Naïve Bayes Classifier for Extracting Bibliographic Information from Biomedical Online Articles

Jongwoo Kim, Daniel X. Le, and George R. Thoma
National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA

Abstract - *A Naïve Bayes classifier has been developed to extract grant numbers, a key piece of bibliographic information, from online, HTML-formatted, biomedical articles for the National Library of Medicine's MEDLINE® database. Grant numbers identify research support from funding organizations, and are part of the MEDLINE citations. 47,362 sentences are collected from articles cited in the MEDLINE database to train and test the classifier, and 4,721 words are identified as suitable features for classification. Experimental results are evaluated using three measures: Precision, Recall, and F-Measure, all of which exceed 98.05%.*

Keywords: Naïve Bayes, Rule-based, Labeling, Grant Number.

1 Introduction

The U.S. National Library of Medicine (NLM) creates and manages MEDLINE®, a bibliographic database of 16 million citations to the biomedical journal literature. Citations are created in two ways. First, journals in paper form are scanned and the bibliographic data automatically extracted by the use of rule-based algorithms. The second is the reception of such data directly from journal publishers. However, these frequently omit certain bibliographic information such as grant numbers, databank accession numbers and funding agencies, requiring operators or expert indexers to search manually for these missing items, a labor-intensive task prone to human error.

To minimize this manual step, a system called Web-based Medical Article Records System (WebMARS) [1, 2] has been developed to automatically extract these items from online articles. One of the modules in WebMARS uses rule-based algorithms to detect text zones containing the required bibliographic information [3, 4]. It works reasonably well in most cases. However, rules for the algorithms are created manually and depend on the existence of combined key words in sentences. As a result, when authors use unusual or ambiguous words to express bibliographic information, the algorithms create over- or under-labeling problems. These algorithms are case sensitive, sensitive to typographic errors, and not robust.

The Naïve Bayes classifier [5-9] is commonly used in text mining/classification and information retrieval since it is fast, simple and efficient. It relies on the occurrence of features which are assumed to be stochastically independent. Since this approach can use any number of words in a document as features (and not just on some key phrases), the Naïve Bayes is more robust than rule-based algorithms.

In this paper, we present a Naïve Bayes classifier to label sentences in documents that contain grant numbers. For training and testing, we collect 47,362 sentences with and without grant numbers from articles cited in existing MEDLINE records. From these sentences, we collect 4,721 words as features. In addition to these, we collect three containing special features by manually analyzing sentences with grant numbers. We finally evaluate the performance of the classifier using Precision, Recall and F-Measure.

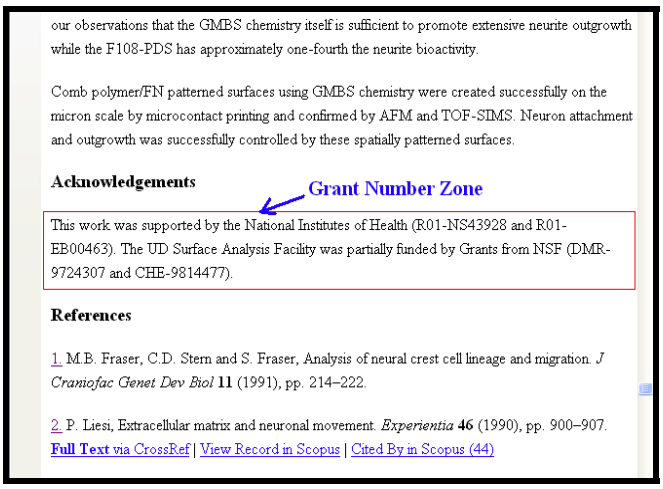
The paper is organized as follows. In Section 2, we define "grant number". In Section 3, we describe our method using the Naïve Bayes algorithm. In Section 4, we show experimental results using two data sets and two feature sets. Conclusions are presented in Section 5.

