Advance Access publication January 29, 2014

Political Analysis (2014) 22:224–242
doi:10.1093/pan/mpt030

Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines

Vito D'Orazio
Institute for Quantitative Social Science, Harvard University
e-mail: dorazio@iq.harvard.edu (corresponding author)

Steven T. Landis
National Center for Atmospheric Research, FL2-2096, 3450 Mitchell Lane, Boulder, CO 80301
e-mail: landis.steven@gmail.com

Glenn Palmer
Department of Political Science, Pennsylvania State University
e-mail: gpalmer@psu.edu

Philip Schrodt
Parus Analytical Systems, 100 N. Patterson St., State College, PA 16801
e-mail: schrodt735@gmail.com

Edited by R. Michael Alvarez

Due in large part to the proliferation of digitized text, much of it available for little or no cost from the Internet, political science research has experienced a substantial increase in the number of data sets and large-*n* research initiatives. As the ability to collect detailed information on events of interest expands, so does the need to efficiently sort through the volumes of available information. Automated document classification presents a particularly attractive methodology for accomplishing this task. It is efficient, widely applicable to a variety of data collection efforts, and considerably flexible in tailoring its application for specific research needs. This article offers a holistic review of the application of automated document classification for data collection in political science research by discussing the process in its entirety. We argue that the application of a two-stage support vector machine (SVM) classification process offers advantages over other well-known alternatives, due to the nature of SVMs being a discriminative classifier and having the ability to effectively address two primary attributes of textual data: high dimensionality and extreme sparseness. Evidence for this claim is presented through a discussion of the efficiency gains derived from using automated document classification on the Militarized Interstate Dispute 4 (MID4) data collection project.

1 Introduction

Large data collection projects require an efficient and systematic method of information retrieval to meet the goal of obtaining knowledge on some specific event of interest. Whether it be voting patterns, political systems, or international conflict events, a researcher often embarks upon this task with a simple idea in mind: "I want to collect information on topic X." However, after making this decision there is often little guidance for taking the next step in the research process. Specifically, how does one identify relevant information on X and, after identifying this information, how can it be efficiently coded into structured data? Although researchers face these questions

Authors' note: The authors would like to thank Emre Haptipoglu, Matthew Lane, and Michael Kenwick for their work on the MID4 project. We would also like to thank the editors of *Political Analysis* and the anonymous reviewers for their insight and constructive comments. Supplementary materials for this article are available on the *Political Analysis* Web site.

each and every time they begin any data collection project, there is often no road map to guide them in their efforts. This article shares some insights on this important but under-discussed step of scientific inquiry.

We treat the initial steps for data collection as an information retrieval problem where we have some information need and our primary goal is to construct a document set that contains sufficient information to meet that need (Rijsbergen 1979). Ultimately, human eyes will read every document, and so our secondary goal is to make the document set as small as possible, while not sacrificing its validity in terms of meeting our information need. In an age where information resources are vast and ever-expanding, meeting these goals in a timely and efficient manner is best accomplished with automated document classification.

This article provides a holistic review of the application of automated document classification for data collection in political science research. In doing so, we discuss the process in its entirety and direct readers to some relevant research examples within both political science and related fields. We recommend that data collectors classify early, often, and creatively. Every data collection project is different, and the appropriate application of classification methods depends very much on a researcher's knowledge of the document set he or she is trying to classify and the relevant document set he or she desires.

Specific to these later stages, we argue that the application of a two-stage support vector machine (SVM) classification process offers advantages over other well-known alternatives in a variety of settings. SVMs offer several important benefits that make them extremely attractive for political science research. Stemming from their nature as a linear, discriminatory classification method, SVMs are capable of handling the two primary attributes of textual data: high dimensionality and extreme sparseness (Joachims 1998; Aggarwal and Zhai 2012). Specific to data collection, SVMs have been shown to perform particularly well when classifying a set of observations where there are relatively few labeled documents and a large number of unlabeled documents (Sindhwani and Keerthi 2006). This is an especially important benefit for data collection where the goal is efficiency and labeling documents is costly, typically requiring a representative sample of both relevant and irrelevant documents from an extensive corpus, as well as labor-intensive human coding.

The advantages of this two-stage SVM method are illustrated by showing how it was used in the most recent update of the Correlates of Wars Militarized Interstate Dispute (MID4) data set. From a set of approximately 1.74 million text documents on potential militarized interstate incidents (MIIs, our event of interest), this system is capable of generating an average reduction of greater than 92%. As a result, human coders read only a fraction of the documents initially collected, greatly improving MID4's efficiency. Through comparing event distributions with previous MID collection efforts, we conclude that the performance equals if not surpasses that of past expansions to the MID database.

Although largely project dependent, the greatest efficiency gains derived from automated document classification are to be had on data collection projects where the social concept being measured is complex and classifying the document set requires a level of pattern recognition that cannot be attained without statistical methods. For example, the 297 militarized interstate disputes that have taken place between 1993 and 2001 are composed of 2119 incidents. These incidents range from verbal threats by state officials, to the seizure of goods on a fishing vessel, to aerial bombardments of infrastructure. Without automated methods, it would be exceedingly difficult to find information on each nuanced type of incident that exists.

The two-stage implementation is particularly attractive in cases where the structure of the document set varies according to some meta-data attribute. Such variations may preclude a uniform training model, and several subset-specific training models may be more desirable. For example, our second set of training models are estimated on yearly subsets of the data to adjust for the changes in language used to describe MIIs over time. One can imagine a comparable task where training by source, speaker, author, or some other field of meta-data is desirable.

As applications of automated text classification increase, it is likely that most data collection projects will move toward increased automation. Doing so will not only improve the quality and quantity of available data, but it also moves the replicability of the scientific undertaking back to

the origin of the data and not just the origin of the analysis, which is common for replication materials today. By ensuring the integrity of the data, confidence in subsequent analyses improves. Furthermore, to facilitate the application of this information retrieval system to other data collection projects, all software used is completely open-source.[1]

2 Collecting Observational Data

In studying political events for quantitative analysis, researchers often find themselves collecting observational data on some concept of interest. The routine method for doing so involves a team of research assistants (RAs), headed by a principal investigator (PI) or data manager, reading and coding some archive of text documents that were generated either electronically or contained within an existing historical archive (e.g., a library). Prior to RA coding, the PI's primary responsibilities center on establishing an archive that is both relevant and manageable—relevant meaning the archive contains sufficient information to populate an unbiased data set, and manageable meaning the archive is small enough for the research team to finish coding within some realistic project deadline. Depending on the scale of the project, such an archive can be something as limited as a few historical sources to something as expansive as *Google News*. In any case, the combination of human coding and source selection ultimately limits which documents are coded and which are not, resulting in an archive with an inability to fully populate the desired data set.

The ideal archive is one in which all relevant information on a topic is included, with each document containing unique and accurate information, with no redundancy across documents, and with no irrelevant documents contained within the archive. Since reality prevents obtaining such an archive, researchers instead seek to maximize both precision and recall in the pursuit of the perfectly relevant and manageable. Recall refers to the ratio of relevant information in the archive to the total amount of relevant information in existence. Precision refers to the ratio of relevant information in the archive to the total amount of information in the archive. Ideally, maximizing both precision and recall would result in the relevant information in the archive equaling the relevant information in existence, with all information contained in the archive being relevant to the data collection effort.

The pursuit of maximizing recall and precision within data collection projects is fraught with difficulties, however. To clarify, let Ω be the set of all accessible documents. The initial subset from Ω is typically based on source selection, a set of keywords, or both. In any effort to collect observational data, the sources selected and the keywords searched represent the largest cut from Ω. Researchers must take these initial decisions seriously and be mindful of the primary and secondary goals put forth above: construct a document set with sufficient information and make it small enough to be efficient.

For subsets of Ω with sufficient recall and poor precision (i.e., an unnecessarily large document set), the concern is efficiency. Funding is always temporary, and having RAs spend their time sorting through stacks of documents is wasteful. Convincing a funding agency of a project's efficiency a priori is a simple way to circumvent common data collection criticisms based on the project's efficiency.

In developing an efficient project (i.e., increasing precision), researchers must be sure not to sacrifice the sufficiency of their recall, lest the resulting data be biased. For example, it is not uncommon to split unnecessarily large document sets by examining a select number of sources. However, doing so requires careful validity assessment because a biased source selection results in a biased data set. Unfortunately, given the large number of available sources and ways in which the concepts of interest may be described, bias is often induced into the data at these very early stages by selecting sources with limited coverage or not constructing a representative set from the start.

