A Method to Automatically Identify the Results from Journal Articles

Henry A. Gabb, University of Illinois at Urbana-Champaign
Ana Lucic, University of Illinois at Urbana-Champaign
Catherine Blake, University of Illinois at Urbana-Champaign

Abstract
The idea of automating systematic reviews has been motivated by both advances in technology that have increased the availability of full-text scientific articles and by sociological changes that have increased the adoption of evidence-based medicine. Although much work has focused on automating the information retrieval step of the systematic review process with a few exceptions the information extraction and analysis have been largely overlooked. In particular, there is a lack of systems that automatically identify the results of an empirical study. Our goal in this paper is to fill that gap. More specifically, we focus on the identification of 1) any claim in an article and on the identification of 2) explicit claims, a subtype of a more general claim. We frame the problem as a classification task and employ three different domain-independent feature selection strategies (χ^2 statistic, information gain, and mutual information) with two different classifiers [support vector machines (SVM) and naïve Bayes (NB)]. With respect to both accuracy and F1, the χ^2 statistic and information gain consistently outperform mutual information. The SVM and NB classifiers had similar accuracy when predicting any claim but NB had better F1 performance for explicit claims. Lastly, we explored a semantic model developed for a different dataset. Accuracy was lower for the semantic model but when used with SVM plus sentence location information, this model actually achieved a higher F1 score for predicting explicit claims than all of the feature selection strategies. When used with NB, the prior model for explicit claims performed better than MI, but the F1 score dropped 0.04 to 0.08 compared with models built on training data in the same collection. Further work is needed to understand how features developed for one collection might be used to minimize the amount of training data needed for a new collection.
Keywords: Machine learning, classification, scientific results, systematic reviews, text categorization, feature selection
Citation: Gabb, H.A., Lucic, A., Blake, C. (2015). A Method to Automatically Identify the Results from Journal Articles. In *iConference 2015 Proceedings*.

Acknowledgements: This research is made possible by a grant from the U.S. Institute of Museum and Library Services, Laura Bush 21st Century Librarian Program Grant Number RE-05-12-0054-12 Socio-Technical Data Analytics (SODA).
Contact: gabbiii2@illinois.edu, alucic2@illinois.edu, clblake@illinois.edu

1 Introduction

The idea of automating systematic reviews has been motivated by both advances in technology that have increased the availability of full-text scientific articles and by sociological changes that have increased the adoption of evidence-based medicine (EBM). A systematic review, which is the cornerstone of EBM can take 5-6 people more than 1000 hours to complete (Petrosino, 1999), so help is urgently needed to reduce the time between when new results are published and when they are integrated into clinical practice. The systematic review process is fundamentally an information and organization problem (Blake & Pratt, 2002) and there have been several efforts to automate the information retrieval (Bekhuis & Demner-Fushman, 2012; Cohen, Ambert, & McDonagh, 2009; Cohen, Demner-Fushman, Iorio, Sim, & Smalheiser, 2013; Cohen, Hersh, Peterson, & Yen, 2006; Kilicoglu, Demner-Fushman, Rindflesch, Wilczynski, & Brian Haynes, 2009; Matwin, Kouznetsov, Inkpen, Frunza, & O'Blenis, 2010) and information extraction (Blake, 2005) stages of the process. In addition to the automated strategies, a manual effort is underway to capture data required in a systematic review (Buckingham Shum, Domingue, & Motta, 2000). Tools are also available to help with writing the manuscript (RevMan). Figure 1 shows how this earlier work relates to the entire information synthesis process that was developed after a manual study of how scientists systematically review the literature (Blake & Pratt, 2002).

Our goal is to automatically identify results (also called findings or claims) from an empirical study. At the same time, we want to situate our experiment within a larger context to show the extent to which decisions that are made during each step of building this system can impact the final result of the experiment, in particular the selection of the data set, the feature selection strategies, and the choice of classification algorithms and their parameters. While the project has a practical component—recognizing results in an empirical study with the ultimate goal of connecting results from several publications or

corpora—it also puts on self-reflexive glasses as it examines the choice of metrics and parameters and their influence on the final analyses and results. The system proposed in the paper would reduce the time required to conduct a systematic review by connecting the retrieval phase of the systematic review process directly to the extraction and analysis phases. However, what is also of interest here is the tight connection that exists between the retrieval and analysis phase that is not always manifest and not always acknowledged when reporting the results of an experiment.

