Automated Identification of Biomedical Article Type Using Support Vector Machines

In Cheol Kim*, Daniel X. Le, George R. Thoma
Lister Hill National Center for Biomedical Communications
National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894

ABSTRACT

Authors of short papers such as letters or editorials often express complementary opinions, and sometimes contradictory ones, on related work in previously published articles. The MEDLINE® citations for such short papers are required to list bibliographic data on these "commented on" articles in a "CON" field. The challenge is to automatically identify the CON articles referred to by the author of the short paper (called "Comment-in" or CIN paper). Our approach is to use support vector machines (SVM) to first classify a paper as either a CIN or a regular full-length article (which is exempt from this requirement), and then to extract from the CIN paper the bibliographic data of the CON articles. A solution to the first part of the problem, identifying CIN articles, is addressed here. We implement and compare the performance of two types of SVM, one with a linear kernel function and the other with a radial basis kernel function (RBF). Input feature vectors for the SVMs are created by combining four types of features based on statistics of words in the article title, words that suggest the article type (letter, correspondence, editorial), size of body text, and cue phrases. Experiments conducted on a set of online biomedical articles show that the SVM with a linear kernel function yields a significantly lower false negative error rate than the one with an RBF. Our experiments also show that the SVM with a linear kernel function achieves a significantly higher level of accuracy, and lower false positive and false negative error rates by using input feature vectors created by combining all four types of features rather than any single type.

Keywords: "Comment-on", "Comment-in", Online biomedical documents, Support vector machine.

1. INTRODUCTION

MEDLINE is the premier bibliographic online database of the U.S. National Library of Medicine (NLM) containing more than 18 million citations from over 5,200 selected biomedical journals, and accessed through NLM's PubMed service. Since the biomedical literature is continually and rapidly growing, there is a strong motivation to develop automated systems to minimize human labor to provide bibliographic data in a timely fashion. The Lister Hill National Center for Biomedical Communications (LHNCBC), a research and development division of NLM has developed an automated system to analyze and extract bibliographic information from online biomedical journal articles to create citations for MEDLINE [1][2].

"Comment-on" (CON) is a field in a MEDLINE citation listing previously published articles commented on by authors of a given paper in a complimentary, or sometimes contradictory manner. We refer to the "commented on" articles as CON articles, and the (usually short) papers in which such opinions are expressed as "Comment-in" (CIN) articles. Generally, CIN articles are short papers such as letters, editorials, or brief correspondence. Full-sized "regular" articles are exempt from this requirement by the indexing conventions at the NLM.

Manually extracting the CON list from a given article is time-consuming and labor-intensive, and relies heavily on human operators' linguistic knowledge and their understanding of scientific expressions and writing styles. In order to minimize such manual effort and to improve accuracy and processing speed, we have developed an automated method that identifies a CON list in a given (CIN) article by recognizing the sentences citing CON articles (called "CON sentences") in its body text, and by analyzing their bibliographical descriptions in the reference section of the CIN article [3]. Here, CON sentences are recognized based on a set of cue phrases, their positions within the body text, and frequency of occurrence of author names of external sources extracted from the reference section. However, similar sentences are often found in "regular" articles as well, thereby generating many false positive errors.

To avoid such false positive errors, we introduce an automated text categorization method using a support vector

machine (SVM) that classifies HTML-formatted online biomedical articles into two categories: CIN or other regular articles. Our strategy is to filter out regular articles in advance, and then to submit only CIN articles to the next step of identifying CON sentences in the body text. Four types of features are employed to create an input feature vector of SVM: 1) word statistics representing how differently a word is distributed in the title of CIN and other regular articles, 2) a set of cue phrases commonly found in CON sentences, 3) CIN-specific article types such as "commentary" and "letter to the editor" generally found in the header section of these documents, and 4) the size of body text. We also implemented two types of SVMs: one with a linear kernel function and the other with a radial basis function (RBF), and compared their performance in terms of accuracy, and false positive and false negative error rates.