Given these issues, our recommendation is to initially construct an extensive document set such that one can say with near certainty that the set contains all relevant information that one can hope

[1]The processing software and documentation are available at D'Orazio et al. (2013). The SVM software is available at http://svmlight.joachims.org/. LexisNexis is a database of copyrighted material and requires a subscription to download documents, but we do provide the search string and source list used in the supplementary materials.

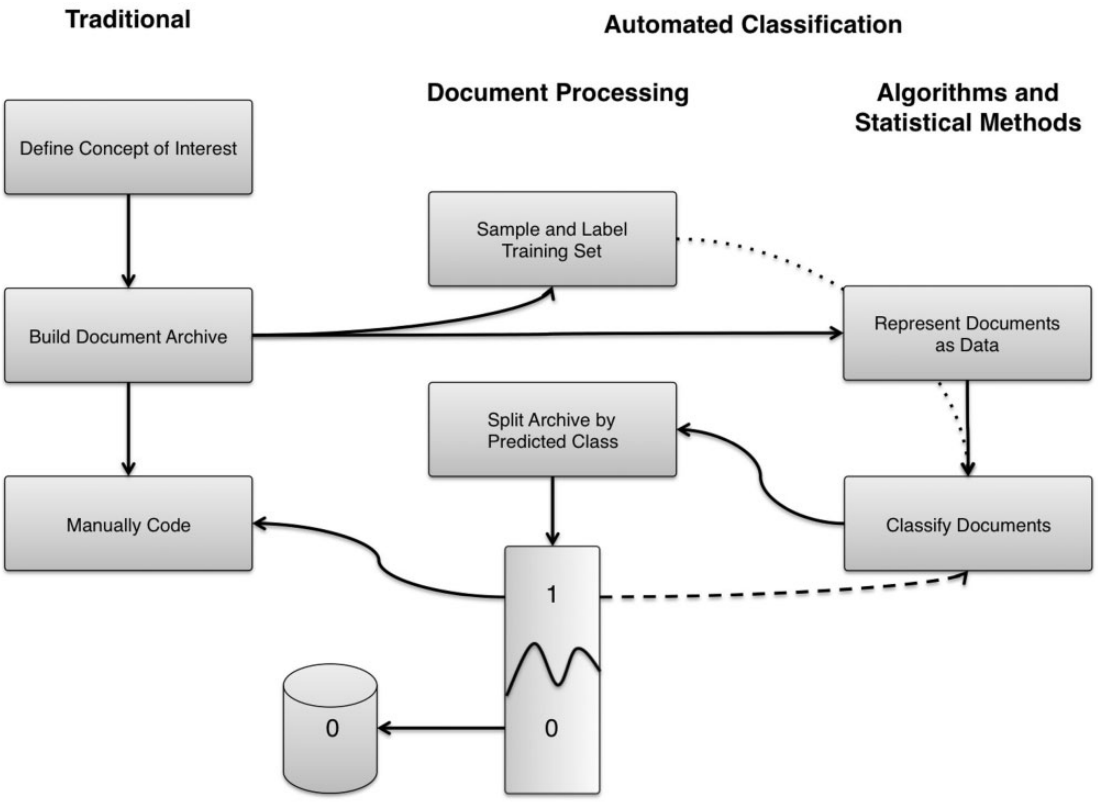

Fig. 1 The data collection process.

to reasonably access. Thus, the goal is to split Ω such that one subset could not conceivably contain information that is both pertinent to the research effort and not contained in the other subset. In a practical sense, we recommend *more* sources, *more* keywords, and *fewer* exclusion parameters in the initial search.

Figure 1 shows the traditional data collection process on the left and introduces the method of automated document classification in the middle and right sections. Because the document archive that is to be manually coded undergoes some number of automated classifications, we can afford to make it exceedingly large to ensure that we have not discarded relevant information. Rather than manual labor, the intent is use advanced methods of pattern recognition to trim this archive down to a codeable size. Should a single iteration of classification not meet the project's definition of a codeable size, the positively classified documents (labeled "1" in the lower-middle object) may be subjected to another phase of classification, whereas the documents classified as irrelevant (labeled "0") are stored accordingly. In the following pages, we discuss these steps in greater detail.

2.1 *Constructing the MID4 Archive*

In the MID application, operational definitions have been established by previous researchers (Gochman and Maoz 1984; Jones et al. 1996); therefore, the goal is not to modify or refine those rules, but to simply populate the database according to those same rules for 2002–2010.[2]

[2]A "militarized incident is defined as a single military action involving an explicit threat, display, or use of force by one system member state towards another system member state" (Jones et al. 1996, 169). Each dispute is composed of a set of events, or MIIs. For our purposes, MIIs are the events of interest.

Many data collection efforts are similar in their goal of expanding existing data sets, and so this example assumes that the operational definition is complete.[3]

Our initial approach for data collection is guided by past MID efforts, notably that of MID3. The issues facing MID3 resemble the general class of issues facing most data collectors in an information-rich world and are not unique to this application. For its time, MID3 employed a relatively new method of informational retrieval using the LexisNexis (LN) Universe for its data collection (Ghosn et al. 2004). Databases such as LN make possible a keyword search on an extremely large number of text documents from a variety of news sources, resulting in volumes of potentially relevant information. Whenever research projects rely on these types of databases that use keyword searches to query large textual archives, inefficiency is a common problem because the retrieval method is imprecise. Given the diversity of word usages in spoken language, keywords often appear in a variety of settings that are unrelated to the research topic. For instance, querying words like "destroy" or "fight" may return stories about MIIs, but they also return stories on other topics such as sporting events and weather reports. Given the focus on obtaining as much information as possible, the larger concern is on improving retrieval efficiency.

Unfortunately, MID3 lacked systematic and clear definitions for extracting documents from LN. The news sources used in its information retrieval were unsystematic and comprised a variety of general, world, and newswire services. Researchers would query regional or country-specific search terms in disjointed fashion—sometimes cities, sometimes geographic features, sometimes specific actors, and sometimes all of them at once. As a result, the information retrieved was researcher-dependent and of considerable variance, which was further complicated by the fact that MID3 was a multi-institutional project, and generally these institutions did not coordinate their approaches on a day-to-day basis. This retrieval method resulted in two main sources of inefficiency, which are relatively common in most data collection efforts.

The first is in the unpredictability of the database query, which can only be solved by using a consistent set of search terms, exclusion parameters, and carefully selected sources. This ensures that documents are extracted from the LN universe in a systematic way, and should any types of MIIs be missed, they will be missed systematically, which eases the recognition and correction of this oversight. The MID4 search string and the sources selected can be found in the supplementary materials, but it is sufficient to know that its goal is to construct an archive based on consistent and known rules that contains all relevant information found in the LN universe and without it limiting the eventual size of the archive.

The second and more serious contributor to inefficiency is the degree of manual document classification. With MID3, every query of LN required researchers to manually classify some number of stories as either "relevant" or "irrelevant." Sorting through information in this manner is problematic for at least three reasons. First, it is costly on a project where decisions are made with respect to efficiency. Second, the high volume of classification leads to coder fatigue, which increases the likelihood of misclassified documents. Third, the classification is researcher-dependent because what constitutes a relevant document for one coder might not qualify as relevant for another. Specific to MID4, the use of a systematic global search string and selected sources returned an initial archive that exceeded 1.74 million news documents, thus precluding manual classification at this phase under any circumstances. Fortunately, machines have proven very capable of accurately classifying documents and resolving many of these issues.

3 Automated Document Classification

Automated document classification has become a permanent fixture of machine learning research (Sebastiani 2002; Kolari et al. 2006; Britt et al. 2008; Aggarwal and Zhai 2012). Notable examples of this method in political science include classifying party affiliation and political opinion (Shulman 2005; Yu et al. 2008). For the purposes of data collection, the specific intent is to eliminate as many irrelevant documents as possible while retaining all of those that contain information

[3]For examples of research on defining social science concepts, see Adcock and Collier (2001), Goertz (2006), or Sartori (1984).

relevant to the topic of interest. Thus, despite the international relations focus of this article, automated document classification is useful across a variety of political science subfields where the goal is to classify observational data derived from text documents. One may imagine that the problems faced here would similarly apply to research topics as diverse as coding the salience of court opinions and agenda setting in American politics (Barid 2004), to coding instances of farmer-herder conflict from police reports in the developing world (Witsenburg and Adano 2009). Overall, when classifying in this manner, the objective is to maximize precision while maintaining sufficient recall.

