Feature Selection

Vito D'Orazio
PhD Candidate
Department of Political Science
Pennsylvania State University
vjd125@psu.edu [*]

June 6, 2013

While it is commonly recognized that the availability of digitized text represents an enormous opportunity for political scientists to access and structure new data, relatively little research on text analysis has been tailored to working with the types of text commonly seen in Political Science. This is especially true in the area of automated document classification, which is used primarily as a method for improving the efficiency of data collection projects and where the peculiarities of Political Science documents may warrant different methods than those found to be successful in the benchmark sets (e.g. Reuters-21578 and RCV1). In this study three feature selection methods specifically tailored for improving the efficiency of document classification for Political Science research are compared to a standardized approach. For the purposes of comparison, all other aspects of classification are kept constant across evaluations. An analysis of the results yields two interesting findings. First, substantial improvements in precision can be had by *not* removing named entities and minor improvements in precision can be had by classifying sources separately. Second, feature selection appears to be compatible with ensemble learning, as the highest precision is found in the intersections of the four classified sets.

[*]

Introduction

One of the fundamental problems with collecting data from digitized sources is the tremendous number of irrelevant documents. These documents eat away at resources as principal investigators pay research assistants to manually sort through them, often to perform a simple binary classification: relevant or irrelevant.

In the past fifteen years, the automated classification of documents has developed into a mainstream branch of research in Information Sciences and has seen numerous applications in other fields, including Political Science (Sebastiani, 2002; Yin et al., 2010; D'Orazio et al., 2012). The binary classification of documents, the most simple form of classification, can yield large dividends to researchers seeking a method for improving that simple relevancy classification. Although automated classification has been approached using a variety of methodologies, it is often the case that researchers developing these methosds are not experimenting with the same types of document sets that we use when gathering data in Political Science.

Observational data in Political Science often relates some actor, such as states, politicians, or some political institution to some type of action, such as a conflict mediation, campaign statements, or a domestic crackdown on protets. Our search for information about these concepts incorporates different types of context-specific information, such as actor names and a variety of sources, and the information is often searched for by filtering through news reports. Ultimately, what is desired is to a perform content analysis on a small but sufficient set of documents relevant to our concept of interest (Krippendorff, 2013).

The following experiments compare and contrast three methods of feature selection to a baseline method that has its roots in the text analysis literature. Feature selection is the process by which one decides which aspects of the text documents to include as features (or variables) in the data. Each method is potentially one that can improve the efficiency of any data collection effort where automated document classification is used by making more appropriate feature selection decisions.

The first method of feature selection that is explored is whether or not named entities should be included in the classification. D'Orazio et al. (2012) strips named entities from the text, but the trade-off associated with doing so has not been empirically tested. Should it be the case that named entity inclusion does not alter or actually increases classification accuracy, incorporating named entities into the potential feature space would be desirable.

Second, our document sets are typically comprised of multiple news or other sources. In these experiments, the concept of interest is joint military exerices (JMEs), an international phenomenon and so four global newsources are used: The Associated Press, Agence France Presse, Xinhua General News Agency, and Interfax News Agency. Each source has different norms about structuring their news and how they go about reporting certain stories, and so classifying these sources separately may improve the system's performance.

Finally, I examine a feature selection method where the features are only selected from the relevant documents in the training set. Researchers commonly know how to find relevant information for their topic of interest, and many seek such information on a regular basis. However, most will not know how to find a representative sample of irrelevant information. Removing the necessity of labeling irrelevant documents, then, may lead to improved efficiency.

These three methods are compared to a baseline method comprised of unigrams, document frequency thresholding, stopword removal, named entity removal, and stemming. Terms are weighted using the term-frequency inverse document-frequency method, and the classification algorithm used throughout is inductive support vector machines.

Results from the experiments yield two interesting findings. First, substantial improvements in precision can be had by *not* removing named entities while minor improvements result from classifying sources separately. Extracting features solely from relevant documents results in a considerably worse classification than the baseline. Second, and potentially important for future research, feature selection appears to be compatible with ensemble learning, as the highest precision is found in the intersections of the four classified sets.

Practical Concerns for Data Collection

An ideal information retrieval system for collecting data is one with *sufficient recall.* Recall is a term that refers to the ratio of relevant returns to the universe of relevant documents. Data collectors do not need *every* relevant document; they only need enough to conduct a valid and reliable content analysis (Krippendorff, 2013). Sufficient recall refers to the number of documents necessary for such an analysis.