2. RELATED WORK

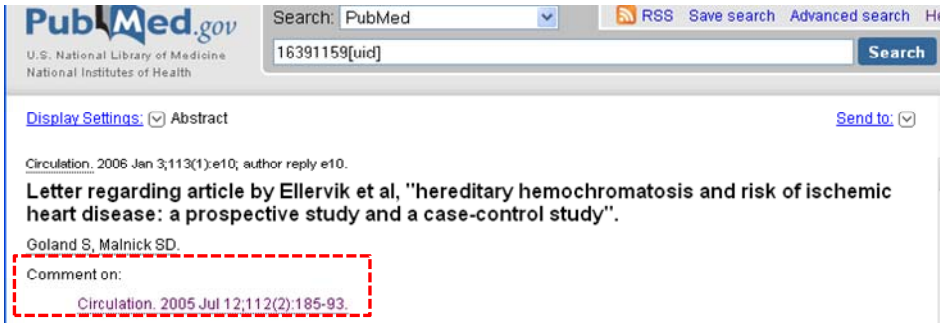
Our task of classifying biomedical articles into one of two pre-defined classes, CIN and "other", is a typical text categorization problem. Text categorization has been addressed by various methods based on statistical theories and machine learning techniques, e.g., Rocchio [4], k-Nearest Neighbor (k-NN) [5][6], Naïve Bayes [7][8], decision tree [9], neural networks [10], and SVMs [11][12]. Among these, the SVM has demonstrated superior classification performance owing to its ability to model text documents which usually have a high dimensional but sparse feature space.

For learning and classification, documents need to be converted to feature vectors. The most widely used is a bag of words [13], a binary vector in which each component is assigned 1 if the corresponding word is found in the document, or 0 otherwise. Thus, the vector size depends on the number of words collected from a training set. Unlike previous research that has mainly focused on the body text of a document to extract a bag of word feature for their task, our method extracts it from the article title. In addition, the size of body text and other clues found from the header section and body text of an HTML document are used as an important feature to separate CIN and full-sized regular articles.

3. AUTOMATED CLASSIFICATION OF "COMMENT-IN" ARTICLES

3.1 Issues on classifying "Comment-in" articles

CIN and CON articles are indicated in MEDLINE citation fields, "Comment in" and "Comment on" respectively, and linked together.



(a)

(b)

Figure 1: (a) "Comment on" and (b) "Comment in" citations in MEDLINE

As an example, Fig. 1(a) is the MEDLINE citation of a CIN article in which a "commented on" article is cited. This CON information, shown enclosed in a dotted box, consists of the journal title, publication year, volume, issue number, and pagination. Conversely, as shown in the dotted box in Fig. 1(b), the MEDLINE citation for this CON article cites the CIN article in which it was mentioned. Thus the reader may get to either citation from the other.

As mentioned earlier, authors of a CIN article cite CON articles as primary external sources on which they express either complimentary or contradictory opinions. The full bibliographical descriptions for these CON articles can usually be found in the reference section of a CIN article. Furthermore, all external sources (journal articles, books, or Web links) listed in the reference section are generally mentioned at least once within sentences ("citation sentences") in the body of the CIN paper. From this observation, a CON list for a given article may be identified by recognizing citation sentences that mention CON articles ("CON sentences") and analyzing the corresponding bibliographic data in the reference section[3]. Figure 2(a) shows an example of a citation sentence citing a CON article (solid underline) and Fig. 2(b) shows the corresponding reference (solid box).

