The use of bibliography enriched features for automatic citation screening

Babatunde Kazeem Olorisade[a,*], Pearl Brereton[a], Peter Andras[a]

[a] *School of Computing and Mathematics*
Keele University, Staffs. ST5 5BG, UK

Abstract

Context
Citation screening (also called study selection) is a phase of systematic review process that has attracted a growing interest on the use of text mining (TM) methods to support it to reduce time and effort. Search results are usually imbalanced between the relevant and the irrelevant classes of returned citations. Class imbalance among other factors has been a persistent problem that impairs the performance of TM models, particularly in the context of automatic citation screening for systematic reviews. This has often caused the performance of classification models using the basic title and abstract data to ordinarily fall short of expectations.

Objective
In this study, we explore the effects of using full bibliography data in addition to title and abstract on text classification performance for automatic citation screening.

Methods
We experiment with binary and Word2vec feature representations and SVM models using 4 software engineering (SE) and 15 medical review datasets. We build and compare 3 types of models (binary-non-linear, Word2vec-linear and Word2vec-non-linear kernels) with each dataset using the two feature sets.

Results
The bibliography enriched data exhibited consistent improved performance in terms of recall, work saved over sampling (WSS) and Matthews correlation coefficient (MCC) in 3 of the 4 SE datasets that are fairly large in size. For the medical datasets, the results vary, however in the majority of cases the performance is the same or better.

[*] Corresponding Author
Email addresses: b.k.olorisade@keele.ac.uk (Babatunde Kazeem Olorisade), o.p.brereton@keele.ac.uk (Pearl Brereton), p.andras@keele.ac.uk (Peter Andras)

Conclusion
Inclusion of the bibliography data provides the potential of improving the performance of the models but to date results are inconclusive.

Keywords: Computing methodologies, citation screening automation, systematic reviews, text mining, feature enrichment

1. Introduction

Systematic review (SR) is a literature review approach that provides for a rigorous, dependable and 'auditable' review methodology with the main goal of building an impartial and complete synthesis of available empirical research evidence on a specific topic; thus, creating a focused platform on which practically useful decisions and conclusions can be made [1, 2, 3]. This approach to review research is being widely used in the medical research and has been a major approach to review in software engineering research since its introduction in 2004 by Kitchenham et al. [1]. It is also commonly used in other disciplines such as education, social science, psychology etc.

Citation screening (CS) is a stage in the SR process, where the reviewers go through the titles and abstracts of retrieved articles (sometimes thousands of them) to determine whether they are relevant to the research focus or not, using prior (clearly) laid out inclusion/exclusion criteria [2]. This is possibly one of the most time consuming and critical stage of the SR process [4, 5]. Consequently, there has been a growing interest in providing support for the CS stage in SR using machine learning (ML) techniques with the majority of the focus on supervised ML approach. The research efforts at providing support for SR or any of its constituent stages through the use of ML/text mining techniques is currently more active within the medical and software engineering research.

Supervised ML algorithms typically learn patterns underlying the example data and project the knowledge to predict similarity or otherwise of new data to the learned example [6]. A major problem in using these algorithms for classification purposes in automatic CS is the small number of relevant (class) examples to learn from; the proportion of relevant to irrelevant class examples is commonly 1%-5% of the total data size. This situation is referred to as class imbalance [7]. Due to the highly imbalanced nature of the data classes in SRs, researchers working on automating the CS stage continue to find ways to make up for the shortage of relevant class examples.

Some of the methods the ML community have proposed to address this situation are:

(a) Cost Assignment: Assignment of different costs or weights to training samples [8].

(b) Data resampling: The repeated sampling of the original data either by over-sampling or under-sampling [9, 10]:

- Over-sampling: This involves including repeated or multiple instances of the minority class samples to make up for its under-representation during training.
- Under-sampling: The process of under-sampling involves reducing the samples of the majority class to create a 'reasonable' representation proportion between the majority and the minority class samples.

(c) SMOTing: Using the synthetic data produced from the Synthetic Minority Over-sampling Technique (SMOTE) [7]. The SMOTE combines both the over-sampling and under-sampling techniques to produce new data samples of both classes in the proportion specified by the user.

(d) Feature enrichment: An approach used in text classification to improve model performance by adding other possibly useful information, sometimes from (external) sources [11, 12]. In the context of automatic CS with text mining techniques it can be said to be the inclusion of other data beyond title and the abstract (that would have been ordinarily assessed by human) e.g. pre-trained embedding (sometimes from external sources), keywords, subject classification data, cited articles etc. to provide more information that could potentially strengthen the probability of identifying similarities or differences between articles [11, 12, 13].

Despite these efforts, there is yet to be an acceptable solution to the problem of stemming the effect of class imbalance in building text mining models to automatically screen citations. In this study, we explore the use of reference information to enrich the dataset.

We hypothesize that the reports of similar studies will include common references and hence common terminologies, authors and journals or conferences. Therefore, using the reference information will introduce some new terms such as the authors' names and technical terminologies from conferences or journals into the input feature which may help distinguish between relevant and irrelevant studies. The goal of this study therefore, is to test the effect of the reference data as a feature enrichment artefact on the performance of text mining models for automatic CS. We are not aware of any study that has previously investigate the effect of using full bibliography data with support vector machine (SVM) classifiers on model performance in automatic CS.

We build SVM models from two different sets of data, one with reference information (TiAbs(MeSH)Ref data) and the other without the reference information (TiAbs(MeSH) data). We used datasets from 4 software engineering (SE) reviews and 15 from clinical reviews. We explore datasets from two different domains because of the following:

(a) we considered the study to be about the ability of predictive models as a support tool for CS in SRs which should not necessarily be sensitive to the discipline of its input but the features of its training data.

(b) the process of the initial study selection based on study titles and abstracts is relatively straight forward and similar across the disciplines.

(c) the model is expected to learn to discriminate the articles from the features of its training (examples) data with no (elaborate) need for explicit inclusion/exclusion criteria.

In addition, the medical field has more labelled data from past SRs studies than does SE at the moment. The SVM was chosen for this study because a previous study has found it to be the most used in the context of automatic CS [14] and it has also been shown that it outperforms other classical ML algorithms good at text classification like Naïve Bayes [15].

Three of the SE datasets - Hall, Wahono and Radjenovic, named after the lead authors of the papers reporting the reviews were generated from existing SE SRs [16, 17, 18], annotated and made available by Zhe Yu [1]. The Kitchenham dataset (also named after the lead author) is the annotated SR data by Kitchenham et al. [19] and made available by Prof. Kitchenham [2].

The 15 clinical review datasets are part of the drug evaluation review program (DERP) conducted by the Oregon Evidence Based Practice Center (EPC) and made available through the Text REtrieval Conference (TREC) 2004 dataset. It is one of the few SR datasets available with inclusion/exclusion annotations. It has also been used in SE oriented research on CS support [13, 20].

In the rest of the paper, a brief overview of the related work is presented in section 2, the conduct of the study is the focus of section 3, the results are presented in section 4 with the discussion in section 5. The possible threats to the validity of the study are presented in section 6 while conclusions are presented in section 7.

2. Related Work

In this section, we highlight existing work relating to enriching features for use in automatic CS and the metrics used in this study to assess the performance of the ML methods.

2.1. Enriching feature for automatic CS

Severe data class imbalance is a common phenomenon in SR datasets. In CS, human researchers, guided by the preset inclusion/exclusion criteria, use the articles' titles and abstracts to decide whether to include or exclude an article for a review.

Imbalanced data classes however impair the performance of classification models in ML. Introducing external data as a way of enriching the base data is one way the community has devised to tackle the situation. This approach attempts to leverage the machine speed and power by increasing the basic (titles and abstracts) textual input data to provide (possibly) more information from each article that could further show which ones are related or not.