This characterization means that when it comes to data collection, the document classification problem is different than it is in other fields of research. With search engines, for example, when users query a search engine they are essentially asking Google, or Yahoo!, or Bing, to provide them with a ranking of web pages (i.e. documents) based on the likelihood that the pages contain relevant information. Users want their information need met quickly and they do not necessarily care about the total number of documents retrieved; users only care about the relevancy of the first page of hits.

Thus, effective search engine algorithms are predominantly based on high precision at the expense of recall. Precision is a term that refers to the ratio of relevant returns to the total number of returns. Search engines want their top hits to contain information relevant to the user and do not necessarily care if they return every page on the Web that contains relevant information. When constructing an information retrieval system for collecting data, precision is thought of differently.

With respect to data collection, precision is important because poor precision will quickly consume an investigator's time and resources. Moving forward as a discipline, the demand for better precision will increase because of two reasons. First, digitized text is proliferating, making available more relevant information on increasingly specific concepts while also increasing the amount of irrelevant information available. Sorting through the information require precise document classification. Second, funding agencies appear increasingly less-likely to fund data collection efforts, especially those which are updating already existing datasets. In this respect, increases in precision are desirable, as they will increase the project's efficiency because researchers can essentially do more with less.

Feature Selection: Common Procedures and Difficulties

Feature selection algorithms can have a substantial impact on the performance of a document classification system (Dasgupta et al., 2007). Generally speaking, feature selection is the decision of which aspects of the text to include as features (or variables) in the classification. It is a primary component of any text analysis in general and of document classification systems in particular.

There are many ways to go about selecting which features of the text to represent as data (Monroe, Colaresi and Quinn, 2008; Lowe, 2008; Forman, 2003; Guyon and Elisseeff, 2003), most of which begin with the bag-of-words model, or unigrams. Unigrams are the individual words that make up the text. The implementation of a unigram approach is computationally simple and is based on splitting a string by its whitespace characters. Using this appraoch, each individual word has the potential to be meaningful while phrases are split into separate features.

Other feature selection approaches begin with Ngrams, which inherently have the potential to hold more semantic content than unigrams, as a functioning Ngram algorithm will extract meaningful phrases as single feature rather than a set of size N of individual features (Papka and Allan, 1998; Banerjee and Pedersen, 2003). However, in any given document although some Ngrams will have meaning as a phrase, most do not. For example, in the preceding sentence, "some Ngrams" might be informative but "as a" is most likely not (at least not without a connection to "meaning" and "phrase"). The process by which meaningful Ngrams are recognized is more computationally expensive and less straight-forward than the process by which we recognize unigrams. For example, "an Ngram might be considered interesting if it occurs more often than would be expected by chance, or has some tendency to predict the occurrence of other phenomena in text" (Banerjee and Pedersen, 2003, p. 370).[1] Because of this, implementation on custom programming scripts becomes more difficult, especially for those interested in using text analysis as a tool and not a research agenda in itself.

In any case, despite some apparent advantages when incorporating Ngrams, the use of individual words as features has been shown to perform quite well in most applications (Grimmer, 2010; Hopkins and King, 2010; Lewis, 1992; Yang and Pedersen, 1997). Feature selection in the bag-of-words approach has taken on many forms, the most simple of which are stopword removal and document frequency thresholding (Forman, 2003; Rijsbergen, 1979). Stopword removal consists of removing a list of commonly used words and document frequency thresholding consists of removing the least commonly seen words across documents. The general idea behind both approaches is that neither will provide useful information for the purposes of classification. For example, the words "it," "the," and "but" 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, words such as "attenuate" and "vinculation" only show up in a few of the documents and again contribute little

[1] Banerjee and Pedersen (2003) have written an open-source software package in Perl called Ngram Statistics Package. Incorporating this or parts of this package into PreText could be promising in the future.

information important for our classification task.[2]

Stopword removal and document frequency thresholding are straight-forward, 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). They have also come under some criticism, in Political Science from Monroe, Colaresi and Quinn (2008), for example. In addition to the method described by Monroe, Colaresi and Quinn, there are a number of more complex feature selection methods that have emerged in the literature of other disciplines, such as information gain approaches (Botsis et al., 2011).

Information gain approaches are based on selecting the features that most improve the classification performance on the training set. A notable example of an information gain approach is that of Das (2001), who uses an iterative feature selection process based on the Adaboost algorithm designed to increase model performance (Freund et al., 2003).

In addition to these selection approaches, the tokens, or the single words that have been extracted from the text, often undergo some type of stemming. 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 also appear identical. The extent and rules for stemming depends on the chosen stemming algorithm.[3] In a basic sense, stemming has the effect of mapping different tokens from the text to the same feature in the dataset. In this same line of thought, synonymy poses a related problem.

