Automatic Document Categorization for Highly Nuanced Topics in Massive-Scale Document Collections: The SPEED BIN Program

March 2011

Cline Center for Democracy
University of Illinois at Urbana-Champaign
Automatic Document Categorization for Highly Nuanced Topics in Massive-Scale Document Collections: The SPEED BIN Program

ABSTRACT

This whitepaper offers a brief introduction to the BIN system of the Social, Political and Economic Event Database (SPEED) project. BIN provides automatic document categorization of highly nuanced topics across massive-scale document archives. The BIN system allows a group of trained human editors to present the computer with a relatively small collection of hand-categorized documents representing a given topic. It uses the semantic characteristics of these documents to develop a statistical model that is capable of identifying other documents on that same topic from the Cline Center global news archive, which contains tens of millions of news reports. Tests have shown that BIN has a false negative (incorrectly discarded relevant documents) rate of 1-4%. This paper outlines the basic premise and motivation behind BIN, its development, and its application to the SPEED project.

CONTRIBUTOR

Kalev Leetaru
# Table of Contents

The SPEED Project ....................................................................................................................... 1  
Automated Searching of a Large Text Collection ................................................................. 2  
Representing a Document ........................................................................................................ 3  
Modeling .................................................................................................................................. 4  
Multi-layered Learning Models ............................................................................................... 7  
Review ..................................................................................................................................... 8
Automatic Document Categorization for Highly Nuanced Topics in Massive-Scale Document Collections: The SPEED BIN Program

The SPEED Project

The Social, Political and Economic Event Database (SPEED) project is a technology-intensive effort to compile a global event database covering all countries in the world from 1946-present using news reports. Using a combination of sophisticated software tools and specialized protocols (Societal Stability, Property Rights, Supremacy of Law, Integrity of Electoral Processes, etc.), a team of trained human coders examines each news report and codifies its information into a series of quantitative database entries. A Nigerian Tribune newspaper article about a labor protest is transformed from a block of text to a structured event entry, capturing scores of derived variables about the event, such as its location, date, group affiliations of actors, etc. Such structured representations can be explored through a wide range of statistical techniques and integrated with economic, legal, and other social science datasets to offer insights into key behavioral patterns and relationships that are valid across countries and over time.

To capture the world’s news across six decades, the SPEED project relies on several key historical news archives. The New York Times provides rich summary coverage of world events over the entire post World War II era, while the CIA Foreign Broadcast Information Service (FBIS) and BBC Summary of World Broadcasts (SWB) are open source intelligence services that provide translations of local print and broadcast news coverage from throughout the world. These historical archives are supplemented with a selected stable of web-based news outlets to capture the contemporary era.

Aggregating even a sample of the global news content from more than 60 years yields an archive of massive size and scope: more than 87 million documents as of 2011. The New York Times alone contributes over 5.9 million articles, while the Summary of World Broadcasts contains nearly 3.9 million. The vast majority of this content, however, has little bearing to the SPEED project’s core areas of interest. Events such as riots, demonstrations, strikes, assassinations, coups, and power transfers, which are central to the Societal Stability Protocol (SSP), are far less common in the average day’s news than routine criminal arrests or the previous evening’s baseball scores. Of the 5.9 million New York Times articles, just 700,000 (12%) are relevant to the SSP, while only 2 million (51%) of Summary of World Broadcast articles are.

Thus, the first step in the process used by SPEED to extract information from these news archives is to “filter” the individual reports and discard those that are not relevant to a particular protocol. The most simplistic approach to this screening task is to use a team of humans to skim each article. Yet, with just the New York Times and Summary of World Broadcasts alone, this would require that more than 7.1 million irrelevant articles be read and discarded. A relatively simple article might take up to a minute for a trained human to read and screen, while a longer

---

and more complex article can take several minutes. Thus, it would take more than a thousand people working non-stop over a year to filter just this small portion of SPEED’s global news archive. Complicating matters, even highly-trained coders have internal biases that lead to inconsistencies in their categorizations; moreover, they can become fatigued over time. Thus, a fully-automated filtering mechanism is needed to cope with the tremendous volume of content that SPEED must process.