In addition to reducing the time needed to conduct a systematic review, the proposed system would enable the systematic review team to focus on the intellectual task of reconciling different results from the studies that have different study design and scope. A similar effort was started for researchers in aging where results from table captions were captured manually (Fuller, Revere, Bugni, & Martin, 2004), but unfortunately the one page article that described how to automate the process did not provide sufficient detail to repeat or evaluate the experiment (Hristovski et al., 2007).

The structured abstract provides a non-automated strategy to identify results, but structured abstracts are not available for all articles and more importantly the abstract does not report all the results. For example, one of the toxicology cohort studies in the collection described here reported 22 different standardized mortality ratios, but only seven of those appeared in the abstract. This finding is consistent with a prior study of biomedical articles which showed that fewer than 8% of claims were made in the abstract (Blake, 2010). From a full-text perspective, one might think that the document structure (i.e., the Results section) would provide another non-automated way to identify the results in a study; but, our analysis shows that authors discuss results throughout an article and that although most sentences in the results section do provide results, on average other sections contain more than half of the total number of reported results.

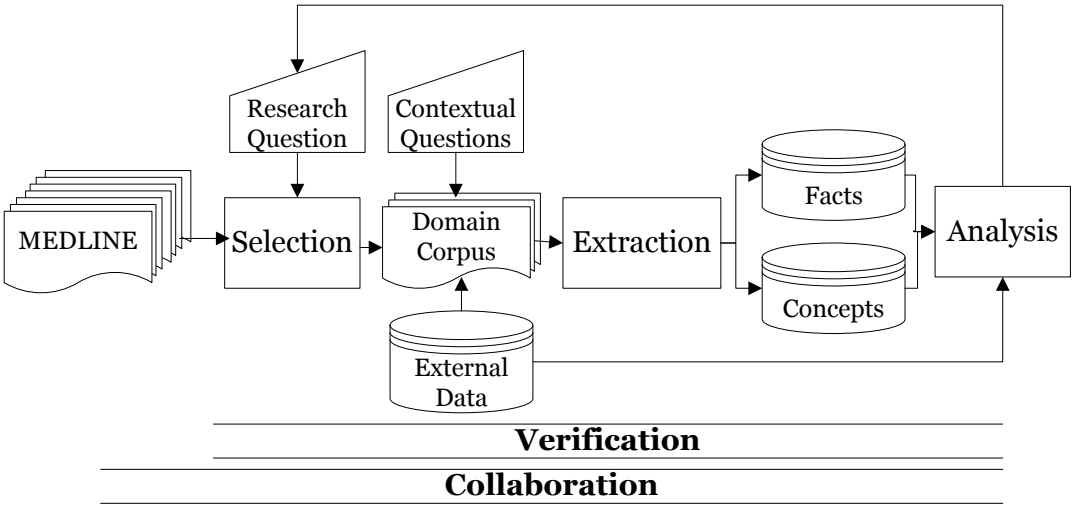

Figure 1. The process used to systematically review literature based on a manual study of scientists (Blake & Pratt, 2002).

Several efforts have focused on identifying results in general scientific literature. In particular, the Create a Research Space model includes counter-claiming and establishing a gap during the 'establishing a niche' phase (Swales, 1990), the Rhetorical Structure Theory includes two results and an evaluation relation (Mann & Thompson, 1988), and a model based on physics articles contains a comparison with experimental results and conclusions template (Kircz, 1991). The ART corpus of 256 full-text biomedical and chemistry articles has been modified and used to train classifiers to recognize five sentence types: hypothesis, motivation, background, objective, and finding (Groza, Hassanzadeh, & Hunter, 2013). Within this annotation scheme, a finding comes closest to the result category that we aim to classify in this paper. Another study used the NICTA-PIBOSO corpus to train models to differentiate between six scientific artefacts: population, intervention, background, outcome, study design, and other (Kim, Martinez, Cavedon, & Yencken, 2011). The outcome artefact is closest to our result category. This

corpus consists of 1000 biomedical abstracts: 500 retrieved from MEDLINE by querying for traumatic brain injury or spinal cord injury and 500 randomly retrieved for various diseases. This study achieves impressive overall results in terms of identifying the six artefacts in their annotation scheme but the corpus uses abstracts rather than full-text articles.

