COMBINATION OF FEATURE SELECTION METHODS FOR TEXT CATEGORIZATION BY USING COMBINATORIAL FUSION ANALYSIS AND RANK SCORE CHARACTERISTIC

ARCHANA B (10Z305)
MANISHA R K (10Z322)
SASIKALA S (10Z340)
SUBHAPRBHA V (10Z349)
LALITHA A (11Z462)

08Z720 PROJECT WORK-I

BACHELOROF ENGINEERING

Branch: COMPUTER SCIENCE AND ENGINEERING
Of Anna University

.3

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

PSG COLLEGE OF TECHNOLOGY

(Autonomous Institution)
COIMBATORE-641 004

Synopsis

PSG COLLEGE OF TECHNOLOGY

(Autonomous Institution)

COIMBATORE – 641 004

COMBINATION OF FEATURE SELECTION METHODS FOR TEXT CATEGORIZATION BY USING COMBINATORIAL FUSION ANALYSIS AND RANK SCORE CHARACTERISTIC

ARCHANA B (10Z305)
MANISHA R K (10Z322)
SASIKALA S (10Z340)
SUBHAPRBHA V (10Z349)
LALITHA A (11Z462)

08Z720 PROJECT WORK-I

BACHELOR OF ENGINEERING

Branch: COMPUTER SCIENCE AND ENGINEERING
of Anna University

October2012

..……

……..……. **Mrs.C.Kavitha**

Dr.R.Venkatesan

Faculty guide Head of the Department

ACKNOWLEDGEMENT

We express our deep sense of gratitude to **Dr. R.RUDRAMOORTHY**, Principal of PSG College of Technology, for having provided the necessary environment to carry out our project successfully.

We profusely thank, **Dr. R.VENKATESAN** , Head of the Department, Computer Science and Engineering and Program Coordinator, **Dr. G.R.KARPAGAM** who has greatly helped in the success of the project, by providing us with the necessary facilities required.

We express our gratitude to **Mrs. C.Kavitha,** our guide, Assistant Professor (Sr.Gr.), Computer Science and Engineering, PSG College of Technology, for guidance to proceed with the project in an effective manner.

We extend our thanks to all our department staff, our friends and Librarian for their timely help and support to complete this project report successfully.

SYNOPSIS

Text categorization is a supervised learning task that assigns the predefined category labels to new documents based on the likelihood derived from a set of labelled training documents. In order to classify documents, each document should be transformed into a model that preserves as much of the original information as possible. The *bag of words* representation is one of the simple and preferred models that represents a document as a set of distinct words by ignoring the order and meaning of words. When the number of words in documents is considered, high dimensionality may become an inevitable problem. Since the data in text categorization are high-dimensional, naturally dimensionality reduction becomes a necessity for efficiency and accuracy.

The comparison is made between the feature selection methods and the varied binary combinations. The performance of five common feature selection methods is studied and then the performance of all possible binary score and rank combinations of these five feature selection methods determine the most appropriate features for classification. Comparing the performance of the individual methods with the performance of the combination methods shows that combining two feature selection methods can significantly improve the performance of the individual methods. In addition, rank combination achieves better performance in the case of global policy while score combination significantly achieves better performance in the case of local policy.

In order to investigate the effectiveness of combining the individual metrics on the performances of text categorization and find the best combinations. It is more likely that combining different feature selection methods obtains more effective performance in text categorization.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

Text Categorization algorithms need effective feature selection methods for improving the efficiency and accuracy .Feature selection aims at removing redundant and irrelevant features from feature space. Feature space which is bag of terms in the document which provides better understanding of the document is needed for improving text mining results. This project explores the possibility of improving overall performance by combining multiple individual feature selection methods which outperforms individual feature selection methods. The proposed method is by combining multiple feature selection methods, by using an information fusion paradigm called Combinational Fusion Analysis.

1.1 Motivation

Text categorization is one of the major problems in text mining and information retrieval. The main task of text categorization is automatically assigning class labels to new document based on its content. In text classification models documents are represented by vector space model which treats the document as bag of terms. The characteristic of this representation is high dimensionality of feature space due to large number of terms which imposes a challenge to the performance of text categorization algorithms. Another problem is that not all features are important of text categorization. Since some features may be redundant or irrelevant.