Simply put, synonyms are words or phrases that have the same meaning. In the context of news stories, this issue is exacerbated since repetition is frowned upon 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 in more detail by (Mohammad and Hirst, 2006; 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

[2] Ask Dr. Monroe about the latter.

[3] PreText comes with the option of including the Porter Stemmer, an aggressive but effective stemming algorithm (Porter, 1980). Other stemming algorithms include the Paice/Husk algorithm (Paice, 1994) and distribution-based Ngram stemming (Mayfield and McNamee, 2003)

matching.[4]

Disambiguation is another potential issue that arises in text analysis, as the same word or words may have different meaning in different contexts (Blair, 1992, 2003). A "battle" between Russia and Georgia should probably be distinguished from a "battle" between the Detroit Lions and the Green Bay Packers.[5] 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 mean to equip 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 (see Hoffart et al. (2011), for example.)

Comparing The Baseline With Three Alternatives[6]

The document set experimented with here has been constructed for the purposes of collecting data on joint military exercises (JMEs). A JME is defined as occurring when the militaries from more than one state interact in such a way as to mutually enhance their ability to carry out military operations. The interaction may revolve around computer simulations or be carried out in the field in the style of a wargame without actual combat.

Four sources, each with global coverage, were selected for the purposes of constructing the initial document set. These sources are The Associated Press, Agence France Presse, Interfax News Agency, and Xinhua News Agency. The range of the data collection is from 1970-2010, but from 1970-1976 all English news sources were included due to a lack of sufficient recall from these four sources. LexisNexis (LN) was queried using the following search string:

((mil! OR war!) AND (exercis! OR train! OR simulat!) AND NOT (sports OR lifestyle

[4] It is important to note that from a computation perspective, 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). This makes the choice of which to implement in future versions of PreText a simple one. Pedersen, Patwardhan and Michelizzi (2004) have written a Perl module based on WordNet, and it may be easily implemented here.

[5] Lions fans can only hope this is a battle and not a slaughter.

[6] All text processing has been done with PreText, original software for text representation written in perl. This software is open-source and available on my website, `vitodorazio.weebly.com`. A copy of the User's Manual and the source code is included in the pages following this chapter of the dissertation.

OR tax cuts OR entertainment OR Wall Stree OR baseball)

This process returns a total of 256,734 documents. Manually reading through so many documents is prohibitively expensive, making automated document classification a necessary alternative for improving precision.

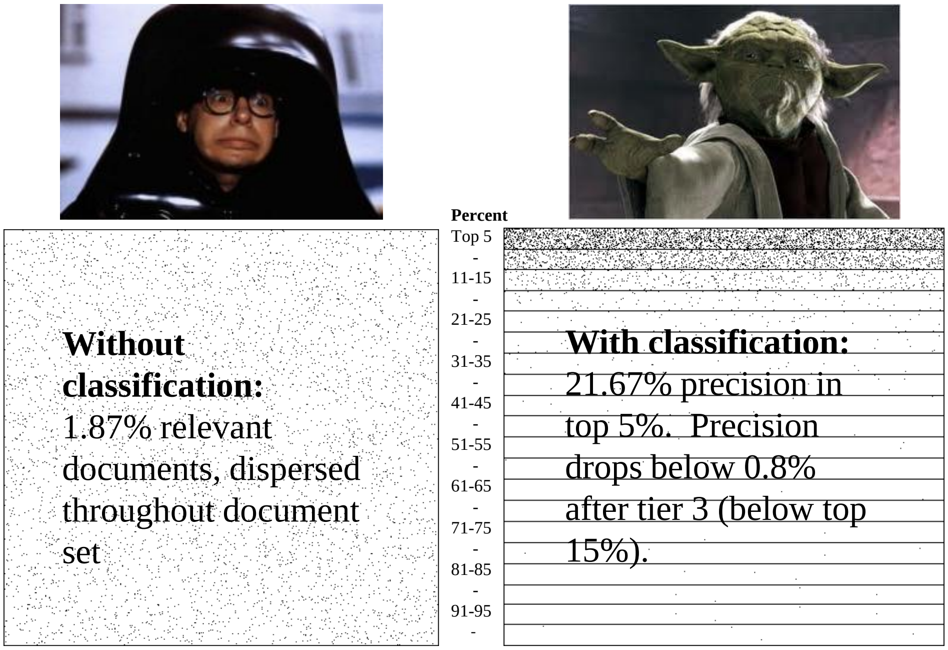
Percent
Top 5
-
11-15
-
21-25
-
31-35
-
41-45
-
51-55
-
61-65
-
71-75
-
81-85
-
91-95
-
Without classification:
1.87% relevant documents, dispersed throughout document set
With classification:
21.67% precision in top 5%. Precision drops below 0.8% after tier 3 (below top 15%).

