INL/CON-15-35266
PREPRINT

Optimal Stop Word Selection for Text Mining in Critical Infrastructure Domain

International Symposium on Resilient Control Systems

Kasun Amarasinghe and Milos Manic
(Virginia Commonwealth University)

Ryan Hruska
(Idaho National Laboratory)

June 2015

The INL is a
U.S. Department of Energy
National Laboratory
operated by
Battelle Energy Alliance

Optimal Stop Word Selection for Text Mining in Critical Infrastructure Domain

Kasun Amarasinghe, Milos Manic
Virginia Commonwealth University
Richmond, Virginia, USA
amarasinghek@vcu.edu, miko@ieee.org

Ryan Hruska
Idaho National Laboratory (INL)
Idaho Falls, Idaho, USA
ryan.hruska@inl.gov

***Abstract*—Eliminating all stop words from the feature space is a standard practice of preprocessing in text mining, regardless of the domain which it is applied to. However, this may result in loss of important information, which adversely affects the accuracy of the text mining algorithm. Therefore, this paper proposes a novel methodology for selecting the optimal set of domain specific stop words for improved text mining accuracy. First, the presented methodology retains all the stop words in the text preprocessing phase. Then, an evolutionary technique is used to extract the optimal set of stop words that result in the best classification accuracy. The presented methodology was implemented on a corpus of open source news articles related to critical infrastructure hazards. The first step of mining geo-dependencies among critical infrastructures from text is text classification. In order to achieve this, article content was classified into two classes: 1) text content with geo-location information, and 2) text content without geo-location information. Classification accuracy presented methodology was compared to accuracies of four other test cases. Experimental results with 10-fold cross validation showed that the presented method yielded an increase of 1.76% or higher in True Positive (TP) rate and a 2.27% or higher increase in the True Negative (TN) rate compared to the other techniques.**

Keywords—Stop word selection; Dimensionality Selection; Text Mining; Genetic Algorithms; Text Classification

I. INTRODUCTION

Text Mining is the process of extracting implicit and potentially useful information from text data [1]. In text mining applications, the widely used text representation methodology is to convert the text information into a numeric representation called the "bag-of-words" matrix [1], which is used as the feature space. Regardless of the application, in creating the feature space, preprocessing of the text corpus has been identified as an extremely crucial step [2]. In the preprocessing phase of text mining, one of the frequently used methods is to eliminate the "stop words" from the feature space [1], [3]. Thus, stop words are considered to have minimal or no additional semantic information. It is shown that elimination of stop words in the text preprocessing stage improves the accuracy of text categorization [4], Optical Character Recognition accuracy [5], font learning and keyword spotting [6].

Stop words are the most used words in the English language which includes but not limited to, pronouns such as *"I, he, she"* or articles such as *"a, an, the"* or prepositions [7]. The concept of stop-words was first introduced in Information Retrieval (IR) systems [1]. It was noticed that a small portion of words in the English language accounted for a significant portion of the text size in terms of frequency of appearance [8]. Further, in [8], it was noticed that the mentioned words were not used as index word to retrieve documents. Thus, it was concluded that such words do not carry significant information about documents. Thus, the same interpretation was given stop words in text mining applications as well [1].

The standard practice of removing stop words from the feature space mainly contributes to reducing the size of the feature space. However, the stop word list that is considered to be removed from the feature space is an application independent generic stop words list. This may have an adverse effect on the text mining application as the semantic importance of a certain word is dependent on the domain and the application [3]. I.e. certain words that are deemed to be stop words in a generic context can prove to be important keywords in certain applications and domains. Thus, studies have been conducted to content specific stop words lists on a language basis. For instance, languages such as Chinese [1], [9] and Arabic [10] has been studied. In [3] the authors have conducted a study to generate a domain specific set of stop words from a pre classified text dataset.