**Automated Searching of a Large Text Collection**

Most users are familiar with the basic keyword query and the use of “Boolean” logic to construct complex queries such as “(riot OR protest OR march) AND (death OR attack OR beating OR police OR army)”. At first glance, it would appear that some of SPEED’s more basic event categories, such as riots, could be identified through keyword searches. However, even a simple search for “volcano” in the New York Times yields a listing in which more than half the returned documents are not accounts of actual volcanic eruptions. A literary flourish comparing a politician’s press conference to a “volcanic eruption” or a book on how a new volcano might erupt in a major US city are all included in such results. Therefore, even for simplistic terms, a keyword query will result in a significant density of irrelevant material. More importantly, many of SPEED’s event types do not easily lend themselves to a keyword representation. For example, what string of keywords could accurately capture “human rights violations”? One attempt to create a list of keywords for just the SSP resulted in a list of nearly a half million terms and could not achieve better than 30% accuracy for many categories.

Obviously a more complex approach is needed, one that can create a statistical model of just what a “human rights abuse” news report looks like. Fortunately, there is a field of research in computer science known as Automatic Text Categorization (ATC) or “document classification” that focuses on how to build statistical representations of particular topics for the purpose of conducting highly nuanced searches across large document archives. ATC systems allow a user to compile a small collection of documents from which it builds a statistical model of what language use within that collection looks like compared with other documents. Once built, this model can be applied to a very large collection of documents to identify “similar” documents efficiently and accurately.

An ATC process begins by having a team of trained human editors compile two collections of documents: a list of “inclusive” documents discussing the topic of interest and a list of “exclusive” documents that do not. Both lists should contain a mix of “clearly” inclusive or exclusive documents and borderline examples that help the computer learn the “fringes” of the topic. For example, an ATC model designed to identify assassinations would include articles about “character assassinations” in its fringe exclusive category to help the machine understand that: (1) the term “assassinations” is used in this context, and (2) such uses of assassination should not be included in the same category as assassinations of political leaders. This is the most important stage of the process, as the quality of the examples given to the ATC modeling

---

Representing a Document

Once the pool of training data has been compiled, an ATC learning algorithm and appropriate tuning parameters must be selected. There are numerous learning algorithms that can be used here, ranging from Naïve Bayesian to Support Vector Machines, Decision Trees, and Neural Networks. Underlying each learning algorithm is a common technique used to represent blocks of text that are processed by the algorithm. This technique involves compiling a list of all unique words used across the entire corpus of documents. The number of times each word appears in relevant and irrelevant documents is then computed, along with the word’s total frequency in the overall corpus. The “weight” or “importance” of each word as a distinguishing characteristic is then calculated through what’s known as a Term Frequency Inverse Document Frequency (TFIDF) formula.

The basic premise of the TFIDF weighting formula is that a word that appears very frequently in the relevant category is likely a strong predictor, unless it also appears with great frequency in the irrelevant category. The word “the” may appear in 99% of relevant documents, but it also likely appears in 99% of the irrelevant documents. Thus, it will have less distinguishing capacity than a word that appears in 10% of relevant documents, but in only 2% of irrelevant documents. There are a large number of adjustments possible with the TFIDF weighting schema, such as normalization and probabilistic frequency adjustments. In the case of the SPEED project, after considerable testing, it was found that word frequency alone, without adjustment by document frequency, yielded the highest accuracy on SSP topics.

The term-weighting stage also allows for “stopwords” (frequently occurring English words like “the” or “and”) to be removed. We found that removing these stopwords had a small negative impact on accuracy. In addition, certain classes of words, such as verbs in English, can be conjugated and thus appear with multiple spellings. A “stemming” process can be used to reduce all words to their root form. Part of speech tagging can also be used to annotate each word in the text with its part of speech to use only noun phrases as input for categorization. In the case of SPEED, all of these modifications proved to have either no impact or a slightly negative impact on accuracy.