Figure 1: Advantages of Document Classification

The general advantages of automated document classification are depicted in Figure 1. In the Dark Helmet plot on the left, the relevant documents are scattered throughout the entire set of documents with a precision of just 1.87%. To improve this precision in an effort to save both time and resources, a training set of 10,000 documents is randomly selected and labeled as relevant or irrelevant. Using inductive support vector machines, a decision rule is learned on the training set and is used to rank the remaining documents in the test set (Vapnik, 1998; Burges, 1998; Joachims, 2002).[7] This ranking is roughly represented by the Yoda plot on the right.

[7] SVMs are generally used as a pure classification algorithm, but a ranking of the documents is possible using

To estimate the precision of the ranked documents, the top twenty percent have been split into four groups, represented by the top four tiers in Figure 1. Five hundred documents have been randomly sampled from each of the top four tiers and labeled so that the precision of the tier may be estimated. The first, or the top five percent, have a precision of about 21.67%. This is a dramatic improvement over the 1.87% in the entire document set. The second tier, or those documents ranked between six and ten percent, has a precision of 12.34%, over 6.5 times the precision of the entire set. The third tier has a precision of 3.35%, and the fourth a precision of 0.8%.[8]

Alternative One: Named Entities

For the baseline method, whose precision is seen above, the top five percent of documents have been extracted and fully labeled.[9] The baseline method utilizes unigrams, Porter stemming, stopword removal, and document frequency thresholding set at ten percent. The features are weighted using the term-frequency inverse document-frequency method, and the classification algorithm is inductive support vector machines.

In addition, named entities are recognized and stripped from the text, removing them from possible inclusion as a feature. The named entity recognition (NER) is accomplished via 1-, 2-, 3-, and 4-gram matching to a dictionary of known actors developed by Phil Schrodt.[10] In general, NER is an active field of machine learning research (Ratinov and Roth, 2009; Lawson et al., 2010) and can be incorporated into the feature selection process in a variety of ways.

In the first experiment, named entities are not stripped from the documents. This removes the necessity of a dictionary, increasing computation efficiency but, more importantly, removing the necessity of a document (`countrycodes.txt`) whose construction and maintenance is quite resource-intensive.

The incorporation of named entities via a dictionary into the text classification process can be handled in a variety of ways. For the baseline data, all named entities that are contained in Schrodt's dictionary, up to and including a 4-gram mention of the entity, have been stripped from the text and are therefore not incorporated into the feature selection.

SVMs by extracting the distance from the separating hyperplane.

[8] The remaining tiers in Figure 1 are visualized at below 0/8%.

[9] These document form the majority of document used to code the JME data.

[10] The dictionary is Schrodt's `countrycodes.txt` file available at `eventdata.psu.edu`.

Alternative Two: Source Selection

Representations of text are intended to capture the characteristics of the documents that distinguish individual documents from one another. In this sample, one of the primary distinctions in the structure of the text is the source. Of the four sources used, each contain a different style of writing and emphasize different aspects of the exercises using often different vocabularies. As such, it is reasonable to believe that the sources should be classified separately. For the baseline, however, all sources were classified together.

In support of this hypothesis, the following three excerpts are taken from the first paragraph about a 2012 military exercise between Thailand and the United States, referred to as Cobra Gold:

> The U.S. Defense Department is open to considering a Thai request to allow Myanmar's military officers to observe a joint military exercise, Pentagon Press Secretary George Little said here on Friday.
>
> —Xinhua General News Service, *U.S. open to inviting Myanmar to observe war game*

> The United States plans to invite Myanmar to a major regional military exercise next year, host country Thailand said Friday, reflecting a dramatic easing of tensions between the former foes.
>
> —Agence France Presse – English, *Myanmar 'set to join US military drills'*

> Myanmar's military, long criticized for human rights abuses, may be invited as an observer at annual U.S.-Thai joint military exercise next year, officials in Washington and Bangkok said Friday.
>
> —The Associated Press, *Myanmar army may get invite to US-Thai exercise*

This example demonstrates at least two differences that have been observed while reading these documents. First, Xinhua is more likely than other sources to refer to U.S. and Western JMEs as "war games" instead of exercises.[11] Second, Xinhua tends to downplay abuses such as those committed by the government of Myanmar, as is evidenced by the lack of any mention of abuses or tensions in the lead paragraph. The Associated Press and Agence France Presse, on the other hand, make note of "human rights abuses" and "tensions," respectively.