Letter Regarding Article by Ellervik et al, "Hereditary Hemochromatosis and Risk of Ischemic Heart Disease: A Prospective Study and a Case-Control Study"

Sorel Goland, MD; Stephen D. Malnick, MSc, MBBS(Lond)

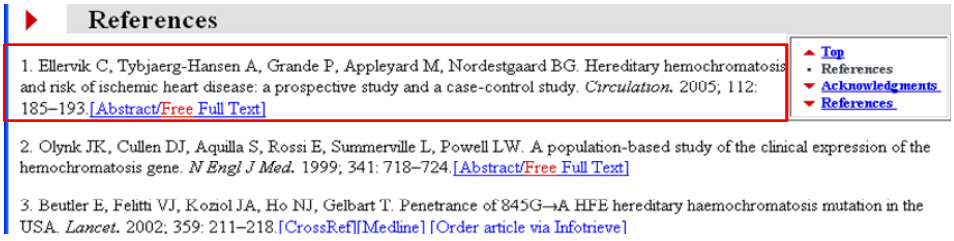
Division of Internal Medicine, Kaplan Medical Center, Rehovot, Israel

To the Editor:

We were interested to read the report of Ellervik et al,[1] which did not find a connection between ischemic heart disease and the presence of the common hemochromatosis mutations. There is, however, no proof of the presence or absence of clinical hemochromatosis in the homozygotic patients.

In a study from Australia of 16 subjects identified from a population screening study who had either elevated transferring saturation or C282Y homozygosity, only half were found to have clinical features of hemochromatosis, and only 3 of 11 had fibrosis on a liver biopsy.[2] Furthermore, in a study from Kaiser Permanente of more than 41 000 patients attending a health

(a)



References

Top
References
Acknowledgments
References

1. Ellervik C, Tybjaerg-Hansen A, Grande P, Appleyard M, Nordestgaard BG. Hereditary hemochromatosis and risk of ischemic heart disease: a prospective study and a case-control study. *Circulation.* 2005; 112: 185–193.[Abstract/Free Full Text]

2. Olynk JK, Cullen DJ, Aquilla S, Rossi E, Summerville L, Powell LW. A population-based study of the clinical expression of the hemochromatosis gene. *N Engl J Med.* 1999; 341: 718–724.[Abstract/Free Full Text]

3. Beutler E, Felitti VJ, Koziol JA, Ho NJ, Gelbart T. Penetrance of 845G→A HFE hereditary haemochromatosis mutation in the USA. *Lancet.* 2002; 359: 211–218.[CrossRef][Medline] [Order article via Infotrieve]

(b)

Figure 2: (a) A citation sentence citing a CON article and (b) its bibliographic description in the reference section

We also observe that CON sentences typically have certain linguistic or contextual clues such as the cue phrases shown in Table 1, and are most likely to be located at the beginning of the body text of a CIN article. In addition, the authors of CON articles are found to be more frequently mentioned in the body text than those of other references. These clues can serve to build a reliable feature to distinguish a CON sentence from other "citation sentences".

Table 1: Examples of cue phrases.

The article (paper, letter, study, research) by …
I (We) read with interest …
In the editorial …
would like to reply (comment) to …
In this issue …
In their recent article (letter, paper, report) **…**

However, citation sentences satisfying these conditions are often also found in full-sized regular articles, from which CON data is not required, thereby generating many false positive errors. These errors can be reduced by eliminating regular articles in advance, and submitting only legitimate CIN articles to the stage of extracting CON sentences. To do this, we propose an SVM-based automated text categorization method that classifies a given article into one of two categories: CIN or "other" regular articles.

3.2 Proposed method

Our method consists of three main steps: 1) extraction of text zones of interest, 2) creation of an input feature vector, and 3) classification of CIN articles by SVMs. Since our method takes advantage of clues from body text, article title, and the header section in a given HTML document, we need to segment the entire article into smaller logical zones, and detect such zones first. We define zones located between an author name/affiliation and the reference section as body text zones (as a result, an abstract zone is also included in the body text), and the zone right above a title as a header section zone. In our research, these text zones of interest are extracted using zoning and labeling modules described in [1] and [14].

3.3 Feature extraction

Input feature vectors for training and testing the SVMs are created by combining four types of features which are experimentally found to be effective to distinguish CIN articles from other articles. These features are extracted from the text zones of interest (header section, article title, and body text).