1.2 Objective

Feature selection aims at removing redundant and irrelevant features from feature space. Thus feature selection should not only reduce the high dimensionality of feature space but also should provide a better understanding of documents in order to improve text mining results. Feature selection can improve efficiency and accuracy of text categorization algorithms. Feature selection methods are classified as supervised or unsupervised depending on the requirements of class label information for training documents. Supervised feature selection methods such as information gain, statistic and mutual information estimates each term based on the occurrence of terms in labelled document frequency. Unsupervised feature selection methods such as term strength and document frequencyestimate each term based

on the distribution of term across the corpus without using the class label information of training documents.

Soft feature selection approach can improve the performance of text classification. In soft feature selection method, unselected features are down-weighted instead of being eliminated from the feature space. The feature selection uses the weight from the linear support vector machines in combination with a learning algorithm such as naïve Bayes classifier and perceptron has better classification performance than other feature weighting methods.

In this project real data sets are illustrated which combines multiple feature selection methods to improve the performance of text categorization.

The different feature selection methods introduce the score function and rank function of feature selection method, score combination and rank combination of multiple feature selection methods, rank function and rank score graph. The five feature selection methods are compared with each other in terms of F-measure value of the classification result of naïve Bayes classifier.

Multiple feature selection methods combine individual feature selection methods for better performance by taking the advantage of individual strength. A rank score function and its associated graph called rank-score graph are adopted to measure the diversity of different feature selection methods. The combination of multiple feature selection methods can outperform a single method only if each individual feature selection method has unique scoring behaviour and relatively high performance. Rank score function and rank score graph are useful in the selection of a combination of feature selection methods.

1.3Software requirement for development

Platform: Windows XP/7

Software Packages: JDK 1.6

Languages Used: Java using NetBeans

CHAPTER 2
LITERATURE SURVEY

2.1 MACHINE LEARNING IN AUTOMATED TEXT CATEGORIZATION:

2.1.1 Fabriziosebastiani [1]

In the last decade of years content based document management tasks have gained a prominent status in the information system field due to the increased availability of documents in the digital form and the ensuring need to access them in flexible ways.

Text categorization is the activity of labelling natural language texts with thematic categories from a predefined set, is one such task. Text categorization dates back to the early ‘90s but it became a major subfield of the information systems. Text categorization is now being applied in many contexts, ranging from document indexing based on a controlled vocabulary ,to document filtering, automated metadata generation, word sense disambiguation, population of hierarchical catalogues of Web resources, and in general any application requiring document organization or selective and adaptive document dispatching.

The arrival of the machine learning methods in the text categorization field is one of the most important factors that accelerate the improvement in this field by strong theoretical motivations. A growing number of machine learning methods have been used for text categorization, including probabilistic classifiers, decision trees, regression methods, nearest neighbour classifiers, Rocchio method, neural networks

2.2A COMPARATIVE STUDY ON FEATURE SELECTION IN TEXT CATEGORIZATION:

2.2.1 YimingYang,JanO.Pedersen [2]

The focus is on aggressive dimensional reduction. Five methods are evaluated which includes term selection based on document frequency, informationgain, mutualinformation, chi-square and term strength. Information gain and chi-square are found to be more effective.Using information gain thresholding with K nearest neighbour classifier on the reutors corpus; removes 98% of unique terms yielded improved classification accuracy. Document frequency thresholding performed similarly but possess high computation measures which are too expensive.

New techniques are proposed to address these challenging tasks involving many irrelevant and redundant variables. There are many potential benefits facilitating feature selection which includes data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance. Effective feature selection methods are important for improving the efficiency and accuracy of text categorization algorithm by removing redundant and irrelevant terms from the corpus and to improve the performance of individual feature selection methods.

2.3AN EXTENSIVE EMPIRICAL STUDY OF FEATURE SELECTION METRICS FOR TEXT CLASSIFICATION:

2.3.1George Forman [3]

An empirical study of twelve feature selection metrics was evaluated on a text classification problem. The primary focus is on obtaining the best overall classification performance regardless of the number of features needed to obtain that performance The metrics excel were analysed, when only a very small number of features is selected, which is important for situations where machine resources are severely limited, fast classification is needed, or large scalability is demanded. The results from each of the perspectives of accuracy, precision, recall, and F-measure were also analysed, since each serves different purposes.