2 Definition of grant number

A grant number is an identifier assigned by a funding organization to a grant that supports the research reported in the article. Grant numbers usually appear in sentences that contain other information such as organizational names and/or words that suggest funding support, e.g., "supported", "funded", "financed", etc. A typical sentence is "This work was supported by National Institutes of Health Grant GM46904", where "GM46904" is the grant number and "GM" stands for the *National Institute of General Medical Sciences (NIGMS)*, the funding agency. As in this example, for grants issued by one of the institutes of the National Institutes of Health (NIH), the grant number includes a two-letter identifier. This is also shown in the following examples. Figure 1(a) shows NIH grant numbers R01-NS43928 and R01-EB00463 where "NS" and "EB" stand for *National Institute of Neurological Disorders and Stroke (NINDS)* and *National Institute of Biomedical Imaging and Bioengineering (NIBIB)*, respectively. Figure 1(b) shows NIH grant number 5R01AI20451-18 where "AI" stands for *National Institute of Allergy and Infectious Diseases Extramural Activities (NIAID)*.

The grant number consists of six parts as shown in Table 1, each having a distinct meaning explained with an example in Table 2.

A more detailed description about grant numbers is available in [10].


our observations that the GMBS chemistry itself is sufficient to promote extensive neurite outgrowth while the F108-PDS has approximately one-fourth the neurite bioactivity.

Comb polymer/FN patterned surfaces using GMBS chemistry were created successfully on the micron scale by microcontact printing and confirmed by AFM and TOF-SIMS. Neuron attachment and outgrowth was successfully controlled by these spatially patterned surfaces.

Acknowledgements

Grant Number Zone

This work was supported by the National Institutes of Health (R01-NS43928 and R01-EB00463). The UD Surface Analysis Facility was partially funded by Grants from NSF (DMR-9724307 and CHE-9814477).

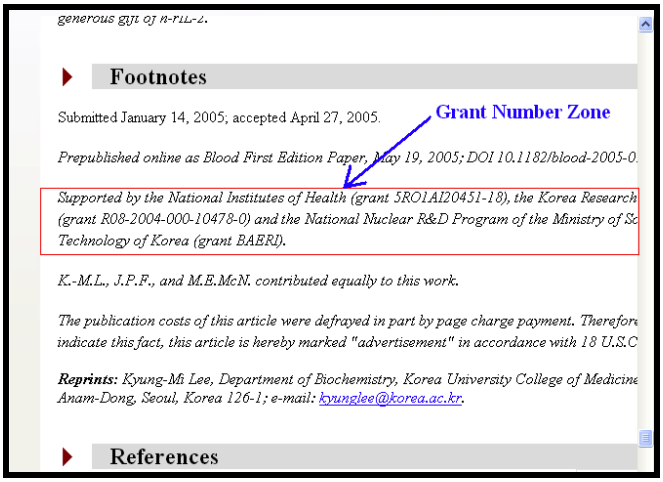
References

1. M.B. Fraser, C.D. Stern and S. Fraser, Analysis of neural crest cell lineage and migration. *J Craniofac Genet Dev Biol* **11** (1991), pp. 214–222.

2. P. Liesi, Extracellular matrix and neuronal movement. *Experientia* **46** (1990), pp. 900–907.

Full Text via CrossRef | View Record in Scopus | Cited By in Scopus (44)

(a)


generous gift of n-rIL-2.

Footnotes

Submitted January 14, 2005; accepted April 27, 2005.

Grant Number Zone

Prepublished online as Blood First Edition Paper, May 19, 2005; DOI 10.1182/blood-2005-0.

Supported by the National Institutes of Health (grant 5RO1AI20451-18), the Korea Research (grant R08-2004-000-10478-0) and the National Nuclear R&D Program of the Ministry of Sc Technology of Korea (grant BAERI).

K.-M.L., J.P.F., and M.E.McN. contributed equally to this work.