We adopted the Claim Framework because it focuses on results and has a rhetorical structure that has been evaluated with respect to biomedical literature using full-text articles (Blake, 2010). The framework comprises four information facets (agent, object, change, and dimension) that are pieced together to form five different types of claims (explicit, implicit, comparison, correlation, and observation). In this paper, we identify the specific results from full-text scientific articles by building models to identify result-containing sentences in empirical studies. We take a supervised machine learning approach where the sentences from an article are manually annotated, and we report the effect on predictive performance of three feature selection strategies, χ^2 statistic (CHI), information gain (IG), and mutual information (MI), in order to establish words that are good predictors of a result. We used two classifiers, support vector machines (SVM) and naïve Bayes (NB), because they are known to work well with text (Yang & Liu, 1999).

2 Method

2.1 **Data Collection and Annotation**

Our goal is to demonstrate a method that works well with full text scientific articles that could be used with the marked up text that is increasingly becoming available through services such as PubMed Central. In this study, the topic area constrained the scope of articles, and the articles were only available in PDF (i.e., XML or HTML was not available). We used automated tools where possible to convert the PDF to plain text and corrected any errors manually. In addition to the text, we kept information about the section, paragraph and sentence. A total of seventeen full-text published journal articles were collected with the following PubMed identifiers: 1497384, 2258019, 2340980, 3371457, 7760778, 8013410, 8248722, 8812199, 8901891, 8901897, 9463552, 9463553, 9463554, 9472338, 9624276, 9789948, and 10613399. These articles were selected to give a mix of experimental designs.

Four people, including one of the present authors (C.B.), were provided with sentences (in sequential order) from the articles, and sentences that reported a result were annotated using the Claim Framework. These annotations were subsequently checked for consistency and accuracy by two of the present authors (C.B. and H.G.). Sentences can have more than one claim, and in some cases more than one claim type. A summary of the five different claim types is included below:

- Explicit – A change is stated directly with a clear causative agent.

  *Exposure to **CCl4**[Agent] caused **dosage-dependent**[Change Modifier] **increases**[Change Direction] in **relative liver**[Object Modifier] **weight**[Object]…[PMID=2258019]*

- Implicit – A change is stated but the causative agent is implied.

  *As at 48 hr, 2 distinct patterns of **injury**[Object] were **apparent**[Change] in the **livers**[Object Modifier] at **72 hr after dosing**[Agent].[PMID=8248722]*

- Comparison – The statement establishes a comparative relationship but does not specify a causal relationship.

  ***Maternal BW**[Dimension] were **equivalent**[Change Direction] **between**[Change] **groups**[Agent-Object] on pnd 0 (immediately prior to the first dose).[PMID=10613399]*

- Correlation – The statement establishes a correlative relationship between two concepts but does not specify a causal relationship.

  ***Cancer risk**[Object] was found to **increase**[Change Direction] with **age**[Agent] for **all**[Object Modifier] of the **LHC groups**[Object Modifier]…[PMID=8901896]*

- Observation – The statement simply states that a change has occurred without specifying a change agent.

Meiosis[Object] in **stage XIV**[Object Modifier] was **normal**[Change]·[PMID=3371457]

2.2 Analysis

Although the annotations provide context, we were curious about how different classifier and feature selection strategies would impact prediction accuracy. To explore this question, we built several classifiers to discern sentences from full-text scientific articles that report a result (of any claim type) from those that do not. The automated feature selection strategies could draw from three sets of features: the individual words from each sentence; the section in which the sentence was located (abstract, introduction, method, result or discussion); and whether the sentence is the first or last sentence of a paragraph, as results are frequently reported at the beginning or end of a paragraph. Note that words were normalized to their base forms and stop words were removed before feature selection. The SPECIALIST lexicon from the Unified Medical Language System provides morphological (e.g.; increase, increases, increased, and increasing become increase; laryngeal becomes larynx) and orthographic (e.g., aortae becomes aorta) normalization (McCray, Srinivasan, & Browne, 1994). A customized set of 298 stop words was used. These are mainly determiners, auxiliary verbs, pronouns, prepositions, etc.

A recent study (Groza et al., 2013) takes a similar machine learning approach and a combination of linguistic and distribution features that are corpus-independent. Our approach differs in that complex annotation is not required. Only sentence location and result/no result annotation of sentences are required. Also, while our approach is corpus-dependent, i.e., the calculation of the top CHI, IG, and MI terms depends on lexical choices within a particular corpus, it is domain-independent in that we do not use any domain-specific features. The CHI, IG, and MI metrics used for feature selection as described in Yang & Pedersen (Y. Yang & J. P. Pedersen, 1997) have been used for a variety of classification tasks. Although these metrics are not new, to the best of our knowledge, they have not been used to build a classifier to distinguish between result and non-result sentences. Given that these metrics are easy to compute and do not use any additional linguistic or rule-based features, our approach has the advantage of simplicity while delivering comparable predictive accuracy. While the choice of feature selection metrics is limited (e.g., we could have selected IG ratio or Gini ratio, tf-idf, or any other feature selection metric, or maximum rather than average CHI, IG, and MI values), it serves a purpose of showing the connection and dependence between the data set used in an experiment, the feature selection strategy, and the analysis applied to this data set. Put differently, the experimental results that will be shown in the next section demonstrate this connection between these three seemingly disparate tasks that inform each other and draw from each other.