[1] at https://doi.org/10.5281/zenodo.837298

[2] This dataset has also been made available at https://doi.org/10.5281/zenodo.837298

One of the earliest attempts at feature enrichment in automatic CS research was the addition of Medical Subject Headings (MeSH) and the 'Medline' publication type data to the title and abstract [21]. A number of studies have used a similar approach. In [22], the authors mentioned using metadata alongside title and abstract while the authors of [23] mentioned using MeSH and natural language processing (NLP) features. Felizardo et al. used the mapping of citations to the contents that contain them to create article clusters for the identification of relevant citations [24].

Khabsa et al. used co-citation and clustering features to improve the feature quality of 15 SRs dataset in [13]. For co-citation, they worked on the assumption that if two articles are cited together in a third article, then both articles are likely to be on a similar subject. Therefore, either of the two articles that was not initially included in the dataset to be classified is retrieved and included as a positive sample. They further used the brown clustering algorithm [25] to create word clusters containing related words. With the cluster, each word is represented with a code that refers to a cluster of similar words which might have appeared in the training corpus.

2.2. Metrics

Several measures have been used to assess the performance of ML models than can be covered in this study. For this study, we assess our models with recall, precision, Matthews correlation coefficient (MCC), work saved over sampling (WSS) measures and the number of support vectors. These metrics are briefly defined below and in Table 1.

(a) Recall: the ratio of the correctly classified positive class examples given the total positive class examples in the test corpus [26]. Its value ranges from 0–1.

$$recall = \frac{TP}{TP + FN}$$

(b) Precision: the ratio of actual positive class examples and the total positive prediction [26]. Its value ranges from 0–1.

$$precision = \frac{TP}{TP + FP}$$

Table 1: Confusion matrix

Actual data Class	Classified as positive	Classified as negative
positive candidate	True positive (TP)	False negative (FP)
negative candidate	False positive (FN)	True negative (TN)

(c) MCC: a measure of the quality of classifications of a two-class classification model [27]. Its value ranges from −1− (+1).

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

(d) WSS: given a certain recall level (rl), the WSS is the percentage of the articles initially returned by the literature search which the researcher would not have to read because they have been screened out by the model [21].

$$WSS@rl = \frac{(TN + FN)}{N} - 1 + \frac{TP}{TP + FN}$$

$$WSS@rl = \frac{(TN + FN)}{N} - 1 + recall$$

where *rl* refers to the recall level for which the measure is taken.

3. Method

We prepared two datasets from each of the 19 review studies used. The first set - TiAbs(MeSH), contained title and abstract, and an additional MeSH feature for the 15 clinical review datasets. The second set - TiAbs(MeSH)Ref, contained the first set and the full reference list for each candidate article (where available, accessible and retrievable). The number of candidate articles for each review and their class distribution is shown in Table 2.

We had access to the data required for set one and wrote scripts to obtain the reference list to make up set two. We used the full article link provided for the four SE review studies to search for the articles in the publishers' websites and where possible and available, automatically extracted the reference list for each of the candidate articles. In the case of the 15 medical review studies, we used the provided PMID to search the pubmed database for information on the publisher(s) providing access to the full article. We extracted this information (if one was found), visited the publisher's site and attempted to retrieve the desired reference list for each of the candidate articles. The retrieved reference texts were initially cleaned of html tags and any url information before being merged with TiAbs(MeSH) data to form the set two data - TiAbs(MeSh)Ref. The number of references found per review is shown in Table 3 with detailed distribution according to the classes in Table 4.

Afterwards, each of the datasets underwent the same process from preprocessing to model assessment. In developing the models for this study, the following steps were followed:
Step 1 - data loading: Each data set was loaded into memory and shuffled.
Step 2 - data partitioning: The shuffled data was split into train and test sets using the ratio 7:3.
Step 3 - feature representation: In [21] Cohen et al. found that the binary feature representation produced better performance than the tfidf. But the same

Table 2: Size and class distribution for each review

Review	No. of candidate studies	Negative class	Positive class
Kitchenham	1704	1659	45
Hall	8911	8805	106
Wahono	7002	6940	62
Radjenovic	6000	5962	48
ACEinhibitor	2544	2503	41
ADHD	851	831	20
Antihistamines	310	294	16
AtypicalAntipsychotics	1120	974	146
BetaBlockers	2072	2030	42
CalciumChannelBlockers	1218	1118	100
Estrogens	368	288	80
NSAIDs	393	352	41
Opioids	1915	1900	15
OralHypoglycemics	503	367	136
ProtonPumpInhibitors	1333	1282	51
SkeletalMuscleRelaxants	1643	1634	9
Statins	3465	3380	85
Triptans	671	647	24
UrinaryIncontinence	327	287	40

Table 3: Number of references retrieved per study

Review	Not found	Found
Kitchenham	60	1644
Hall	408	8503
Wahono	313	6689
Radjenovic	347	5653
ACEinhibitor	1533	1011
ADHD	484	367
Antihistamines	192	118
AtypicalAntipsychotics	707	413
BetaBlockers	1182	890
CalciumChannelBlockers	770	448
Estrogens	206	162
NSAIDs	223	170
Opioids	1123	791
OralHypoglycemics	295	205
ProtonPumpInhibitors	802	531
SkeletalMuscleRelaxants	1079	564
Statins	2040	1425
Triptans	367	304
UrinaryIncontinence	178	149

Table 4: Class distribution of retrieved references

Review	Positive class		Negative class	
	Not found	Found	Not found	Found
Kitchenham	2	43	58	1601
Hall	1	105	407	8398
Wahono	2	60	311	6629
Radjenovic	1	47	346	5616
ACEinhibitor	24	27	1509	994
ADHD	7	13	477	354
Antihistamines	12	4	180	114
AtypicalAntipsychotics	77	69	630	344
BetaBlockers	15	27	1167	863
CalciumChannelBlockers	44	56	726	392
Estrogens	41	39	165	187
NSAIDs	20	21	203	149
Opioids	8	7	1116	784
OralHypoglycemics	68	68	230	138
ProtonPumpInhibitors	23	28	779	503
SkeletalMuscleRelaxants	5	4	1074	560
Statins	47	38	1993	1387
Triptans	7	17	360	287
UrinaryIncontinence	18	22	160	127

has not been established for the SE datasets. So, we started by running a grid search of four feature types (binary, term frequency (tf), term frequency inverse document frequency (tfidf) and the average word Word2vec) against the linear and non-linear SVM kernels and other SVM parameters to select the feature and the best SVM parameters combination that produced the highest recall. In our opinion, given the evidence gathering goal of SRs, recall is the simplest metric that can convey to a systematic reviewer how many of the relevant articles have been correctly identified by a model. This is important in SRs (particularly in the medical domain where it may be critical to ensure all available evidence is retrieved). We evaluated the models using other metrics but determining how many candidate articles have been correctly included may not be directly interpreted from them. Thus, we made recall the primary metric to select the model parameters.

Step 4 - data pre-processing: During the feature representation, we removed the English stopwords and any word that appeared in more than 80% of the corpus or less than 20% of the corpus. In binary representation of the features, a document term matrix is returned in the form:

$$f(x) = \begin{cases} 1, & \text{if } x \in d; \\ 0 & \text{otherwise} \end{cases}$$

where x represents each word from the vocabulary built from the training documents D and d is a document in D.

Step 5 - Feature selection/Dimensionality reduction: This step focussed on dimensionality reduction. In [21] Cohen et al. reported the results of top features for the 15 DERP datasets, using 0.05 α value with χ^2 technique to rank and select the top 5% features. There is no such reference for the SE datasets so we used the same χ^2 method and explored values between 1% and 50% to determine the most viable value for the top percentile to use for the reduction of the dimension of the resulting sparse feature vector.