The publication costs of this article were defrayed in part by page charge payment. Therefore indicate this fact, this article is hereby marked "advertisement" in accordance with 18 U.S.C

***Reprints:** Kyung-Mi Lee, Department of Biochemistry, Korea University College of Medicine Anam-Dong, Seoul, Korea 126-1; e-mail: kyunglee@korea.ac.kr.*

References

(b)

Figure 1. Examples of HTML-formatted articles showing text zones containing grant numbers. (a) Grant numbers are R01-NS43928 and R01-EB00463. (b) Grant number is 5R01AI20451-18.

Funding organizations other than NIH express grant numbers in other formats, as shown in Table 3. In this table, "#" stands for Arabic number, "+" for any symbol such as space, "-", "_", "/", etc., "*" and "A" for alphabetic character, and "B" for Arabic number or alphabetic character. In these formats, AA identifies an Administering Organization and its subdivision. Each subdivision (e.g., an institute at NIH) has its own Administering Organization identified by a two-letter code, as shown in the third column.

Table 4 shows a collection of triplets of {an institution name, a subdivision name, Administering Organization Code} from eight institutions belonging to the Public Health Service (PHS), each of which has a number of subdivisions. The first row means National Library of Medicine (NLM) uses "LM" as its Administering Organization. Therefore, a research grant from NLM starts with "LM" followed by a five or six-digit number.

Funding agencies outside the U.S. government, such as the British "Wellcome Trust", have their own grant number formats.

3 Our approach

Existing rule-based algorithms [3, 4] are based on the established formats of grant numbers. They are easy to implement and manage, and work reasonably well in general cases.

However, since the rules are dependent on the existence of combined key words in sentences, they are not flexible and not robust to typographic errors. Thus, over- or under-labeling errors of the algorithms occur frequently. Since the Naïve Bayes classifier is based on statistics and depends on several words in a sentence, it usually can accommodate complicated situations such as typographic errors. Therefore, we analyze the ability of the Naïve Bayes to resolve such problems.

TABLE 1
OFFICIAL FORMAT OF GRANT NUMBER

Part	**Application Type**	**Activity Code**	**Administering Organization**	**Serial Number**	**Suffix Grant Year**	**Suffix Other**
Example	3	R01	CA	12329	04	S1A1

TABLE 2
PARTS OF A GRANT NUMBER DEFINED

Part	**Explanation**	**Example 3 R01 CA 12329 04S1A1**
Application Type	A single-digit code identifying the type of application received and processed.	3 (a supplemental request for additional fund)
Activity Code	A three-digit code identifying a specific category of extramural activity.	R01 (Research Project)
Administering Organization	A two-letter code identifying the first major-level subdivision.	CA (National Cancer Institute)
Serial Number	A five (or six)-digit number assigned sequentially to a series with an institute, center, or division.	12329
Suffix Grant Year	A two-digit number indicates the actual segment or budget period of a project.	04 (grants in their fourth year)
Suffix Other	A four digit code composed of Composed of Supplement (S), Amendment (A), or Allowance(X).	S1A1

TABLE 3
FORMATS OF GRANT NUMBERS

Organization	Formats	Example
Public Health Service	#+*##+AA+#####+##+###, #+*##+AA+######+##+###, **+#####, **+######	3 R01 CA 12329 04S1A1, GM46904
Agency for Health Care Policy and Research	###+##+####, ###+##+###	347-29-7834, 235-67-396
Centers for Disease Control and Prevention	BB+###+###, ###+##+####, ###+##+###, ***+######	45-297-364, CDC-53497
Wellcome Trust	######+B+BB+B, #####+B+BB+B, ######, #####	057321/3/Z2/4, 76345

TABLE 4
GRANTING ORGANIZATIONS AND A REPRESENTATIVE SUBDIVISION BELONGING TO THE U.S. PUBLIC HEATH SERVICE.