Each of the three feature selection strategies use the category label, so we selected a stratified sample that included 10% of the sentences to be used for model evaluation with the remaining 90% of annotated sentences used for feature selection and model development. The selected features were then used by the SVM and NB classifiers, as implemented in the Oracle Data Miner 11g (ODM release 1), which is part of Oracle SQL Developer (version 3.2). Experiments reported in this paper use the ODM default settings (Table 1). The ODM parameter space is quite large and the optimal settings for a given modeling problem are data-dependent. A more detailed study of the ODM parameter space, the effects of various parameters on these models, and the difficulties posed to reproducibility is presented elsewhere (Blake & Gabb, 2014).

Parameter	Classifier	Default
Singleton Threshold	NB	0
Pairwise Threshold	NB	0
Kernel Function	SVM	Linear
Tolerance Value	SVM	0.001
Complexity Factor	SVM	System
Active Learning	SVM	On

Table 1. Classifier settings used in these experiments.

In addition to the claim/no claim category, we conducted experiments to see if the features developed in a prior study of explicit claims (Blake, 2010) would work for the articles in this collection. In these experiments, 69 terms annotated as either a change or a change direction in the Claim Framework

were used as the feature set. These experiments test whether the lexical semantics of *change* can reduce the dimensionality of the feature space while improving the discernment of explicit claims.

3 Results and Discussion

Of the 2556 sentences, 965 reported a result (38%). On average, articles contained 150.3 ± 47.5 sentences of which 56.8 ± 21.8 reported a result. As one might expect, the majority of sentences in the results section (63%) provide details of the experimental findings, but focusing exclusively on the results section would miss more than half of the reported results (Table 2). Much of the current work in text mining explores abstracts but only 12% of the results from these articles are reported in the abstract, which is slightly higher than the earlier report of 7.84% (Blake, 2010). The distribution of claims in the main body of these articles also differed from the earlier study. Specifically, this collection was much less likely to have a result sentence in the introduction section (3% vs. 29% of total claims reported in (Blake, 2010)) and the discussion section (33% vs. 43%), whereas a greater proportion of claim sentences were in the results section (52% vs. 23%). Neither this study nor the previous study found many results in the methods section.

	Average Result (Min, Max, Std. Dev)				**Average Not a Result** (Min, Max, Std. Dev)			
Section	**Number**		**Percentage**		**Number**		**Percentage**	
Abstract	6.8	(3, 15, 3.1)	0.5	(0.3, 0.7, 0.1)	6.6	(3, 11, 2.1)	0.5	(0.3, 0.7, 0.1)
Introduction	1.9	(0, 7, 2.1)	0.1	(0, 0.4, 0.1)	16.7	(6, 31, 6.6)	0.9	(0.6, 1, 0.1)
Method	0.6	(0, 10, 2.3)	0.0	(0, 0.2, 0.1)	49.1	(13,87,22.8)	1.0	(0.8, 1, 0.1)
Result	28.9	(7,66,15.6)	0.6	(0.3, 0.8, 0.1)	15.1	(5, 34, 6.2)	0.4	(0.1, 0.7, 0.1)
Discussion	18.6	(4, 37, 9.0)	0.5	(0.2, 0.9, 0.2)	21.8	(3, 71, 17.3)	0.5	(0.1, 0.8, 0.2)

Table 2. Distribution of result-containing sentences per article section.

Most of the 965 sentences that reported a result were explicit (302, 31.3%) or implicit (302, 31.3%) claims, followed by comparisons (208, 21.5%), observations (206, 21.3%), and correlations (28, 2.9%). This distribution is very different to the earlier study which reported 77, 3, 5, 10 and 5 percent of explicit, implicit, comparisons, observations, and correlation claims, respectively.