Step 6 - model training and assessment: This step combines the model training and assessment. The SVM model was built using the Python's sklearn (SVC) implementation of the SVM based on libsvm. The whole dataset was fitted with the selected parameters using a stratified 5x2-fold cross validation on the 15 DERP datasets as previous studies have used the same approach on this dataset [28] and stratified 2x5-fold cross validation on the four SE reviews datasets. The stratification ensured that as much as possible the positive - negative class distribution in the original dataset was maintained in the train/test partitions used for the cross validation. The difference in the cross validation folding was necessary because the SE review datasets are larger than the DERP datasets, so they can use more data for model training and still have a substantial amount to test on. In n x k folds cross-validation the dataset was divided into k equal parts called folds, the data model was trained using one part of the data for testing and the rest ($k-1$ part) for training (note that the testing data is not seen during training). This is continued until each fold has been used as the test portion while the model was built from the rest of the data from scratch, with no information from previously trained models. This training and testing procedure was repeated n times to get more robust estimates of the generalisation error of the trained data models. The cross-validation error is the average error calculated across all tests. The average values of recall, precision, MCC and WSS were recorded for each model.

Step 7: The datasets were converted into a Word2vec feature representation which was also found in step 3 to perform better with both linear and non-linear SVM kernels. We removed English stopwords as in step 4, set maximum word length to 10, with a context window of 15 and with the number of feature equal to the size of the dimensions obtained from the χ^2 operation in step 5. In Word2vec, a corpus was converted to an embedding, where each unique word or phrases are mapped to vectors of real numbers [29]. This process involves the training of a two-layer shallow neural network to reconstruct the linguistic contexts of words. Word vectors in the resulting vector space were organized in such a way that similar words end up being closer to each other [30]. Then step 6 was repeated.

The details required for the reproducibility of this study are provided in appendix Appendix A.

4. Results

Results from steps 3, 4 and 5 in Section 3, that correspond to key activities in the text mining process, are presented here.

4.1. Feature representation

The binary features produced good results only with the non-linear kernels of the SVM. The Word2vec feature representation on the other hand showed comparable performance with both the linear and non-linear kernels of the SVM. Therefore, models were built from the *binary-non-linear, Word2vec-linear and Word2vec-non-linear* feature representation-SVM kernel combinations.

4.2. Dimensionality reduction

Setting $\alpha = 0.05\%$ for χ^2 technique to select the top 5% of ranked features did not result in the same values as given in [21] for the DERP datasets, this in addition to the fact that there exists no similar information on the SE datasets that we are aware of, informed our decision to explore different values to confirm which value would result in a reduced vector dimension with highest 'acceptable recall' value. By 'acceptable recall' value we imply the highest possible value for recall where the model still exhibited some discriminatory power over the dataset. It is interesting to find that better recall values can be obtained for the datasets at values of α other than 0.05.

We started with the binary feature representation for the TiABs(MeSH) data and found that each of the datasets performed best for different α-values with the χ^2 method used. However, the majority of the datasets seemed to start recording high ($\geq$ 90%) recall performance at around $\alpha = 5\%$ top percentile value. The performance around the 5% percentile is consistent with the findings reported in [21]. The different alpha values used and their corresponding feature size for the TiAbs(MeSH) data is presented in Table 5a. The TiAbs(MeSH) data recall performance results are used as a benchmark to search for the appropriate reduced dimension of the TiAbs(MeSH)Ref data that will produce equal, close enough or better recall performance than was initially observed in the TiAbs(MeSH) feature SVM models. The resulting feature sizes and their corresponding α values are shown in Table 5b.

Table 5: χ^2 selected top features

(a) TiAbs(MeSH) data

Reviews	Initial size	α value	final size
Kitchenham	5730	4	227
Hall	11834	8	947
Wahono	11137	6	668
Radjenovic	10165	5	508
ACEinhibitor	4933	5	246
ADHD	3017	4	122
Antihistamines	1570	2	31
AtypicalAntipsychotics	3237	3	98
BetaBlockers	4724	4	192
CalciumChannelBlockers	3462	4	138
Estrogens	1861	18	339
NSAIDs	1790	21	376
Opioids	4661	1	46
OralHypoglycemics	2112	10	211
ProtonPumpInhibitors	3299	5	165
SkeletalMuscleRelaxants	4826	1	48
Statins	6150	5	308
Triptans	2372	5	118
UrinaryIncontinence	1691	30	5075

(b) TiAbs(MeSH)Ref data

Reviews	Initial size	α value	final size
Kitchenham	20095	3.8	763
Hall	44302	5	2215
Wahono	41800	2	836
Radjenovic	33929	1	339
ACEinhibitor	3808	1.8	248
ADHD	7680	4	307
Antihistamines	2983	3.5	105
AtypicalAntipsychotics	7920	3	236
BetaBlockers	14510	4	580
CalciumChannelBlockers	90332	6	542
Estrogens	4780	10	478
NSAIDs	4457	21	936
Opioids	12034	0.9	116
OralHypoglycemics	5050	8	404
ProtonPumpInhibitors	8251	2.5	206
SkeletalMuscleRelaxants	11723	1	118
Statins	19454	3	584
Triptans	4969	3	150
UrinaryIncontinence	3634	15	545

4.3. Model assessment

With the binary feature, the TiAbs(MeSH)Ref data exhibited better recall values than the TiAbs(MeSH) data in 12 review topics, equal in four and less in two. The models could not produce any useful results for the 'SKeletalMuscleRelaxants' data despite the fact that it is larger than some other datasets in the collection. This might be due to the fact that it has the smallest number of positive candidates (9 compared to the negative class size of 1634, see Table 2).

Table 6: Binary feature non-linear kernel

(a) TiAbs(MeSH) data

Reviews	Mean Performance					Support vectors		configuration
	precision	recall	accuracy	WSS	MCC	neg	pos	parameters[3]
Kitch.	0.04 ± 0.01	0.93 ± 0.04	0.44 ± 0.16	0.35 ± 0.11	0.12 ± 0.03	1352 ± 1	25 ± 1	rbf, 1.0
Hall	0.33 ± 0.03	0.93 ± 0.04	0.98 ± 0.00	0.90 ± 0.04	0.55 ± 0.03	2238 ± 152	48 ± 2	sigmoid, 1.0
Wahono	0.19 ± 0.02	0.91 ± 0.09	0.97 ± 0.00	0.86 ± 0.09	0.41 ± 0.05	1947 ± 84	38 ± 2	sigmoid, 1.0
Radje.	0.13 ± 0.03	0.77 ± 0.11	0.96 ± 0.01	0.72 ± 0.11	0.31 ± 0.05	1961 ± 96	28 ± 1	sigmoid, 1.0
ACEIn.	0.14 ± 0.02	0.84 ± 0.06	0.91 ± 0.02	0.74 ± 0.05	0.32 ± 0.03	898 ± 58	15 ± 1	sigmoid, 1.0, .001
ADHD	0.13 ± 0.02	0.95 ± 0.05	0.85 ± 0.02	0.78 ± 0.05	0.32 ± 0.03	316 ± 33	10 ± 0	rbf, 1.0, .001
Antihistamines	0.06 ± 0.02	0.59 ± 0.37	0.46 ± 0.35	0.04 ± 0.07	0.02 ± 0.04	139 ± 16	8 ± 0	rbf, 10, .001
Atypical.	0.22 ± 0.02	0.81 ± 0.04	0.59 ± 0.06	0.32 ± 0.04	0.25 ± 0.03	465 ± 16	39 ± 2	rbf, 1.0, auto
BetaBl.	0.05 ± 0.02	0.91 ± 0.10	0.63 ± 0.17	0.52 ± 0.11	0.17 ± 0.04	1009 ± 15	14 ± 2	sigmoid, 1.0, .001
Calcium.	0.23 ± 0.03	0.77 ± 0.07	0.76 ± 0.04	0.49 ± 0.06	0.33 ± 0.04	441 ± 30	29 ± 2	rbf, 1.0, auto
Estrogens	0.36 ± 0.03	0.97 ± 0.03	0.61 ± 0.05	0.38 ± 0.04	0.41 ± 0.04	141 ± 2	28 ± 2	rbf, 1.0, auto
NSAIDs	0.33 ± 0.04	0.94 ± 0.06	0.79 ± 0.03	0.64 ± 0.04	0.48 ± 0.04	165 ± 9	15 ± 2	rbf, 10, .0001
Opioids	0.06 ± 0.05	0.81 ± 0.22	0.55 ± 0.45	0.36 ± 0.32	0.13 ± 0.12	904 ± 136	6 ± 1	sigmoid, 1.0, .001
OralHy.	0.29 ± 0.02	0.97 ± 0.05	0.33 ± 0.07	0.05 ± 0.05	0.09 ± 0.07	184 ± 0	40 ± 4	sigmoid, 1.0, .001
Proton.	0.08 ± 0.02	0.88 ± 0.08	0.61 ± 0.11	0.46 ± 0.07	0.19 ± 0.03	624 ± 30	16 ± 1	rbf, 1.0, .001
Skeletal.	0.00 ± 0.00	0.00 ± 0.00	0.99 ± 0.00	0.00 ± 0.00	−0.00 ± 0.00	735 ± 103	4 ± 0	rbf, 1.0, .001
Statins	0.06 ± 0.01	0.87 ± 0.08	0.67 ± 0.10	0.52 ± 0.04	0.18 ± 0.02	1584 ± 114	27 ± 2	sigmoid, 1.0, .001
Triptans	0.11 ± 0.04	0.81 ± 0.14	0.68 ± 0.21	0.47 ± 0.11	0.22 ± 0.06	307 ± 27	10 ± 1	rbf, 1.0, .001
UrinaryIn.	0.25 ± 0.01	0.88 ± 0.12	0.55 ± 0.25	0.35 ± 0.18	0.28 ± 0.15	143 ± 2	17 ± 1	sigmoid, 1.0, auto