The first feature is words suggesting the *article type* found in the header section. By analyzing a training data set consisting of several thousand articles, we find that the header section of many CIN articles contains words suggesting a specific article type such as "letter to the editor", "editorial", or "correspondence", whereas most full-sized regular articles do not. Thus these words can serve as a good feature to distinguish CIN articles from others. In our study, we extracted 13 such words shown in Table 2 that are frequently found in header section of CIN articles, and converted them into a 13-bit binary vector of which each component is set to 1 if the corresponding word is found in the header section of an input article, or 0 otherwise.

Table 2: Words suggesting a CIN article

letter	editor	comment	discussion	correspondence
reply	response	reflection	controversy	preview
perspective	mailbox	viewpoint		

Another feature is article size. CIN articles are generally letter-like short papers and thus are smaller than regular articles, though as seen in Table 3 there is an overlap in the range of sizes. Table 3 shows statistics on the size of CIN articles and other articles estimated from our training dataset. Here, the article size is simply the number of characters in a body text, excluding all HTML tags, figures, and tables. To build an input feature vector, the size of each article is first normalized by the sum of mean and standard deviation of the regular article to a real value ranging between 0 and 1. A size larger than the sum of mean and standard deviation of the regular articles is set to 1. The normalized real-value of an article size is then converted to a 10-bit binary vector for SVM (i-th bit position corresponding to real values between $i/10$ and $(i+1)/10$).

Table 3: Statistics of article size

	CIN	Others (regular)
Min.	126	5,332
Max.	78,317	120,645
Mean	5,568	34,489
Std Dev.	4,616	12,220

Next, we adopt a bag of words, a vector of words, as another feature. Titles of CIN articles often have an explicit expression of commenting on other articles (s) or answering/responding to the questions or opinions from other article (s) on authors' previous article, as shown in Fig. 3. Thus, unlike other studies, words are collected not from body text but from an article title. In our research, a total of 39,505 words excluding stop words were collected from the titles of our training article set.

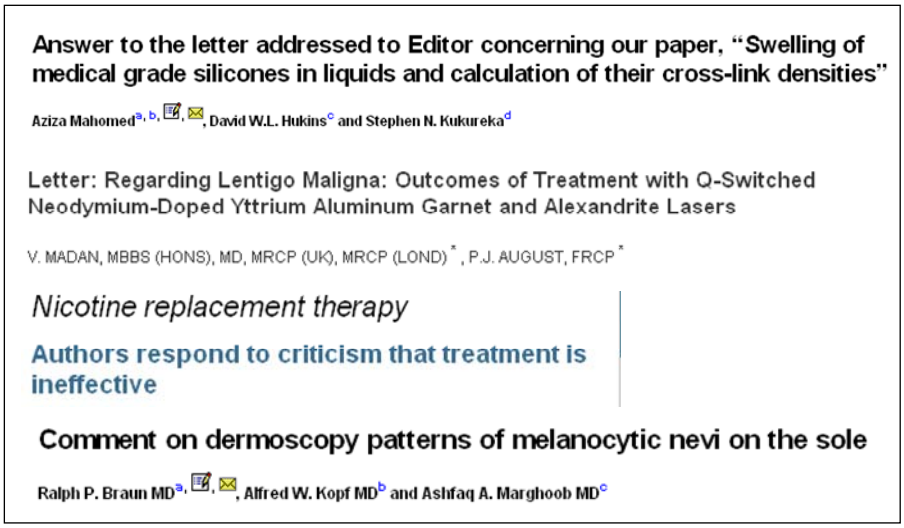


Figure 3: Examples of titles showing explicit expressions of commenting on or answering/responding to another article.