3 Because of the importance of generating high quality training data a great deal of effort has been devoted to this task within the SPEED project. For example, in the case of the Societal Stability Protocol, which has an event ontology of nearly 100 categories pertaining to various manifestations of civil unrest, an initial pool of over 30,000 documents was categorized into “relevant” and “irrelevant” categories.
Some representation models allow for the ordering of words to be considered, whereas the traditional approach is known as a “bag of words” and simply represents each document as a frequency list of the words it contains, with no information on their ordering within the text. We adopted this approach SPEED because it offered slight accuracy improvements when coping with the heavy typographical error found in digitized historical material.

This stage of the ATC pipeline generates a large table with an entry for every distinct word used across the entire corpus of documents, along with frequency counts of their occurrences in both the relevant and irrelevant categories. In a large set of training documents this list could easily contain hundreds of thousands of words. Thus, a “feature selection” process is run that selects the top 25% of the words with the greatest term-weights and discards the rest. Each document is then converted from a block of freeform text to a word probability table that can be evaluated by the recognition model.

**Modeling**

Once the word probability table has been computed, the modeling phase of the ATC process begins. At this stage a learning algorithm must be chosen to build a model of how much each word in the table contributes to the document classification process. Numerous modeling algorithms exist, such as Naïve Bayesian, Support Vector Machines, Decision Trees, and Neural Networks. There is currently no clear consensus in the literature as to which algorithm works the best as their respective accuracies are highly dependent on the nuances of the underlying data. Nevertheless, Bayesian approaches have been found to enjoy high accuracy across a wide range of genres.

The Bayesian algorithm used in document classification is known as a “naïve” implementation in that it assumes complete independence of all word probabilities across the corpus. A human editor would recognize that the joint occurrence of both “riot” and “killed” in a document increases the probability that the document is about a violent protest. Unfortunately, the use of joint probabilities vastly increases the computational needs of an automated classifier; thus a Naïve Bayesian classifier only makes use of the independent probabilities of each word.

---


After extensive testing, the Naïve Bayesian classification algorithm was found to outperform both Support Vector Machines and Decision Trees with respect to SPEED’s Societal Stability Protocol by about 30%. Moreover, it was nearly three times faster than the SVM implementation and only slightly slower than the Decision Tree implementation. There are a large number of parameters that can be adjusted to fine-tune the Bayesian learning algorithm. Two were found to have a significant impact on accuracy with respect to the SSP: the size of the relevant and irrelevant pools of examples. Rather than give the system all 30,000 categorized articles as input, adjusting the relative sizes of the relevant and irrelevant input queues allowed the model’s accuracy to be carefully controlled along the dimensions of recall (the percentage of relevant articles the machine correctly identifies) and precision (of all the documents identified by the system as relevant, the percentage that are actually relevant). The combined collection of relevant and irrelevant documents comprises the learning algorithm’s sole knowledge about what belongs in the target category. Thus, adjusting the ratio of these two collections allows the biasing of the model towards greater accuracy in recognizing relevant articles versus greater accuracy in identifying irrelevant articles.

The massive size of SPEED’s article store is such that articles discarded as irrelevant by the system could not be manually reviewed. Thus, the classification system’s false negative rate (the percentage of relevant articles incorrectly discarded as irrelevant) has to be as low as possible – even if this means having a false positive rate (irrelevant articles incorrectly identified as relevant) that is higher than desired. In the SPEED project all articles identified as relevant are manually reviewed by a human, so false positives impose less of a cost in terms of accuracy. A high false positive rate does, however, impose costs in terms of efficiency since they must be discarded manually.

To enhance the accuracy of the classification process with respect to SSP-related documents, a third category of human training data was added after the first round of testing, known as “Relevant to Irrelevant” documents. This category is composed of documents that the learning algorithm had initially categorized as relevant, but which subsequent human review determined were irrelevant. This pool of examples added discriminatory power because it helped in identifying irrelevant fringe cases that contained much of the language of a relevant article.

