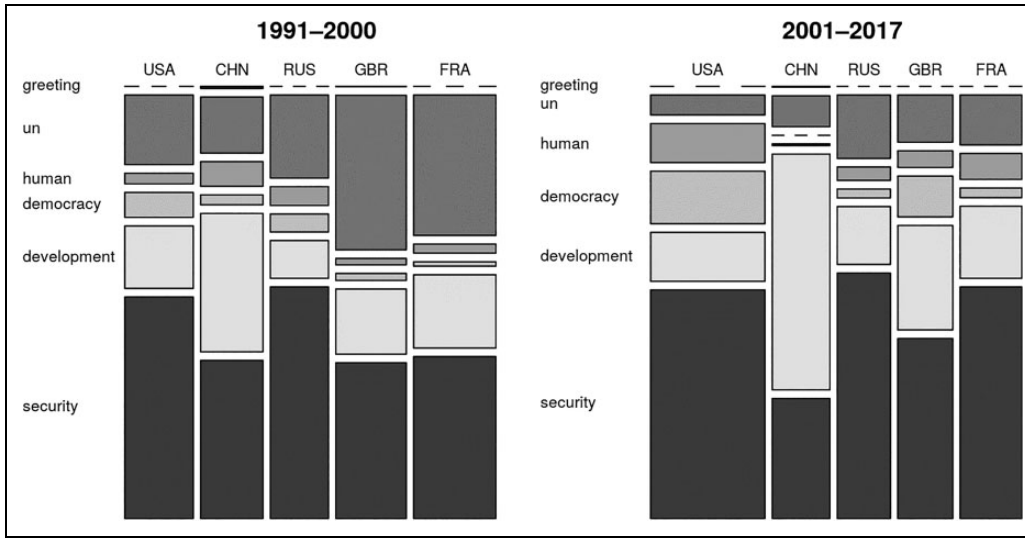

Figure 8. Topic in speeches by the Security Council permanent members before and after the 9/11 attack.
Note. Width indicates length of speeches in number of sentences.

selection). These seed words can be collected from a list of most frequent words (term or document frequency) in the corpus but they must be semantically equivalent to the knowledge-based seed words.

Seed words that increase seed word coverage usually have great impact on the classifier's performance either positively or negatively. In order to minimize the risk of including false seed words, researchers should monitor changes in AFE relative to the value when only the knowledge-based seed words are included. If a frequency-based seed word increases AFE, researchers should check carefully if the seed word occur only in relevant documents using a keyword-in-context tool before adding them to the dictionary.

Example

We classified all the sentences of speeches by the five permanent members of the UN Security Council to demonstrate a possible application of the techniques described above (Figure 8). First, the figure reveals that the proportion of "UN" have decreased significantly because the peacekeeping and Security Council reforms, which have been became prevalent issues after the end of the Cold War, lost their importance in more recent years. Second, "human rights" and "democracy" became more visible in the United States, Britain, and France in the latter period, but they almost disappeared in China. This result might challenge the conventional wisdom that economic development can cause democratization (Huber et al., 1993). Third, "security" has increased considerably in the UK and France after 2011 because the 9/11 attack raised European countries' security concerns (Nesser, 2008).

Conclusions

Compared to unsupervised topic models and simple dictionary analysis, the advantage of the semi-supervised classification is clear in IR research: (1) Researchers can define categories of classification using seed words to make their analysis consistent with existing theoretical frameworks and

(2) they only need to choose a small number of seed words for each category to classify documents correctly. In the experiments, we have demonstrated that Newsmap can identify topics of more than half of sentences with only a few seed words for each topic and its accuracy exceeded 60% when more seed words were added. In contrast, the simple dictionary analysis could classify only 25% of sentences and the seeded LDA only 40% of setences correctly, highlighting the difficulty in classifying short texts using existing tools.

The seeded LDA underperformed Newsmap, presumably because the model has inappropriate assumption for sentence classification and excessive complexity considering the small number of seed words available. LDA models assume documents to have multiple topics but sentences of speeches usually have only one topic. The complexity of the model also demands a greater amount of weak supervision to sufficiently learn. The greater cost of training the seeded LDA is clear as it performs better with frequency-based seed words than knowledge-based seed words. Nevertheless, we admit that further investigation is required to make conclusive remarks on the seeded LDA models. The model would outperform Newsmap when documents are longer or it is trained with a sufficiently large number of seed words.