In text classification problems, there is typically a substantial class distribution skew, and it worsens as the problem size scales up. For example, in selecting news articles that best match one's personalization profile, the positive class of interest contains many fewer articles than the negative background class. If the background class contains all news articles posted on the Internet worldwide. For multi-class problems, the skew increases with the number of classes.

High class skew presents a particular challenge to induction algorithms, which beat the high accuracy achieved by simply classifying everything as the negative majority class. The hypothesis that feature selection should then be relatively more important

in difficult, high-skew situations. The study contrasts the performance under high-skew and low-skew situations, validating this hypothesis.

Finally, which feature selection metric or combination is most likely to obtain the best performance for the single given dataset at hand, supposing their text classification problem is drawn from a distribution of problems. The results on datasets show that the Information Gain metric is a good choice.

2.4 TEXT CATEGORIZATION WITH SUPPORT VECTOR MACHINES: LEARNING WITH MANY RELEVANT FEATURES:

2.4.1Joachims, T [4]

With the rapid growth of online information, text categorization has become one of the key techniques for handling and organizing text data. Text categorization techniques are used to classify news stories, and interesting information on the WWW, also to guide a user's search through hypertext. Since building text classifiers by hand is difficult and time-consuming.

The benefits of SVM were introduced and it provides both theoretical and empirical evidence that SVMs are very well suited for text categorization. The theoretical analysis concludes that SVMs acknowledge the particular properties of text: high dimensional feature spaces, few irrelevant features and sparse instance vectors.

Reducing dimensionality is another critical issue in text categorization. Feature selection is one of the effective methods that improves the efficiency and accuracy of the classifiers by selecting only more discriminative terms in a dataset as features. Various feature selection methods have been presented and analysed.

2.5TEXT CLASSIFICATION BASED ON MULTI-WORD WITH SUPPORT VECTORE MACHINE

2.5.1 Wen Zhang, Taketoshi Yoshida, Xijin Tang[5]

Multiword is a newly exploited feature for text representation in the field of information retrieval and text mining. The multi-word extraction is implemented based on the syntactical structure of the noun multiword phrases. For the sake of reduction on computation, repletion pattern identification is proposed to be extracted from sentences firstly and then use the extracted repetition patterns for regular expression matching to extract the multi-words. In order to use the multi-words for representation, two strategies are developed based on the different semantic level of the multiwords: the first is the decomposition strategy using general concepts for representation and the second is combination strategy using subtopics of the general concepts for representation. IG method was employed as a scale to remove the multiword from the feature set to study the robustness of the classification performance. Finally, a series of text classification tasks were carried out with SVM in linear and non-linear kernels, respectively, to analyze the effect of different kernel functions on classification performance. That is, to study the problem of what kind of vector mapping method is more preferred to project the document vector space to the category space.

The combination strategy for multiword representation outperforms the decomposition strategy, and linear kernel outperforms non-linear kernel with SVM. In addition, it also appears that the combination strategy has poorer robustness than the decomposition strategy when the low IG value features are removed from the feature set. Nevertheless, the effect of using different representation strategies is greater than using different kernel functions in SVM on the classification performances.

The benefits of multiword representation include at least three aspects. Firstly, it has lower dimension than individual words but its performance is acceptable. Secondly, multiword is easy to acquire from documents by corpus learning without any support of thesaurus, dictionary or ontology. Thirdly, multiword includes more semantics and is a larger meaningful unit than individual word.

2.6IMPROVING TEXT CATEGORIZATION PERFORMANCE BY COMBINING FEATURE SELECTION METHODS

2.6.1 EceOzbilen [6]

Feature selection is one of the well-known processes that reduce the dimensionality by ranking all features according to their importance estimated by a method and then selecting ones with the highest values. Feature selection not only reduces time and storage requirements but also improves the efficiency and accuracy of the classifiers. Feature selection makes applying classifiers on data more efficient by reducing the size of the effective features. In addition, feature selection often improves classification accuracy by eliminating noise features that are non-informative or misleading for classification and lead to incorrect generalization, overfitting from the training documents.

