Abstract

As the availability of digitized text proliferates, so does our ability to collect detailed information on events of interest. Sorting through the volumes of available information, however, can be a time-consuming task. This paper presents the automated document classification system used by the Militarized Interstate Dispute (MID) 4.0 research project for updating the database from 2002-2010. By using global search parameters and fifteen international news sources, we collected a set of over 1.74 million documents from LexisNexis. Care was taken to create an all-inclusive set of search parameters as well as a sufficient and unbiased list of news sources. We then represent these documents as data and classify them into relevant and irrelevant classes using two types of support vector machines (SVMs). Using inductive SVMs and a single training set, we remove 90.2% of documents from our initial set. Then, using year-specific training sets and transductive SVMs, we further reduce the number of human-coded stories by an additional 21.6%. The resulting classifications contain 10,215 to 19,834 documents per year.

1 Introduction

The importance of information retrieval and document classification is a routinely overlooked and under-valued aspect of social scientific inquiry. Take, for example, the perspective of Hopkins and King: "The manager of a congressional office would find useful an automated method of sorting individual constituent letters by policy area... In contrast, political scientists would be interested primarily in tracking the proportion of mail in each policy area" (Hopkins and King 2010, p. 230). This statement seems to suggest political scientists would not care to be able to classify documents based on their content: We beg to differ.

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 we face these questions each and every time we begin any data collection project, we often have no road map to guide us in our efforts. In this article we offer some insight on this important but under-discussed step of scientific inquiry.

This paper provides an in-depth look at the information retrieval system used in the Militarized Interstate Dispute 4 (MID4) data collection project. We begin by creating a set of just over 1.74 million documents that we believe might contain information on reported militarized interstate incidents (MIIs) for the years 2002 through 2010. This archive is created by downloading documents from LexisNexis (LN) using an inclusive search string and a selection of news sources that we show to be sufficient for retrieving

all relevant information on reported MIIs. We then implement two phases of automated text classification using support vector machines (SVMs) for classifying the documents as relevant or irrelevant. Our results yield a yearly average of 14,724 documents - a reduction of more than 90% in the number of documents that must be evaluated by human coders - which are then human-evaluated and coded according to the MID coding rules.

Schrodt, Palmer and Haptipoglu (2008) demonstrated that such an information retrieval system was able to at least equal the quality of data produced by the MID3 project, which did not use automated document classification. Our results exceed these initial experiments. By automating the process of deciding which documents to read, we use just a fraction of the resources used by MID3 and preliminary results suggest that we have coded no fewer than and up to 40% more incidents per year than MID3. Furthermore, the process is entirely reproducible, allows for transparent data collection, and ensures the integrity of the data, all elements of data collection that provide confidence in the validity of our data.

As applications of automated text classification increase, we expect most data collection projects to move toward some version of this automated process. Doing so will not only improve the quality and quantity of available data, it moves the replicability of the scientific undertaking back to the origin of the data, not just the origin of the analysis as is commonly done with replication materials today. By ensuring the integrity of our data, our confidence in our analyses improves. To facilitate the application of this information retrieval system to other data collection projects, all software used is completely open-source.[1]

[1]The processing software and replication instructions are available at `http://eventdata.psu.edu/`. 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 we used.

2 Data Collection in International Conflict Research

Since its founding as a field, international relations has been centrally concerned with questions of war and peace. The Correlates of War Project has been the most important and the only sustained effort to collect systematic data on wars and a set of independent variables (such as capability, alliance formation, diplomatic importance) going back to 1816. In the last twenty years, these data have been supplemented by the collection of data on militarized interstate disputes (MIDs) from 1816-2001.

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, Bremer and Singer 1996, 169). Each dispute is comprised of a set of events – militarized interstate incidents (MIIs) – which further permit researchers to study the specific actions of conflict, their geographic location, and the frequency of violence, among other factors.

The standard method for data collection in IR involves a team of research assistants (RAs) reading through and coding some archive of documents. The method for creating this archive limits which documents are coded and which are not, heavily influencing the quality of the data. Prior to coding, the principal investigator's (PI) primary responsibility is to create an archive that simultaneously contains sufficient relevant information to code the topic of interest *and* is small enough for the project to be finished in a timely manner. Depending on the scale of the project, such an archive can be something as limited as the CIA World Factbook or something as expansive as Google News.