"Sufficient recall" is slightly less strict than "perfect recall." Perfect recall implies that *all* relevant documents are obtained, which is usually neither desirable nor feasible. Sufficient recall, on other hand, refers to a situation where one has all relevant documents that contribute *unique* information and are necessary for ensuring data validity. For example, in the context of news reports, it is often the case that events are initially reported in a very terse manner. Future reports on the same event build upon previous ones, containing all the information mentioned earlier plus additional facts that have been discovered. In a system designed for perfect recall, *all* these reports would be labeled as relevant. In addition to the manual coding inefficiency, the multiple reports that are delivered in this fashion are less valuable for ensuring data validity than are documents delivered from other sources or written independent of the initial reports. For these reasons, sufficient recall is the objective.

During MII processing, each year is classified separately, with each annual news archive containing approximately 200,000 documents. The classification process is carried out in three steps. First, there is a preprocessing and feature selection process, which is leveraged at points to discard additional documents. The second step is applying the training model developed by Schrodt et al. (2008) and subsequent classification using inductive SVMs. Finally, there is the novel implementation of year-specific training models and classification using transductive SVMs. Taken together, this MID-specific application provides an example of a successful, labor-saving application of automated document classification to data collection efforts.

3.1 *Preprocessing and Feature Selection*

Prior to classification, structured data representing the document set is required.[4] To obtain such data, the documents are formatted identically using preprocessing (or filtering) software, which is customized for the various formats found within the LN downloads.[5] These filters identify meta-data such as the headline, the news source, and the date, as well as distinguishing the beginning and end of an individual document and the location of the textual report within the document.

The desired meta-data are a function of the specific project and need not be limited to, or only include, the meta-data extracted from news reports. For example, if one's goal is to classify legislator speeches, the meta-data might include the legislator's name, district, and the time of the speech. If one's objective is to classify US Supreme Court opinions, the extracted meta-data might differ yet again. In this stage, what is vital for text representation purposes is that each document has a unique key and a specific set of characters that delimit the start and end of each textual report.

Figure 2 shows a screenshot of a news report after preprocessing. One will notice that the document is delimited from others by a line of dashes. The document key, found two rows below the delimiter, is easily locatable for future identification purposes. Other meta-data, specifically the headline, date, and source, are also included.

With the documents stacked in semi-structured form, one can now select meaningful textual features and represent the text as fully structured data by coding each document according

[4]Replication for the preprocessing and feature selection can be found in D'Orazio et al. (2013). For new applications, we suggest using PreText, an original, open-source software package developed for LN document processing and representation. See http://vitodorazio.weebly.com/pretext.html for more details.

[5]We used, with modifications, filters originally developed and made available by the pilot MID4 project (Schrodt et al. 2008).

```
---------------------------------------------------------------
Tajik border guards exchange fire with alleged Afghan drug smugglers
20060217--0012-Feb17_2006_LN_NP1.TXT-files.list
February 17, 2006 Friday 10:49 AM GMT
(c) Associated Press Worldstream

Tajik border guards seized about 60 kilograms (132 pounds) of heroin after an
exchange of fire with a group of alleged Afghan smugglers, an official said
Friday.

One of the Afghans was detained Thursday and the others fled back to Afghan
territory, said Faizullo Gadoyev, a department chief at the Drug Control Agency.

He said the smugglers opened fire after being spotted by border guards. No
one was hurt, he said.

The impoverished former Soviet republic in Central Asia is a major route for
narcotics traffickers from Afghanistan.

---------------------------------------------------------------
```

Fig. 2 Filter output example.

to those features.[6] There are many ways to go about selecting features for text, and a comparison of these methods is beyond the scope of this article (for a variety of comparisons, see Forman 2003; Guyon and Elisseeff 2003; Lowe 2008; Monroe et al. 2008). Rather, MID4 employs a basic method of feature selection by using unigrams (single words), removing all common stopwords, utilizing document frequency thresholding, and removing all proper nouns. The resulting data set is an N by K structured data set where N is the number of documents and K is the number of unique features.[7]

Stopword removal and document frequency thresholding have been standard approaches in natural language processing literature for some time (Luhn 1958; Rijsbergen 1979; Dasgupta et al. 2007) and are found in political science applications (Hopkins and King 2010; Spirling 2012), although some have made arguments against these practices (Monroe et al. 2008). The remaining two components of the feature selection—the removal of proper nouns and the use of unigrams—are less standard.

Generally speaking, researchers will want to strip proper nouns when it is believed they will bias the classification algorithm. In our application, proper nouns associated with nation-states are removed as a way of increasing the possibility that stories will be classified based on their content and not any state-specific criteria. For example, consider the following three statements:

1. Israeli aircraft violated Lebanon's airspace.
2. Brazilian aircraft violated Uruguay's airspace.
3. The prime ministers of Israel and Lebanon held a meeting.

As per the definition of an MII, Items 1 and 2 are to be classified as relevant and Item 3 to be classified as irrelevant. Because "Israel" and "Lebanon" are commonly found unigrams in documents relevant for our purposes, if these unigrams are not stripped from our feature set, the worry would be that *all* documents containing "Israel" or "Lebanon" (or their variants) would be classified as relevant, regardless of the context. Similarly, it is not very common for "Brazil" and "Uruguay" to be found in a document pertaining to an MII, but it is certainly possible for these two states to engage in a militarized dispute.

[6] In this context, semi-structured refers to the fact that entire documents, meta-data, and the textual report are delimited in consistent fashion across the entire document set.

[7] PreText has a variety of built-in feature selection methods from which users can choose the most appropriate. Alternative software options for feature selection include the tm package in R and the Natural Language Toolkit in Python.

The need to remove specific features is a common problem that exists elsewhere because textual documents are largely similar. For example, classifying instances of state repression in Latin America may require the removal of country names, whereas classifying US Supreme Court opinions may require the removal of the Justices' names or the words "plantiff" and "defendant." Overall, these specific features of text are best ignored in an effort to focus the classifier on what is most important: the content of the story.

The removal of proper nouns entails a basic named entity recognition (NER) task.[8] Generally, NER is accomplished by either using algorithmic or statistical methods for recognition or matching *n*-grams in the text to a dictionary of known named entities (for a review, see Nadeau and Sekine 2007). MID4 employs the latter by utilizing an XML database of country-specific information and matching *n*-grams in the text to those in the database.[9]

Recognized named entities constitute meta-data that researchers can use to leverage their substantive knowledge of the task at hand to improve their project's efficiency and classification performance. Our procedure uses the two most frequently mentioned countries as a "best guess" for identity of the dyad of interest, a step which makes the human coding considerably more efficient. Additionally, any report that does not mention at least two nation-states is removed—by our definition, a relevant document must mention at least two countries in order to be an MII.

The removal of documents from further processing based on the presence of multiple words, specific phrases, or combinations of words is commonly referred to as the RIPPER technique (Cohen 1996; Basu et al. 1998). Keyword searching may be thought of as a special case of the RIPPER method whereby the presence of a word is sufficient for its inclusion. However, the RIPPER technique also applies to the presence of common word stems, variants of named entities (i.e., United States or America), and word frequencies, among others. This technique has a wide array of applications in text classification and has been shown to be a consistently effective method (Cohen and Singer 1999).

For some data collection projects, the RIPPER technique may itself provide sufficient recall and an acceptable level of precision. For example, if one were to collect data on drug-related violence in a specific geographic location, it may suffice to extract only those documents that mention the geographic location and contain some number of references to words such as "drugs" or "dealers." Alternatively, if one is researching opinions pertaining to US capital punishment, there are a limited number of ways one can report a story about the death penalty, and a creative researcher could develop a set of rules that encompass the varieties. Some of these stories can be extracted using keyword searches but, as in cases where named entities are involved, occasionally one may need more customizable approaches.

Appropriate features for representing the documents can be limited to individual words (unigrams) or be composed of some mixture of words and phrases (*n*-grams). In this application, the features are limited to individual words and therefore cannot contain multi-word phrases. This approach induces some information loss because phrases consisting of multiple words hold more semantic content than individual words (Papka and Allan 1998; Spirling 2012). However, the use of individual words as features for text classification has been shown to work in general settings (Lewis 1992; Yang and Pedersen 1997; Hopkins and King 2010) as well as in an MID-specific setting (Schrodt et al. 2008), and so we opt for the more simplistic unigram approach.