Organization Name	Subdivision	Administering Organization Code
National Institutes of Health (NIH)	National Library of Medicine	LM
Health Resources and Services Administration (HRSA)	Division of Disadvantaged Assistance	MB
Food and Drug Administration (FDA)	Center for Biological Evaluation and Research	BA
Centers for Disease Control and Prevention (CDC)	National Center for Injury Prevention and Control	CE
Office of the Assistant Secretary of Health (OASH)	Office of Family Planning	FP
Substance Abuse and Mental Health Services Administration (SAMHSA)	Office of the Administrator	OA
Agency for Health Care Policy and Research (AHCPR)	Agency for Health Care Policy and Research	HS

3.1 Naïve Bayes

Assume that we have a binary feature vector from a sentence $\boldsymbol{x}=(x_1, x_2, x_3,\ldots, x_m)$ where m is the dimension of the vector and x_i= 0 or 1 means absence or presence of the ith feature (words in our case) in the vector. Assume there are two classes C_r and C_n: relevant and non-relevant classes. The discrete distribution form of the Bayes' Theorem is expressed as

$$P(C_i \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid C_i)P(C_i)}{P(\boldsymbol{x})}, \text{ i} = r, n,$$

where $P(C_i)$ is the prior probability of C_i.

The decision function can be written as

$$P(\boldsymbol{x}|C_r)\, P(C_r) > P(\boldsymbol{x}|C_n)\, P(C_n) \qquad (1)$$

Assume that features x_i in feature vector $\boldsymbol{x}=(x_1, x_2,\ldots, x_m)$ are stochastically independent. Let us define p_i as the probability of occurrence of a word (suitable as a feature) in a sentence that is in a relevant class, and q_i as the probability of such a word in a non-relevant sentence. This is expressed as:

$$p_i = P(x_i=1|C_r) \qquad (2)$$

$$q_i = P(x_i=1|C_n) \qquad (3)$$

Then, $P(\boldsymbol{x}|C_i)$ can be rewritten as

$$P(\boldsymbol{x} \mid C_r) = \prod_{i=1}^{m} p_i^{x_i}(1-p_i)^{1-x_i} \qquad (4)$$

$$P(\boldsymbol{x} \mid C_n) = \prod_{i=1}^{m} q_i^{x_i}(1-q_i)^{1-x_i} \qquad (5)$$

When we insert Equations (4) and (5), take logs in Equation (1), and move the right term to the left, we have the linear decision function G(x) as follows.

$$G(\boldsymbol{x}) = \sum_{i=1}^{m} \log\frac{p_i(1-q_i)}{q_i(1-p_i)}x_i + \sum_{i=1}^{m} \log\frac{(1-p_i)}{(1-q_i)} + \log\frac{P(C_r)}{P(C_n)} \qquad (6)$$

When G($\boldsymbol{x}$) is positive, $\boldsymbol{x}$ belongs to C_r. If not, $\boldsymbol{x}$ belongs to C_n.

To decide on feature selection, the following equation is used [11, 12].

$$\left|\log\frac{p_i(1-q_i)}{q_i(1-p_i)}\right| \geq t \qquad (7)$$

When a feature candidate x_i satisfies the above criterion (greater than or equal to the threshold t) in Equation (7), we choose x_i as one of features in $\boldsymbol{x}=(x_1, x_2, x_3,\ldots, x_m)$. We use t=1 in our experiment, though this may be varied in future work.

In this paper, x_i stands for a word (a frequently occurring one) selected from sentences with and without grant numbers.

3.2 Analysis of sentences with grant number

An analysis of several thousand sentences leads to the four types shown in Table 5.

Sentences containing grant numbers may be recognized by three attributes: Granting Organization, Support Word, and

Grant Format. The first type of sentence shown in Table 5 mentions the Granting Organization (NIH), a Support Word "supported", and a correctly formatted grant number. Two of the attributes, Granting Organization and Support Word are collected as features for the Naïve Bayes classifier, as is an additional feature called Grant Word. Table 6 summarizes these three "special features" used in addition to "general features" discussed in the next section.

3.3 Performance evaluation measures

We use three measures, Precision, Recall, and F-Measure, to evaluate the performance of the proposed Naïve Bayes classifier. The measures are expressed as follows:

$$Precision = TP/(TP+FP)$$

$$Recall = TP/(TP+FN)$$

$$F\text{-}Measure = 2\times Precision\times Recall/(Precision+Recall)$$

Where *TP*, *FP* and *FN* mean "number of true-positives", "number of false-positives", and "number of false-negatives", respectively.