The semisupervised classification technique is particularly useful when researchers apply quantitative text analysis in new fields because training sets for full-supervised models are very expensive to produce and keyword dictionaries are often unavailable outside comparative politics or political communication and nearly nonexistent in non-European languages. While it is infeasible for underresourced researchers to create a large keyword dictionary from scratch, they can easily create a seed word dictionary for semisupervised models because the number of words required for a seed word dictionary is a fraction of a keyword dictionary.

Semisupervised document classification is also an alternative approach to multilingual text analysis, which is becoming increasingly important in recent years. Multilingual analysis can be achieved by translating a keyword dictionary (Proksch et al., n.d.) or a corpus (De Vries et al., 2018), but it can also be done only by translating a seed word dictionary if a semisupervised model is employed. In fact, Newsmap performs multilingual geographical classification of news articles using seed word dictionaries in more than 10 European and non-European languages.[8]

However, we recognized that the difficulty in making a proper seed word dictionary is an obstacle for researchers to employ semisupervised models. For this reason, we created the AFE statistic as a diagnostic tool that does not require manually coded documents as the gold standard. We argue that seed word dictionaries should be constructed based primarily on theory, but they should also include frequent words to produce good classification results. Inclusion of frequent false seed words to a seed word dictionary hugely damages the classifier's performance, but researchers can reduce this risk by monitoring relative changes in AFE.

The AFE statistic allowed us to achieve a higher classification accuracy by removing many of the frequency-based seed words in "greeting," "human rights," and "democracy" in our experiment, but it is still likely that some of them were true seed words that could improve the classifier's performance if included. This did not occur in our experiment probably because our knowledge-based seed words were not entirely appropriate, and some of the true frequent seed words were wrongly identified as risky seed words. This shows how difficult it is to operationalize abstract concepts such as human rights and democracy using words, but we believe it is possible as a collective endeavor: Researchers who employ semisupervised models should publish their seed word dictionaries to allow others to refine them by adding or removing words; words that are commonly used in these seed word dictionaries will be the knowledge-based seed words that are independent of corpora.

Contextual smoothing of topic probability is an auxiliary technique in this study, which is not directly related to the semisupervised models, but we believe that this technique can be useful in many studies. We expect that the contextual smoothing technique would improve classification

accuracy of sentences in various machine learning models without changing model assumptions and specifications. This is the primary reason that we decided to train the Newsmap model on individual sentences instead of temporal groups of sentences, although training on both labeled sentences and surrounding unlabeled sentences is closer to semisupervised learning techniques in computer science (Zelikovitz & Hirsh, 2000).

Nevertheless, contextual smoothing does not always perform well as we saw in the experiment: The classifier's performance deteriorated when it was applied to relatively short speeches. In fact, we found a positive correlation between the number of sentences and the changes in the F1 scores ($r = .38$), which suggest that the window size should be smaller depending on the length of the speech that sentences compose. More importantly, contextual smoothing has negative impact on the classification results when topics changes between sentences in any length of speech. Although we did not develop algorithms to adjust window sizes to tackle this problem, we expect that common words in adjacent sentences will offer useful information to achieve this in future research.

Another limitation of this study is that we only tested our approach on the corpus of the UN General Assembly speeches. We believe that the proposed seed word selection procedure holds conceptually in wider research settings, but the AFE statistic would behave differently depending on the types of texts and the categories for classification. Since AFE is created for classification of sentences, which usually have only one topic each, its behavior can be different when it is applied to entire documents. For this reason, we recommend users of AFE to segment documents into the smallest unit (e.g., sentences or paragraphs) in seed word dictionary making. After a seed word dictionary has been created, they can apply the dictionary to train a Newsmap model on entire documents. This seems possible because Newsmap has already demonstrated its ability to classify longer texts correctly (Lankina & Watanabe, 2017; Watanabe, 2017) despite its geographical dictionary was created for news summaries originally. Relatedly, classification categories should also be mutually exclusive for AFE because it gauges co-occurrences of words with seed words in different categories to detect risky seed words. To avoid this problem, we recommend users to define smallest categories in a dictionary and aggregate them after classification. Newsmap takes similar approach in classification of oversea territories of European countries (e.g., British Virgin Islands and French Guiana).

Finally, we hope the readers of this article understand the strengths and weaknesses of semisupervised models vis-à-vis popular unsupervised topic models and simple dictionary analysis. Use of semisupervised models requires knowledge of both methodology and substance, but there will be an increasing number of young political scientists who understand both aspects. We hope our work will encourage them to embark on quantitative analysis of documents in understudied subjects and languages on the horizon of quantitative text analysis.