Using words as an input feature requires a very high dimensional feature space (39,505 dimensions in our case). Although SVM can manage (lead to a convergence) such a high dimensional feature space, many have suggested the need for word selection or dimension reduction to employ other conventional learning methods, to reduce the computational cost, to improve the generalization performance, and to avoid the over-fitting problem. A typical approach for word selection is to sort out words according to their importance. Many functions have been proposed to measure the importance of a word, including term frequency (TF), inverse document frequency (IDF), χ^2 statistics, and simplified χ^2 ($s\chi^2$) statistics [15]. The use of $s\chi^2$ has been reported as delivering the best performance since it removes redundancies, and emphasizes extremely rare features (words) and rare categories from χ^2 [16].

In our task, $s\chi^2$ of word t_k for CIN articles (class c_0) and "other" articles (class c_1) can be defined as follows;

$$s\chi^2(t_k, c_i) = P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i) \qquad i = 0, 1 \qquad (1)$$

where $P(t_k, c_i)$ denotes the probability that, for a random article title *x*, word t_k occurs in *x*, *x* belongs to class c_i, and is estimated by counting its occurrences in the training set. The importance of word t_k is finally measured as follows;

$$s\chi^2_{max}(t_k) = \max_i s\chi^2(t_k, c_i) \qquad i = 0, 1 \qquad (2)$$

Accordingly, the more differently a word is distributed in CIN and other classes the higher its $s\chi^2_{max}(t_k)$.

Our 39, 505 words are sorted according to their $s\chi^2_{max}$ and a bag of words feature is created by selecting words scoring highest $s\chi^2_{max}$. A series of experiments to investigate the influence of word reduction and to discover the number of words showing the best classification performance is also performed. These experiments are described in Section 4. A bag of words feature is also converted to a binary vector; each vector component is assigned to 1 if the corresponding word is found in the title of a given article, or 0 otherwise.

The last input feature we employed is based on cue phrases frequently found in CON sentences. Basically, there are no linguistic or contextual differences between CIN articles and regular ones in their body text. As a result, we do not expect to find words distributed very differently in these two article classes. Therefore, instead of extracting a bag of words feature from the body text based on $s\chi^2_{max}$, we collected 8 groups of 180 cue phrases consisting of multiple words from several thousand ground-truth samples of "CON sentences". Examples of these cue phrases can be seen in Table 1. A cue phrase feature is converted to an 8-bit binary vector. Once one or more cue phrases are found, according to the cue phrase group to which they belong, the corresponding bit in the feature vector is set to 1.

Finally, all these feature vectors are concatenated to build an input feature vector for the SVM-based training and categorization tasks.

3.4 CIN categorization using SVM classifiers

SVM [17] was originally introduced as a supervised learning algorithm based on the structural risk minimization principle for solving a two-class problem, though it can be easily extended to handle multi-class problems. Owing to its consistently superior performance compared to other existing methods, SVM has been widely used in many text categorization tasks. Recognizing CIN articles is such a two-class text categorization problem; articles are categorized into two classes, "CIN" and "Others".

The basic idea of using SVM to solve a non-linear pattern recognition problem is to map a non-linear separable input space to a linear separable higher dimensional feature space using a predefined kernel function, and to find the optimal hyperplane that maximizes the margins between the classes in that feature space. We implemented two types of SVMs: one with a linear kernel function and the other with an RBF. These two kernel functions, defined in equations (3) and (4) below, respectively, have been commonly used in SVM-based pattern recognition applications. We evaluate their recognition performance using real online biomedical journal articles.

$$K(x_i, x_j) = (x_i^T \cdot x_j) \tag{3}$$

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2) \tag{4}$$

4. EXPERIMENTS

4.1 Dataset

The dataset for our classification experiments consists of 5,691 HTML-formatted CIN articles and 12,097 regular articles from those indexed in MEDLINE in 2006.

8,000 articles from two classes (4,000 CIN + 4,000 others) are randomly selected to create input feature vectors for SVM learning. The statistics ($s\chi^2_{max}$) of words in the titles are also estimated from this training set. The remaining 9,788 articles (1,691 CIN + 8,097 others) are used as the test set to evaluate SVM performance.