4 Experimental results

We select 15,211 sentences to train the algorithm from articles published in 2006. 5,142 of these sentences contain grant numbers (relevant class) and 10,069 sentences do not (non-relevant class). We obtain the 10,069 sentences using a random sampling technique from a large number of sentences. We also collected the 4,844 most frequently occurring words in these sentences as features for the Naïve Bayes classifier using the Equation (7). To test the trained algorithm, we also collect 23,862 sentences which have 5,142 sentences with grant numbers and 18,718 sentences without. Table 7 shows the performance of training and testing results using Precision, Recall, and F-Measure.

Training results in the first row show above 98.37% accuracy in all three measures. However, testing results in the second row show poor Precision and F-Measure performance, due to several false-positives. Table 8 shows examples of the Naïve Bayes result. The first two rows show examples of true-positive cases and the other rows show examples of false-positive cases. These false positives are sentences containing institutional affiliations misrecognized as grant number sentences. When we compare words used in the first and third rows, we find several words in common, such as "National", "Institutes", and "Health". We also find the same words in the second and fourth rows such as "Wellcome" and "Trust". It means that sentences containing affiliations share many common words with grant number sentences.

To resolve the problem, we collect about 8,000 affiliation sentences for the non-relevant class and add them to the existing training set which now total 23,500 sentences. 5,142 of these sentences have grant numbers and 18,538 do not.

TABLE 5
TYPES OF SENTENCES WITH GRANT NUMBERS

Type	Grant Organization	Support Word	Grant Format	Example of a Sentence
1	Yes (NIH)	Yes (supported)	Correct	This work was **supported** by funds from the **National Institutes of Health** (Grant **R01-NS43928** and **R01-EB00463**.).
2	Yes (NIH)	Yes (funded)	Incorrect	This work is **funded** by **NIH 23756**.
3	No	Yes (supported)	Correct	This research was **supported** by grants **5 RO1 AI29471**, **RO1 AI40297**, and Research contract **NO1 AI45251**.
4	Yes (PHS)	No	Correct	**Public Health Service R01AI 47736**.

TABLE 6
SPECIAL FEATURES USED IN THE NAIVE BAYES CLASSIFIER FOR GRANT NUMBER

Word lists	Words in the list
Support Word	supported, funded, granted, financed, etc.
Grant Word	grant, fund, scholarship, etc.
Granting Organization	NIH, FDA, CDC, OASH, SAMHSA, etc.

TABLE 7
PERFORMANCE OF THE NAÏVE BAYES CLASSIFIER RESULTS

Data Set	Precision (%)	Recall (%)	F-Measure (%)
Training	98.37	99.69	99.03
Testing	30.53	99.96	46.77

TABLE 8
EXAMPLES OF THE NAÏVE BAYES CLASSIFIER RESULTS.

Class	NB Result	Sentence
Relevant	True-Positive	This work was supported by the National Institutes of Health (R01-NS36834) and the Canada Foundation for Innovation. Result(1.000000)
Relevant	True-Positive	This research was supported by the United Kingdom Medical Research Council (grant G9803180),EUROMALVAC I (QLK2-CT-1999-01293) and The Wellcome Trust (grant 057270)
Non-Relevant	False- Positive	Cell and Molecular Biology Section, Pediatric Oncology Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA
Non-Relevant	False- Positive	Wellcome Trust/Cancer Research UK Gurdon Institute of Cancer and Developmental Biology,University of Cambridge Tennis Court Road,Cambridge,CB2 1QN,UK
Non-Relevant	False- Positive	Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland 4072, Australia.

We also collect 6,870 of the most frequently occurring words in these sentences, and select 4,721 words as features using the criterion expressed in Equation (7). We refer to these as "general features" from now on. In addition, we use the three "special features" shown in Table 6.