Although we do not deal directly with plurals, tenses, synonyms, or disambiguation, a brief discussion is warranted, as these methods are often applied to feature selection processes.[10] Stemming consists of representing the root of a word as the feature. When stemmed, plurals, for example, appear identical to the same word in singular form. Past, present, and future tenses may

[8] In a special issue of the *Journal of Information Technology and Politics*, Cardie and Wilkerson (2008) introduce several articles that discuss related NER tasks in political science research.

[9] The original database has been extended into the 32,000-line file called CountryInfo.txt, available at http://eventdata.psu.edu/software.dir/dictionaries.html, and includes major cities, lists of national leaders, and variants of country names.

[10] For the MID classifications, we did not stem because it had not yet been implemented in our software. The current version of PreText comes with the Porter algorithm and has the option for adding other stemming algorithms.

also appear identical. The extent and rules for stemming depend on the chosen stemming algorithm. Examples include the Porter Stemmer (Porter 1980), the Paice/Husk algorithm (Paice 1994), and distribution-based *n*-gram stemming (Mayfield and McNamee 2003). In a basic sense, stemming has the effect of mapping different tokens from the text to the same feature in the data set; synonymy poses a related problem.

Synonyms are words or phrases that have similar meanings. In the context of news stories, this issue is exacerbated since repetition is discouraged and synonyms are intentionally used to make the document read smoother. In an ideal case, and similar to how plural and singular form tokens are mapped to the same feature, when selecting features synonyms should be treated as equivalent entries. This problem is discussed at length by Mohammad and Hirst (2006) and Bikel and Castelli (2008), and methods of synonym recognition are evaluated by Wang and Hirst (2012). Such methods include distributional methods, which evaluate texts for the appearance of terms in nearly identical contexts, and dictionary-based methods, which use variations of thesaurus-style matching. From a computational perspective, it is important to note that not only are dictionary-based shown to be at least as good as distributional methods, but they are much more efficient (Wang and Hirst 2012).

Disambiguation is another potential issue that arises in text analysis, as the same word or words may have different meanings in different contexts (Blair 1992; Blair 2003). A "battle" between Russia and Georgia should be distinguished from a "battle" between rival sports teams. A related need for disambiguation can arise from the use of homonyms. For example, "forearm" can be used to refer to the length of arm between the wrist and elbow, or it can be used to refer to equipping oneself with weapons in advance of some event. In practice, however, this sort of context-based disambiguation is very difficult to achieve, and the payoffs from such disambiguation may not be worth the computational costs. Research in the area of disambiguation has, for the most part, been limited to the disambiguation of named entities (Hoffart et al. 2011).

With the documents processed and features selected, every document is coded into a representation known as the vector space model. Here, each document in the archive (S) corresponds to a single observation or row (S_n) and each feature to an individual variable or column (S_k). Based on the earlier experiments in Schrodt et al. (2008), we use the normalized term frequencies to weight the features in each document.[11] That is, each cell ($S_{n,k}$) is given a value equal to the number of times term S_k appears in document S_n divided by the total number of terms in document S_n. With each of the $N \times K$ cells coded, the fully structured data are ready for classification.

3.2 *Classification Algorithms*

Sparse vectors and high dimensionality are the two primary attributes of text-based data worthy of primary consideration when deciding on an appropriate classification algorithm. Depending on the size of the training set and whether or not one may wish to classify all documents simultaneously, the fully structured MID4 data can have an N of more than 1.74 million and, after document frequency thresholding, a K of up to 11,672. Most news reports are five hundred words or less, meaning that *at most* such observations contain data for five hundred out of a possible 11,672 variables.

More explicitly, Fig. 3 shows the count of unique terms per document for the years 2004 and 2010. For effect, the distribution is truncated at two hundred but there exists a considerably long right-tail. The number of observed features hovers between 25 and 100 for most documents and extends beyond 200 for comparatively few, creating an enormous level of sparsity.[12] Consequently, methods for automated document classification have developed alongside the various applications, with several particular algorithms becoming relatively common as research has progressed.

The classification method used here is SVMs. Other popular methods suitable for classifying text include *k*-nearest neighbors (kNN), random forests, and naive Bayes (for further discussion, see

[11] See Salton and Buckley (1988) for a summary of traditional methods for weighting terms. Lowe (2008) and Monroe et al. (2008) discuss more current weighting methods.

[12] Specifically, from the 2004 data the truncation drops 7594 documents; from the 2010 data the truncation drops 8150.

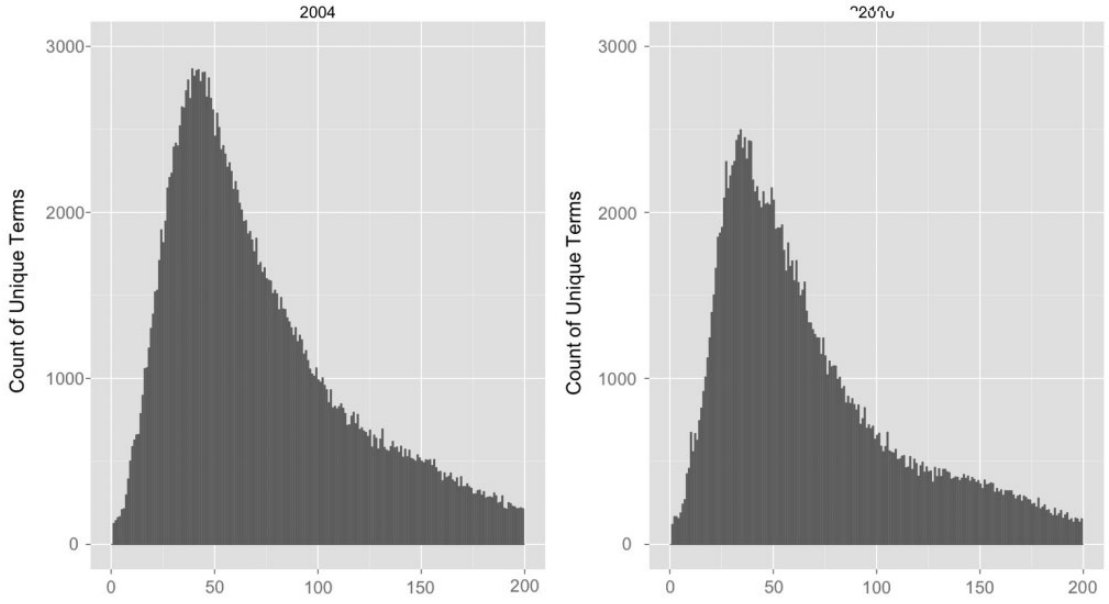

Fig. 3 Unique terms per document for 2004 and 2010.

Aggarwal and Zhai 2012). Since the purpose of this article is to present the general methodology of automated document classification, accompanied with a set of guidelines derived from the MID4 data collection effort, we provide only a brief discussion of some of these more popular methods, describing why they are less appropriate for MII classification than SVMs.

When selecting a classification method, researchers should be mindful of the specific attributes of their data problems. One primary concern when using these advanced methods of classification for MID4 data collection is that the document set has already been classified several times. Thus, it is a nonrandom sample where documents share commonalities stemming from a common source selection, keyword search, and RIPPER techniques. The goal is to utilize more complex classifiers at this stage for the purposes of recognizing patterns that are beyond the human ability to recognize.

Another concern is that the many ways in which MIIs are described is unknown; thus, one cannot confidently describe the structure of the latent classes. For example, depending on the source, the geographic location of the incident, the moment in time that the incident occurred, and the author of the original news report, two incidents that both qualify as a "border violation" may be described using considerably different terms.[13] Furthermore, a border violation is drastically different from a verbal threat by a foreign minister, which may also be described in any number of illustrative ways. These concerns guide the following discussion.

Clustering is a way to group objects, in this case text documents, based on a definition of similarity that is applied to it (Hastie et al. 2009, 502). The Rocchio algorithm and kNN are two examples of clustering methods commonly used for classifying text because of their ability to cluster quantitative data based on easily identifiable classes (Rocchio 1971; Joachims 1996). The goal of the Rocchio algorithm is to define a prototype for each class and to classify documents based on their distance to the prototype. Learning vector quantization is a comparable clustering method where the prototype is defined algorithmically (Kohonen 2001). kNN is similar, with the primary difference being that kNN classifies based on the class of labeled documents closest in proximity to the unlabeled document.