In text categorization there are two main policies to apply feature selection: local policy and global policy. The local policy, where a different set of features is selected from each class independent from other classes, gives equal weight to each class. Thus, it tends to optimize the classification performance on frequent and infrequent classes by selecting the most important features for each class. On the other hand, the global policy, where a single set of features is selected from all classes, provides a global view of the entire dataset by extracting a single global score from the local scores. Thus, the global policy tends to penalize the infrequent classes in highly skew datasets by selecting the most important features for the entire dataset

One of the main drawbacks in text categorization is imbalance data distribution since the rare classes are dominated by common classes. As the performances of the classifiers are directly affected by the skewness of the datasets, the performance is commonly measured by two different alternatives: micro-averaged and macro-averaged F-measures. First gives equal weight to each document and therefore it tends to be dominated by the classifier's performance on common categories while reflects the overall accuracy better. On the other hand, second gives equal weight to each category regardless of its frequency and thus it is influenced more by the classifiers performance on rare categories.

2.7AN INTRODUCTION TO VARIABLE AND FEATURE SELECTION:

2.7.1 Isabelle Guyon [7]

This potential benefit of variable and feature selection includes facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance.

The issue mainly on constructing and selecting subsets of features that areuseful to build a good predictor. This contrasts with the problem of finding or ranking all potentially relevant features. Selecting the most relevant features is usually suboptimal for building a predictor, particularly if the features are redundant. Conversely, a subset of useful variables may exclude many redundant, but relevant, variables.

The problem of supervised learning is treated more extensively than that of unsupervised learning; classification problems serve more often as illustration than regression problems, and only vectorial input data is considered. Complexity is progressively introduced throughout the sections: The first section starts by describing filters that select variables by ranking them with correlation coefficients subset selection methods.Thewrapper methods that assess subsets of variables according to their usefulness to a given predictor. The problem of feature construction, whose goals include increasing the predictor performance and building more compact feature subsets. All of the previous steps benefit from reliably assessing the statistical significance of the relevance of features.

2.8COMBINING SVMs WITH VARIOUS FEATURE SELECTION STRATEGIES:

2.8.1Yi-Wei Chen and Chih-Jen Lin [8]

The performance of combining support vector machines (SVM) and various feature selection strategies. Some of them are filter type approaches which include general feature selection methods independent of SVM, and some are wrapper-type methods which include modifications of SVM which can be used to select features and choose important features as well as conduct training and testing. Support Vector Machine (SVM) is an effective classification method, but it does not directly obtain the feature importance.

The combination SVM with various feature selection strategies and investigate its performance. Most of them are independent of the classifier used. This work preliminary study that for an SVM package what feature selection strategies should be included.

2.9 WEIGHTING AND SELECTION OF FEATURES:

2.9.1 WłodzislawDuch and Karol Grudzinski [9]

Features are excluded and ranked according to their contribution to the classification accuracy in the cross validation tests. Weighting factors used to compute distances which are optimized using global minimization procedures or search-based methods. The experiments show that, for some datasets, these methods give much better results than classical nearest neighbour methods.

It is important to stress that the establishment of statistical association by means of chi-square does not necessarily imply any causal relationship between the attributes beingcompared, but it does indicate that the reason for the association is worth investigated.

2.10FEATURE SELECTION USING IMPROVED MUTUAL INFORMATION FOR TEXT CLASSIFICATION:

2.10.1 Jana Novovi·cova, Antonin Malik and PavelPudil [10]

In text classification, usually a document is represented using a bag of words. This representation scheme leads to very high dimensional feature space. Feature selection is a very important step in textclassification because irrelevant and redundant words often degrade the performance of classification algorithms both in speed and classification accuracy. Methods for feature subset selection for text document classification task were used for evaluation function that is applied to a single word. All words are independently evaluated and sorted according to the assigned criterion. Then a predefined number of the best features are taken to form

the best feature subset. Scoring of individual words is be performed using some of the measures which include document frequency, term frequency, mutual information, information gain, chi square and term strength. The information gain is a very simple frequency measures which were reported to work well on text data.

The major characteristic of text document classification problem is extreme high dimensionality of text data. Two algorithms were presented for feature selection for the purpose of text classification. Sequential forward selection methods based on improved mutual information were introduced for non-textual data.

The improved mutual information has not yet been applied in text classification as a criterion for reducing vocabulary size. The simple but effective naïve Bayes classifier is used based on multinomial model.