[3] Parameter — kernel, C, gamma

(b) TiAbs(MeSH)Ref data

Reviews	Mean Performance					Support vectors		configuration
	precision	recall	accuracy	WSS	MCC	neg	pos	parameters[4]
Kitch.	0.03 ± 0.01	0.94 ± 0.0	0.24 ± 0.21	0.16 ± 0.16	0.06 ± 0.05	1327 ± 1	28 ± 1	rbf, 1.0
Hall	0.36 ± 0.04	0.93 ± 0.0	0.98 ± 0.00	0.90 ± 0.08	0.24 ± 0.57	2002 ± 244	37 ± 2	sigmoid, 1.0
Wahono	0.13 ± 0.04	0.94 ± 0.00	0.94 ± 0.02	0.87 ± 0.06	0.33 ± 0.06	2408 ± 491	32 ± 3	sigmoid, 1.0
Radje.	0.09 ± 0.01	0.85 ± 0.01	0.93 ± 0.01	0.78 ± 0.11	0.26 ± 0.03	1642 ± 237	15 ± 1	sigmoid, 1.0
ACEIn.	0.06 ± 0.02	0.84 ± 0.11	0.70 ± 0.24	0.53 ± 0.19	0.17 ± 0.06	1250 ± 3	13 ± 2	sigmoid, 1.0, .001
ADHD	0.12 ± 0.03	0.91 ± 0.08	0.82 ± 0.06	0.71 ± 0.04	0.28 ± 0.04	348 ± 50	8 ± 1	rbf, 1.0, .001
Antihi.	0.06 ± 0.02	0.65 ± 0.33	0.42 ± 0.34	0.05 ± 0.1	0.03 ± 0.06	144 ± 10	8 ± 0	rbf, 10, .001
Atypical.	0.15 ± 0.03	0.95 ± 0.07	0.27 ± 0.17	0.10 ± 0.13	0.09 ± 0.1	487 ± 1	42 ± 3	rbf, 1.0, auto
BetaBl.	0.04 ± 0.02	0.90 ± 0.14	0.46 ± 0.27	0.34 ± 0.19	0.11 ± 0.05	1015 ± 0	17 ± 2	sigmoid, 1.0, .001
Calcium.	0.12 ± 0.03	0.92 ± 0.08	0.41 ± 0.17	0.26 ± 0.11	0.17 ± 0.06	556 ± 7	37 ± 2	rbf, 1.0, auto
Estrogens	0.23 ± 0.01	0.99 ± 0.01	0.29 ± 0.05	0.07 ± 0.05	0.13 ± 0.05	143 ± 2	31 ± 2	rbf, 1.0, auto
NSAIDs	0.35 ± 0.02	0.96 ± 0.04	0.81 ± 0.02	0.67 ± 0.04	0.50 ± 0.03	160 ± 4	18 ± 0	rbf, 10, .0001
Opioids	0.06 ± 0.08	0.83 ± 0.23	0.38 ± 0.46	0.21 ± 0.27	0.1 ± 0.14	909 ± 124	6 ± 1	sigmoid, 1.0, .001
OralHy.	0.28 ± 0.02	0.97 ± 0.05	0.32 ± 0.03	0.05 ± 0.05	0.06 ± 0.07	181 ± 1	48 ± 4	sigmoid, 1.0, .001
Proton.	0.08 ± 0.03	0.90 ± 0.09	0.48 ± 0.26	0.35 ± 0.19	0.15 ± 0.08	633 ± 16	16 ± 2	rbf, 1.0, .001
Skeletal.	0.00 ± 0.00	0.1 ± 0.03	0.89 ± 0.3	0.00 ± 0.00	−0.00 ± 0.00	583 ± 191	4 ± 0	rbf, 1.0, .001
Statins	0.06 ± 0.01	0.87 ± 0.08	0.66 ± 0.09	0.52 ± 0.06	0.17 ± 0.02	1602 ± 108	28 ± 3	sigmoid, 1.0, .001
Triptans	0.09 ± 0.05	0.86 ± 0.15	0.47 ± 0.36	0.30 ± 0.25	0.14 ± 0.12	318 ± 18	9 ± 1	rbf, 1.0, .001
UrinaryIn	0.21 ± 0.05	0.90 ± 0.007	0.55 ± 0.15	0.35 ± 0.12	0.27 ± 0.08	143 ± 2	15 ± 1	sigmoid, 1.0, auto

[4] Parameter — kernel, C, gamma

Tables 7a and 7b show the results of the SVM linear models for the TiAbs(MeSH) and TiAbs(MeSH)Ref Word2vec features respectively. The tables show that the TiAbs(MeSH)Ref data has higher recall in nine reviews, lower in seven reviews and equal to the TiAbs(MeSH) data in three reviews.