The perfect archive is one where all relevant information is included, every document contains unique and trustworthy information, there is no redundancy, and there are no irrelevant documents. Since we cannot create a perfect archive, we instead seek to identify a set of documents with both high recall and high precision. Recall refers to the ratio of relevant information in the archive to the total amount of relevant information available.

In other words, perfect recall means that all relevant information on the topic of interest is included in the archive. Precision refers to the percent of relevant information in the archive to the archive's total amount of information. In other words, perfect precision means that every document in the archive contains relevant information.

The problem of balancing precision and recall has become increasingly important with the expansion of the sources available for coding. Prior to the introduction of the World Wide Web, international relations coding projects typically focused on a small number of sources of information, typically "archival" sources such as *Facts on File* or a small number of newspapers. This limited the number of stories that needed to be considered, and frequently had already gone through a process of editorial filtering. In contrast, contemporary news sources, whether a proprietary database such as LN or general news aggregators such as Google News and European Media Monitor, track thousands of news sources. This potentially provides a nearly-comprehensive source of information, but the quantity of stories means that it is no longer possible to rely solely on human judgment to determine what is important.

As data collectors, we like to *think* our archive has perfect recall. Increasing the number of documents in the set should give us better recall, but generally at the expense of precision - there will be more irrelevant documents for our RAs to read. This tradeoff is always present when collecting data, and it is problematic for traditional approaches as well as for machine-assisted approaches: when the archive is too large, the project never gets finished, but when the archive is too small, we will have missed information of interest. Fortunately, with advances in automated document classification techniques, we can address the central problem of extracting relevant documents from a corpus that has become too large to read.

2.1 The MID Document Set

The MID4 pipeline begins with the definition of a MID and ends with data on all instances where a MID is observed from 2002 through 2010. While these aspects of the MID projects have been nearly identical during different phases of this 30-year project, how we get from the definition of a MID to data on MIDs has changed.

MID1 was released for widespread use in the mid-80s and contained data for disputes from 1816-1976 (Gochman and Maoz 1984). The collection of the first updated and extended version of the data set began a few years after that, and was released as MID2 in the mid 1990s (Jones, Bremer and Singer 1996); that version of the data brought the years covered up through 1992. Both MID1 and MID2 utilized *Keesing's World Events* supplemented with *New York Times* articles as the primary sources of information.

MID3 employed a new method of informational retrieval using the LN Universe for its data collection (Ghosn, Palmer and Bremer 2004). News databases such as LN make possible a keyword search on an extremely large number text documents from a variety of news sources, making available volumes of information not previously available to MID researchers. Although MID3 took advantage of this source of information, the complexity of MIDs in terms of the variety of events which constitute an MII render a keyword search inefficient. Very specific keyword searches miss events of interest, and generalized keyword searches retrieve too much irrelevant information: the recall/precision problem we discussed earlier.

MID3 lacked a systematic and efficient approach for dealing with this tradeoff. The news sources used in the information retrieval were unsystematic and comprised a variety of general, world, and newswire services. Researchers would query regional or country-specific search terms in a scattered fashion – sometimes including cities, sometimes geographic features, sometimes specific people or actors. In the simplest terms, the project had extremely poor precision, inconsistent recall, and a very high degree of manual doc-

ument classification.

In moving forward with the MID project, we recognized two main sources of inefficiency in MID3. The first is in the randomness of the database query, which can be improved 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; should any types of MIIs be missed, they will be missed systematically and thus we will have an easier time recognizing and correcting the oversight.

The second and more serious contributor to inefficiency is the manual document classification. Because MID was originally designed for human coding and requires a number of assessments about the dispute as a whole, rather than focusing solely on single interactions (the approach of event data, which is now routinely coded using automated methods), the elimination of human coding for this type of data is neither possible nor desirable. However, manual document classification takes time. Improving the efficiency of the MID project necessarily involves reducing the number of irrelevant documents read by RAs.