The main emphasis of this work is to propose an optimal domain specific stop word set identification methodology for a text mining task in the critical infrastructure domain. The paper presents the first steps of creating a critical infrastructure knowledge framework in the form of a text classification problem. The presented text mining algorithm excludes the step of stop word removal in the preprocessing phase. Thus, when creating the feature vector all the stop words are retained. Then, the presented methodology utilizes a dimensionality selection methodology for finding the optimal set of dimensions which results in the best classification accuracy. The presented methodology was tested on a corpus of open source news articles related to critical infrastructure hazards. The experimentation was carried out on the said corpus to classify the text content into two classes: 1) text content with geographic location mentions and 2) text content without geographic location mentions. The geographic locations are extracted from the text to mine the geographical dependencies between critical infrastructures. The performance of the

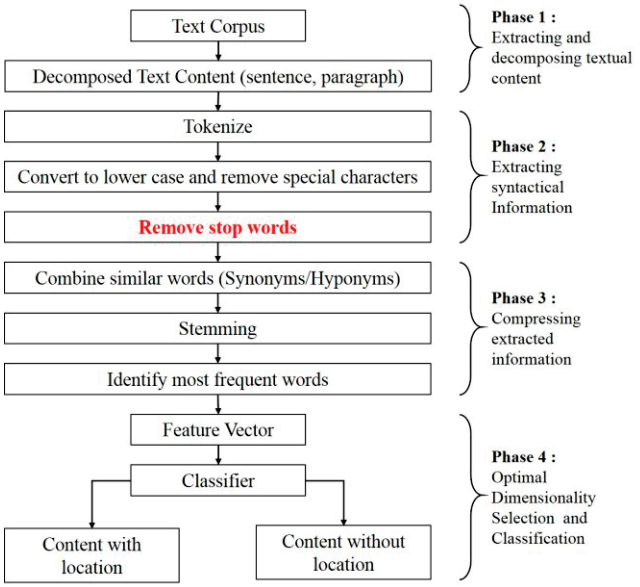

Fig. 1. Standard text mining process and the complete text classification methodology. The critical step of removing stop words is shown in red.

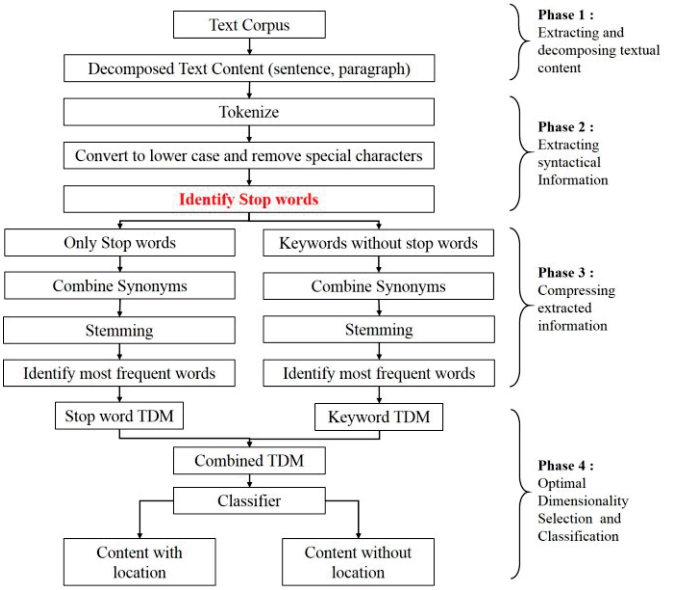

Fig. 2. Presented optimal stop word selection methodology

presented method was compared against four other test cases: **Test Case A:** classification with standard stop word elimination and without dimensionality selection, **Test Case B:** classification with standard stop word elimination and with dimensionality selection, **Test Case C:** classification with all the stop words included and dimensionality selection methodology applied only to keywords, and **Test Case D:** classification with all the keywords and stop words included without any dimensionality selection. Experimental results showed that the presented optimal stop word selection methodology outperformed the above methods consistently.

The rest of the paper is organized as follows. Section II elaborates the presented domain specific optimal stop word selection methodology. Section III describes the used dataset and implementation details of the presented method. Section IV details the conducted experiments and the experimental results obtained. Finally, Section V presents the conclusions and further research directions.

II. Presented Optimal Stop Word Selection Methodology

This section presents the details of the presented optimal stop word selection methodology. This section will first describe the overall classification process and then discuss the text mining techniques used in detail.