To test this hypothesis, the documents are grouped by source and each group is classified separately. The training set for the baseline model has been split by source, yielding 3,534 labeled

[11] Xinhua calls them "war games" in the title of the article.

documents by The Associated Press, 3,219 by Agence France Presse, 363 by Interfax, and 2,076 by Xinhua.[12] The test set has also been split by source: 85,664 from The Associated Press, 79,728 from Agence France Presse, 8,951 from Interfax, and 51,469 from Xinhua).

Four decision rules were learned, one for each source, and used to classify the test set of each source. From each of these four classifications, the top five percent have been extracted and combined to form the set of documents that is compared to the baseline.

Alternative Three: Feature From Relevant Documents

For the baseline and the previous two alternatives, the features have been selected from both the positively and negatively-labeled documents in the training set. This experiment is designed to test the necessity of using features from both labels and instead uses only features derived from positively-labeled documents. This is a supervised form of feature selection, as opposed to the others which do not require us to know anything about the relevance of the document set in order to select the features.

This experiment may prove useful in that it is common for researchers to be able to locate a representative sample of documents related to their concept of interest. They may know of key examples from previous research or key terms that are associated with the concept, making it easier to identify and locate.

All relevant documents in the training set have been extracted, and the dictionary is comprised of features selected solely from these documents. All documents in the training set – both positive and negative ones – are used to learn a decision rule to then classify the test set. As in the other experiments, the top five percent are randomly sampled and labeled in an effort to measure the precision of this method.

Results and Discussion

Figure 2 illustrates some of the results in a non-proportional Venn Diagram. The baseline, named entities, and sources sets have a large intersection: 7,032 documents, or almost 55% of all documents in the baseline. However, each set also has a substantial number of documents that are not contained

[12] I had hoped Interfax would have had better coverage of the Soviet Union and post-Soviet states, but it turns out that its inclusion provided a marginal improvement at best.

in any other set: 1,302, 2,344, and 2,810 for baseline, sources, and named entities, respectively. In contrast, the relevants set is by far the most unique set with 10,760 documents not intersecting with any other.