To show the relative occurrence of these features in sentences containing grant numbers and those that do not, we compute the probability figures (p_i and q_i) listed in Table 9, which are derived from a frequency analysis of the training set. For example, the word "national" occurs in about 66% of the sentences in the relevant class (i.e., containing grant numbers), while it appears in less than 2% of the sentences in the non-relevant class.

We conduct two experiments, the first using the general features alone ("Without Special Features") and the second with special features as well ("With Special Features"). Tables 10, 11, and 12 show the comparable training, testing, and performance results of the two experiments.

As shown in Tables 10 and 11, "Without Special Features" shows fewer false-negative errors than "With Special Features". However, "Without Special Features" shows more false-positive errors. In total, "Without Special Features" shows more errors than "With Special Features".

Table 12 shows performance results using the three measures. In training set, all three measures exceed 97.83% and 98.56% for "Without Special Features" and "With Special Features", respectively. "With Special Features" shows better performance than "Without Special Features" in Precision and F-Measure, but "With Special Features" shows lower Recall performance. In testing set, we have a similar result. All three measures exceed 97.01% and 98.05% in "Without Special Features" and "With Special Features", respectively. "With Special Features" shows better performance than "Without Special Features" in Precision and F-Measure. However, Recall is lower with "With Special Features".

The two experiments show that "With Special Features" is comparable overall to "Without Special Features", though slightly better in Precision and F-Measure.

A journal article usually has more than one hundred sentences. Among these, one or two sentences (less than 1%) belong to the relevant class and the other sentences (more than 99%) belong to the non-relevant class. That is, the Naïve Bayes classifier has inputs from the non-relevant class ninety-nine times the number of inputs from the relevant class. Therefore, the Precision measure is more important than Recall, since minimizing false-positive error is more important than minimizing false-negative error. For this reason we will use special features in future work.

Table 13 shows examples of misclassification. In the false-negative error examples, there are a granting organization and a correctly formatted grant number. However, there is no "Support Word" or "Grant Word", which presumably causes the algorithm to make a wrong decision. In the false-positive error examples, one of the NIH institutes is named correctly though no grant number appears. We assume the correct organization name causes the algorithm to make an error.

TABLE 9
SOME WORD FEATURES AND CORRESPONDING P_I AND Q_I

Feature Type	Feature	p_i	q_i
Special	Granting Organization	0.86192143	0.01160257
Special	Support Word	0.89478802	0.00103497
Special	Grant Word	0.91579152	0.00076261
General	national	0.66297161	0.01803029
General	supported	0.81777518	0.00119839
General	grant	0.47394010	0.00054472
General	health	0.54453520	0.03791263
General	work	0.53753403	0.00114392
General	institutes	0.46499417	0.00544722
General	research	0.32360949	0.06710971
General	grants	0.41423571	0.00032683

TABLE 10
TRAINING RESULTS WITH/WITHOUT SPECIAL FEATURES

		Without			**With**	
Sentence (Total:23,500)		**True**	**False**		**True**	**False**
Relevant (5142)		5,099	43		5,070	72
Non-Relevant (18,538)		113	18,245		74	18,284

TABLE 11
TEST RESULTS WITH/WITHOUT SPECIAL FEATURES

		Without			**With**	
Sentence (Total:23,862)		**True**	**False**		**True**	**False**
Relevant (5,144)		5127	17		5,120	24
Non-Relevant (18,718)		158	18,560		102	18616

TABLE 12
PERFORMANCE OF THE NAÏVE BAYES CLASSIFIER WITH/WITHOUT SPECIAL FEATURES (PERCENTAGE)

	Without				**With**		
Data Set	**Precision**	**Recall**	**F-Measure**		**Precision**	**Recall**	**F-Measure**
Training	97.83	99.16	98.50		98.56	98.60	98.58
Test	97.01	99.67	98.32		98.05	99.53	98.78

TABLE 13
EXAMPLES OF MISCLASSIFICATION BY THE NAÏVE BAYES CLASSIFIER

Class	Naïve Bayes Result	Sentence
Relevant	False-Negative	A B was partially covered by NIH 1R15CA113331-01.
Relevant	False-Negative	Support for the course at Woods Hole was provided by MH-062204
Non-Relevant	False- Positive	The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA, USA.
Non-Relevant	False- Positive	Requests for reprints:P Andrew Futreal,Cancer Genome Project,Wellcome Trust Sanger Institute,Hinxton CB10 1SA,United Kingdom.