A. Overall Text Classification Process

This paper presents an optimal stop word selection methodology for a text classification task. In order to perform the classification text information has to be represented in a numeric feature vector. Fig. 1 shows the standard text preprocessing and classification process. Fig. 2 shows the presented optimal stop word incorporated classification process. In both methodologies, the overall process can be divided into four phases.

In **Phase 1,** the text documents are decomposed into the content granularity specified. For instance, the text can be decomposed into paragraphs or sentences depending on the application. Then, each paragraph or sentence will be considered as one input pattern (text pattern) to the classifier.

In **Phase 2,** the decomposed text content is processed to extract syntactical information. The methods of extracting these information is discussed further in section B. After extracting the syntactic information the decomposed text patterns are represented as vectors of unique words. Each unique word is called a keyword. The syntactic information retrieval phase removes formatting and symbols from text. Further, the standard method removes the stop words from the keyword set. The presented methodology identifies the stop words and separates them from the other keywords.

In **Phase 3**, the extracted keyword vectors are compressed. I.e. the number of keywords that represent the text contents are reduced. This step reduces the size of feature space, which results in reducing the sparseness of the feature vector as well as making the process less processor intensive. This phase uses text processing techniques to combine keywords with similar meanings so that the most general set of keywords that best represent the text corpus is obtained. The techniques used in this phase is discussed further in Section II.B. Further, this phase keeps track of the frequency in which each keyword appears in the text corpus, so that the highest occurring key words can be identified. In the presented methodology the compression step is applied to stop words and keywords separately.

Finally in **Phase 4**, the extracted keywords are used to create a feature space that represent the text corpus. Each dimension of the feature space represents a unique keyword or a set of keywords with similar meaning. This feature matrix is then sent to the classifier to perform the classification. In the presented method, a stop word feature matrix and a keyword

feature matrix is created separately and combined to create the final matrix.

B. Details of Text Mining Steps

This section elaborates the main text processing techniques utilized in **Phase 2** and **Phase 3**. Both Phases consist of three different stages in both the standard method and the presented method.

In the syntactical information retrieval phase **(Phase 2)**, the first step is to tokenize the text content. Each decomposed text component is represented as a *bag-of-words*. Thus, all semantic information is lost. In this phase, each unique keyword that appears in the text corpus is stored. Then for each decomposed text content or text input pattern, the frequency in which each unique keyword appears is stored. This representation is also known as the *term-frequency* representation [11]. Therefore, a text pattern P_i in a feature space with n unique words can be expressed as,

$$P_i = \{f_{i1}, f_{i2}, ..., f_{ij}, ..., f_{in}\} \tag{1}$$

Where f_{ij} is the frequency in which the j^{th} feature or keyword appears in the i^{th} document. The bag-of-words representation disregards all the formatting, punctuations and special characters that appear in text.

In the standard practice, stop words, are removed from the feature space (See Fig. 1 step shown in red). In the presented methodology, this step is replaced. (See Fig. 2) The stop words are identified and separated from the rest of the keywords. The text mining techniques specified in **Phase 3** are applied to the stop word set and the keyword set separately as Fig. 2 shows.

Thus, with the term-frequency representation, for N patterns in a feature space with M unique keywords. The information can be presented in an $N \times M$ term-frequency matrix:

$$TDM = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1M} \\ f_{21} & f_{22} & \cdots & f_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ f_{N1} & f_{N2} & \cdots & f_{NM} \end{bmatrix} \tag{2}$$

This $N \times M$ matrix is called the Term-Document Matrix (TDM). In the TDM, each row represents a text input pattern, which, in this instance, is a decomposed text portion such as a sentence or a paragraph. Each column of the TDM represent a unique keyword, the f_{ij} values represent the frequency in which j^{th} keyword appears in the i^{th} text portion. In the presented methodology, a separate TDM is created for the stop words (Stop-word TDM) and one for the regular keywords (keyword TDM).