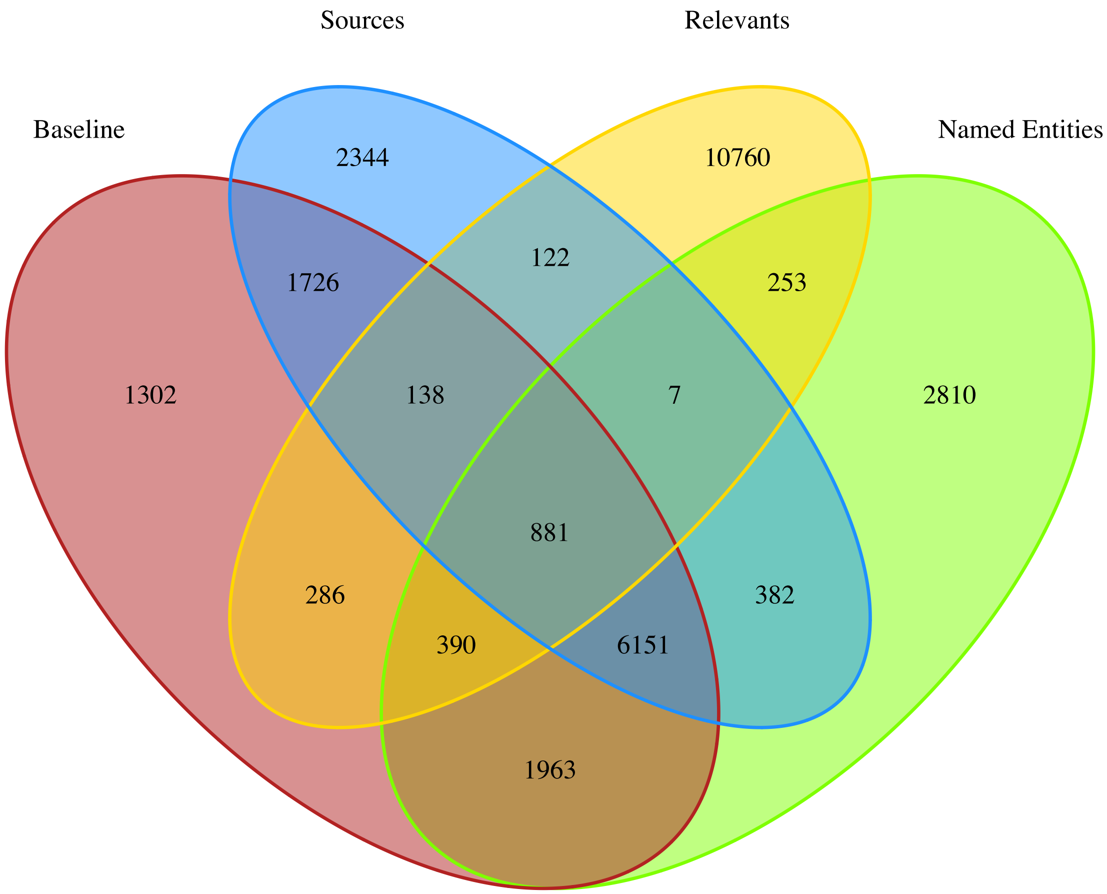

Figure 2: Unlabeled Classification Results

Of primary interest here is the overall performance of the feature selection methods. However, because of resource constraints, not every document has been labeled. Of the four sets, only the baseline set has been completely labeled.[13] From the remaining three sets a random sample of between 200 and 500 documents have been drawn and labeled. Although estimating each set's

[13] The baseline set is what has been used to code the JME data. As these experiments progressed, however, I have updated the dataset using information from the experimental sets.

precision from these numbers is simple, it requires some clarification.

With the exception of the baseline, each set has been separated into five subsets: the four that intersect with the baseline and the the relative complement of the baseline with respect to the experimental set. The precision is estimated by weighting the precision of each of the five subsets by the number of documents, summing over the five sets, and then dividing by the total number of documents. For example, for the sources set, the equation is:

$$\frac{2855 \times .16 + 6151 \times .2974 + 881 \times .3291 + 138 \times .0362 + 1726 \times .0707}{11751} = 23 \tag{1}$$

The precision of the various subsets are reported in Table 1. Comparing the three experiments here to the baseline method, which has a precision of 21.67%, it appears that the named entities set significantly outperforms the others with an estimated precision of 28.69%. The sources method marginally outperforms the baseline with a precision of 23%. With a precision of 3.8%, the relevants method is by far the worst performing method of feature selection.

Theoretically, using two-class inductive SVMs may not be an appropriate methodological choice for the relevants method. The nature of inductive SVMs is to learn a decision rule – a multidimensional hyperplane – separating the positive and negative-labeled documents in the training set. In the case where features are selected based solely on relevant documents, a more appropriate classification algorithm would be one designed to recognize similarities in a single class. A clustering algorithm, such as k-means or learning vector quantization, might be appropriate here. One-class SVMs exist and have been shown to be effective in document and other classification tasks (Scholkopf et al., 2001; Manevitz and Yousef, 2002; Munoz-Mari et al., 2010), and may also be more effective.

Interestingly, the intersections of the baseline and each of the three experimental sets exhibits a higher level of precision than the entire baseline set. For example, the named entities set contains 9,385 that are in the baseline set, 6,827 of which are negative and 2,558 of which are positive – a precision of 27.26% in the intersection. The sources set has an intersection with the baseline set of 8,896 document, 6,650 of which are negative and 2,246 of which are positive.[14] The resulting

[14] Note that the sources experiment does not have the same number of documents in the top five percent as the baseline and other experiments because these sources alone did not provide sufficient coverage in the early 1970s, and thus other sources were included from 1970-1976. The difference is 1,086 documents.

Table 1: Precision Results

B=Baseline; NE=Named Entities; S=Source; R=Relevants		
Subset	Documents	Precision
B	12,837	21.67
NE	12,837	28.69
S	11,751	23
R	12,837	3.8
$B^C \cap NE$	(2810+253+7+382)	32.6
$B^C \cap S$	(2344 + 122 + 7 + 382)	16
$B^C \cap R$	(10760 + 253 + 122 + 7)	< 1
$B \cap NE \cap S \cap R$	(881)	32.91
$B \cap S \cap NE^C \cap R^C$	(1726)	7.07
$B \cap S \cap R \cap NE^C$	(138)	3.62
$B \cap S \cap NE \cap R^C$	(6151)	29.74
$B \cap NE \cap S^C \cap R^C$	(1963)	19.15
$B \cap NE \cap R \cap S^C$	(390)	16.15
$B \cap R \cap S^C \cap NE^C$	(286)	6.64
$B \cap NE^C \cap S^C \cap R^C$	(1302)	5.91

precision is of the baseline and sources intersection is 25.25%. The intersection of the relevants set and baseline set is small with just 1,695 documents. However, even here 1,318 are negative and 377 are positive – a precision of 22.24%.