To address these concerns, we increased the size of the archive to whatever size was deemed necessary to be confident that all relevant information was included. MID4 accomplishes this by keyword searching over a number of sources which we have identified to be sufficient with a set of global, all-inclusive search parameters.

2.2 MID4: Search Parameters and Source Identification

Unlike MID3, which relied on a dyadic/region specific identification of MID-relevant news stories, MID4 uses a consistent and uniform approach for establishing the initial set of documents. We use the approach discussed in Schrodt, Palmer and Haptipoglu (2008) for constructing global search parameters, which allows researchers to search across multiple reporting news agencies, providing global coverage for each day of the year. To do so, we

compiled the following MID3 related search words in a comprehensive, MID-specific, LN boolean search string:

> (air base OR air strike OR airbase OR aircraft OR airstrike OR alert OR antiaircraft OR armed OR armo! OR arms OR army OR artillery OR attack OR batteries OR battery OR battle OR battleship OR block! OR bomb OR border OR buildup OR carrier OR casualties OR casualty OR cease OR ceasefire OR cease-fire OR clash! OR combat OR conflict OR crisis OR cruiser OR damage OR declare war OR defence OR defense OR defensive measures OR defian! OR deploy! OR destroy OR detained OR dispatch! OR display of force OR dispute! OR embargo OR erupt! OR fight! OR fire OR fired OR forc! OR fortification OR hit OR hostile OR incursion! OR infantry OR interstate OR invasion OR jet OR kill! OR launch! OR liberate OR line of control OR maneuver OR milit! OR missile! OR mobiliz! OR mortar OR naval OR nuclear OR occup! OR offensive OR operation OR patrol! OR peace declaration OR pullback OR radar OR raid! OR recon! OR reinforcement OR reprisal OR retali! OR rocket OR security OR seiz! OR shell! OR shoot OR shot down OR show of force OR shrapnel OR skirmish OR soldier! OR squadron OR stronghold OR subside! OR target OR tension! OR territ! OR threat! OR trade fire OR troop OR truce OR ultimatum OR USS OR vessel OR violat! OR violence OR vows to OR war OR warn! OR warplane OR warship OR weapon! OR weapons OR withdraw!)

To improve the initial precision from LN, every query in this part of the data collection process also includes a set of "AND NOT" exclusion parameters. Their purpose is to improve the "true" to "false" ratio of returns by systematically removing stories with the following words contained in the headline:

> AND NOT (sports OR business OR lifestyle OR tax cuts OR entertainment OR Wall Street OR budget OR baseball OR food OR weather OR health OR natural disasters)

Though these parameters are intended to be all-inclusive, source selection remains an important concern. Including too many sources can lead to redundancy and loss of precision. However, too few sources can lead to loss of recall due to the sources' scope of coverage and various types of reporting biases. Our goal is to use the fewest number of sources capable of collecting all MID-relevant information.

We begin our source selection process with all news sources that are found to contain MID-relevant information in MID3. Sources were initially kept or rejected based on

Table 1: MID4 News Sources

Associated Press	Deutsche Presse Agentur	London Times
United Press International	Japan Economic Newswire	New York Times
Agence France Presse	ITAR-TASS News Agency	Interfax
British Broadcasting Corporation	Montreal Gazette	AFX News
Xinhua General News Service	Jerusalem Post	CNN
(All sources are English versions)		

temporal coverage. Because the LN database contains subscriptions to news agencies on a year-to-year basis, some sources in the initial list were removed because they did not have the required nine year coverage. This validation process results in a pared down list of thirty different candidate sources with coverage that matched the temporal domain of the project.

We then group sources based on geographic coverage, here referred to as Lists A, B, and C. List A contains the sources with the most international coverage, comprising of agencies across multiple continents. Examples include the Associated Press and British Broadcasting Corporation; the complete list of sources is found in Table 1. List B is more restrictive, containing agencies confined to a continent or hemisphere. Examples include the *Boston Globe* and *Toronto Star*. List C is the most specific, having agencies with coverage constrained to a particular region of a continent or area of the world. Examples include *Straits Times* (Singapore) and *New Straits Times* (Malaysia).