The main drawback of the created TDMs is, when the number of text patterns *(N)* increases, the number of unique words *(M)* will increase as well. This results in a very large and sparse matrices which leads to increased computational resource and time consumption.

In order to overcome the said drawback, further text mining techniques are applied to compress the feature space to combine redundant information that might appear in the TDM. I.e. these techniques are aimed at reducing the dimensionality of the TDM. This process is carried out in **Phase 3**.

The compression process is carried out in 3 stages. All the stages are applied to the stop word TDM and the keyword TDM separately. The process is explained for a general TDM. The first stage includes combining of synonyms in the feature space to reduce the dimensionality. Synonyms are keywords which have similar or near similar meanings. In this step, such keywords are combined into one dimension. Thus, a keyword m_i with r synonyms can be expressed as one dimension as given below,

$$m_i = \{a_{i1}, a_{i2}, ..., a_{ir}\} \tag{3}$$

where a_{ik} is the k^{th} word in the dimension m_i. Before identifying synonyms the number of words in each dimension, r is 1 for all dimensions.

Let "*word a is a synonym of word b*" be represented as $a \approx b$, then, for the two keyword sets m_i and m_j represented by:

$$m_i = \{a_{i1}, a_{i2}, ..., a_{iA}\} \tag{4}$$

$$m_j = \{b_{j1}, b_{j2}, ..., b_{jB}\} \tag{5}$$

$$m_i = m_j \;\; if \;\; any \;\; a_{ik} \approx b_{jl} \vee b_{jl} \approx a_{ik} \;\; \forall \; k,l \tag{6}$$

Further, if $m_i = m_j$:

$$m(syn)_{i,j} = m_i \bigcup m_j \tag{7}$$

and m_i and m_j is deleted from the TDM and $m(syn)_{i,j}$ is added to the TDM.

This process is iterated over the complete set of identified words M in the TDM, until there are no more dimensions that satisfy $m_i = m_j$. Thus, the combining of synonyms effectively reduces the dimensionality of the TDM.

The second stage in the compression phase (**Phase 3**) is deconstructing words into their base forms and combining similar words. The deconstruction of words into their base form is known as stemming. Stemming deconstructs words that have been transformed by pluralizing or by adding a gerund, into their basic form. This enables identification of transformed words as similar to their base words.

Genetic Algorithm

1: Initialize the population with random solutions (individuals)
2: Evaluate population (Calculate fitness of each individual)
3: Repeat until termination criterion is met
  3.1: Select parents (Selection)
  3.2: Recombination pairs of parents for new offspring (Crossover)
  3.3: Mutate offspring (Mutation)
  3.4: Evaluate new population

Fig. 3 Pseudo-code of the Genetic Algorithm (GA) flow.

The process of identifying similar stemmed words and combining them is similar to the process described above for synonyms. Therefore, dimensionality of the TDM is further reduced with minimal loss of information.

The third and final stage of the compression step is identifying the most frequently used keyword sets in the TDM. I.e. selecting the keywords that most appear in the text corpus. Typically, keyword sets that appear in less than *P%* of the documents are removed from the TDM. *P* can be defined at the experimenter's discretion. This, not only reduces the dimensionality of the TDM but also identifies the most general set of words that describe the text corpus.

C. *Genetic Algorithm based optimal dimension selection and classification*

This section elaborates the classification and the process of selecting the optimal set of dimensions for classification.

Once the feature vector is created using the text mining techniques discussed above, it is fed in to the classification and optimal dimensionality selection phase of the methodology **(Phase 4)**. Even though Phase 3 of the process compressed the information, still the resultant TDM is sparse and very high dimensional. Furthermore, in the presented methodology, the third stage of Phase 1; removing stop words (shown in red in figure 1), is excluded. As mentioned, stop words are most frequently used words in the English language. Therefore, this results in the dimensionality of the feature vector being even higher than when following the standard method.

Thus, it is extremely important to use a dimensionality selection methodology in order to select the subset of dimensions in the feature space which results in the highest classification accuracy. Further, this feature selection methodology is important to identify the best subset of stop words to be used in conjunction with the regular keywords to obtain the highest classification accuracy. Theoretically, any optimal dimensionality selection methodology can be applied in this step, such as genetic algorithms, random selection.