The precision of the intersection of the four sets is 32.91% – the highest of any subset of documents. Although there are exceptions, in general the intersections have a higher precision than other comparable subsets. For example, the precision of the 1,302 documents contained in the baseline set and not in any other of the three is just 5.91%, lower than any intersection involving the baseline. This includes the 286 documents where only the baseline and relevants sets classified the documents as positive.

One explanation for the difference in precision is simply that I have rediscovered the benefits of ensemble learning. Ensemble learning is quite simple: multiple models, independent of one another, are used to estimate some outcome, and then some function is used to combine the predicted outcomes to generate a single prediction. For example, if we are predicting a dichotomous outcome variable using an ensemble of five models, we might rule by majority and only predict a 1 if three (or more) of the five models predict a 1.

Typically, ensemble learning is accomplished using different classification algorithms, or different features, while holding the feature selection method constant. For example, in a text application we might use SVMs, naive Bayes, regularized logistic regression, and random forests while keeping our feature selection method the same. In these experiments, I have reversed this: the classification algorithm is held constant while the feature selection method is varied.[15] Based on these results, it appears to be an effective performance enhancer.

Although not the original purpose of this paper, this finding is unexpected and warrants further examination because of its potential for impacting feature selection methods. The purpose of feature selection is to reduce the dimensionality of the data and, in this application, to measure and extract features of the text that best enable accurate classification. However, there is no "correct" method for selecting features and the general practices followed are based on what has been shown to work well.

Future research could explore an algorithm where various feature selection methods are used to classify a set of unlabeled documents, as is done in the experiments here. Then, when deciding which documents to extract and read in search for relevant information, the documents that are in the intersection of the most sets are used. This is in place of using SVMs, or some other algorithm to rank the documents and then use some arbitrary cutoff (here at five percent) to select those for further analysis.

Various intersections may also be weighted differently depending on their performance in in-sample tests. For example, if the named entities method performs best on the training data, intersections of the named entities set could be given more weight than intersections involving the relevants set, which would perform the worst on in-sample data.

Conclusions

Feature selection is an integral component of text analysis. As it relates to measuring concepts of interest in Political Science research, it is embedded in the document classification process, an increasingly important tool for locating relevant information. As political scientists, we have an application of automated document classification that is different than its traditional applications. Specifically, we construct initial document sets to have *sufficient recall*, i.e. enough documents to conduct a valid and reliable content analysis. Once this initial document set has been constructed,

[15] As a caveat, these methods may not be that divergent from one another. In this sense, perhaps the results are due more to selecting different features than actually selecting different feature selection methods.

the goal becomes to maximize precision.

In these experiments, three feature selection methods based on practical improvements geared towards improving the efficiency of data collection have been examined alongside a baseline method. The features selection methods are compared in terms of their precision in the classification of the initial document set, which has been downloaded from LN and consists of 256,734 documents from four news sources.

The results suggest that the highest level of precision results from leaving the named entities in the data, as opposed to the baseline method which systematically removes named entities found in Schrodt's `countrycodes` files. A marginal improvement over the baseline method is had by classifying each of the four sources separately. Finally, selecting features based purely on relevant documents is shown to be a very unproductive feature selection method. However, the low precision of the relevants method could be due to an inappropriate classification algorithm. Future research could look towards using clustering algorithms of various types when features are selected based on a set of relevant documents as opposed to a set of relevant and irrelevants.

Future research in this area should also explore the finding that the precision appears to be highest when the various feature selection methods agree on the classification. This is a basic example of ensemble learning, but instead of varying the classification algorithm, as is commonly practiced, the feature selection method itself is varied. Future research should explore this finding using more divergent feature selection methods, including variations on the weighting, stemming, and Ngram usage.

References

Banerjee, Satanjeev and Ted Pedersen. 2003. The Design, Implementation, and Use of the Ngram Statistics Package. In *Computational Linguistics and Intelligent Text Processing*, ed. Alexander Gelbukh. Vol. 2588 of *Lecture Notes in Computer Science* Springer Berlin Heidelberg pp. 370–381.

Bikel, Daniel M. and Vittorio 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: Association for Computational Linguistics pp. 145–148.

Blair, David C. 1992. "Information Retrieval and the Philosophy of Language." *The Computer Journal* 35(3):200–207.