For MII classification purposes, the primary issue with clustering algorithms is that latent class attributes are not well defined and known in advance. That is to say, the basis by which one assigns

[13]A border violation is one type of MII.

similarity between two or more observations is determined by a set of related attributes; however, in the case of MIIs, these attributes are usually disparate words and phrases. Hence, a relevant document can involve relations that may otherwise seem unrelated to clusters of other labeled documents. The implication of this ambiguity is that in order to use clustering effectively one must know (or accurately anticipate) how many latent clusters exist (in the supervised setting) or adequately distinguish relevant clusters from irrelevant ones (in the unsupervised setting).

Another attractive option for text classification is naive Bayes, or in the case of most text classification multinomial naive Bayes (MNB), which is a probabilistic classification method where the probability of a document belonging to a latent class is related to the presence of individual words that are independent of one another. According to Hastie, Tibsharini, and Friedman (2009, 210), naive Bayes is an appropriate classifier when the dimensionality of the feature space is high. Given that MIIs have high dimensionality, classifying via naive Bayes may seem attractive because documents can be predicted to belong in one class or another based on the presence of individual words and not combinations of individual words. However, when classifying MIIs there is a tendency for various MID-related words to appear with regular frequency in news stories. As a result, classifying via the presence of individual words is inappropriate for classifying MIIs, because accurately identifying MIIs often requires evaluating both the presence of individual words *and* their combinations.

For instance, "missile," "troops," "fight," and "war" may all be present in a text document reporting a "military skirmish." Thus, the very characteristics of these news reports—the fact that these subject words have dependencies with each other and often accompany word phrases (e.g., "war games" or "missile battery")—means that MNB's inherent assumption of independence is violated, which results in biasing the weighting scheme toward preferring one class over another.[14] As Rennie et al. (2003) note, double weighting can unintentionally occur with certain word phrases, leading to disproportionate weighing selections given the term frequency in the word dependencies of written and spoken word. Thus, a story containing one instance of the word "war" will receive half the weight of a story containing one instance of the phrase "war games" (i.e., counting "war" and "games" as independent weight contributors).[15] Furthermore, in addition to violating this independence assumption, scholars have also found that MNB is problematic when class sizes are highly skewed, implying that mislabeling documents is dependent on the number of words in the document and the number of training examples in each class (Rennie et al. 2003; Frank and Bouckaert 2006). Due to the fact that news reports of certain types of MIIs are rare (e.g., declarations of war), but of otherwise equal importance to more frequent types (e.g., border violations), this is a concern because MNB may mislabel these rare events as irrelevant.

Finally, random forests represent another classification method that may have served as an attractive classifier for MIIs (Breiman 2001; Liaw and Wiener 2002). Random forests have a unique ensemble-based approach that is built from basic classification and regression trees (CART). The idea behind CART is to select a variable from the data and classify the training data with minimum error using just the selected variable. This initial classification forms the first node of the tree, and subsequent nodes are formed in similar fashion. CART models, however, are very unstable and highly dependent on the order of the selected variables (Gey and Poggi 2006). Random forests address this instability by bootstrapping the training data and growing a tree by randomly selecting small (e.g., 3) groups of variables at each node. Some large (e.g., 500+) number of trees are grown, with each tree being grown from a newly iterated bootstrap of the training data. Test data are classified with each tree and assigned the modal classification.

Due to their iterative sampling and ensemble nature, random forests have the capability to handle very high dimensional feature spaces, which is one of the primary attributes in textual data. Research has shown them to be an effective classifier in a variety of settings, including that of document classification (Koprinska et al. 2007). However, a limitation of random forests is that

[14]Taskar et al. (2001) experiment with probabilistic methods that sidestep the IID assumption in MNB, and thus present an effective alternative.

[15]Rennie et al. (2003) use the words "Boston" and "San Francisco" to illustrate the notion of double weighting that occurs when using MNB.

their iterative sampling process is inherently nonlinear. Specifically, at each node in each tree, the feature space is sampled and a rule is assigned based on the optimal split of the training data per the sampled features. For text classification, which often contains linearly related features, random forest sampling might overlook these relationships, leading to the misclassification of documents. Finally, random forests may also be problematic for a data set as large as MID4.

With respect to SVMs, there are fewer theoretical causes for concern. SVMs commonly outperform other text classifiers, including those discussed here, and have general properties that make them well suited for text classification (Dumais et al. 1998; Joachims 1998; Zhang and Oles 2001; Joachims 2002). Additionally, SVMs have a history of successfully and accurately classifying MIIs, as demonstrated by Schrodt et al. (2008). Subsequent evaluations of SVM classification for MID4 support this intuition.

3.3 *Phase I: SVM Inductive Classification*

MID4 uses Joachims's (2002) SVM^{light} software, version 6.02.[16] SVM^{light} operates on the data structures in Fig. 4, where the observations (individual documents) are the rows, the first column is the document label (0 if unclassified, +1 if relevant, −1 if irrelevant), and in each term the number to the left of the colon is the feature index—the specific word—and the number to the right of the colon is the normalized term frequency for feature S_k in document S_n. In Fig. 4, the 0 in the first columns indicates that this is a sample of unlabeled documents. Each of these documents contain the term indexed by 1, as seen to the left of the colon in the second column.[17] Moving from left to right, the columns are arranged such that the feature indices correspond to terms that appear with increasing frequency.

SVMs work for inductive classification by using a hyperplane that separates the two classes of labeled training documents with the maximum margin of separation (Vapnik 1995; Burges 1998; Vapnik 1998). In this application, the documents on one side of the hyperplane are those labeled as relevant, and on the other are the documents labeled as irrelevant. New documents are then classified according to which side of the hyperplane they are on.

SVMs have been built for two-category problems, which works well for most data collection projects since the purpose of using automated classification is to increase the precision of the document set by discarding irrelevant documents. However, situations may arise where multiple categories of documents are labeled and one needs to classify the unlabeled documents accordingly. In such cases, SVMs are usually applied through some combination of binary classifications (Duan and Keerthi 2005). Rubin et al. (2012) experiment with generative methods, which include naive Bayes and topic modeling, for multi-class document classification and find that they have a roughly comparable performance to that of discriminative methods (SVM).

```
0    1:0.0306    3:0.0044    4:0.0044    8:0.0044   14:0.0044
0    1:0.0421    3:0.0077    4:0.0077    5:0.0077    7:0.0038
0    1:0.0381    2:0.0095    3:0.0095    4:0.0095   10:0.0286
0    1:0.0395    4:0.0132    7:0.0132   10:0.0132   26:0.0132
0    1:0.0491    2:0.0035    3:0.0035    5:0.0070    8:0.0070
0    1:0.0349    3:0.0044    4:0.0087    5:0.0087    7:0.0044
0    1:0.0500    2:0.0125    3:0.0125    4:0.0250    5:0.0125
0    1:0.0194    4:0.0097    9:0.0097   10:0.0097   11:0.0097
0    1:0.0659    2:0.0220    7:0.0220    8:0.0330   22:0.0110
0    1:0.0196    3:0.0131    4:0.0261    8:0.0196    9:0.0131
```

Fig. 4 SVM^{light} input example.

[16] SVM^{light}, as well as a complete description of the required format for input, is available at http://svmlight.joachims.org/. R has packages implementing the SVM algorithm as well, notably *svmpath*, *kernlab*, *e*1071, and *klaR*. See Karatzoglou et al. (2006) for a discussion of these packages and some other SVM implementations.

[17] Term 1 is "said," which commonly appears in news reports.

For obtaining a better intuition about how inductive SVMs work and how the decision rule is learned, we simplify the problem to cases where we have training data that can easily be separated into two classes by a hyperplane without classification error. This is known as perfect separability. More specifically, we are looking for a set of hyperplanes, $h(x)$, as seen in equation (1):

$$h(x) = \text{sign}\{\beta_0 + \beta X\} = \begin{cases} +1 & \text{if } \beta_0 + \beta X > 0 \\ -1 & \text{else,} \end{cases} \quad (1)$$

which are subject to the constraints in equation (2).

$$\begin{aligned} \beta_0 + \beta X &\geq +1 \quad \text{if } y_i = +1 \\ \beta_0 + \beta X &\leq -1 \quad \text{if } y_i = -1 \end{aligned} \quad (2)$$

Of the set of hyperplanes that satisfies these constraints, $h^*(x)$ is the hyperplane with the maximum margin of separation. The margin is defined as 2δ, where δ is the perpendicular distance from $h^*(x)$ to the closest training example (Vapnik 1995, 1998). The closest training example is also known as a support vector.