These lists were then evaluated to assess two distinct concerns. The first is the cost, in terms of unnecessary additional stories, of using all thirty potential sources, rather than only the global sources on List A. To assess this, we queried the LN database using our search string on fifteen randomly generated dates between the years 2003-2004 to examine the number of returns A, B, and C yield. The averaged results of these tests are displayed below:

- List A: Daily average of 2,365
- List B: Daily average of 266

- List C: Daily average of 320

Lists B and C would add an additional 550-600 stories per day, which we then evaluated to determine whether these contained new reports relevant to the coding. Ten MIIs from MID3 were selected at random for the year 2001. Using our global search parameters and querying on the day *after* each MII began, we assessed each list on its ability to catch these incidents. The List A candidates caught six of ten, B one of ten, and C zero. Next, we looked at dates before, on, *and* after the start of the ten randomly chosen MIIs. This subsequent test improved A's performance to ten of ten, with B and C catching no additional MIIs. Based on these results, we concluded that List A is sufficient.

This result runs counter to the intuition of many researchers that local sources contain more detail on local events. That may well have been the case historically, prior to the advent of low-cost electronic news sources, but in the contemporary environment, there are two reasons that the international sources are likely to be more comprehensive. First, the major international sources now have arrangements with the local papers – or individual "stringers" reading these to pick up any information likely to be of interest (and MIIs would almost always be in this category). Second, the Web-based version of local papers is often the printed version of that paper, which is subject to space limitations which are not found in the all-electronic wire services. The absence of significant value-added that we found in our experiments was consistent with the findings of the Integrated Conflict Early Warning System project (O'Brien 2010).

As may be expected with such a general search string and a large number of sources, the LN queries return massive numbers of documents. The average number of stories per year is about 200,000, for a total number of over 1.74 million from 2002 through 2010 – far too many for manual classification. We therefore proceed by automating the classification process.

3 Automated Document Classification

Automated document classification has become a permanent fixture of machine learning research (Britt et al. 2008; Kolari, Finin and Joshi 2006; Sebastiani 2002; Aggarwal and Zhai 2012). In political science the method has been applied, for example, to classify party affiliation and political opinion (Shulman 2005; Yu, Kaufmann and Diermeier 2008). For the purposes of this project, we use classification to eliminate as many irrelevant documents as possible while retaining all of those which might contain information relevant to our topic of interest.

In our 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 the pre-processing and feature selection process. Second, we use the training model developed in Schrodt, Palmer and Haptipoglu (2008) and classify using inductive support vector machines (SVMs). Third, we develop year-specific training models and classify using transductive SVMs.

3.1 Pre-Processing and Feature Selection

The first task in automated classification is to represent the document in a standardized form, then split the data into a training set and a test set. To do so, we need to pass the documents through pre-processing steps so that they are formatted identically. This is done using pre-processing or filtering software which is customized for the various formats found in the LN downloads; we used, with occasional modifications, filters originally developed and made available by the preliminary MID4 project (Schrodt, Palmer and Haptipoglu 2008). The filters identify meta-data such as the headline, the news source, and the date, as well as distinguish the beginning and end of a document and the location of the textual report within the document.

The next step involves selecting and representing text as features, and coding each

document for those features. There are many ways to go about selecting features for text (Monroe, Colaresi and Quinn 2008; Lowe 2008; Forman 2003; Guyon and Elisseeff 2003). We use unigrams (single words), remove all common stopwords, utilize document frequency thresholding and remove all proper nouns. The resulting dataset is an N by K structured dataset where N is the number of documents and K is the number of features.

Stopword removal consists of removing the most common words in the archive and document frequency thresholding consists of removing the least commonly seen words across documents. For example, the words "and," "the," and "that" are so common that they will be found in relatively high frequency in every document and so will not help to distinguish any particular document. Conversely, rare words, such as "non-peripheral" and "lackluster" only show up in a few of the documents and again contribute little information important for our classification task. Stopword removal and document frequency thresholding have been standard approaches in natural language processing literature for some time (Dasgupta et al. 2007; Luhn 1958; Rijsbergen 1979) and are found in political science applications (Hopkins and King 2010; Spirling 2012), although some have made arguments against these practices (Monroe, Colaresi and Quinn 2008).