Data and Software Availability

The United Nations General Debate Corpus is available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/0TJX8Y

Data and R syntax are available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FGSDZNV

We have used the following packages in this study. The AFE measure is also made available as part of Newsmap package: Quanteda (https://cran.r-project.org/web/packages/quanteda/), Seeded-LDA (https://github.com/koheiw/quanteda.seededlda), Newsmap (https://cran.r-project.org/web/packages/newsmap/), Ldatune (https://cran.r-project.org/web/packages/ldatuning/)

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The International Component for Unicode library segments texts with acronyms with dots erroneously, but this problem was solved in its Version 56. Although this new functionality was not available when this study was conducted, the number of errors is small in our data because the *quanteda* package has special handling for titles (e.g., "Mr." and "Dr.").
2. We found that 20 is optimal number of topics for latent Dirichlet allocation using methods developed by Cao et al. (2009) and Deveaud et al. (2014) through the *ldatune* package.
3. The F1 score is a harmonic mean of precision and recall. Precision measures the proportion of items classified as X are true X; recall measures the proportion of items classified as X in all true X.
4. The seeded LDA model is available in the R package, *topicmodels* (Grün & Hornik, 2011).
5. The average feature entropy function is implemented in the *newsmap* package (available on CRAN).
6. An example of such sentences without a clear indication of topic is "The most immediate area of concern for the international community is the situation in Yugoslavia" (Speech by Australia in the UNGA corpus, 1991). This sentence can be classified into "security" only by looking at its previous sentence "We are all acutely conscious of how newly emergent nationalism within the borders of many existing countries around the world . . . ." Another example is "We once again underline the importance of those values in spreading security, peace and stability . . . and sustainable development in a spirit of optimism and confidence in a better future" (Speech by Tunisia in the UNGA corpus). This sentence relates to multiple topics, but it should be classified as "security" because it is the most salient.
7. We set pseudocounts for the seed words to 500 (roughly .1% of the total number of sentences), but the choice of the value did not affect the result much.
8. Languages currently available are English, German, Spanish, French, Italian, Russian, Arabic, Hebrew, Japanese, and Chinese.

References

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *The American Political Science Review*, *95*(3), 529–546.

Albugh, Q., Sevenans, J., & Soroka, S. (2013). *Lexicoder topic dictionaries, June 2013 versions*. McGill University. http://www.lexicoder.com/docs/LTDjun2013.zip

Banerjee, S., Ramanathan, K., & Gupta, A. (2007). *Clustering short texts using Wikipedia*. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 787–788. https://doi.org/10.1145/1277741.1277909

Banta, B. (2013). Analysing discourse as a causal mechanism. *European Journal of International Relations*, *19*(2), 379–402.

Baturo, A., Dasandi, N., & Mikhaylov, S. J. (2017). Understanding state preferences with text as data: Introducing the UN general debate corpus. *Research & Politics*, *4*(2), 2053168017712821. https://doi.org/10.1177/2053168017712821

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*, 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blum, A., & Mitchell, T. (1998). *Combining labeled and unlabeled data with co-training*. Proceedings of the Eleventh Annual Conference on Computational Learning Theory, 92–100. https://doi.org/10.1145/279943.279962

Brunn, S. D. (1999). The worldviews of small states: A content analysis of 1995 UN speeches. *Geopolitics*, *4*(1), 17–33.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Chapelle, O., Schölkopf, B., & Zien, A. (Eds.). (2010). *Semi-supervised learning*. The MIT Press.

Däubler, T., Benoit, K., Mikhaylov, S., & Laver, M. (2012). Natural sentences as valid units for coded political texts. *British Journal of Political Science*, *42*(4), 937–951. https://doi.org/10.1017/S0007123412000105

Deveaud, R., Sanjuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Revue Des Sciences et Technologies de l'Information—Série Document Numérique*, 61–84. https://doi.org/10.3166/DN.17.1.61-84

De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that Google translate works for comparative bag-of-words text applications. *Political Analysis*, *26*(4), 417–430. https://doi.org/10.1017/pan.2018.26

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, *18*(1), 1–35. https://doi.org/10.1093/pan/mpp034

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 1–31. https://doi.org/10.1093/pan/mps028

Grün, B., & Hornik, K. (2011). Topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), 1–30.

Gurciullo, S., & Mikhaylov, S. (2017). Topology analysis of international networks based on debates in the United Nations. *ArXiv:1707.09491 [Cs, Math, Stat]*. http://arxiv.org/abs/1707.09491

Hansen, L. (2006). *Security as practice: Discourse analysis and the Bosnian war*. Routledge.

Holzscheiter, A. (2014). Between communicative interaction and structures of signification: Discourse theory and analysis in international relations. *International Studies Perspectives*, *15*(2), 142–162.