4.2 Experimental results

First, we investigated the influence of word selection, i.e., varying word dictionary size in recognizing CIN articles. As previously mentioned, many pattern recognition studies have suggested the need for word selection to reduce the computational cost and to improve the recognition performance of SVMs, even though SVMs are known to handle high dimensional feature spaces.

Figure 4 shows accuracy, and false positive and false negative error rates as functions of the size of the word dictionary. A false positive error means that a regular article is misclassified as a CIN article. A false negative error is the reverse of the above. When only the bag of words feature ($s\chi^2_{max}$) is used as the input feature vector, the SVM with an RBF kernel function provides higher accuracy and lower false positive error rate than the SVM with a linear kernel function, for large dictionary size (> 200). However, the performance of SVM with an RBF is found to be unreliable with respect to a high dimensional feature vector because its false negative error rate increases unacceptably, as can be seen from Fig. 4(c).

On the other hand, the SVM with a linear kernel function shows reasonably consistent and reliable performance with respect to word selection; its accuracy and false positive error rate do not significantly vary with dictionary size, except

for the smallest size (= 50). Moreover, it shows a significantly lower false negative error rate than SVM with an RBF. In this study, false negative errors are considered much more serious than false positive errors, since the latter may be corrected by an operator at the final verification stage. We therefore conclude that SVM with a linear kernel function is a more appropriate scheme for classifying CIN articles.

Next, we evaluated the performance of the SVM with a linear kernel function for input feature vectors that are created by combining all four features ($s\chi^2_{max}$ + article type + size of body text + cue phrases). The dimension of the combined input feature vector then corresponds to the sum of size of the word dictionary created based on $s\chi^2_{max}$, 13-bit article type, 10-bit body text size, and 8-bit cue phrase. Our experiments show that with the combined feature vector, the SVM with a linear kernel function achieves a significant improvement of performance in terms of accuracy, and false positive and false negative error rates, as shown in Fig. 4.