An example of a model with two features and perfect separability is shown in Figs. 5 and 6. The data alone are shown in Fig. 5; dark data points correspond to observations labeled as relevant and light data points to irrelevant. The separating hyperplane is shown in Fig. 6. All observations in the test data are classified based on whether or not they lie above or below the hyperplane.

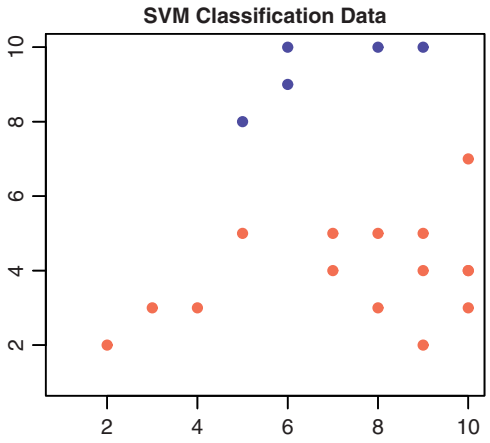

Fig. 5 Training data.

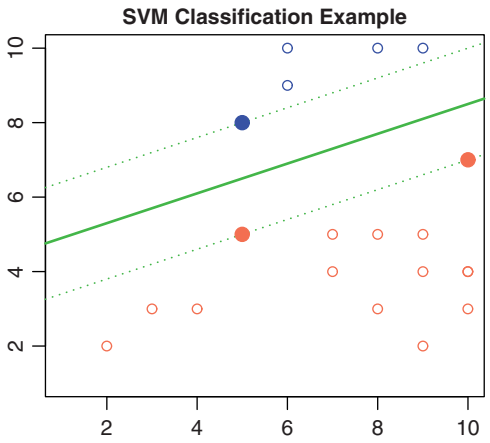

Fig. 6 Decision rule.

In application, however, the training data are not always perfectly separable. In the inductive setting, SVMs may deal with such cases in two general ways: soft-margin SVMs and nonlinear kernels.[18] For our purposes, we utilize linear soft-margin SVMs, which have been shown to outperform nonlinear kernels for many text classification problems (Manning et al. 2008, 333–34). Furthermore, as noted in Aggarwal and Zhai (2012, 199), "the general consensus has been that the linear versions of these methods work very well, and the additional complexity of non-linear classification does not tend to pay for itself, except for some special data sets. The reason for this is perhaps because text is a high dimensional domain with highly correlated features and small non-negative values on sparse features." Aggarwal and Zhai (2012, 199) also note that although various linear classifiers—for example regression and linear discriminant analysis—were developed independently, "they are surprisingly similar at a basic conceptual level . . . [and] . . . the main difference is in terms of the details of the objective function optimized, and the iterative approach determining the optimum direction of separation."

A soft-margin SVM, as opposed to a hard-margin SVM, allows some amount of training error and trades this off for model complexity. In such cases, the algorithm introduces slack variables, ξ_i, and a regularization parameter, C. A slack variable is essentially some training error for each misclassified training document. The regularization parameter trades off this error for model complexity, or the number of slack variables (Joachims 2002, 40).

For establishing this intuition, imagine a case where there exists no $h^*(x)$ that perfectly separates the data. But, we have two potential decision rules, call them $h_1(x)$ and $h_2(x)$, that do separate the data with some minimal error. Assume that all misclassified observations produce the same amount of training error. $h_1(x)$ has a larger margin but misclassifies three observations in the training data. $h_2(x)$ has a smaller margin but only misclassifies one observation. The choice of regularization parameter, C, determines which of these two decision rules becomes $h^*(x)$. If C is sufficiently large, $h_2(x)$ will be chosen because a large C means model complexity is more desirable than training error. If C is sufficiently small, $h_1(x)$ will be chosen because a small C means error is more desirable than complexity.

Since the value of the regularization parameter is subjective and may, in fact, take on many different values, some objective but appropriate criterion should be used for choosing C. Here, that criterion is the decision rule that maximizes the precision/recall break-even point (PRBEP) for S_{train}.[19] The PRBEP is the point where precision equals recall.

In order to develop the training set for the MID4 project, 24,042 news stories from 1994 to 2001 were sampled from LN downloads using the search parameters given in the supplementary materials. These documents were labeled by manually reading and classifying them as either relevant or irrelevant depending on whether document contained information about an MII. In the first phase of SVM classification, the same SVM training model is used for each year, 2002 through 2010.

In Phase I of the MID4 classification, all documents classified as irrelevant are removed from the archive. MID4 has been able to classify approximately 90% of the documents as irrelevant from Phase I alone, leaving the coders with anywhere from 13,339 (in 2008) to 26,640 (in 2006) documents.

3.4 *Phase II: SVM Transductive Classification*

Although there is success in using a common decision rule for Phase I, it is known that reports of international conflict vary in both time and space and thus may be described in different ways. The year 2001, for example, had a disproportionate number of documents dealing with the attacks of 9/11/2001—which were not an MID because the attackers were not nation-states—and the subsequent NATO coalition attack on Afghanistan (which was an MID). The year 2003 contained

[18]Nonlinear kernels map the original feature space to some new space where the separating hyperplane functions as intended. This may consist of a transformation such as including the square or cube of a certain feature. When the squared feature is included, the algorithm learns an $h^*(x)$ such that the training data are perfectly separated in this space. More on nonlinear kernels can be found in Manning et al. (2008).

[19]More precisely, SVM^{light} maximizes the arithmetic mean of precision and recall, an approximation of the PRBEP.

a very different set of reports dealing with the war in Iraq. Consequently, we have added year-specific transductive SVMs in a second phase of automated classification.

The logic of transductive classification is quite similar to that of inductive classification, except that transductive SVMs use the information in the training data *and* the test data to formalize the decision rule. It begins with a decision rule created as if one were classifying inductively, then incorporates information from the test data to try to improve the classification by iteratively weighting the rule based on out-of-sample observations that fall *within* the margin.[20]

For example, assume one has a decision rule that perfectly separates the training data. However, many test observations fall within the margin, which suggests that the rule may have a high classification error for the test set. The transductive SVM algorithm will adjust the rule, for example by altering the βs, so that fewer test observations fall within the margin. Eventually, the iterative process converges on a decision rule that is then used to classify the test data.

In developing the MID4 classifier, for each year we randomly sample approximately 250 of the positively classified documents from Phase I for the transductive training set in Phase II. Each document is manually labeled as relevant or irrelevant. As Joachims (2002, 140) notes, transductive SVMs improve performance "most substantially for small training samples and large test sets." Our training samples of approximately 250 observations, and our test sets of roughly 15,000, fit this criterion and are comparable to other experiments where TSVMs are shown to outperform inductive ones (Joachims 2002).

In the transductive setting, the ratio of positive-to-negative classified observations is an input parameter. For our purposes, we use the ratio of true:false in the transductive training set (the randomly drawn 250 stories) as the ratio of documents we want classified as true:false in the test set. Due to this feature of transductive SVMs, this labeling of the training set is very conservative. Only documents that are *clearly* false are labeled as such. As a result, many documents that are not MIIs, such as those about civil conflicts or drug trafficking conflicts, are coded as relevant in this step because stories that contain many of the same features may in other instances be an MII.