Table 7: Word2vec feature with linear SVM kernel

(a) TiAbs(MeSH) data

Reviews	Mean Performance					Support vectors		configuration
	precision	recall	accuracy	WSS	MCC	neg	pos	parameters[5]
kitch.	0.06 ± 0.01	0.91 ± 0.08	0.59 ± 0.04	0.48 ± 0.08	0.16 ± 0.02	957.0 ± 79.0	11.0 ± 1.0	100
Hall	0.11 ± 0.01	0.97 ± 0.04	0.91 ± 0.01	0.86 ± 0.03	0.31 ± 0.02	1732.0 ± 242.0	12.0 ± 1.0	1
Wahono	0.07 ± 0.01	0.96 ± 0.05	0.89 ± 0.01	0.84 ± 0.05	0.25 ± 0.02	1533.0 ± 119.0	9.0 ± 1.0	1
Radje.	0.05 ± 0.01	0.92 ± 0.1	0.87 ± 0.02	0.78 ± 0.08	0.2 ± 0.02	1442.0 ± 185.0	10.0 ± 1.0	1
ACEIn.	0.08 ± 0.02	0.96 ± 0.04	0.8 ± 0.05	0.74 ± 0.04	0.24 ± 0.03	590.0 ± 101.0	7.0 ± 1.0	1
ADHD	0.08 ± 0.0	0.96 ± 0.08	0.75 ± 0.02	0.68 ± 0.06	0.24 ± 0.02	252.0 ± 33.0	4.0 ± 1.0	1
Antihi.	0.06 ± 0.0	0.9 ± 0.11	0.21 ± 0.11	0.07 ± 0.04	0.04 ± 0.03	140.0 ± 7.0	5.0 ± 1.0	40
Atypical.	0.18 ± 0.01	0.9 ± 0.04	0.45 ± 0.05	0.24 ± 0.03	0.2 ± 0.02	417.0 ± 13.0	28.0 ± 1.0	1000
BetaBl.	0.05 ± 0.0	0.91 ± 0.06	0.64 ± 0.04	0.53 ± 0.03	0.16 ± 0.01	683.0 ± 58.0	8.0 ± 1.0	1
Calcium.	0.13 ± 0.01	0.92 ± 0.04	0.47 ± 0.04	0.32 ± 0.03	0.19 ± 0.02	461.0 ± 26.0	20.0 ± 1.0	100
Estrogens	0.3 ± 0.02	0.93 ± 0.04	0.52 ± 0.05	0.26 ± 0.04	0.3 ± 0.03	125.0 ± 8.0	12.0 ± 1.0	1000
NSAIDS	0.15 ± 0.01	1.0 ± 0.0	0.39 ± 0.03	0.28 ± 0.03	0.21 ± 0.02	158.0 ± 2.0	6.0 ± 0.0	1
Opiods	0.03 ± 0.01	0.8 ± 0.12	0.78 ± 0.06	0.57 ± 0.1	0.13 ± 0.03	469.0 ± 65.0	4.0 ± 1.0	1
OralHy.	0.28 ± 0.01	0.99 ± 0.01	0.3 ± 0.02	0.02 ± 0.02	0.07 ± 0.04	183.0 ± 1.0	34.0 ± 3.0	10000
Proton.	0.06 ± 0.01	0.94 ± 0.05	0.44 ± 0.09	0.35 ± 0.06	0.15 ± 0.02	545.0 ± 56.0	9.0 ± 1.0	1
Skeletal.	0.01 ± 0.0	0.64 ± 0.28	0.55 ± 0.14	0.2 ± 0.21	0.03 ± 0.03	581.0 ± 136.0	4.0 ± 1.0	1
Statins	0.05 ± 0.01	0.93 ± 0.03	0.56 ± 0.05	0.46 ± 0.03	0.15 ± 0.01	1252.0 ± 73.0	15.0 ± 1.0	1
Triptans	0.06 ± 0.01	0.94 ± 0.12	0.45 ± 0.11	0.36 ± 0.07	0.15 ± 0.02	283.0 ± 27.0	4.0 ± 1.0	1
UrinaryIn.	0.2 ± 0.03	0.94 ± 0.07	0.5 ± 0.11	0.33 ± 0.08	0.26 ± 0.05	128.0 ± 16.0	5.0 ± 1.0	100

[5] Parameter — C

(b) TiAbs(MeSH)Ref data

Reviews	Mean Performance					Support vectors		configuration
	precision	recall	accuracy	WSS	MCC	neg	pos	parameters[6]
Kitch.	0.04 ± 0.01	0.88 ± 0.12	0.44 ± 0.13	0.3 ± 0.06	0.1 ± 0.02	1204.0 ± 112.0	14.0 ± 2.0	1
Hall	0.16 ± 0.01	0.97 ± 0.03	0.94 ± 0.01	0.9 ± 0.03	0.38 ± 0.02	1492.0 ± 250.0	12.0 ± 1.0	1
Wahono	0.11 ± 0.01	0.98 ± 0.03	0.93 ± 0.01	0.9 ± 0.02	0.31 ± 0.01	1255.0 ± 150.0	9.0 ± 1.0	1
Radje.	0.07 ± 0.01	0.96 ± 0.07	0.9 ± 0.01	0.85 ± 0.05	0.25 ± 0.01	1197.0 ± 123.0	8.0 ± 2.0	1
ACEIn.	0.08 ± 0.02	0.91 ± 0.07	0.8 ± 0.06	0.7 ± 0.07	0.23 ± 0.04	607.0 ± 123.0	7.0 ± 1.0	1
ADHD	0.09 ± 0.01	0.97 ± 0.05	0.75 ± 0.03	0.7 ± 0.03	0.25 ± 0.01	236.0 ± 33.0	4.0 ± 1.0	1
Antihi.	0.05 ± 0.0	0.84 ± 0.19	0.2 ± 0.19	0.0 ± 0.05	0.0 ± 0.04	136.0 ± 22.0	5.0 ± 1.0	10
Atypical.	0.2 ± 0.01	0.83 ± 0.05	0.53 ± 0.03	0.28 ± 0.02	0.22 ± 0.02	359.0 ± 19.0	33.0 ± 3.0	1000
BetaBl.	0.04 ± 0.01	0.86 ± 0.05	0.58 ± 0.07	0.43 ± 0.06	0.13 ± 0.02	793.0 ± 74.0	10.0 ± 1.0	1
Calcium.	0.13 ± 0.01	0.93 ± 0.03	0.48 ± 0.04	0.34 ± 0.04	0.21 ± 0.03	484.0 ± 19.0	21.0 ± 2.0	10
Estrogens	0.36 ± 0.02	0.94 ± 0.04	0.62 ± 0.04	0.38 ± 0.03	0.4 ± 0.03	104.0 ± 7.0	12.0 ± 1.0	100
NSAIDS	0.14 ± 0.01	1.0 ± 0.01	0.37 ± 0.06	0.26 ± 0.05	0.2 ± 0.03	168.0 ± 5.0	7.0 ± 1.0	1
Opiods	0.04 ± 0.0	0.82 ± 0.17	0.82 ± 0.03	0.64 ± 0.14	0.15 ± 0.02	424.0 ± 35.0	5.0 ± 1.0	1
OralHypoglycemics	0.33 ± 0.02	0.91 ± 0.03	0.47 ± 0.04	0.16 ± 0.03	0.23 ± 0.04	165.0 ± 8.0	30.0 ± 2.0	10000
Proton.	0.05 ± 0.0	0.92 ± 0.08	0.37 ± 0.08	0.27 ± 0.03	0.11 ± 0.01	576.0 ± 44.0	11.0 ± 2.0	1
Skeletal.	0.0 ± 0.0	0.26 ± 0.23	0.69 ± 0.11	−0.06 ± 0.17	−0.01 ± 0.03	530.0 ± 111.0	4.0 ± 0.0	1
Statins	0.04 ± 0.0	0.94 ± 0.04	0.5 ± 0.04	0.42 ± 0.03	0.13 ± 0.01	1365.0 ± 65.0	16.0 ± 2.0	1
Triptans	0.06 ± 0.01	0.94 ± 0.08	0.47 ± 0.12	0.38 ± 0.07	0.15 ± 0.02	269.0 ± 37.0	6.0 ± 1.0	1
UrinaryIncontinence	0.17 ± 0.01	0.95 ± 0.09	0.44 ± 0.06	0.28 ± 0.05	0.23 ± 0.03	122.0 ± 13.0	7.0 ± 1.0	10

[6] Parameter — C

With the Word2vec feature representation and SVM non-linear kernels, the TiAbs(MeSH)Ref data show higher recall in six reviews (Table 8b), lower recall values in nine reviews and equal recall values in four reviews compared to the TiAbs(MeSH) data (Table 8a).