The remaining two components of our feature selection, the removal of many proper nouns and the use of single words – unigrams – is less standard. We remove proper nouns associated with nation-states as a way of increasing the possibility that stories will be classified based on their content and not simply their geographic location. For example, we do not want to give extra weight to stories about states such as Israel, Syria and Lebanon and less weight to stories about states such as Brazil and Chile simply because the former are more likely to experience an MII than the latter.

To remove proper nouns, we begin with a database of country names and match n-grams in the text to the names in the database.[2] This information is also used to

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

remove any report that does not mention at least two nation-states – by definition, an MII cannot be simply an internal dispute. The two most frequently mentioned states are saved in the data and serve as a "best guess" for the dyad, a step which makes the human coding considerably more efficient since the coders are looking at a chronology of related stories. Less common proper nouns indicating geographical location, such as "Bekaa" or "Spratly" generally do not make it past the document frequency thresholding cutoff and do not appear to have significantly affected the classification.

Unigrams drop some information 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 (Hopkins and King 2010; Lewis 1992; Yang and Pedersen 1997) as well as in a MID-specific setting (Schrodt, Palmer and Haptipoglu 2008).

At this point, every document can be coded for the remaining features, a representation known as a vector space model. From the perspective of our structured dataset (S), each document in the archive is a single observation (S_n) and each unique term in the archive that has not been removed is a variable (W_k). Based on the earlier experiments in Schrodt, Palmer and Haptipoglu (2008), we use the normalized term frequencies to weight the features in each document.[3] This is simply the number of times term W_k appears in document S_n divided by the total number of terms in document S_n, giving a value in the interval [0,1] for every cell in our N by K dataset.

Our feature selection process produces an enormous but very sparse dataset. To put it in perspective, N is just over 1.74 million and K is 11,672. Sparse vectors and high dimensionality, in addition to the fact that our training sample (S_{train}) contains roughly 24,024 examples and the test set (S_{test}) contains the remaining 1.72 million observations, are the primary concerns when selecting a method for classifying text. While it is technically possible to use classical statistical methods in this environment,

[3]See Salton and Buckley (1988) for various methods for weighting terms.

Figure 1: SVM^{light} Input Example

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

inversion of extremely large sparse matrices is potentially problematic, and consequently we turned instead to machine learning methods, notably support vector machines, which have been developed and have been shown to work for precisely this sort of task (Dumais et al. 1998; Joachims 1998, 2002).

3.2 Phase I: SVM Inductive Classification

While there are many methods for classifying text, we have opted to use support vector machines (SVM), a relatively common approach to document classification in machine learning (Dumais et al. 1998; Joachims 1998, 2002). Our decision to use SVM stems from the findings in Joachims (1998, 2002), from the success of the method for correctly classifying MIIs in Schrodt, Palmer and Haptipoglu (2008), and from subsequent evaluations as we progressed through the years coded.

We use Joachims (2002) SVM^{light} software, version 6.02.[4] SVM^{light} operates on the data structures in Figure 1, where the observations 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.

The general intuition behind SVMs for inductive classification is to find a hyperplane

[4] 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, Meyer and Hornik (2006) for a discussion of these packages and some other SVM implementations.

that separates the two classes of labeled training documents with the maximum margin of separation (Burges 1998; Vapnik 1995, 1998). In our applications, the document on one side of the hyperplane are those labeled as relevant; on the other are the documents labeled as irrelevant. New documents are then classified according to which side of the hyperplane they are on.

To gain a better intuition about how inductive SVMs work and how the decision rule is learned, let us simplify the problem to the case 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),

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

that are subject to the constraints in Equation (2).

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

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.

Graphically, an example of a model with two features and perfect separability is shown in Figures 2 and 3. The data alone are shown in Figure 2; blue dots correspond to observations labeled as relevant and red dots to irrelevant. The separating hyperplane is shown in Figure 3. All observations in the test data would be classified based on whether or not they lie above or below the hyperplane.