Naive Bayes Classifier

According to the bag of words representation, the document d_i can be represented by a feature vector consisting of one feature variable for each word w_t in the given vocabulary V= {w_1,..., w_n} containing *n* distinct words. Let C = {c_1,...., $c_{|c|}$} be set of |C| classes. A new document d, the probability that *d* belongs to class c_j is given by Bayes rule,

$$P(c_j|\mathbf{d}) = P(c_j)P(d|c_j)/P(d)$$

If the task is to classify a new document into a single class, simply select the class *c** with the highest posterior probability.

Best individual features

Best individual features methods evaluate all the n words individuallyaccording to a given criterion, sort them and select the best k words. Since the vocabulary has usually several thousand or tens of thousands of words, the BIF methods are

popular in text classification because they are rather fast, efficient and simple. However, it evaluates each word separately and completely ignores the existence of other words and the manner how the words work together. It has been proven that the best pair of features needs not contain the best single features.

Scoring of individual features can be performed using some of the measures for instance document frequency, term frequency, mutual information, information gain, statistic or term strength. The above methods prove that information gain and X^2 statistic are most effective in word selection.

CHAPTER 3
FEATURE SELECTION METHODS

Feature selection is one of the effective methods that improves the efficiency and accuracy of the classifiers by selecting only more discriminative terms in a dataset as features. Five methods are used which involves term goodness criterion threshold to achieve a desired degree of term elimination from the full vocabulary of document corpus.

3.1 DOCUMENT FREQUENCY

The issue of term specificity, given a pair of terms/concepts that have been found to be related, to determine which is the more specific concept/term an approach used is document frequency.Document frequency is the number of documents in which the term has occured. Document frequency for each unique term is calculated and those terms with the document frequency less than predetermined threshold are removed from feature space. The assumption is that rare terms are non-informative in category prediction or non influential in global performance. In either case removal of rare terms reduces the dimentionality of the feature space. Improvement in categorization also happens when rare items happen to be noise items.

Document frequency is the simplest technique for vocabulary reduction. It easily scales up to a very large corpora of data with approximately linear computational complexity in number of training documents. It is uaually considered an ad hoc technique to improve efficiency and not a principle approach to select predictive features.Also, Document frequency is not used in aggressive term removal because of the widely received assumption in information retrieval that low –valued Document frequency terms are relatively informative and therfore should not be removed aggressively.

3.2 INFORMATION GAIN

Information gain is frequently addressed as a term goodness criterion in the field of machine learning. It is the measure of number of bits of information available for categorical prediction based on the presence or absence of a term in the document. It is measure the decrease in entropy by existence or absence of the term in a document. Information gain score will be null for two independent variables and it will be high because of the dependence between two variables. Let C_i be the categorical set in the target space, Information gain of term can be calculated as

$$G(t) = - \sum^{m}_{i=1} P(C_i) \log P(C_i) + P(t) \sum^{m}_{i=1} P(C_i|t) \log P(C_i|t)$$

Usage of more general formula is because the text categorization problem involves m-ary category space and we need to calculate the goodness of the term globally with respect to all categories on average.

Given a training corpus, for each unique term computed the information gain , and removed those terms whose information gain is less than some predefined threshold. Information gain feature selection method selects the terms with the highest information gain scores which contains much information about the classes

3.3 MUTUAL INFORMATION

In probability theory and information theory, the mutual information (sometimes known by the archaic term transinformation) of two random variables is a quantity that measures the mutual dependence of the two random variables. The most common unit of measurement of mutual information is the bit, when logarithms to the base 2 are used. Mutual information is a criterion commonly used in statistical modelling of word association and related application. If on considering contingency table of term t, and category C, Mutual information is given by

$$I(t , c) = \log \frac{P(t , c)}{P(t)P(c)}$$

where $p(t,c)$ is the joint probability distribution function of X and Y, and p(C) and p(t) are the marginal probability distribution functions of X and Y respectively. Intuitively, mutual information measures the information that X and Y share: it measures how much knowing one of these variables reduces uncertainty about the other. For example, if X and $Y$are independent, then knowing X does not give any information about Y and vice versa, so their mutual information is zero. At the other extreme, if X and Y are identical then all information conveyed by X is shared with Y: knowing X determines the value of Y and vice versa. As a result, in the case of identity the mutual information is the same as the uncertainty contained in Y (or X) alone, namely the entropy of Y (or X: clearly if X and Y are identical they have equal entropy). Mutual information is a measure of the inherent dependence expressed in the joint distribution of X and Y relative to the joint distribution of X and Y under the assumption of independence. If X and Y are independent, then $p(x,y) = p(x)\ p(y)$, Determines how much information a term contains about a class.