Considering the MCC, which is a measure that takes all the four basic model performance measures (TN, FN, TP and FP) into account, the TiAbs(MeSH)Ref data records higher values in 11 of the 19 reviews compared to the TiAbs(MeSH) data with the non-linear kernel of the SVM and Word2vec feature (Tables 8b and 8a). For the binary feature the TiAbs(MeSH) feature (Table 6a) has better MCC values than the TiAbs(MeSH)Ref feature (Table 6b) in all the reviews. With the linear SVM kernel and Word2vec feature however, the TiAbs(MeSH)Ref data (Table 7b) show higher MCC values than the TiAbs(MeSH)

Table 8: Word2vec feature non-linear kernel

(a) TiAbs(MeSH) Features

Reviews	Mean Performance					Support vectors		configuration
	precision	recall	accuracy	WSS	MCC	neg	pos	parameters[7]
Kitch.	0.04 ± 0.01	0.98 ± 0.04	0.36 ± 0.13	0.32 ± 0.11	0.11 ± 0.03	1299.0 ± 28.0	11.0 ± 1.0	rbf, 1000, 0.001
Hall	0.11 ± 0.01	0.97 ± 0.04	0.91 ± 0.01	0.86 ± 0.03	0.31 ± 0.02	1732.0 ± 242.0	12.0 ± 1.0	sigmoid, 1000, 0.001
Wahono	0.07 ± 0.01	0.96 ± 0.05	0.89 ± 0.01	0.84 ± 0.05	0.25 ± 0.02	1533.0 ± 119.0	9.0 ± 1.0	sigmoid, 1000, 0.001
Radje.	0.03 ± 0.0	0.96 ± 0.07	0.76 ± 0.02	0.72 ± 0.06	0.15 ± 0.01	2560.0 ± 170.0	9.0 ± 1.0	sigmoid, 100, 0.001
ACEIn.	0.09 ± 0.02	0.92 ± 0.06	0.83 ± 0.05	0.74 ± 0.04	0.25 ± 0.03	486.0 ± 103.0	7.0 ± 1.0	rbf, 1000, 0.001
ADHD	0.09 ± 0.01	0.95 ± 0.08	0.76 ± 0.02	0.69 ± 0.06	0.24 ± 0.02	210.0 ± 28.0	4.0 ± 1.0	rbf, 1000, 0.001
Antihi.	0.06 ± 0.0	0.92 ± 0.1	0.18 ± 0.11	0.06 ± 0.04	0.05 ± 0.02	141.0 ± 7.0	4.0 ± 1.0	sigmoid, 1000, auto
Atypical.	0.15 ± 0.01	0.96 ± 0.03	0.29 ± 0.06	0.14 ± 0.04	0.14 ± 0.02	466.0 ± 12.0	24.0 ± 2.0	sigmoid, 10000, 0.001
BetaBl.	0.07 ± 0.01	0.82 ± 0.05	0.76 ± 0.05	0.57 ± 0.07	0.19 ± 0.03	469.0 ± 54.0	9.0 ± 1.0	sigmoid, 1000, auto
Calcium.	0.12 ± 0.01	0.93 ± 0.03	0.45 ± 0.04	0.31 ± 0.03	0.19 ± 0.02	472.0 ± 26.0	20.0 ± 1.0	sigmoid, 10000, auto
Estrogens	0.24 ± 0.02	0.98 ± 0.03	0.32 ± 0.08	0.08 ± 0.06	0.13 ± 0.08	141.0 ± 4.0	12.0 ± 1.0	sigmoid, 10000, auto
NSAIDS	0.17 ± 0.01	1.0 ± 0.01	0.51 ± 0.03	0.4 ± 0.03	0.28 ± 0.02	145.0 ± 4.0	5.0 ± 1.0	sigmoid, 1000, auto
Opiods	0.02 ± 0.0	0.98 ± 0.05	0.6 ± 0.06	0.57 ± 0.06	0.1 ± 0.01	736.0 ± 41.0	4.0 ± 1.0	sigmoid, 10, auto
OralHy.	0.27 ± 0.0	1.0 ± 0.0	0.27 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	184.0 ± 0.0	45.0 ± 5.0	sigmoid, 1000, auto
Proton.	0.04 ± 0.0	0.98 ± 0.04	0.13 ± 0.09	0.07 ± 0.06	0.04 ± 0.03	639.0 ± 4.0	8.0 ± 1.0	sigmoid, 100, 0.001
Skeletal.	0.01 ± 0.0	0.9 ± 0.2	0.31 ± 0.2	0.21 ± 0.08	0.04 ± 0.01	740.0 ± 103.0	4.0 ± 1.0	rbf, 100, 0.001
Statins	0.03 ± 0.0	0.98 ± 0.02	0.26 ± 0.06	0.22 ± 0.06	0.08 ± 0.02	1647.0 ± 27.0	14.0 ± 1.0	sigmoid, 100, 0.001
Triptans	0.06 ± 0.01	0.94 ± 0.12	0.42 ± 0.12	0.33 ± 0.07	0.14 ± 0.02	288.0 ± 25.0	5.0 ± 0.0	sigmoid, 100, auto
UrinaryIn.	0.17 ± 0.04	0.93 ± 0.07	0.37 ± 0.2	0.19 ± 0.16	0.15 ± 0.12	131.0 ± 15.0	6.0 ± 1.0	rbf, 10000, auto

[7] Parameter — kernel, C, gamma

(b) TiAbs(MeSH)Ref data

Reviews	Mean Performance					Support vectors		configuration
	precision	recall	accuracy	WSS	MCC	neg	pos	parameters[8]
Kitch.	0.03 ± 0.01	0.93 ± 0.09	0.21 ± 0.22	0.12 ± 0.15	0.04 ± 0.05	1308.0 ± 35.0	13.0 ± 2.0	rbf, 100, 0.001
Hall	0.16 ± 0.01	0.97 ± 0.03	0.94 ± 0.01	0.9 ± 0.03	0.38 ± 0.02	1492.0 ± 250.0	12.0 ± 1.0	sigmoid, 1000, 0.001
Wahono	0.11 ± 0.01	0.98 ± 0.03	0.93 ± 0.01	0.9 ± 0.02	0.31 ± 0.01	1255.0 ± 150.0	9.0 ± 1.0	sigmoid, 1000, 0.001
Radje.	0.09 ± 0.01	0.96 ± 0.07	0.92 ± 0.01	0.87 ± 0.05	0.27 ± 0.01	942.0 ± 103.0	10.0 ± 2.0	rbf, 1000, 0.001
ACEIn.	0.08 ± 0.02	0.91 ± 0.07	0.8 ± 0.06	0.7 ± 0.07	0.23 ± 0.04	607.0 ± 123.0	7.0 ± 1.0	sigmoid, 1000, 0.001
ADHD	0.09 ± 0.01	0.97 ± 0.05	0.75 ± 0.03	0.7 ± 0.03	0.25 ± 0.01	236.0 ± 33.0	4.0 ± 1.0	sigmoid, 1000, 0.001
Antihi.	0.05 ± 0.01	0.85 ± 0.19	0.2 ± 0.19	0.01 ± 0.07	0.01 ± 0.05	136.0 ± 22.0	5.0 ± 1.0	sigmoid, 1000, auto
Atypical.	0.16 ± 0.01	0.95 ± 0.04	0.34 ± 0.07	0.18 ± 0.05	0.17 ± 0.03	459.0 ± 18.0	26.0 ± 2.0	sigmoid, 10000, 0.001
BetaBl.	0.05 ± 0.01	0.83 ± 0.05	0.65 ± 0.07	0.47 ± 0.05	0.14 ± 0.02	708.0 ± 83.0	10.0 ± 1.0	sigmoid, 1000, auto
Calcium.	0.13 ± 0.01	0.93 ± 0.03	0.48 ± 0.04	0.34 ± 0.04	0.21 ± 0.03	484.0 ± 19.0	21.0 ± 2.0	sigmoid, 10000, 0.001
Estrogens	0.29 ± 0.02	0.96 ± 0.02	0.48 ± 0.04	0.25 ± 0.04	0.29 ± 0.03	132.0 ± 3.0	13.0 ± 1.0	sigmoid, 10000, 0.001
NSAIDS	0.17 ± 0.01	0.99 ± 0.02	0.48 ± 0.04	0.37 ± 0.03	0.26 ± 0.02	153.0 ± 9.0	7.0 ± 1.0	rbf, 1000, 0.001
Opiods	0.02 ± 0.0	0.99 ± 0.04	0.63 ± 0.03	0.61 ± 0.04	0.11 ± 0.01	726.0 ± 42.0	5.0 ± 1.0	rbf, 100, 0.001
OralHy.	0.28 ± 0.01	0.95 ± 0.03	0.33 ± 0.04	0.04 ± 0.03	0.06 ± 0.08	179.0 ± 3.0	32.0 ± 2.0	rbf, 1000, 0.1
Proton.	0.05 ± 0.0	0.92 ± 0.08	0.37 ± 0.08	0.27 ± 0.03	0.11 ± 0.01	576.0 ± 44.0	11.0 ± 2.0	sigmoid, 100, 0.01
Skeletal.	0.01 ± 0.0	0.86 ± 0.2	0.16 ± 0.16	0.01 ± 0.06	0.01 ± 0.01	806.0 ± 28.0	4.0 ± 0.0	sigmoid, 100, 0.001
Statins	0.03 ± 0.0	0.98 ± 0.03	0.22 ± 0.07	0.18 ± 0.05	0.07 ± 0.01	1655.0 ± 26.0	15.0 ± 1.0	rbf, 100, 0.001
Triptans	0.04 ± 0.0	1.0 ± 0.0	0.07 ± 0.05	0.03 ± 0.04	0.02 ± 0.03	324.0 ± 0.0	5.0 ± 1.0	sigmoid, 100, 0.001
UrinaryIn.	0.17 ± 0.01	0.95 ± 0.09	0.44 ± 0.06	0.28 ± 0.05	0.23 ± 0.03	122.0 ± 13.0	7.0 ± 1.0	sigmoid, 10000, 0.001