3.1 Feature Selection Strategies for Result/Non-Result Categories

With respect to the feature selection strategies, Figure 2 shows that increasing the number of features typically leads to increased predictive accuracy, but the extent of the increase varies with respect to both the feature selection strategy and the classifier. Regardless of the classifier or the feature selection strategy, including the sentence location information always improved classification performance.

The CHI feature selection strategy outperformed the other two strategies. These results are consistent with prior work in other genres (Y. Yang & J. O. Pedersen, 1997). Of the three selection strategies, MI gave the lowest performance, and unlike the other models which generally improved with more features, the performance dropped for the NB classifier as the number of features increased.

3.2 Classifier Performance for Result/Non-Result Categories

For the proposed approach to be widely adopted in the systematic review community, the system must achieve good recall performance (i.e., all the result sentences must be identified even if non-result sentences are sometimes returned). With respect to the classifiers, the SVM classifier generally outperforms NB. However, the feature selection strategy plays a greater role in determining overall accuracy than the choice of classifier (Figure 2). This is consistent with Duda and Hart's observation that a "completely optimal feature extractor can never be anything but an optimal classifier." (p. 248) (Duda & Hart, 1973) The limits of the classifier performance reflects the limits of the features used and their discriminatory power. The highest accuracy using SVM is achieved with the top 200 ranked features from each feature selection strategy. However, the accuracy values vary from one feature selection strategy to another: 84% for CHI plus sentence location features, 75% for IG plus sentence location features, and 68% for MI plus sentence location features. Using NB with the top 200 ranked features gives the following accuracies: 79% for CHI plus sentence location features, 77% for IG plus sentence location features, and 77% for MI plus sentence location features.

The accuracy graph also reflects the rapid saturation that occurs in most strategies as the number of features increases. Increasing the number of features beyond 150 generally results in little or no improvement and can sometimes degrade performance.

Figure 2 also shows the precision/recall tradeoff of the SVM and NB classifiers and the various feature selection strategies. Using the top 100 term features plus sentence location features, the SVM classifier achieves a precision of 0.67 and 0.54 for CHI and MI, respectively, at the highest attainable recall (0.9). At the highest attainable recall for IG (0.8), SVM achieves a precision of 0.66. Although the MI selection strategy typically gave lowest predictive performance, when combined with sentence location features, it gives comparable accuracy to SVM and IG for the NB classifier. The decision tree (DT) and the generalized linear model (GLM) classifiers were also tested in this study. DT testing was stopped due to consistently inferior performance relative to the other classifiers. GLM results were comparable to SVM and NB (results not shown).