In the case of continuous random variables, the summation is replaced by a definite double integral.

I(t,C)= zero when t and C are independent of each other. Mutual Information reaches its maximum value if the term is a perfect indicator for class membership.

The weakness of MI is that it is score is strongly influenced by marginal probability of terms. Therefore two terms with same conditional probability, rare term will have higher score than common term.

3.4 χ^2- STATISTICS (chi square statistics)

The chi square statistics is used to measure the degree of independence between term t and category C. Investigates whether distributions of categorical variables differ from one another. It is the Normalised value of Mutual information hence chi square values are comparable across terms of same category. Using contingency table of term t and category C ,chi square can be calculated as

$$\chi 2=\Sigma i\{(O_i-E_i)/2E_i\}$$

where O_i is the Observed count and E_i is the expected count. The sum of the squared difference between observed (*o*) and the expected (*e*) data (or the deviation, *d*), divided by the expected data in all possible categories

If chi-square score of a term tk is low value, this means tk is independent from the class ci and if chi-square score of a term tk is high value, this means tk is dependent of the class ci . Thus the chi-square feature selection method selects the terms with the highest chi-square score which are more informative for classification.

The chi-squared statistic provides a test of the association between two or more groups, populations, or criteria. The chi-square test can be used to test the strength of the association between term and Category.

In experimental sciences, chi-square statistics is frequently used to measure how the observation results differ from the expected results. In other words, it measuresthe independence of two random variables.

3.5TERM STRENGTH

This technique is used to find how commonly a term is likely to occur in the closely-related documents and thus measures the term importance.

It uses training set of documents to determine the document pairs whose similarity is above the threshold value. Term strength is then calculated from conditional probability that a term occurs in the first half of the related document when it has occurred in the second half of the document.

It is based on the document clustering assuming that documents with shared words are highly related and that the terms in heavily overlapping area of related document is relatively informative.

This method is not task specific that is it does not depend on term-category association in the way it is similar to Document frequency and different from chi-square statistics and Mutual Information. The parameters in Term strength calculation is the threshold on the Document similarity values that is to determine how related a pair of documents are , on which the term importance test is to be conducted.

3.6. SCORE AND RANK COMBINATIONS OF FEATURE SELECTION METHODS

The previous studies show that the success of the combination and the number of feature selection method involved in combination are inversely related. As the number of the feature selection method in the combination increases, the performance of the combination decreases. In addition, it states that the best performances are achieved by the combination of two feature selection methods. Thus consider combining two distinct feature selection methods. In the study, evaluate the performance of all possible binary-combinations (2-combinations) of five feature selection methods.

The principle of the feature selection methods can separate into two steps. First is scoring that is giving higher scores to the terms which considered more informative for classification and second is selection that is selecting the terms with the highest scores. In order to combine the outputs of varied feature selection methods in scoring step, scores of each term from the varied feature selection methods are normalized using the maximum and minimum scores according to the below formula:

Score = (s1, s2, ...,sn) where si score of the ith term, n is the total number of term.

$score'$=($Score - min\ Score$) / ($max\ Score - min\ Score$)

By normalization, the scores fall in the same range [0, 1] and scores of the terms from the varied feature selection methods are represented equally which is allowing for meaningful comparisons between the methods.

3.6.1. SCORE COMBINATION

Score combination is averaging the normalized term scores of the different feature selection methods.

$$C_{score} = \sum^{M}{}_{I=1} (S_{core} / M)$$

where M is the number of feature selection methods which is 2 for this study.

3.6.2. RANK COMBINATION

Rank Combination is averaging the term ranks of the different feature selection methods obtained from the term scores.