In this paper, genetic algorithms (GA), are used to perform the dimensionality selection. GAs are a type of evolutionary algorithm. Evolutionary algorithms are inspired by Darwin's theory of evolution. Simulated biological evolution has been translated into an effective tool for global optimization [12]. The underlying idea of GAs is that the algorithm maintains a set of individuals where each individual encodes a candidate solution to the problem. The strength of each individual is evaluated based on a predefined application specific fitness function. Parents for the next generation are then selected using selection methodologies. From the parents, new offspring are produced by recombination/crossover operators and the off spring are randomly altered by mutation operators. This main cycle is iterated for a specified number of iterations or until another convergence criterion is met, The general pseudo-code of GA is summarized in Fig. 3. Specifics of the implemented GA is provided Section III.

The classifier performs the final classification depending on the feature vector that is fed into it. I.e., the classifier assigns each text component (sentence or paragraph) into their class. In this paper, the classifier determines whether a certain text content contains geographical location information or not. Similar to the dimensionality selection algorithm, the classification can be performed using different algorithms. Details about the used classifier in the implementation is provided in Section III.

III. IMPLEMENTATION

This section elaborates the implementation details of the presented optimal stop word selection methodology. First, the used text corpus will be introduced, and then the implementation specifics of the evolutionary dimensionality selector and classification algorithms will be discussed.

A. *Text Dataset*

This section will describe the text dataset that was used in the implementation process of the presented methodology.

The dataset used for the text classification task described in this paper consisted of a corpus of open source online news articles gathered from various news sources. The data set was gathered for autonomous creation of a critical infrastructure knowledge framework via text mining. The data set consisted of 55 such articles and only the text content of the main body was considered for the classification process.

As mentioned in Section II, the text corpus was decomposed into text components before processing. In the implementation the granularity of decomposing was considered to be a sentence. Thus, the classification was carried out to determine the sentences which contained information about geographical locations.

The document corpus was decomposed to 1462 sentences. The corpus contained 682 sentences with geographic location mentions and 780 sentences without geographic location mentions.

B. *Implementation details*

As mentioned in Section II, each decomposed sentence is considered as one input pattern. First, in **Phase 2** the sentences are tokenized and special characters were removed.

TABLE I. IMPLEMENTATION DETAILS OF THE GENETIC ALGORITHM

Algorithm	Generational Genetic Algorithm
Population size	100
Selection method	Tournament Selection with a Tournament size of 10
Elitism	3 Copies of the best individual retained
Crossover method	Uniform Crossover
Crossover rate	50%
Mutation method	Randomly swapping values in dimensions with a probability of 0.50
Mutation rate	30%

In **Phase 3,** from the tokenized words, the generic stop word list was used to extract the stop words. As shown in Fig.2 2, the sequential steps in **Phase 3**, are applied to the set of stop words and the set of key words sans the stop words separately. For both sets, the synonyms combination process is carried out using the English lexical database Wordnet [13]. Stemming to the basic form of the words were implemented using the porter stemming [14] method.

Thus, a TDM for stop words and a TDM for key words was created after **Phase 3.** The final TDM was considered as the combination of the key word TDM and the stop word TDM. The combined TDM consisted of 376 keywords and was used in the dimensionality selector.

A generational GA was implemented as the dimensionality selection methodology in the implementation. An individual was encoded as a binary string as mentioned in Section II. The length of the binary string was equal to the number of keywords remaining in the combined TDM. If the keyword was selected for the classification process, the binary value corresponding to the keyword was set to "1" and "0" otherwise. A population was considered to have 100 individuals. The implemented GA consisted of operations, elitism, selection, mutation and crossover. The details about the implemented GA is given in Table I.