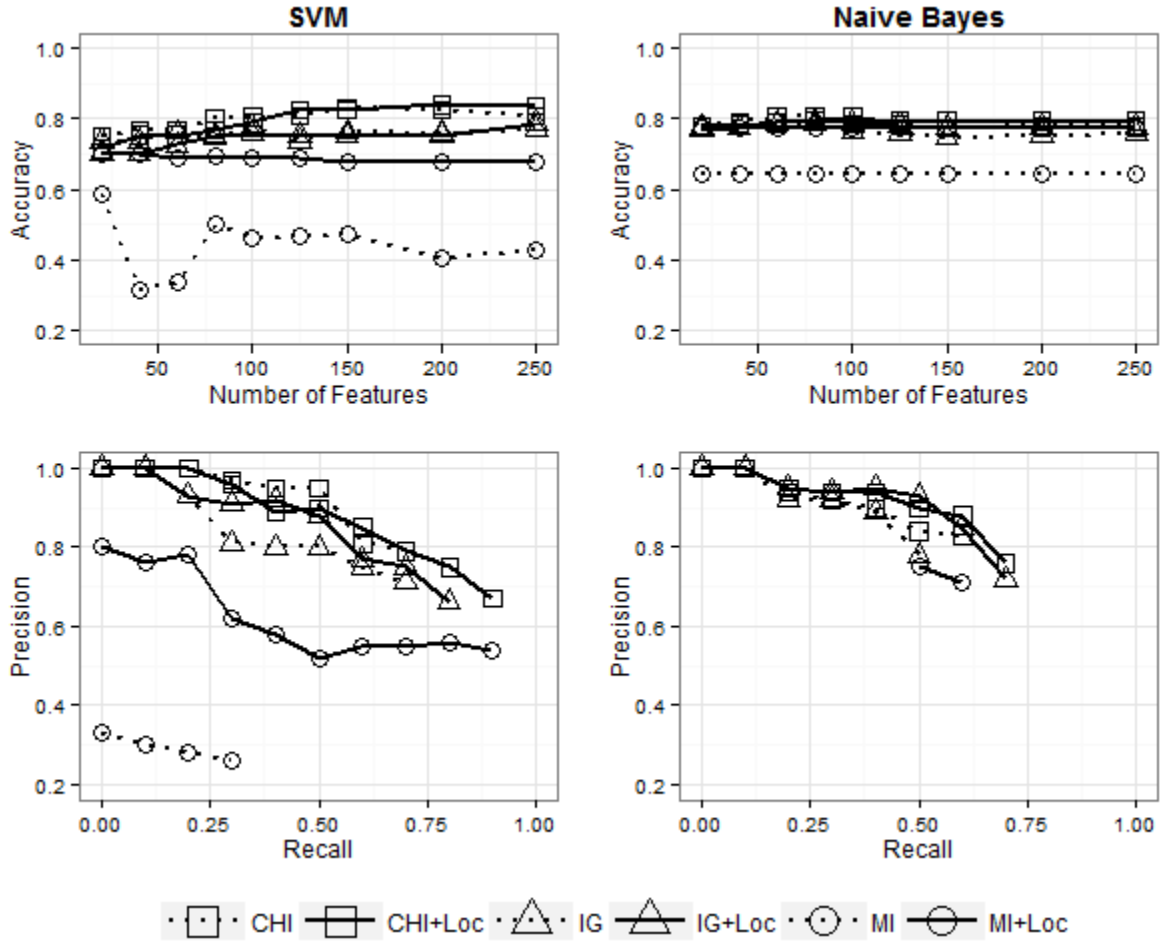

Figure 2. Accuracy and precision-recall curves for SVM and NB classifier results. The two top charts show the effect of feature selection strategy and the number of features on model accuracy. The two bottom charts show the tradeoff between precision and recall for each feature selection strategy (top 100 term features with or without sentence location features).

3.3 Term Overlap between Feature Selection Strategies

Both the CHI and IG feature selection strategies achieved good predictive performance, so we were curious to see which terms were identified by both strategies. The top 10 terms identified by both strategies were: *increase*, *hr*, *ccl4*, *ppm*, *observe*, *group*, *significant*, *significantly*, *decrease*, and *weight*.

Some of these terms capture the topic of the articles (e.g., ccl4), which suggests that the models may have over-fitted to the training data, but the remaining terms are what one might expect to see in a general result sentence from any scientific discipline.

Figure 3 shows that more than half (137, 54.8%) of the terms with the highest CHI scores also had the highest IG scores from the 250 highest ranked terms in both sets. Interestingly, there were no terms common to all three feature selection methods.

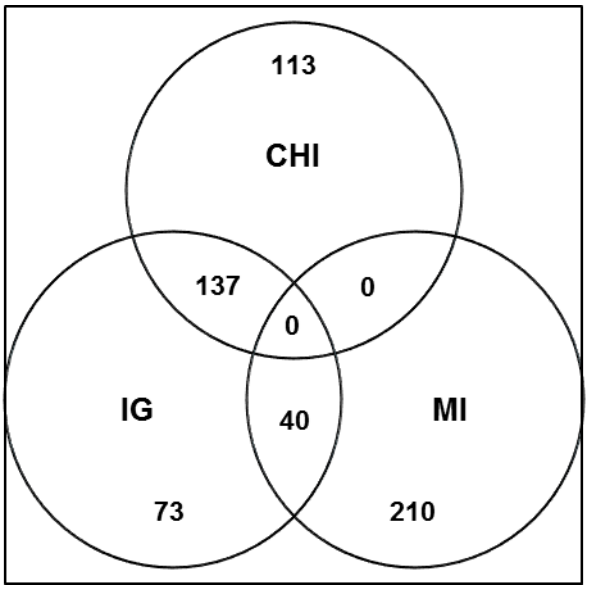

Figure 3. Overlap between the 250 highest ranked terms from each feature selection strategy.

3.4 Feature Selection Strategies for Explicit Claims

Having examined the ability of the system to predict whether or not a sentence contains a claim, we tested how well the feature selection strategies performed when the class of interest was a specific type of claim. More specifically, we were interested in how well the system could predict whether a sentence contains an explicit claim. The number of explicit claims is much smaller than the total number of claims (302 vs. 965), which corresponds to approximately 12% of the total number of sentences. Thus, the classification task to predict an explicit claim should be more difficult than predicting general result sentences. However, Table 3 shows that for the same number of features, explicit claims are predicted more accurately. This suggests that the vocabulary of explicit claims is more narrow and information-rich than the general case. As with our previous results, CHI and IG are similar and outperform MI. With that said, MI performance with the NB classifier is quite good when predicting explicit claims.