Huber, E., Rueschemeyer, D., & Stephens, J. D. (1993). The impact of economic development on democracy. *Journal of Economic Perspectives*, *7*(3), 71–86. https://doi.org/10.1257/jep.7.3.71

Jackson, R. (2005). *Writing the war on terrorism: Language, politics and counter-terrorism*. Manchester University Press.

Jagarlamudi, J., Daumé, H. III, & Udupa, R. (2012). *Incorporating Lexical priors into topic models*. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 204–213. http://dl.acm.org/citation.cfm?id=2380816.2380844

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

Lankina, T., & Watanabe, K. (2017). 'Russian spring' or 'spring betrayal'? The media as a mirror of Putin's evolving strategy in Ukraine. *Europe-Asia Studies*, *69*(10), 1526–1556. https://doi.org/10.1080/09668136.2017.1397603

Lu, B., Ott, M., Cardie, C., & Tsou, B. K. (2011, December 11). *Multi-aspect sentiment analysis with topic models*. 2011 IEEE 11th International Conference on Data Mining Workshops, 81–88.

Lundborg, T., & Vaughan-Williams, N. (2015). New Materialisms, discourse analysis, and international relations: A radical intertextual approach. *Review of International Studies*, *41*(1), 3–25.

Milliken, J. (1999). The study of discourse in international relations: A critique of research and methods. *European Journal of International Relations*, *5*(2), 225–254.

Nesser, P. (2008). Chronology of Jihadism in Western Europe 1994–2007: Planned, prepared, and executed terrorist attacks. *Studies in Conflict & Terrorism*, *31*(10), 924–946. https://doi.org/10.1080/10576100802339185

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). *Learning to classify short and sparse text and web with hidden topics from large-scale data collections*. Proceedings of the 17th International Conference on World Wide Web, 91–100. https://doi.org/10.1145/1367497.1367510

Proksch, S.-O., Lowe, W., Wäckerle, J., & Soroka, S. (n.d.). Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*. https://doi.org/10.1111/lsq.12218

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., Albertson, B., & Rand, D. (2014). Structural topic models for open ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082.

Sartori, R., & Pasini, M. (2007). Quality and quantity in test validity: How can we be sure that psychological tests measure what they have to? *Quality & Quantity*, *41*(3), 359–374. https://doi.org/10.1007/s11135-006-9006-x

Schoenfeld, M., Eckhard, S., Patz, R., & van Meegdenburg, H. (2018). Discursive landscapes and unsupervised topic modeling in IR: A validation of text-as-data approaches through a new corpus of UN Security Council speeches on Afghanistan. *ArXiv:1810.05572 [Cs]*. http://arxiv.org/abs/1810.05572

Schönhofen, P. (2009). Identifying document topics using the Wikipedia category network. *Web Intelligence and Agent Systems*, *7*(2), 195–207. https://doi.org/10.3233/WIA-2009-0162

Smith, C. B. (2006). *Politics and process at the United Nations: The global dance*. Lynne Rienner.

Suzuki, S. (2015). The rise of the Chinese "Other" in Japan's construction of identity: Is China a focal point of Japanese nationalism? *The Pacific Review*, *28*(1), 95–116.

Watanabe, K. (2017). The spread of the Kremlin's narratives by a western news agency during the Ukraine crisis. *The Journal of International Communication*, *23*(1), 138–158. https://doi.org/10.1080/13216597.2017.1287750

Watanabe, K. (2018a). *Conspiracist propaganda: How Russia promotes anti-establishment sentiment online?* ECPR General Conference.

Watanabe, K. (2018b). Newsmap: A semi-supervised approach to geographical news classification. *Digital Journalism*, *6*(3), 294–309. https://doi.org/10.1080/21670811.2017.1293487

Zanotti, L. (2005). Governmentalizing the post–Cold War international regime: The UN debate on democratization and good governance. *Alternatives*, *30*(4), 461–487.

Zelikovitz, S., & Hirsh, H. (2000, June 29). *Improving short text classification using unlabeled background knowledge to assess document similarity*. Proceedings of the 17th International Conference on Machine Learning, Stanford, CA, USA.

Author Biographies

Kohei Watanabe is senior assistant professor at the Digital Science Centre and the Political Science Department of the University of Innsbruck (Austria). He is also a visiting researcher at Waseda University (Japan) and the London School of Economics (UK). He studies international political communication employing quantitative text analysis tools that he develops as open-source software.

Yuan Zhou is a PhD student at Kobe University. He specializes in international relations and political communication, focusing on China. He also works in quantitative research methods, especially quantitative text analysis.