[8] Parameter — kernel, C, gamma

data (Table 7b) in nine reviews and equal values in one review.

The TiAbs(MeSH)Ref data appeared to be saving more work over random sampling in 15 out of the 19 reviews (see WSS in Table 8b and 8a). Given the binary feature and SVM non-linear model the TiAbs(MeSH)Ref (Table 6a) has higher WSS value in four reviews and equal values in four. In Word2vec based linear kernel SVM models the TiAbs(MeSH)Ref data (Table 7b) has higher MCC values in 10 of the reviews than the TiAbs(MeSH) data(Table 7a).

5. Discussion

The DERP datasets where the average reference retrieval rate is relatively lower (generally below 50%) did not present any particular pattern to establish the effect of the additional reference features in the performance of the models. If however, the SE datasets where the average reference retrieval rate are approximately 95% is considered in isolation, it can be seen from Tables 6a and 6b that there is an improvement in the recall with the TiAbs(MeSH)Ref data in three of the four datasets, where the recall performances were equal. In Table 7, the TiAbs(MeSH)Ref data showed equal or higher recall in three reviews. This pattern was repeated with the non-linear kernel in Table 8. In Table 6, the TiAbs(MeSH)Ref data (Table 6b) showed higher WSS performance in two reviews, equal performance in one and lower in one. However, in the Word2vec representation, the TiAbs(MeSH)Ref data (Tables 7b and 8b) recorded higher WSS values in three of the four reviews. On closer inspection, the dataset where the WSS values were less in these two cases is the Kitchenham review which is the smallest among the datasets. This pattern was also noticed with the MCC where the TiAbs(MeSH)Ref data had higher MCC values in three of the four SE datasets except the Kitchenham (see Table 7 and Table 8). The MCC performance of the TiAbs(MeSH)Ref data was however the other way around for the binary representation where the TiAbs(MeSH) data had higher values in all four datasets.

The TiAbs(MeSH)Ref data used fewer support vectors across the SE datasets except the Kitchenham as shown in Table 6b, Table 7b and Table 8b than the TiAbs(MeSH) data. There is variation across the rest of the dataset on the number of support vectors used for both sets of the data. It was also observed that the three relatively large SE datasets used on average, only about 30% of their training data as support vectors against an average of about 90% in other datasets. This shows that the models from these datasets are likely to be more robust and less complex than those from the smaller datasets. It further emphasizes the importance of data volume in the learning of the algorithms during the training phase. In SVM, the smaller the ratio of the support vectors used, the better the model learns from the data pattern and thus, the better it can generalize over other examples.

It is not clear yet whether adding the reference information to the datasets increases the performance of a text mining model for automatic CS. More work needs to be put into investigating the factors that contributed to the improved performance in some cases and not in others. Nevertheless, this study has shown that the chances of sustaining or recording an improvement in a model's performance by adding the bibliography information is higher than the chances of recording a lower one.

We noted in the retrieved bibliography data that (usually) only one of the authors' names is fully spelled out. This may result in loss of information that may be vital to the establishment of an association between articles that might have cited similar authors due to a common subject since the initials are removed during preprocessing leaving only one name each from the authors. Access to the full author names in the databases could have aided more, the discrimination of the documents. The same situation affects abbreviation of journal names which could have contributed to linking articles with similar journal names.

If the results of the 19 review datasets are considered as a whole, no consistent trend could be established at this stage on the effect of adding the bibliography data to the input data. This may be due to a number of reasons part of which may be the low reference retrieval rate in the DERP datasets, the relatively small size of the datasets or the severity of the class imbalance. The results of the SE datasets indicated that larger data size with high number of reference inclusion could actually lead to an improvement in the performance of models.

The results of this study however was compared to part of the results presented by Khabsa et al. [13] where they build SVM models with unigram features using the article title, abstract and citations. The Recall and WSS results from the enriched features of this study (Binary-NL, W2vec-NL and W2vec-L) is presented in Table 9 alongside similar results from [13] tagged 'Khabsa-uniCite'.

Table 9: Recall and WSS comparison with existing study

Reviews	Binary-NL [9]		W2vec-NL		W2vec-L		Khabsa-UniCite	
	Recall	WSS	Recall	WSS	Recall	WSS	Recall	WSS
ACEIn.	0.84	0.53	0.91	0.7	0.91	**0.7**	0.986	0.469
ADHD	0.91	0.71	0.97	0.7	**0.97**	**0.7**	0.98	0.447
Antihi.	0.65	0.05	0.84	0	**0.85**	0.01	0.75	0.03
Atypical.	0.95	0.1	0.83	0.28	**0.95**	**0.18**	0.96	0.199
BetaBl.	0.90	0.34	0.86	0.43	0.83	0.47	0.947	0.361
Calcium.	0.92	0.26	0.93	0.34	0.93	0.24	0.98	0.287
Estrogens	**0.99**	0.07	0.94	**0.38**	0.96	0.25	0.983	0.18
NSAIDS	0.96	**0.67**	1.0	0.26	**0.99**	0.37	0.995	0.404
Opiods	0.83	0.21	0.82	0.64	**0.99**	**0.61**	0.933	0.455
OralHy.	**0.97**	**0.05**	0.91	0.16	0.95	0.04	0.971	0.074
Proton.	0.90	0.35	0.92	0.27	0.92	0.27	0.976	0.288
Skeletal.	-	-	0.26	0	**0.86**	0.01	0.822	0.371
Statins.	0.87	0.52	0.94	**0.42**	**0.98**	0.18	0.941	0.4
Triptans	0.86	0.30	0.94	**0.38**	**1.0**	0.03	0.958	0.312
UrinaryIn.	0.90	0.35	**0.95**	0.28	**0.95**	0.28	0.94	0.411

[9] NL - Non-Linear; L - Linear; W2vec - Word2vec

The points where this study exhibited similar or improved performance compared to the Khabsa et al.'s study table shows that results from this study are shown in bold fonts. This study exhibited similar or improved recall performance in 11 of the 15 reviews. Nine of these 11 cases is from linear kernel models with Word2vec (W2vec-L) features with six cases presenting improved performance.

6. Validity threats

Though the datasets we used in this study cut across two fields - SE and healthcare, there is still not enough evidence to generalize the findings. We have only used four reviews from SE and three of them address similar topics while the medical review datasets are relatively small in size. There is an indication that including the bibliography data may improve model performance but the retrieval rate for the medical review datasets is too low to explain or establish the noted difference. The performances observed is limited to SVM models and features used, it is not necessarily generalizable.

7. Conclusions

We have studied the effect of including bibliography data with titles and abstracts on the performance of text mining models for automatic CS. We prepared two sets each of the four SE and 15 medical review datasets, one with title, abstract and optional MeSH terms and the other with title, abstract, optional MeSH and bibliography data. We represented each set as binary and average word Word2vec features and developed SVM models from the features.