The results from the MID4 classification process are shown in Table 1. Due to the fact that the number of documents that are classified as relevant in the transduction phase is the ratio of true:false in the randomly selected 250, there is considerable variability in the percent of documents removed via transduction. In 2003, the process removed 37.25%, whereas in 2008 just 6.29% were removed. Under certain circumstances, a high level of removal variability may be a cause for concern. According to Yu, Kaufmann, and Diermeier (2008, 42), the performance of text classification methods can break down when the distribution that generates the data is not fixed. Here, we are relatively certain that this is the case. However, one *expects* this to be the case because the

Table 1 Automated classification results

Year	*Documents I*	*Inducted (%)*	*Documents II*	*Transducted (%)*	*Documents III*
2002	225,598	89.60	23,462	30.70	16,245
2003	248,010	89.65	25,658	37.25	16,098
2004	222,454	90.72	20,643	9.64	18,591
2005	183,320	92.68	13,419	20.15	10,715
2006	230,662	88.45	26,640	25.54	19,834
2007	173,865	89.68	17,936	14.43	15,373
2008	126,136	89.42	13,339	6.29	12,498
2009	161,527	91.34	13,997	27.02	10,215
2010	172,945	91.97	13,884	6.76	12,946
Total	1,744,517	90.2	168,978	21.6	132,515

[20]See Joachims (2002, 167–9) for a detailed explanation of how transductive SVMs formalize a decision rule.

data-generating process for international conflict events *is* stochastic, making the events not independently and identically distributed across spatial or temporal domains.

However, in other contexts removal variability may be undesirable. For example, if one wanted to classify US Congressional speeches in an effort to predict voting behavior, a high level of removal variability may imply a serious loss of relevant information (for further discussion, see Poole and Rosenthal 1991a, 1991b; Poole and Rosenthal 2007). The reason for this is that any corpus of data with a stable, pre-existing record with consistent temporal patterns based on a specific set of variables—that is to say one can guess with a reasonable degree of accuracy how a US Congressman will vote based on his or her past Congressional speeches—makes automated classification quite consistent and the subsequent removal variability low. Thus, in these instances, a high removal variability would be suggestive of serious flaw in the method. Unlike US Congressional voting, however, we argue that it is unrealistic to assume consistent removal variability when coding MIIs. In fact, the variability in our Phase II step is quite advantageous for our purposes.

First, it ensures that we can classify our data based on content rather than news sensationalism and geographic bias by calibrating transductive removal on a yearly basis to remove high-profile events that produce a large number of irrelevant, redundant news reports. For instance, the 2006 US troop surge in Iraq produced several thousand reports, none of which were relevant to the project because of the MID4 coding rules—Iraq was an ongoing conflict and these additional troops did not affect how this MID was coded. Because we could calibrate our Phase II removal method on a yearly basis, we were able to remove the irrelevant reports before we even began human coding. In these instances, a high degree of knowledge about certain international events can be a valuable asset, particularly given the incredible number of world events and the tendency of over-reporting. This flexibility allows for the inclusion of expert opinion into the automated classification process, which ultimately improves the algorithm's decision rule.

Second, it addresses the importance of time and temporal bias. Hopkins and King (2010, 242) recommend that "if we are studying documents over a long period of time, where the language used to characterize certain categories is likely to change, it would not be advisable to select the labled test set only from the start period." Thus, using yearly specific training models in Phase II is an appropriate and efficient complement to the rigor and uniformity imposed in the Phase I, because it allows us to address concerns of temporal bias as world events change.

By way of illustration, we alleviate any concerns that we have removed relevant information in the Phase II classification by conducting a complete post-hoc evaluation of the 2003 classification. The year 2003 was chosen because it contains the largest number of removed documents, in both raw total and percentage. We reason that if the process is successful here, it is safe to assume it will be successful elsewhere. Our post-hoc evaluation criteria is that some false negatives are acceptable, but a newly identified MID is unacceptable and represents a flaw in the classification process that would need to be addressed.

Of the 7745 documents removed from the 2003 reports by transduction, only 519 (6.7%) contain information that might have been relevant to an MII. We compared these 519 stories to the coded MIIs from 2003 and found that 58 of the 519 reports contained information on forty-seven incidents that *had not* been previously coded. Of these forty-seven new incidents, twenty-nine are between India and Pakistan, six are between India and Bangladesh, and six are between Azerbaijan and Armenia. These forty-one incidents all take place in the context of highly militarized and intense situations that had already been coded as MIDs involving many other comparable incidents. Similarly, each of the remaining six incidents is confined to a previously coded MID consisting of several other incidents.

Thus, despite the loss of a few MIIs in our Phase II classification, none of the reports would lead to a significant change in the coding of any ongoing MID. Furthermore, none of the reports contain information about an incident or set of incidents that would constitute a new, previously uncoded MID. Therefore, the results from the post-hoc evaluation of the 2003 documents demonstrate that no vital information has been lost through the implementation of this method.

The improvements in efficiency by conducting a second SVM classification using year-specific training models are obvious through our analysis. Over the entirety of the project, our coders manually classified 36,463 fewer documents because of the Phase II classification. Put another

way, if each year on average contains about 12,000 documents, then the Phase II classification saves us three years' worth of news stories.

We recommend that data collectors adopt similar subsetting techniques when the structure of the document varies based on some field of meta-data. Here, the argument is that reporting on MIIs changes over time, and thus necessitates yearly training models. One can imagine a comparable task where training by other fields of meta-data, such as the source, is more appropriate.

4 Conclusion

With the growth of easily accessible information over the past decade and a half, data collectors in political science often find themselves in a situation of having too much information at their disposal. For harnessing this information into structured data for quantitative analysis, methods of information retrieval and document classification should be more widely applied. This article encourages this through its review of the data collection process using a successful document classification method as it applied to the MID4 database update.

By representing the data collection process as an information retrieval system, a new perspective is provided for addressing this common task in quantitative research. Data collection is time-consuming and costly, but often the process is made even more inefficient by the choice of classification method. It is likely that there are many instances of information and efficiency loss in the plethora of political science data collection efforts based on observational data. In attempting to make the resulting data as unbiased as possible, we have moved toward a system that retrieves information on nearly all *reported* MIIs. By querying LN using a set of global search parameters and fifteen news sources for the years 2002–2010, we collected a set of over 1.74 million documents. These documents were filtered, formatted, and represented as data so that they may be classified using two phases of SVM automated document classification. Using inductive SVMs, Phase I of the automated classification process removed 90.2% of documents from this set. Using transductive SVMs, Phase II reduced this number by an additional 21.6%. The resulting classifications contain anywhere from 10,215 to 19,834 stories per year.

The results from the MID4 information retrieval system and its method of automated document classification have been quite successful and should serve as an example for comparable collection efforts. Although MID4's incident coding is ongoing, preliminary results suggest that we have captured up to 40% more incidents per year than MID3, which is all the more remarkable given that the 2002–2010 period coded in MID4 coincided with a decline in international conflict worldwide (Human Security Report Project 2012). Furthermore, the MID4 retrieval method has reduced the level of required resources to the point where further updates can be accomplished by a single institution, which also makes this method quite desirable as it relates to management costs and project overhead.

This retrieval system is reproducible and open-source, which enables maximum transparency and allows others to adapt this software for their purposes. As data collection projects continue to move toward increased automated classification for making their efforts more efficient and accurate, the efficiency gains should lead to more frequent updates. Although real-time updating may not be desirable for all data collection efforts, annual updating (at the very least) is not only desirable, but becomes tractable when automated methods are successfully applied to optimally address researchers' needs.

Funding

National Science Foundation (SES-0719634, SES-0924240).

References

Adcock, R., and D. Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review* 95(3):529–46.

Aggarwal, C. C., and C. X. Zhai. 2012. A survey of text classification algorithms. In *Mining text data*, eds. C. C. Aggarwal and C. X. Zhai, 77–129. New York: Springer.

Barid, V. A. 2004. The effect of politically salient decisions on the U.S. Supreme Court's agenda. *Journal of Politics* 3(66):755–72.

Basu, C., H. Hirsh, and W. Cohen. 1998. Recommendation as classification: Using social and content-based information in recommendation. *AAAI/IAAI* 714–20.

Bikel, D. M., and V. Castelli. 2008. Event matching using the transitive closure of dependency relations. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, HLT-Short '08, Stroudsburg, PA, USA, 145–48. Association for Computational Linguistics.

Blair, D. C. 1992. Information retrieval and the philosophy of language. *Computer Journal* 35(3):200–207.

———. 2003. Information retrieval and the philosophy of language. *Annual Review of Information Science and Technology* 37(1):3–50.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1):5–32.

Britt, B. L., M. W. Berry, M. Browne, M. A. Merrell, and J. Kolpack. 2008. Document classification techniques for automated technology readiness level analysis. *Journal of the American Society for Information Science and Technology* 59(4):675–80.

Burges, C. J. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2(2):121–67.