Creating the training data for the classifier is time consuming, so we explored a "prior" feature set which includes 69 change and change direction terms that were used in earlier work (Blake, 2010). Although the terms were developed for an entirely different collection, their predictive performance with SVM and NB is good but generally inferior to the automated feature selection methods using labeled training data from the same corpus.

	SVM				NB			
Accuracy	**CHI**	**IG**	**MI**	**Prior**	**CHI**	**IG**	**MI**	**Prior**
Any result – Top 100 terms	0.79	0.76	0.64		0.80	0.76	0.62	

Any result – Top 100 terms + location	0.80	0.80	0.77		0.82	0.79	0.71	
Explicit – Top 100 terms	0.87	0.89	0.59	0.75	0.88	0.89	0.86	0.71
Explicit – Top 100 terms + location	0.87	0.88	0.86	0.73	0.87	0.89	0.86	0.71

Table 3. Accuracy for each classifier and feature selection strategy when predicting explicit claims. "Prior" refers to a lexico-semantic feature selection strategy developed from a different collection.

The skewed distribution of sentences (i.e., most sentences are not results) containing explicit claims can obscure the results when considering accuracy alone. Table 4 shows F1 measures when trying to predict explicit claims from the top 100 features. The F1 measure depends on the combination of classifier and feature selection strategy. CHI feature selection performs best when the NB classifier is used but its performance suffers with the SVM classifier. With SVM, IG performs best. MI performs poorly regardless of classifier. When used with the SVM model, the feature set derived from prior work performed better than all selection strategies when used with location and better than the CHI and MI methods without location (and only 0.01 worse with MI). When used with NB, the prior model performed better than MI, but resulted in a 0.02 to 0.04 drop for IG and a 0.07 and 0.08 drop in F1 measures when used with the CHI and IG models without and with location.

	SVM				**NB**			
F1 Measure	**CHI**	**IG**	**MI**	**Prior**	**CHI**	**IG**	**MI**	**Prior**
Explicit – Top 100 terms	0.38	0.45	0.15	0.44	0.50	0.45	0.00	0.43
Explicit – Top 100 terms + location	0.27	0.42	0.00	0.43	0.51	0.47	0.00	0.43

Table 4. F1 measures for each classifier and feature selection strategy when predicting explicit claims. "Prior" refers to the lexico-semantic feature selection strategy developed from a different collection.

3.5 **Limitations**

The approach presented here does have some limitations. First, given the large and continuous parameter space of the ODM, it is hard to know whether optimal modeling parameters were used. Parameter tuning for these models and the difficulty posed by the large parameter space are discussed in more detail elsewhere, but hundreds of modeling experiments were performed to find reasonable parameters for the ODM classifiers (Blake & Gabb, 2014). Also, the feature selection strategy has a greater impact on prediction accuracy than the classifier. Second, the text corpus is fairly small (17 articles, 2556 sentences) because of the requirement for human annotation of the sentences. However, accepted practices were used when building models. The result/non-result sentences were randomly divided into training and evaluation sets, and no term or sentence features were derived from the evaluation set. The accuracy, precision, and recall data shown in Figure 2 indicate that the training set contains sufficient information for our modeling purposes. Third, the corpus is confined to a single scientific domain (i.e., toxicology). Different domains may have peculiarities in the way results are reported, however the experiments using terms from a prior study suggest that some terms do generalize amongst empirical studies. Further experimentation is need to see how far these models will generalize to other scientific domains. Finally, each classifier was tested in isolation. It may be possible to improve performance by aggregating classifier predictions and employing a voting scheme similar to (Groza et al., 2013).

4 Conclusion

Our goal in this study was to develop a method that would automatically identify result sentences from full-text journal articles for the purposes of reducing the time required to conduct a systematic review. We framed this goal as a classification activity and explored the accuracy of two different classifiers (SVM and NB) and three different domain-independent feature selection strategies (CHI, IG, and MI). Our results suggest that the feature selection strategy plays a greater role in predictive performance than the classifier, which is consistent with the notion that feature extraction followed by classification is a theoretically artificial task (Duda & Hart, 1973). SVM and NB give similar predictive performance. CHI and IG outperformed MI in terms of accuracy, precision, and recall, which is consistent with prior work (Y. Yang & J. O. Pedersen, 1997).