Rank = (r1, r2, ...,rn) where *ri* rank of the ith term, n is the total number of term.

$$C_{rank} = \sum_{i=1}^{m} (Rank_i / M)$$

There are various ways for assigning rankings such as standard competition ranking, modified competition ranking, dense ranking, ordinal ranking and fractional ranking. In this study, we rank the terms according to the descending order of their scores with standard competition ranking strategy. In competition ranking *("1, 2, 2, 4" ranking),* terms that have the same score get the same ranking number and then a gab is left in the ranking numbers.

CHAPTER 4

PROTOTYPE IMPLEMENTATION

Prototype implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The work flow of our project is discussed in this chapter.

4.1 MUTUAL INFORMATION

The document which is to be trained is specified and training is performed.

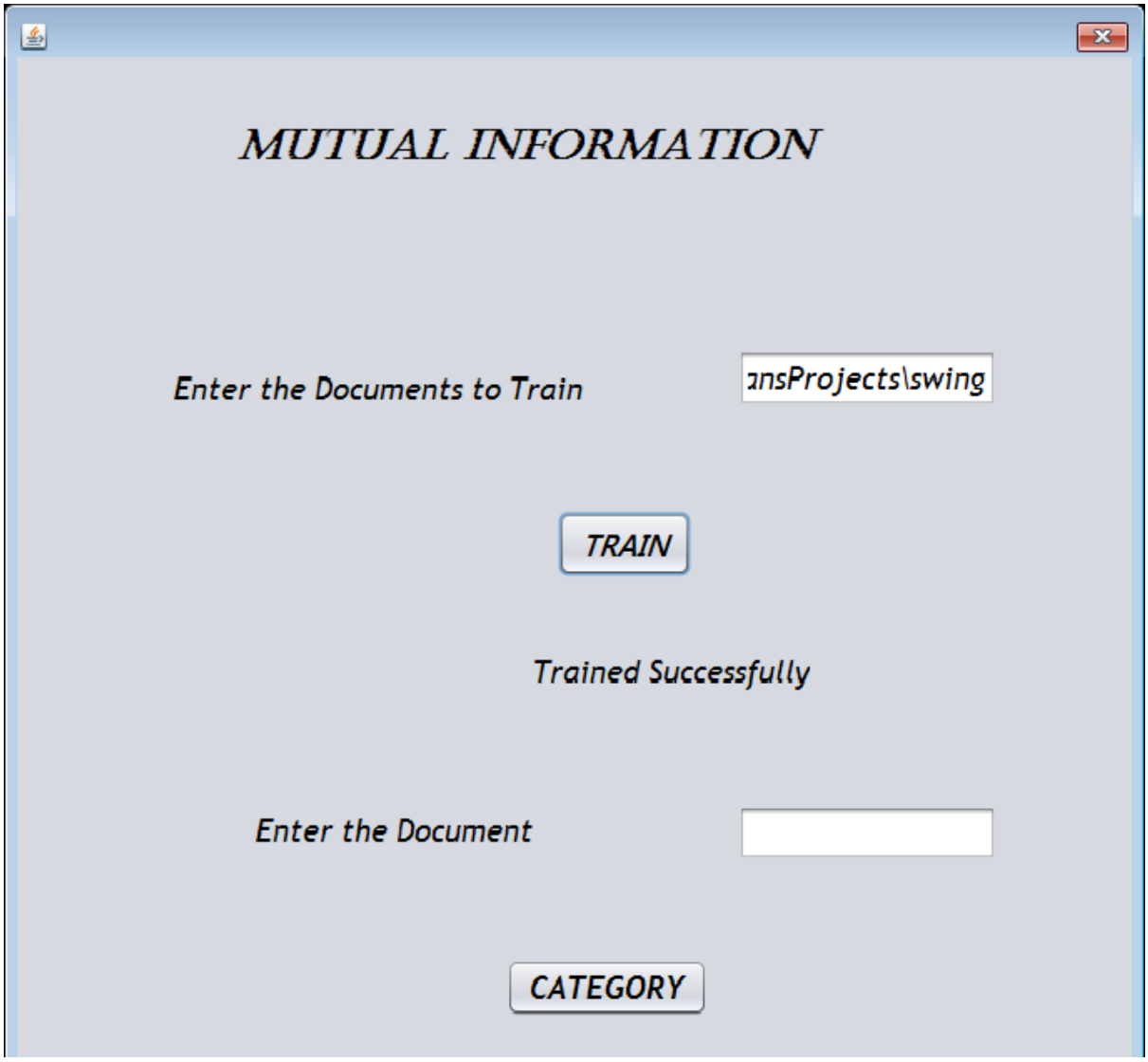


Figure 4.1 Training document

4.1.2 TEST DOCUMENT

Enter the document to be tested and performance document test.