Cardie, C., and J. Wilkerson. 2008. Text annotation for political science. *Journal of Information Technology and Politics* 5(1):1–6.

Cohen, W. W. 1996. Learning rules that classify e-mail. In *AAAI Spring Symposium on Machine Learning in Information Access*, (18): 25. California.

Cohen, W. W., and Y. Singer. 1999. Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems (TOIS)* 17(2):141–73.

Dasgupta, A., P. Drineas, B. Harb, V. Josifovski, and M. W. Mahoney. 2007. Feature selection methods for text classification. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

D'Orazio, V., S. T. Landis, G. Palmer, and P. Schrodt. 2013. *Replication data for: Separating the wheat from the chaff: Applications of automated document classification using support vector machines*. IQSS Dataverse Network. V1.

Duan, K.-B., and S. S. Keerthi. 2005. Which is the best multiclass SVM method? An empirical study. In *Multiple classifier systems*, eds. N. C. Oza, R. Polikar, J. Kittler, and F. Roli, Volume 3541 of *Lecture Notes in Computer Science*, 278–85. Berlin, Heidelberg: Springer.

Dumais, S., J. Platt, D. Heckerman, and M. Sahami. 1998. Inductive learning algorithms and representations for text categorization. In Proceedings of the Seventh International Conference on Information and Knowledge Management.

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research* 3:1289–305.

Frank, E., and R. R. Bouckaert. 2006. Naive bayes for text classification with unbalanced classes. In *Knowledge Discovery in Databases: PKDD 2006*, 503–10. Springer.

Gey, S., and J.-M. Poggi. 2006. Boosting and instability for regression trees. *Computational Statistics and Data Analysis* 50(2):533–50.

Ghosn, F., G. Palmer, and S. A. Bremer. 2004. The mid3 data set, 1993–2001: Procedures, coding rules, and description. *Conflict Management and Peace Science* 21(2):133–54.

Gochman, C. S., and Z. Maoz. 1984. Militarized interstate disputes, 1816–1976: Procedures, patterns, and insights. *Journal of Conflict Resolution* 28(4):585–616.

Goertz, G. 2006. *Social science concepts: A user's guide*. Princeton, NJ: Princeton University Press.

Guyon, I., and A. Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3:1157–82.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. New York: Springer.

Hoffart, J., M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. 2011. Robust disambiguation of named entities in text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, Stroudsburg, PA, 782–92. Association for Computational Linguistics.

Hopkins, D., and G. King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54(1):229–47.

Human Security Report Project. 2012. *Human security report: Sexual violence, education, and war: Beyond mainstream narrative*. Vancouver: Human Security Press.

Joachims, T. 1996. A probabilistic analysis of the Rocchio algorithm with tfidf for text categorization. Technical report, DTIC Document.

———. 1998. Text categorization with support vector machines: Learning with many relevant features. In Tenth European Conference on Machine Learning.

———. 2002. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Norwell, MA: Kluwer Academic Publishers.

Jones, D. M., S. A. Bremer, and J. D. Singer. 1996. Militarized interstate disputes, 1816–1992: Rationale, coding rules, and empirical patterns. *Conflict Management and Peace Science* 15(2):163–213.

Karatzoglou, A., D. Meyer, and K. Hornik. 2006. Support vector machines in R. *Journal of Statistical Software* 15(9):1–28.

Kohonen, T. 2001. Learning vector quantization. In *Self-organizing maps,* Volume 30 of *Springer Series in Information Sciences*, 245–61. Berlin, Heidelberg: Springer.

Kolari, P., T. Finin, and A. Joshi. 2006. SVMs for the blogosphere: Blog identification and splog detection. In American Association for Artificial Intelligence Spring Symposium on Computational Approaches to Analyzing Weblogs.
Koprinska, I., J. Poon, J. Clark, and J. Chan. 2007. Learning to classify e-mail. *Information Sciences* 177(10):2167–87.
Lewis, D. D. 1992. *Representation and learning in information retrieval*. PhD thesis, University of Massachusetts.
Liaw, A., and M. Wiener. 2002. Classification and regression by random forest. *R News* 2(3):18–22.
Lowe, W. 2008. Understanding wordscores. *Political Analysis* 16(4):356–71.
Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2:159–65.
Manning, C. D., P. Raghavan, and H. Schutze. 2008. *Introduction to information retrieval*. Cambridge, MA: Cambridge University Press.
Mayfield, J., and P. McNamee. 2003. Single *n*-gram stemming. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2003, New York, 415–16. ACM.
Mohammad, S., and G. Hirst. 2006. Distributional measures of concept-distance: A task-oriented evaluation. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP 2006, Stroudsburg, PA, 35–43. Association for Computational Linguistics.
Monroe, B. L., M. P. Colaresi, and K. M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16(4):372–403.
Nadeau, D., and S. Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30(1):3–26.
Paice, C. D. 1994. An evaluation method for stemming algorithms. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1994, New York, 42–50. New York: Springer.
Papka, R., and J. Allan. 1998. Document classification using multiword features. In Proceedings of the Seventh International Conference on Information and Knowledge Management, CIKM 1998, New York, 1–8. ACM.
Poole, K. T., and H. Rosenthal. 1991a. On dimensionalizing roll call votes in the U.S. Congress. *American Political Science Review* 85(3):955–76.
———. 1991b. Patterns of Congressional voting. *American Journal of Political Science* 35(1):228–78.
———. 2007. *Ideology and Congress*. New Brunswick, NJ: Transaction Publishers.
Porter, M. F. 1980. An algorithm for suffix stripping. *Program* 14(3):130–37.
Rennie, J. D., L. Shih, J. Teevan, and D. R. Karger. 2003. Tackling the poor assumptions of naive bayes text classifiers. Proceedings of the Twentieth International Conference on Machine Learning, Washington, DC.
Rijsbergen, C. v. 1979. *Information retrieval*. London: Butterworth-Heinemann Press.
Rocchio, J. J. 1971. Relevance feedback in information retrieval. In *The SMART retrieval system: Experiments in automatic document processing*, ed. G. Salton. Englewood Cliffs, NJ: Prentice-Hall.
Rubin, T. N., C. America, P. Smyth, and M. Steyvers. 2012. Statistical topic models for multi-label document classification. *Machine Learning* 88(1–2):157–208.
Salton, G., and C. Buckley. 1988. Term weighting approaches in automatic text retrieval. *Information Processing and Management* 24(5):513–23.
Sartori, G. 1984. *Social science concepts: A systematic analysis*. Beverly Hills, CA: Sage.
Schrodt, P. A., G. Palmer, and M. E. Haptipoglu. 2008. Automated detection of reports of militarized interstate disputes: The SVM document classification algorithm. Presented at the Annual Meeting of the American Political Science Association, Toronto, Canada.
Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34(1):1–47.
Shulman, S. W. 2005. E-rulemaking: Issues in current research and practice. *International Journal of Public Administration* 28(7–8):621–41.
Sindhwani, V., and S. S. Keerthi. 2006. Large scale semi-supervised linear SVMs. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, 477–484.
Spirling, A. 2012. U.S. treaty making with American Indians: Institutional change and relative power, 1784–1911. *American Journal of Political Science* 56(1):84–97.
Taskar, B., E. Segal, and D. Koller. 2001. Probabilistic classification and clustering in relational data. In International Joint Conference on Artificial Intelligence, Vol. 17, 870–78. Lawrence Erlbaum Associates LTD.
Vapnik, V. N. 1995. *The nature of statistical learning theory*. New York: Springer.
———. 1998. *Statistical learning theory*. New York: John Wiley and Sons.
Wang, T., and G. Hirst. 2012. Exploring patterns in dictionary definitions for synonym extraction. *Natural Language Engineering* 18:313–42.
Witsenburg, K. M., and W. R. Adano. 2009. Of rain and raids: Violent livestock raiding in northern Kenya. *Civil Wars* 4(11):514–38.
Yang, Y., and J. O. Pedersen. 1997. A comparative study on feature selection in text categorization. In Proceedings of the Fourteenth International Conference on Machine Learning, ICML 1997, San Francisco, CA, 412–20. Morgan Kaufmann Publishers.
Yu, B., S. Kaufmann, and D. Diermeier. 2008. Classifying party affiliation from political speech. *Journal of Information Technology and Politics* 5(1):33–48.
Zhang, T., and F. J. Oles. 2001. Text categorization based on regularized linear classification methods. *Information Retrieval* 4(1):5–31.