Figure 2: Training Data

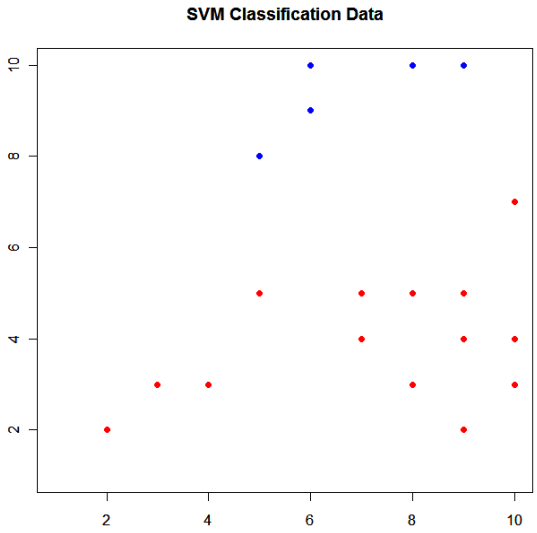
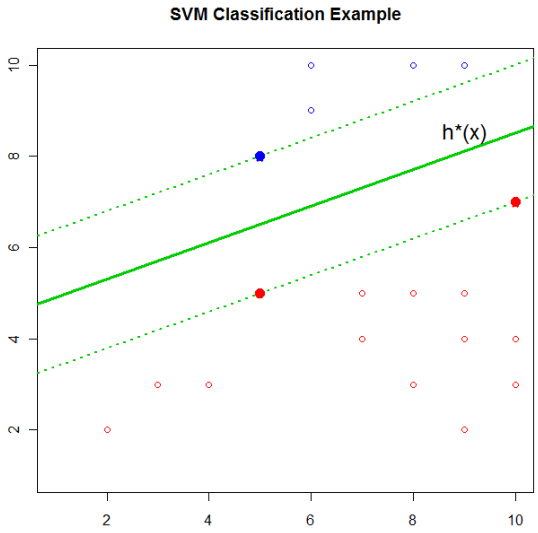


Figure 3: Decision Rule



In real applications, 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 non-linear kernels.[5] For our purposes, we utilize linear soft-margin SVMs, which have been shown to outperform non-linear kernels for many text classification problems (Manning, Raghavan and Schutze 2008, p. 333-334). Furthermore, as noted in Aggarwal and Zhai (2012), "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" (pg. 199). Aggarwal and Zhai (2012) also note that while various linear classifiers—for example regression and linear discriminant analysis—were developed independently, "they are surprisingly similar at a basic concep-

[5]Non-linear 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 non-linear kernels can be found in Manning, Raghavan and Schutze (2008).

tual 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, p. 40).

For 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 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 criteria should be used for choosing C. Here, that criteria is the decision rule that maximizes the precision/recall break-even point (PRBEP) for S_{train}.[6] Specifically, the precision is equal to the proportion of true positives to all positively classified documents. Recall is equal to the proportion of true positives to all true documents. The PRBEP is the point where precision equals recall.

To develop the training set for the MID4 project, 24,042 news stories from 1994-2001 were sampled from LN downloads using the aforementioned search parameters. These

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

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. Typically, we have been able to classifying about 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.

3.3 Phase II: SVM Transductive Classification

While we have had success in using a common decision rule for Phase I, we know that reports of 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 a MID because the attackers were not states—and the subsequent NATO coalition attack on Afghanistan (which was a MID). The year 2003 contained 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 we 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.[7]

For example, assume we have a decision rule that perfectly separates the training data. However, many test observations fall within the margin, which suggests that rule