Blair, David C. 2003. "Information Retrieval and the Philosophy of Language." *Annual Review of Information Science and Technology* 37(1):3–50.

Botsis, Taxiarchis, Michale D. Nguyen, Emily Jane Woo, Marianthi Markatou and Robert Ball. 2011. "Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection." *Journal of American Medical Information Association* 18:631–638.

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

Chieu, Hai Leong and Hwee Tou Ng. 2002. Named entity recognition: a maximum entropy approach using global information. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1.* COLING '02 Stroudsburg, PA, USA: Association for Computational Linguistics pp. 1–7.

Das, Sanmay. 2001. Filters, wrappers and a boosting-based hybrid for feature selection. In *Proceedings of the 18th annual international Machine Learning.* San Francisco, CA: Kaufmann pp. 74–81.

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.*

D'Orazio, Vito, Steven T. Landis, Glenn Palmer and Philip Schrodt. 2012. "Separating the Wheat from the Chaff: Applications of Automated Document Classification to MID.". Presented at the MidWest Political Science Meeting, 2011. Available at `http://vitodorazio.weebly.com/papers.html`.

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.

Freund, Yoav, Raj Iyer, Robert E. Schapire and Yoram Singer. 2003. "An efficient boosting algorithm for combining preferences." *Journal of Machine Learning Research* 4:933–969.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18(1):1–35.

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

Hoffart, Johannes, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* EMNLP '11 Stroudsburg, PA, USA: Association for Computational Linguistics pp. 782–792.

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

Krippendorff, Klaus. 2013. *Content Analysis: An Introduction to its Methodology.* Third ed. Thousand Oaks, CA: Sage.

Lawson, Nolan, Kevin Eustice, Mike Perkowitz and Meliha Yetisgen-Yildiz. 2010. Annotating large email datasets for named entity recognition with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.* CSLDAMT '10 Stroudsburg, PA, USA: Association for Computational Linguistics pp. 71–79.

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.

Manevitz, Larry M. and Malik Yousef. 2002. "One-class svms for document classification." *J. Mach. Learn. Res.* 2:139–154.

Mayfield, James and Paul McNamee. 2003. Single n-gram stemming. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.* SIGIR '03 New York, NY, USA: ACM pp. 415–416.

Mohammad, Saif and Graeme 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 '06 Stroudsburg, PA, USA: Association for Computational Linguistics pp. 35–43.

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.

Munoz-Mari, J., F. Bovolo, L. Gomez-Chova, L. Bruzzone and G. Camp-Valls. 2010. "Semisupervised One-Class Support Vector Machines for Classification of Remote Sensing Data." *Geoscience and Remote Sensing, IEEE Transactions on* 48(8):3188–3197.

Paice, Chris 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 '94 New York, NY, USA: Springer-Verlag New York, Inc. pp. 42–50.

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

Pedersen, Ted, Siddharth Patwardhan and Jason Michelizzi. 2004. WordNet::Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*. HLT-NAACL–Demonstrations '04 Stroudsburg, PA, USA: Association for Computational Linguistics pp. 38–41.

Porter, Martin F. 1980. "An Algorithm for Suffix Stripping." *Program* 14(3):130–137.

Ratinov, Lev and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. CoNLL '09 Stroudsburg, PA, USA: Association for Computational Linguistics pp. 147–155.
URL: *http://dl.acm.org/citation.cfm?id=1596374.1596399*

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.

Scholkopf, Bernhard, John C. Platt, John Shawe-Taylor, Alex J. Smola and Robert C. Williamson. 2001. "Estimating the support of a high-dimensional distribution." *Neural Computation* 13(7):1443–1471.

Schrodt, Philip A. 2009. *TABARI: Textual Analysis By Augmented Replacement Instructions*. `http://eventdata.psu.edu/tabari.html`.

Sebastiani, Fabrizio. 2002. "Machine Learning in Automated Text Categorization." *ACM Computing Surveys* 34(1):1–47.

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*

Tjong Kim Sang, Erik F. and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*. CONLL '03 Stroudsburg, PA, USA: Association for Computational Linguistics pp. 142–147.

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

Wang, Tong and Graeme Hirst. 2012. "Exploring patterns in dictionary definitions for synonym extraction." *Natural Language Engineering* 18:313–342.

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.

Yin, Lanlan, Guixian Xu, Manabu Torii, Zhendong Niu, Jose M. Maisog, Cathy Wu, Zhangzhi Hu and Hongfang Liu. 2010. "Document classification for mining host pathogen proteinprotein interactions." *Artificial Intelligence in Medicine* 49(3):155 – 160.