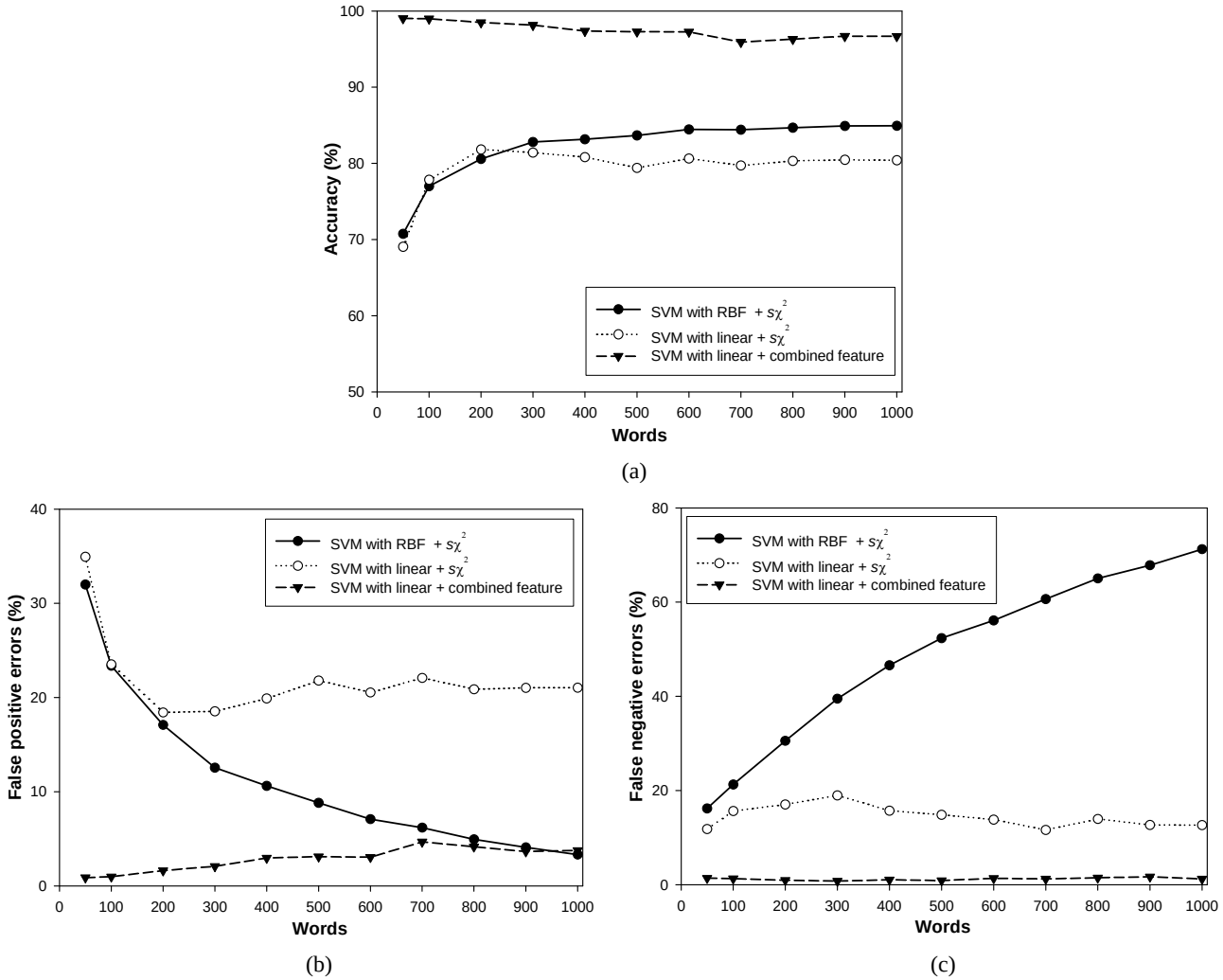


Figure 4: (a) accuracy, (b) false positive and (c) false negative error rates plotted against different word dictionary sizes

Classification errors do occur, however. Table 4 shows examples of these from the SVM with a linear kernel function when combined input feature vectors are used. The false negative error shown in Table 4(a) is caused by the size of the given article being close to the average size of the regular articles, and because neither a CIN specific article type nor cue phrase is found. The false negative error shown in Table 4(b) is due to the lack of three of the four features; 1) words differently distributed in the title of CIN and other articles are not found ("brain" and "heart" are commonly found in

biomedical literature), 2) the body text is bigger in size than the average of CIN articles, and 3) CIN-specific article type is not found in the header section. The false positive error shown in Table 4(c) results from the relatively small body size and several cue phrases strongly suggesting the existence of a CON citation in the body text of the given regular article.

Table 4: Error examples (a) and (b) show false negative errors, and (c) a false positive error

Article title:	Right of the living dead? Consent to experimental surgery in the event of cortical death
Body size:	**32417**
Article type:	**None**
Cue phrase:	**None**

(a)

Article title:	The **brain** and the **heart**
Body size:	**14699**
Article type:	**None**
Cue phrase:	in this issue

(b)

Article title:	RNA trafficking and local protein synthesis in dendrites: an overview
Body size:	**16581**
Article type:	**None**
Cue phrase:	**in this issue** | **the accompanying article by**

(c)

5. CONCLUSIONS

CON ("Comment-on") is a MEDLINE citation field showing previously published articles commented on by authors of a given article ("Comment-in" or CIN) as primary external sources on which they may express complimentary or contradictory opinions. CIN articles, such as editorials or correspondence, are generally shorter than regular full-sized articles. MEDLINE conventions require CON data only from CIN articles, not from the typical regular articles. We identify CON data in a given article by recognizing sentences that contain such information based on cue phrases, sentence positions, and the frequency of occurrence of author names, and then analyzing the corresponding bibliographic data in the article's reference section. However, this approach can result in many false positive errors since similar citation sentences are often also found in full-sized regular articles.