The fitness of an individual was evaluated at each iterations from the classification results. The fitness function was selected by experimenting with different combinations of the True Positive (TP), True Negative (TN) and Bayesian Detection (BD) rates. The three mentioned parameters were used to improve classification accuracy. The fitness was calculated using the following function.

$$F_i = \left(\frac{3TP_i + TN_i}{4} \right) + 10BDR_i \tag{8}$$

Where, F_i is the fitness of the i^{th} individual, TP_i and TN_i are the True Positive and True Negative rates achieved by the classifier for the i^{th} individual and BDR_i is the Bayesian Detection Rate of the classifier for the i^{th} individual. BDR_i is the probability of a sentence classified as a sentence with a location mention being actually a sentence with a location mention [15]. Thus a higher BDR means the number of sentences without location mentions being classified wrongly as sentences with location mentions, is low. The Bayesian Detection Rate can be calculated as follows.

TABLE II. CONFUSION MATRIX

		Classified as	
		Location	No location
Actual Class	Location	True Positives (TP)	False Negatives (FN)
	No location	False Positives (FP)	True Negatives (TN)

TABLE III. CLASSIFICATION RESULTS FOR THE PRESENTED METHODOLOGY AND TEST CASES

Methodology	TP Rate	TN Rate	BD Rate
Optimal stop word selection	81.86%	87.47%	0.86
Test Case A	69.41%	83.34%	0.81
Test Case B	80.10%	85.20%	0.84
Test Case C	71.88%	80.51%	0.78
Test Case D	30.64%	82.30%	0.63

$$BDR_i = \frac{TP_i}{TP_i + FP_i} \tag{9}$$

Where BDR_i and TP_i represents the same entities as described in Eq. (10) and FP_i stands for the False Positive rate achieved by the classifier for the i^{th} individual.

Naïve Bayes Multinomial (NBM) classifier was utilized to perform the final text classification [16]. The classifier was selected because of its high interpretability, high dimensional data classification capability and their fast learning capability

NBM is based on the Bayes theorem of conditional probability and it calculates the conditional probability of a pattern belonging to a class given a set of keywords [16]. Unlike the similar classifier Naïve Bayes, the NBM classifier considers the frequency in which the keyword appears in a text pattern, not just whether a keyword appears or not in the text pattern [16], [17].

IV. EXPERIMENTAL RESULTS

This section presents the conducted experiments and the results obtained from the experiments. The presented methodology was tested on the data set presented in Section III. The implementation of the presented methodology was compared against four other cases. Test Cases A through D, are elaborated below.

In **Test Case A,** the standard text preprocessing methodology was followed and all the stop words were eliminated from the feature space. The TDM was considered as is without any dimensionality selection methodology for the classification.

In **Test Case B,** the standard preprocessing methodology was followed and the stop words were removed from the feature space, but the GA based dimensionality selection methodology was applied to the TDM.

In **Test Case C,** all the stop words were included in the combined TDM and the GA based dimensionality selection methodology was applied only to the keyword TDM.

In **Test Case D,** all the stop words and keywords were included in the combined TDM. No dimensionality selection methodology was used.

Classification in all four test cases and in the presented methodology, were performed with 10-fold cross validation. Classification results are shown in terms of True Positives (TP), True Negatives (TN) and BD rates. The confusion matrix is given in Table II.

Table III presents the final classification results obtained from the presented methodology and the 4 test cases listed above. For experiments with GA based dimensionality selection (Test Cases A, B and presented methodology for optimal stop word set selection), the experiment was repeated 10 times and the averaged values of the 10 runs are given in Table III. It can be seen that the presented methodology for incorporating optimal set of stop words to the final feature space produces the best classification results. It was shown that the presented methodology showed at least a 1.76% increase in the TP rate and a 2.27% increase in the TN rate when compared to the other test cases.

V. Conclusions and Future Work

This paper presented a methodology for selecting the optimal set of stop words for a text mining task in the critical infrastructure domain. A text classification process was conducted to classify the content of a text corpus of open source news articles in to two classes: 1) text content with geo-locations and 2) text content without geo-locations. The presented methodology uses an evolutionary dimensionality selection methodology to identify the optimal set of stop words and Naïve Bayes Multinomial Classifier for classification.