For each of the three categories one quarter of the documents was set aside as an “evaluation pool;” the remaining documents were queued for training the model. Having a separate set of pre-annotated evaluation documents provided for automated testing of alternative models. That is, the system was able to evaluate the accuracy of large numbers of models rapidly by testing them on the evaluation pool and comparing their results against the human-provided annotations. This capacity is important because ATC models tend to have very non-linear response curves: a small adjustment of one input, such as the number of relevant documents, can yield a substantial change in the accuracy of the model. This nonlinearity makes it more difficult to find the ideal model configuration, since accuracy is not improved through simple addition of documents to the input queues. Thus, all possible permutations of the number of documents from each of the three categories had to be tested to identify the ideal configuration.

To iterate across the entire range of possible configurations, where a “configuration” is a particular combination of the number of relevant, irrelevant, and fringe documents, a parameter
sweep was constructed. This parameter sweep consisted of three nested loops, with the outermost loop iterating over the number of relevant documents, the second loop over the irrelevant ones, and the innermost over the fringe documents. Step sizes of 100 were used for all three loops, which were found to yield good representation of the execution space, while yielding a reasonable number of total configurations to be tested. For each configuration, a model was built with that number of relevant, irrelevant, and fringe documents, and tested using the evaluation pool documents. Its true positive, false positive, true negative, and false negative rates were recorded for each configuration in a logfile. While computationally expensive, this process allowed all possible permutations of the input ratios to be evaluated to find the model that maximized the percent of relevant documents identified by the system, while also maximizing its ability to discard irrelevant documents.

Figure 1 illustrates the final accuracy curve for this parameter sweep, sorted in descending order of accuracy at recognizing relevant documents (true positive rate). The X axis represents each of 6,000 of the configurations tested, while the Y axis represents the accuracy of each configuration with respect to the percent of relevant documents correctly identified as relevant and the percent of irrelevant documents correctly discarded. Figure 1 clearly indicates the diffuse behavior of the model’s accuracy at detecting irrelevant documents. An overall inverse relationship between relevant and irrelevant detection accuracy is clear: a model that is better at identifying relevant documents will also yield a higher percentage of irrelevant documents. The configuration found to yield the highest accuracy was a ratio of 0.55 irrelevant documents and 0.5 fringe documents for every relevant article, using 12% of the total 30,000 training documents selected at random. This yielded 99% identification of relevant documents and 65% identification of irrelevant documents on both the training and evaluation document pools.
Multi-layered Learning Models

Every statistical model has a fixed “learning capacity” in terms of the variance it can accurately induce from the underlying data. Most applications of automatic document categorization target relatively simplistic topics with high discriminatory power from other topics (such as “sports” versus “political” news) and are able to develop a single learning model capable of encapsulating the topical space. However, the complexity of the various SPEED event ontologies, particularly the one embedded in the SSP, are well beyond the capacity of a single model.

We therefore make use of a two-layer learning model, with the procedure described above used as a “prefiltering” pass to discard the majority of “easy” irrelevant documents. With respect to the SSP event ontology, it correctly kept nearly 99% of relevant documents, but discarded only 65% of the irrelevant documents. Documents classified as relevant by the initial model are passed through a second model that acts as a more fine-grained filter. This model was trained by taking one thousand documents classified as relevant by the first filter and having them manually reviewed by trained human editors to create a second set of relevant and irrelevant training data. A second parameter sweep process was used to train this model using the 1000 new training documents. Here, the ideal ratio was found to be 0.62 irrelevant documents for each relevant article, using a pool of just 650 documents. This yielded 99% accuracy at identifying relevant documents, and 99% accuracy at identifying irrelevant documents on the testing data. Since the majority of the obviously irrelevant documents are
discarded by the first model, the second model is able to focus its learning capacity on identifying the nuances of more complex cases.