5 Conclusions

This paper describes a Naïve Bayes classifier to automatically label sentences containing grant numbers in HTML-formatted articles.

We conduct two experiments, one using general features and the other using both general and special features. Both experiments show the Naïve Bayes classifier has above 97.01% labeling accuracy in all the three measures. "With Special Features" shows a little better performance than "Without Special Features" in Precision and F-Measure, and a little less performance in Recall.

Since the classifier receives inputs from the non-relevant class ninety-nine times more than from the relevant class, Precision is more important than Recall. Therefore, we intend to use special features with the general features in the future.

The Naïve Bayes classifier is based on statistics and depends on the several words in the zones. Therefore, it usually generates reasonable results that overcome situations such as typographic errors. However, it also shows problems in training for cases that occur rarely. Therefore, as future work, we need to combine the Naïve Bayes classifier with rule-based algorithms (such as Decision Tree and Random Forest that generate rules automatically) to compensate for problems caused by the other. In addition, we seek to refine the feature set (add other features such as formats of grant numbers, etc.) to further improve the accuracy of the classifier.

6 Acknowledgment

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

7 References

[1] G.R. Thoma, D.X. Le "Automating data entry for online biomedical databases", *Proc. 14th National Conference on Integrated Online Library Systems IOLS´99*", Medford, NJ, pp. 121-128, May 1999.

[2 D.X. Le, L.Q. Tran, et. al., "Automated Medical Citation Records Creation for Web-Based On-Line Journals," *14th IEEE Symposium on Computer-Based Medical Systems*, Bethesda, MD, pp. 315-320, July 2001.

[3] J. Kim, D. Le, and G. Thoma, **"**Automated labeling of bibliographic data extracted from biomedical online journals," *Proc. SPIE Electronic Imaging*, Vol. 5010, January, pp. 47-56, 2003.

[4] J. Kim, D. Le, and G. Thoma, "Automated Labeling of Biomedical Online Journal Articles**,"** *Proc. 9th World Multiconference on Systemics, Cybernetics and Informatics*, July, Orlando, FL, Vol. 3, pp. 406-411, 2005.

[5] D. D. Lewis, "Naïve (Bayes) at Forty: The Independence Assumption in Information Retrieval," *ECML*, The Tenth European Conference on Machine Learning, pp.4-15, 1998.

[6] A. McCallum and K. Nigam, "A Comparison of Event Models for Naïve Bayes Text Classification," *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, pp.577, 1998.

[7] Y. Shen and J. Jiang, "Improving the Performance of Naïve Bayes for Text Classification," *cs224n Spring. Technical report*, Stanford University, 2003.

[8] S. Eyheramendy and D.D. Lewis, and D. Madigan, "On the Naïve Bayes Model for Text Classification," *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, pp.332-339, 2003.

[9] D. Madigan, "Statistics and the war on spam," *Statistics:A Guide to the Unknown*, 4th Ed. (R. Peck, G. Casella, G. Cobb, R. Hoerl, D. Nolan, R. Starbuck and H. Stern, eds.), Thomson Brooks/Cole, Belmont, CA, pp.135–147, 2005.

[10] NIH, *Activity Codes, Organization Codes, and Definitions Used in Extramural Programs*, July, 2007. Available: http://grants.nih.gov/grants/funding/ac.pdf).

[11] J.S. Milton and J.C. Arnold, Introduction to Probability and Statistics, McGraw-Hill, pp. 71-75.

[12] S. Sohn, W.K. Kim, D.C., et. al., "Optimal Training Sets for Bayesian Prediction of MeSH Assignment," *Journal of the American Medical Informatics Association*, 2008, (Accepted).