[7]See (Joachims 2002, 167-169) for a detailed explanation of how transductive SVMs formalize a decision 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 about 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) points out, transductive SVMs improve performance "most substantially for small training samples and large test sets" (140). Our training samples of about 250 observations and our test sets of roughly 15,000 fit this criteria and are comparable to other experiments where TSVMs are shown to out-perform 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, our 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 2. 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, we have considerable variability in the percent of documents removed via transduction. For example, in 2003 we removed 37.25% while in 2008 we removed just 6.29%. Under certain circumstances, a high level of removal variability may be a cause for concern. According to Yu, Kaufmann and Diermeier (2008), the performance of text classification methods can break down when

the distribution that generates the data is not fixed (42). Here, we are relatively certain that this is the case. However, we *expect* this to be the case because the 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 (see Poole and Rosenthal (1991*b*,*a*, 2007) for further discussion). The reason for this is that any corpus of data with a stable, preexisting 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 their 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 was relevant to our project because of our coding rules; Iraq was an on-going conflict and the 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 enormity of world events and the tendency of overreporting. This flexibility allows

Table 2: 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

us to include the role of expert opinion into our automated classification process, which ultimately improves our 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 compliment 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. We chose 2003 because it contains the largest number of removed documents, both in 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 aceptable, but a newly identified MID is unacceptable and represents a flaw in the classification process that would need to be addressed.

Of the 7,745 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 47 incidents that *had not* been previously coded. Of these 47 new incidents, 29 are between India and Pakistan, six involve India and Bangladesh, and six are between Azerbaijan and Armenia. These 41 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 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 demonstrates that no vital information has been lost through the implementation of our method.

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.

4 Conclusion

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 have been filtered, formatted, and represented as data so that they may be classified using two phases of 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. While MID4 has not finished coding all the incidents for each year, preliminary results suggest that we have captured up to 40% more incidents per year than MID3. Furthermore, we have reduced the level of effort to the point where it can be done at a single institution, which reduces the management costs and increases the level of consistency across region. Additionally, our system is reproducible and open-source, enabling transparency in our work and allowing others to adapt our software for their purposes, as well as providing an easily verifiable record of why any given story was or was not chosen for coding.

By representing this data collection project as an information retrieval system, we have taken a new look at a common task in quantitative IR research. Data collection is time-consuming and costly, but often the process is made even more inefficient by the choice of method. As happened in MID3 (and, we suspect, in other projects) there is often an efficiency loss at the point where PIs make the initial decisions about the archive to be coded. We have demonstrated how MID4 has implemented a method of automated document classification at this point in the project, leading to vast improvements in both efficiency and accuracy.

In attempting to make our source list and the resulting data as unbiased as possible, we have moved towards a system that retrieves information on nearly all *reported* MIIs. However, there is a systemic bias in reporting that cannot be undone if strictly published news sources are used. For example, regardless of what sources are chosen, Africa will still be systematically under-reported and Europe systematically over-reported. Furthermore, some MIIs will just never be reported in published sources. While there are major issues with using non-published sources such as social media outlets, these sources may contain information on MIIs that we could never capture otherwise. Moving forward, we hope researchers continue developing technology to harness this potential information so that

data collection projects can readily tap into these non-published sources.

Our hope is that data collection projects, both in conflict studies and elsewhere, will move towards methods of automated classification to make their efforts more efficient and accurate. The ideal of near-real-time updates can only be realized if automated methods are successfully applied, and we have taken a step in this direction. We recommend that when researchers set out to collect data, they take advantage of this and other technology available to improve the quality, scope, and timeliness of their efforts.

References

Aggarwal, Charu C. and ChengXiang Zhai. 2012. A Survey of Text Classification Algorithms. In *Mining Text Data*, ed. Charu C. Aggarwal and ChengXiang Zhai. New York: Springer chapter 6, pp. 77–129.

Britt, Barry L., Michael W. Berry, Murray Browne, Mary Ann Merrell and James Kolpack. 2008. "Document Classification techniques for Automated Technology Readiness Level Analysis." *Journal of the American Society for Information Science and Technology* 59(4):675–680.
URL: *http://dx.doi.org/10.1002/asi.20770*

Burges, Christopher J.C. 1998. "A Tutorial on Support Vector Machines for Pattern Recognition." *Data Mining and Knowledge Discovery* 2(2):121–167.

Dasgupta, Anirban, Petros Drineas, Boulos Harb, Vanja Josifovski and Michael W. Mahoney. 2007. Feature Selection Methods for Text Classification. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Dumais, Susan, John Platt, David Heckerman and Mehran Sahami. 1998. Inductive Learning Algorithms and Representations for Text Categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*.