Thus, the final system workflow is as seen in Figure 2, where all documents are processed by the first ATC model, which acts as a coarse filter, keeping 99% of relevant documents, but discarding only 65% of irrelevant documents. Documents deemed relevant by the first model are passed to the second model, which provides more refined filtering of false positives and yields 99% accuracy on irrelevant documents and 99% accuracy at identifying irrelevant documents. The final documents that have been identified by both filters as relevant are output by the system into a document store that is accessible by human coders for the purpose of information extraction that is structured by a topic-specific protocol.

**Figure 2 - System Diagram**

![System Diagram]

**Review**

To review the final output of the system for the SSP, the complete population of six million New York Times news articles, 1946-2005, were run through the full categorization pipeline. Then 1000 relevant and 1000 irrelevant articles were selected from the output at random and given to trained human editors to review. Less than 1% of articles in the irrelevant category were identified by a human editor as being relevant (incorrectly discarded by the machine). The editors knew that they were reviewing documents that had been categorized by the computer as irrelevant so, in this respect it was not a true blind test. But this notwithstanding, the low false negative rate, in conjunction with the fact that the false positives were truly marginal events, demonstrates the system’s capacity to not discard relevant events.

Since ATC systems are built around word frequency distributions, they tend to become hypersensitive to the specific language use of their training source: a symptom known as “overlearning” their input data. They perform very well on subsequent material from the same source, but do poorly on documents from other sources that use a different writing style. For example, a model trained on New York Times content would be expected to perform less well on Summary of World Broadcasts content, which is sourced from outlets across the world and translated from local languages into English. The system was run on all three million Summary of World Broadcasts articles and, as before, trained human editors were asked to review the
output. This time, only the false negative rate was examined, since that was of greatest interest to the SPEED project. Despite running on an entirely novel corpus drawn from translated content from across the world, the system’s false negative rate was less than 4% on a collection of 1,000 randomly-selected articles discarded by the system as irrelevant, with most of the misses being borderline cases.

In practice, the system’s false positive rate can be significantly higher than that suggested by its training data and initial editorial checks. The largest source of this error is with the complexity of the SPEED project’s coding rules. The ATC system operates only on the text of each document; it does not have access to higher-level semantic operators in the text, such as dates, locations, or actor names. For example, within SPEED’s SSP, an event with a location that is not at the country, region, or city level, is discarded, as are events with date information that is not resolvable to a particular timeframe. Thus, an article on mass state-sponsored killings “in Africa over the last 100 years” would be identified by the ATC system as relevant to the SSP event ontology, but would be discarded due to imprecise location and date information. Another key source of error is that the SPEED project has a large list of secondary rules and heuristics that are used to determine whether a given event should be coded. For example, a “politically motivated attack” is a codeable event within the SSP, but only if it does not occur during a recognized period of interstate war. Thus, of two articles identified by the ATC system as containing descriptions of attacks, one might ultimately be deemed “irrelevant” because it occurred during a recognized war. Events occurring outside of the SPEED project’s 1946-present analysis period are also not codeable, so an article about an attack in 1936 would ultimately be discarded. Finally, the SSP has very precise definitions of what kinds of attacks fall under its taxonomy. Large-scale criminal activity, such as mass killings, assassinations, and bombings due purely to criminal drug activity, are not codeable events, yet they contain the same types of activities that are found in codeable events.

The current incarnation of the BIN system is a highly effective fully automated filtering system, capable of rapidly processing tens of millions of documents and discarding irrelevant documents with high accuracy, vastly reducing the load required of human coders. It was specifically trained to minimize the false negative at the expense of having a slightly higher false positive rate. It has a much larger false positive rate in production (65%) but this is largely due to a lack of access to the secondary heuristics and rule sets that introduce a great deal of nuance to the relevance determination process. Work is ongoing to address these issues. Yet, despite its false positive rate, the BIN system is able to achieve very high accuracy at discarding irrelevant information rapidly and at scale, offering a powerful filtering mechanism for large content analysis projects.