The TiAbs(MeSH)Ref set shows higher or equal recall, MCC and WSS in the three larger SE datasets with the different feature representations and model kernels. The performance varies when it comes to the smaller medical review and one SE dataset however, there are more instances of higher or equal performances than lower.

Given the pattern established in this study, it is clear that the inclusion of the bibliography information is more likely to improve or sustain the CS model performance than impair it. The study also showed that chances of performance improvement is more certain if the size of the dataset is relatively larger with references included for most of the articles in the dataset.

In the future, we will work to further investigate and characterize possible factors responsible for the variability in the performance of smaller datasets. Also, we want to test the hypothesis on more data with better size and class representations than the DERP dataset since the SE datasets show promising and consistent improvement in results. We will also attempt to change the metric for selecting the model parameters from recall to either MCC and see how this affects the models' performance.

Authors contribution

BKO contributed to the concept, design, data linkage, corpus construction model development, conduct of the study and draft of the initial manuscript. PB and PA validated the design, data linkage, conduct of the study and critically revise the manuscript. All authors have approved the final version of the manuscript. BKO is the corresponding author.

Acknowledgement

Babatunde Olorisade would like to thank the National Information Technology Development Agency, Nigeria (NITDA) for providing funds to sponsor his

PhD. The authors also wish to thank Prof. Barbara Kitchenham for providing us with the dataset we tagged ‘Kitchenham’ and Zhe Yu for making the ‘Hall’, ‘Radjevovic’ and ‘Wahono’ datasets available online.

Conflict of interest

None.

Summary points

What was already known on the topic:

- Text mining based CS for SRs models suffers performance setback due to class imbalance
- Feature enrichment can be explored to improve feature quality and reduce the effect of class imbalance
- MeSH features have been explored and shown to lead to improved model performance

What this study has added:

- Explore mitigating the effect of class imbalance on model performance using full bibliography features
- The bibliography features has the ability to improve the performance quality of text mining models for CS support for SRs
- The performance of models has more chances of being improved or at worst maintained rather than impaired with the addition of bibliography features
- The larger the dataset and number of retrieved references, the higher the chances of improvement in model performance

References

[1] B. A. Kitchenham, T. Dyba, M. Jorgensen, Evidence-based software engineering, in: Proceedings of the 26th international conference on software engineering, IEEE Computer Society, 2004, pp. 273–281.

[2] B. A. Kitchenham, D. Budgen, P. Brereton, Evidence-Based Software engineering and systematic reviews, Vol. 4, CRC Press, 2015.

[3] J. P. Higgins, S. Green, Cochrane handbook for systematic reviews of interventions, Vol. 4, John Wiley & Sons, 2011.

[4] E. Hassler, J. C. Carver, N. A. Kraft, D. Hale, Outcomes of a community workshop to identify and rank barriers to the systematic literature review process, in: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, ACM, 2014, p. 31.

[5] J. C. Carver, E. Hassler, E. Hernandes, N. A. Kraft, Identifying barriers to the systematic literature review process, in: Empirical Software Engineering and Measurement, 2013 ACM/IEEE International Symposium on, IEEE, 2013, pp. 203–212.

[6] K. P. Murphy, Machine learning: a probabilistic perspective, MIT press, 2012.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[8] P. Domingos, Metacost: A general method for making classifiers cost-sensitive, in: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 1999, pp. 155–164.

[9] N. Japkowicz, The class imbalance problem: Significance and strategies, in: Proc. of the Intl Conf. on Artificial Intelligence, 2000.

[10] M. Kubat, S. Matwin, et al., Addressing the curse of imbalanced training sets: one-sided selection, in: ICML, Vol. 97, Nashville, USA, 1997, pp. 179–186.

[11] J. Hu, L. Fang, Y. Cao, H.-J. Zeng, H. Li, Q. Yang, Z. Chen, Enhancing text clustering by leveraging wikipedia semantics, in: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2008, pp. 179–186.

[12] P. Wang, C. Domeniconi, Building semantic kernels for text classification using wikipedia, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2008, pp. 713–721.

[13] M. Khabsa, A. Elmagarmid, I. Ilyas, H. Hammady, M. Ouzzani, Learning to identify relevant studies for systematic reviews using random forest and external information, Machine Learning 102 (3) (2016) 465–482.

[14] B. K. Olorisade, E. de Quincey, P. Brereton, P. Andras, A critical analysis of studies that address the use of text mining for citation screening in systematic reviews, in: Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering, ACM, 2016, p. 14.

[15] S. Choi, B. Ryu, S. Yoo, J. Choi, Combining relevancy and methodological quality into a single ranking for evidence-based medicine, Information Sciences 214 (2012) 76–90.

[16] R. S. Wahono, A systematic literature review of software defect prediction: Research trends, datasets, methods and frameworks, Journal of Software Engineering 1 (1) (2015) 1–16.

[17] T. Hall, S. Beecham, D. Bowes, D. Gray, S. Counsell, A systematic literature review on fault prediction performance in software engineering, IEEE Transactions on Software Engineering 38 (6) (2012) 1276–1304.

[18] D. Radjenović, M. Heričko, R. Torkar, A. Živkovič, Software fault prediction metrics: A systematic literature review, Information and Software Technology 55 (8) (2013) 1397–1418.

[19] B. Kitchenham, R. Pretorius, D. Budgen, O. P. Brereton, M. Turner, M. Niazi, S. Linkman, Systematic literature reviews in software engineering–a tertiary study, Information and Software Technology 52 (8) (2010) 792–805.

[20] B. K. Olorisade, P. Brereton, P. Andras, Reporting statistical validity and model complexity in machine learning based computational studies, in: Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering, ACM, 2017, pp. 128–133.

[21] A. M. Cohen, W. R. Hersh, K. Peterson, P.-Y. Yen, Reducing workload in systematic review preparation using automated citation classification, Journal of the American Medical Informatics Association 13 (2) (2006) 206–219.

[22] T. Bekhuis, D. Demner-Fushman, Towards automating the initial screening phase of a systematic review., in: MedInfo, 2010, pp. 146–150.

[23] A. M. Cohen, Optimizing feature representation for automated systematic review work prioritization, in: AMIA annual symposium proceedings, Vol. 2008, American Medical Informatics Association, 2008, p. 121.

[24] K. R. Felizardo, G. F. Andery, F. V. Paulovich, R. Minghim, J. C. Maldonado, A visual analysis approach to validate the selection review of primary studies in systematic reviews, Information and Software Technology 54 (10) (2012) 1079–1091.

[25] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, J. C. Lai, Class-based n-gram models of natural language, Computational linguistics 18 (4) (1992) 467–479.

[26] S. Kim, J. Choi, Improving the performance of text categorization models used for the selection of high quality articles, Healthcare informatics research 18 (1) (2012) 18–28.

[27] B. W. Matthews, Comparison of the predicted and observed secondary structure of t4 phage lysozyme, Biochimica et Biophysica Acta (BBA)-Protein Structure 405 (2) (1975) 442–451.

[28] A. M. Cohen, An effective general purpose approach for automated biomedical document classification, in: AMIA Annual Symposium Proceedings, Vol. 2006, American Medical Informatics Association, 2006, p. 161.

[29] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.

Appendix A. Reproducibility Information

Software environment details are shown in Table A.10. Other details needed for study reproduction are:

- Initial dataset shuffle: ”seed”: 29
- Fold split seed: {”seed”: 37, 71}.
- SVM parameters:
  - gamma: auto
  - random state: {”seed”: 55}
  - sample_weight: {1:4}
  - class_weight: ”balanced”
- Word2vec model

Table A.10: Software information

S/N	Software packages	Version
1	Python	3.5.2 64bit
2	Ipython	5.1.0
3	Scipy	0.18.1
4	Numpy	1.11.1
5	Sklearn	0.18.1
6	Pandas	0.18.1
7	NLTK	3.2.1
8	Gensim	1.0.1
9	Matplotlib	1.5.3

- num_of_features: as in Table 2.
- min_word_count: 10; context_window: 15