Forman, George. 2003. "An Extensive Empirical Study of Feature Selection Metrics for Text Classification." *Journal of Machine Learning Research* 3:1289–1305.

Ghosn, Faten, Glenn Palmer and Stuart A. Bremer. 2004. "The MID3 Data Set, 1993-2001: Procedures, Coding Rules, and Description." *Conflict Management and Peace Science* 21(2):133–154.

Gochman, Charles S. and Zeev Maoz. 1984. "Militarized Interstate Disputes, 1816-1976: Procedures, Patterns, and Insights." *Journal of Conflict Resolution* 28(4):585–616.

Guyon, Isabelle and Andre Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research* 3:1157–1182.

Hopkins, Daniel and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54(1):229–247.

Joachims, Thorsten. 1998. Text Categorization with Support Vector Machines: Learning with many Relevant Features. In *Tenth European Conference on Machine Learning*.

Joachims, Thorsten. 2002. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Norwell, MA: Kluwer Academic Publishers.

Jones, Daniel M., Stuart A. Bremer and J. David Singer. 1996. "Militarized Interstate Disputes, 1816-1992: Rationale, Coding Rules, and Empirical Patters." *Conflict Management and Peace Science* 15(2):163–213.

Karatzoglou, Alexandros, David Meyer and Kurt Hornik. 2006. "Support Vector Machines in R." *Journal of Statistical Software* 15(9):1–28.

Kolari, Pranam, Tim Finin and Anupam 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*.

Lewis, David D. 1992. Representation and Learning in Information Retrieval PhD thesis University of Massachusetts.

Lowe, Will. 2008. "Understanding Wordscores." *Political Analysis* 16(4):356–371.

Luhn, Hans Peter. 1958. "The Automatic Creation of Literature Abstracts." *IBM Journal of Research and Development* 2:159–165.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schutze. 2008. *Introduction to Information Retrieval*. Cambridge, MA: Cambridge University Press.

Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16(4):372–403.

O'Brien, Sean. 2010. "Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research." *International Studies Review* 12(1):87–104.

Papka, Ron and James Allan. 1998. Document Classification Using Multiword Features. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*.

Poole, Keith T. and Howard Rosenthal. 1991*a*. "On Dimensionalizing Roll Call Votes in the U.S. Congress." *American Political Science Review* 85(3):955–976.

Poole, Keith T. and Howard Rosenthal. 1991*b*. "Patterns of Congressional Voting." *American Journal of Political Science* 35(1):228–278.

Poole, Keith T. and Howard Rosenthal. 2007. *Ideology and Congress*. New Brunswick, NJ: Transaction Publishers.

Rijsbergen, C.J. van. 1979. *Information Retrieval*. London: Butterworth-Heinemann Press.

Salton, Gerard and Christopher Buckley. 1988. "Term Weighting Approaches in Automatic Text Retrieval." *Information Processing and Management* 24(5):513–523.

Schrodt, Philip A., Glenn Palmer and Mehmet Emre 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, Fabrizio. 2002. "Machine Learning in Automated Text Categorization." *ACM Computing Surveys* 34(1):1–47.

Shulman, Stuart W. 2005. "E-Rulemaking: Issues in Current Research and Practice." *International Journal of Public Administration* 28(7-8):621–641.

Spirling, Arthur. 2012. "U.S. Treaty Making with American Indians: Institutional Change and Relative Power, 1784–1911." *American Journal of Political Science* 56(1):84–97. **URL:** *http://dx.doi.org/10.1111/j.1540-5907.2011.00558.x*

Vapnik, Vladimir N. 1995. *The Nature of Statistical Learning Theory.* New York, NY: Springer-Verlag.

Vapnik, Vladimir N. 1998. *Statistical Learning Theory.* New York, NY: John Wiley and Sons.

Yang, Yiming and Jan O. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. In *Machine Learning: Proceedings of the Fourteenth International Conference.* pp. 412–420.

Yu, Bei, Stefan Kaufmann and Daniel Diermeier. 2008. "Classifying Party Affiliation from Political Speech." *Journal of Information Technology and Politics* 5(1):33–48.