We therefore first distinguish CIN articles from regular ones by an automated text categorization method using SVMs, and submit only the articles classified as CIN to the next stage of extracting CON data. We have implemented and tested two types of SVMs, one with a linear kernel function and the other with an RBF. Input feature vectors for these SVMs are created by combining four types of features: a bag of words extracted from the article title, words found in the header section suggesting the article type, size of body text, and cue phrases.

From our experiments, we find that SVM with a linear kernel function yields consistent and reliable performance in terms of accuracy and false positive error rate when a bag of words is used as the input feature vector. Moreover, it shows a significantly lower false negative error rate than the SVM with an RBF. Our experiments also show that SVM with a linear kernel function, when using all four types of features as input, achieves a significant improvement of performance in terms of accuracy, false positive error, and false negative error rates.

ACKNOWLEDGEMENT

This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

REFERENCES

1. J. Kim, D.X. Le, and G.R. Thoma, "Naïve bayes classifier for extracting bibliographic Information from biomedical online articles," *Proc. 5th Int'l Conf. Data Mining*, II, 373-8, Las Vegas (2008).
2. I. Kim, D.X. Le, and G.R. Thoma, "Hybrid approach combining contextual and statistical information for identifying MEDLINE citation terms," *Proc. 15th SPIE, Document Recognition and Retrieval*, 6815, 68150P (1-9), San Jose (2008).
3. I. Kim, D.X. Le, and G.R. Thoma, "Identification of "comment-on sentences" in online biomedical documents using support vector machines," *Proc. 14th SPIE, Document Recognition and Retrieval*, 6500, 65000O (1-8), San Jose (2007).
4. D.D. Lewis, R.E. Schapire, J.P. Callan, and R. Papka, "Training algorithms for linear text classifiers," *Proc. 19th Annual Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, 298-306, Zurich (1996).
5. Y. Yang, "Expert network: effective and efficient learning from human decision in text categorization and retrieval," *Proc. 17th Annual Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, 13-22, Ireland (1994).
6. S. Tan, "Neighbor-weighted K-nearest neighbor for unbalanced text corpus," *Expert Systems with Applications*, 28(4), 667-671 (2005).
7. A. McCallum and K. Nigam, "A comparison of event models for naïve bayes text classification," *Proc. AAAI'98 Workshop on Learning for Text Categorization*, 41-48, Madison (1998).
8. J.D.M. Rennie, L. Shih, J. Teevan, and D.R. Karger, "Tackling the poor assumptions of naive bayes text classifiers," *Proc. 20th Int'l Conf. Machine Learning*, 616-623, Washington DC (2003).
9. D.D. Lewis and M. Ringuette, "Comparison of two learning algorithms for text categorization," *Proc. 3rd Annual Symposium on Document Analysis and Information Retrieval*, 81-93, Las Vegas (1994).
10. E. Wiener, J.O. Pedersen, and A.S. Weigend, "A neural network approach to topic spotting," *Proc. 4th Annual Symposium on Document Analysis and Information Retrieval*, 317-332, Las Vegas (1995).
11. T. Joachims, "Text categorization with support vector machines: learning with many relevant features," *ECML 1998 LNCS*, 1398, 137-142, Springer, Heidelberg (1998).
12. S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Machine Learning Research*, 2, 45-66 (2002).
13. G. Salton and M.J. McGill, *Introduction to modern information retrieval*, McGraw-Hill, New York (1983).
14. J. Zou, D.X. Le, and G.R. Thoma, "Online medical journal article layout analysis," *Proc. 14th SPIE, Document Recognition and Retrieval*, 6500, 65000V (1-12), San Jose (2007).
15. F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, 34(1), 1-47 (2002).
16. L. Galavotti, F. Sebastiani, and M. Simi, "Experiments on the use of feature selection and negative evidence in to automated Text categorization," *ECDL 2000 LNCS*, 1923, 59-68, Springer, Heidelberg (2000).
17. V. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, New York (1995).