The implementation of the presented methodology was compared against four other test cases. All tests were performed on the same dataset with the same dimensionality selection and classification algorithms. The experimental results showed that the presented methodology consistently produced better classification accuracies when compared against the other four methods. The presented methodology showed a 1.76% or higher increase in the TP rate and a 2.27% or higher increase in the TN rate when compared against other methods. As future work, the presented method will be evaluated on a larger dataset in the same domain. Furthermore, the applicability of the method will be tested on different text mining domains.

Acknowledgment

This work was supported in part by the INL Laboratory Directed Research & Development (LDRD) Program under DOE Idaho Operations Office Contract DE-AC07-05ID14517.

References

[1] Z. Yao, C. Ze-wen, "Research on the Construction and Filter Method of Stop-word List in Text Preprocessing," in *Proc. Intelligent Computation Technology and Automation (ICICTA), 2011 International Conference on* , vol.1, no., pp.217,221, 28-29 March 2011

[2] H. Ahonen, O. Heinonen, M. Klemettinen, A.I. Verkamo (1997). Applying data mining techniques in text analysis. *Report C-1997-23, Dept. of Computer Science, University of Helsinki*.

[3] H. Ayral, S. Yavuz, "An automated domain specific stop word generation method for natural language text classification," *in Proc. Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on*, vol., no., pp.500,503, 15-18 June 2011

[4] C. Silva, B. Ribeiro. "The importance of stop word removal on recall values in text categorization." *Neural Networks, 2003. Proceedings of the International Joint Conference on. Vol. 3. IEEE*, 2003.

[5] T. K. Ho, "Stop word location and identification for adaptive text recognition". *International Journal on Document Analysis and Recognition*, 3(1), 16-26, 2000

[6] T. K. Ho, "Fast identification of stop words for font learning and keyword spotting", In Proc of *Document Analysis and Recognition, Fifth International Conference on (ICDAR). IEEE*; pp. 333-336 Sep. 1999

[7] D. Wijayasekara, M. Manic, M. McQueen, "Vulnerability Identification and Classification Via Text Mining Bug Databases," in *Proc. 40th Annual Conference of the IEEE Industrial Electronics Society, IEEE IECON* 2014, Dallas, TX- USA, Oct. 29 - Nov. 1, 2014. Abstract, PDF, DOI: 10.1109/IECON.2014.7049035

[8] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information." *IBM Journal of research and development*, 1(4), 309-317. (1957)

[9] L. Hao, L Hao, "Automatic identification of stop words in Chinese text classification". *In Proc of Computer Science and Software Engineering, International Conference on* (Vol. 1, pp. 718-722). IEEE. Dec. 2008

[10] R. Al-Shalabi, G. Kanaan, J. M. Jaam, A. Hasnah, E. Hilat, "Stop-word removal algorithm for Arabic language." *In Proc. of 1st International Conference on Information and Communication Technologies: From Theory to Applications, Damascus* ; pp. 545-550, April, 2004

[11] C. A. Martins, M. C. Monard, E. T. Matsubara, "Reducing the Dimensionality of Bag-of-Words Text Representation Used by Learning Algorithms," *in Proc of 3rd IASTED International Conference on Artificial Intelligence and Applications*, pp. 228–233, 2003.

[12] D. E. Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning", *Addison-Wesley Professional*, 1989.

[13] C. Fellbaum, "WordNet: An Electronic Lexical Database", *Cambridge, MA: MIT Press*, 1998

[14] M. F. Porter, "An algorithm for suffix stripping," in *Program*, vol. 14, no. 3, pp. 130−137, 1980.

[15] S. Axelsson, "The base-rate fallacy and the difficulty of intrusion detection," in *ACM Transactions on Information and System Security (TISSEC)*, vol. 3, no. 3, pp. 186–205, Aug. 2000.

[16] A. McCallum, K. Nigam. "A comparison of event models for naive bayes text classification," in *Proc. of AAAI-98 workshop on learning for text categorization*, vol. 752, 1998.

[17] Q. B. Duong, E. Zamai, K. Q. Tran Dinh, "Confidence estimation of feedback information using dynamic bayesian networks," in *Proc. IEEE. Int. Conf. of the Industrial Electronics Society, (IECON)*, pp. 3733-3738, Oct. 2012.