We also built several classifiers to predict whether a sentence contained an explicit claim. In spite of having fewer training cases than the result/non-result models, predictive accuracy was higher than the complete set of claims, most likely due to the narrower vocabulary of explicit claims. We also compared the accuracy and F1 measures achieved with automated feature selection strategies against features from a prior study. Although the accuracy was lower, the prior work resulted in better F1 scores for the SVM and slightly lower F1 scores with the NB model. Further work is needed to explore the extent to which features can be repurposed between collections in order to minimize the need for training data.

5 References

Bekhuis, T., & Demner-Fushman, D. (2012). Screening nonrandomized studies for medical systematic reviews: A comparative study of classifiers. *Artif Intell Med, 55*(3), 197-207.

Blake, C. (2005). *Information synthesis: A new approach to explore secondary information in scientific literature.* Paper presented at the JCDL '05, Denver, CO.

Blake, C. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *J Biomed Inform, 43*(2), 173-189. doi: 10.1016/j.jbi.2009.11.001

Blake, C., & Gabb, H. A. (2014). *Parameter tuning: Exposing the gap between data curation and effective data analytics.* Paper presented at the Proceedings of the 77th ASIS&T Annual Meeting, Seattle, WA.

Blake, C., & Pratt, W. (2002). *Automatically identifying candidate treatments from existing medical literature*. Paper presented at the Mining Answers from Texts and Knowledge Bases, Menlo Park, CA.

Buckingham Shum, S., Domingue, J., & Motta, E. (2000). Scholarly discourse as computable structure. 120-128.

Cohen, A. M., Ambert, K., & McDonagh, M. (2009). Cross-topic learning for work prioritization in systematic review creation and update. *J Am Med Inform Assoc, 16*(5), 690-704.

Cohen, A. M., Demner-Fushman, D., Iorio, A., Sim, I., & Smalheiser, N. R. (2013). Tools for identifying reliable evidence and implementing it in everyday clinical care. *AMIA Jt Summits Transl Sci Proc*, 42-44.

Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc, 13*(2), 206-219.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York, NY: John Wiley & Sons.

Fuller, S. S., Revere, D., Bugni, P. F., & Martin, G. M. (2004). A knowledgebase system to enhance scientific discovery: Telemakus. *Biomed Digit Libr, 1*(2), 1-15.

Groza, T., Hassanzadeh, H., & Hunter, J. (2013). Recognizing scientific artifacts in biomedical literature. *Biomed Inform Insights, 6*, 15-27. doi: 10.4137/BII.S11572

Hristovski, D., Revere, D., Bugni, P. F., Fuller, S. S., Friedman, C., & Rindflesch, T. C. (2007). Towards automatic extraction of research findings from the literature. *AMIA Annu Symp Proc*, 979.

Kilicoglu, H., Demner-Fushman, D., Rindflesch, T. C., Wilczynski, N. L., & Brian Haynes, R. (2009). Towards automatic recognition of scientifically rigorous clinical research evidence. *J Am Med Inform Assoc, 16*(1), 25-31.

Kim, S. N., Martinez, D., Cavedon, L., & Yencken, L. (2011). Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics, 12*(Suppl 2), S5.

Kircz, J. G. (1991). Rhetorical structure of scientific articles: The case for argumentational analysis in information retrieval. *Journal of Documentation, 47*(4), 354-372.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8*(3), 243-281.

Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O., & O'Blenis, P. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *J Am Med Inform Assoc, 17*(4), 446-453.

McCray, A. T., Srinivasan, S., & Browne, A. C. (1994). *Lexical methods for managing variation in biomedical terminologies.* Paper presented at the Proc Annu Symp Comput Appl Med Care.

Petrosino, A. (1999). Lead authors of Cochrane reviews: Survey results. Preliminary draft II: Report to the Campbell collaboration. University of Pennsylvania.

RevMan. (2014). Review Manager (RevMan) (Version 5.3): The Cochrane Library.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*: Cambridge University Press.

Yang, Y., & Liu, X. (1999). *A re-examination of text categorization methods.* Paper presented at the SIGIR '99, Berkeley, CA USA.

Yang, Y., & Pedersen, J. O. (1997). *A comparative study of feature selection in text categorization.* Paper presented at the Fourteenth International Conference on Machine Learning (ICML '97).

Yang, Y., & Pedersen, J. P. (1997). *A Comparative Study on Feature Selection in Text Categorization.* Paper presented at the Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97).

Table of Figures

Table of Tables