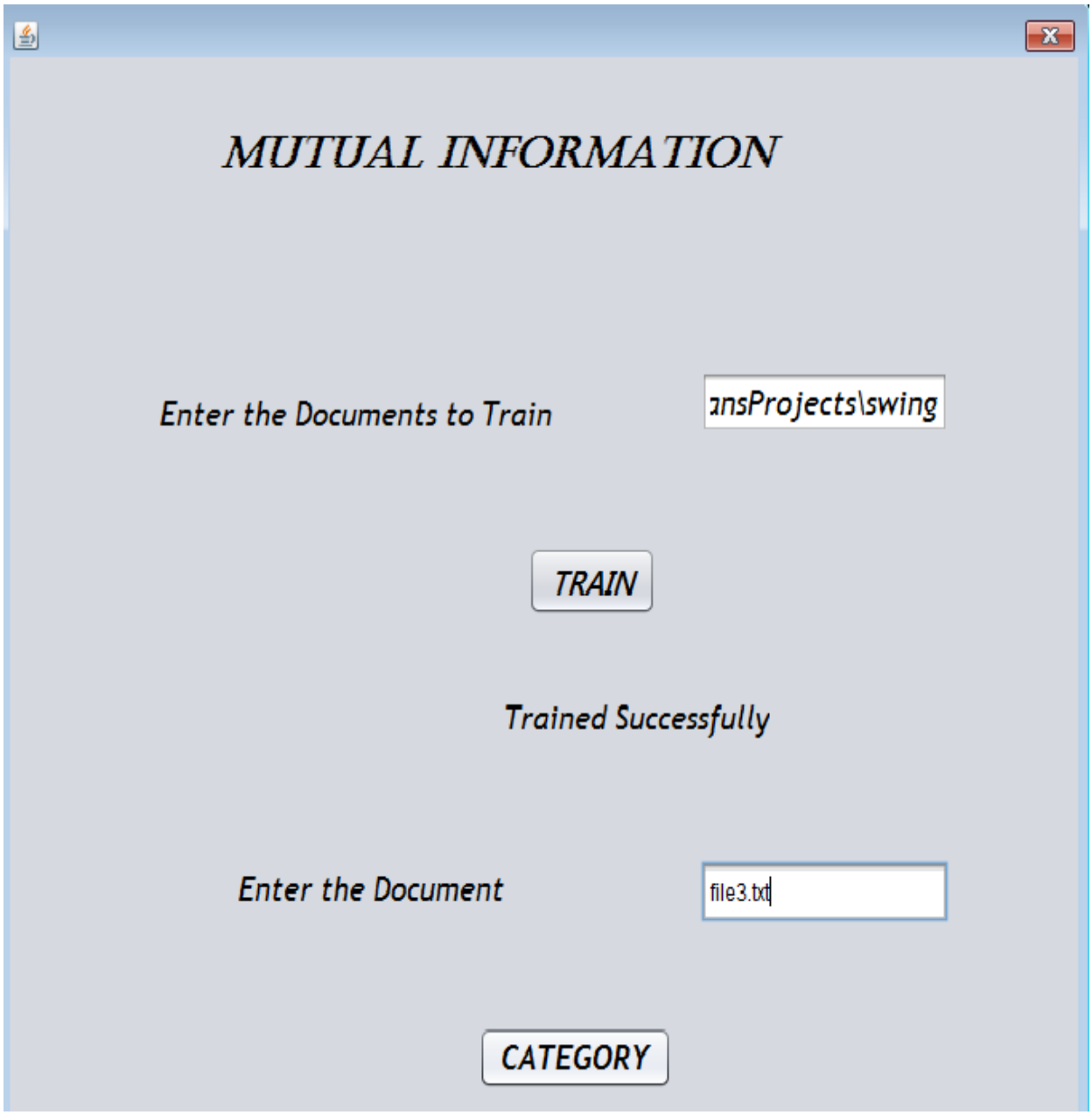


Figure 4.1.2 Test document

The output is shown below

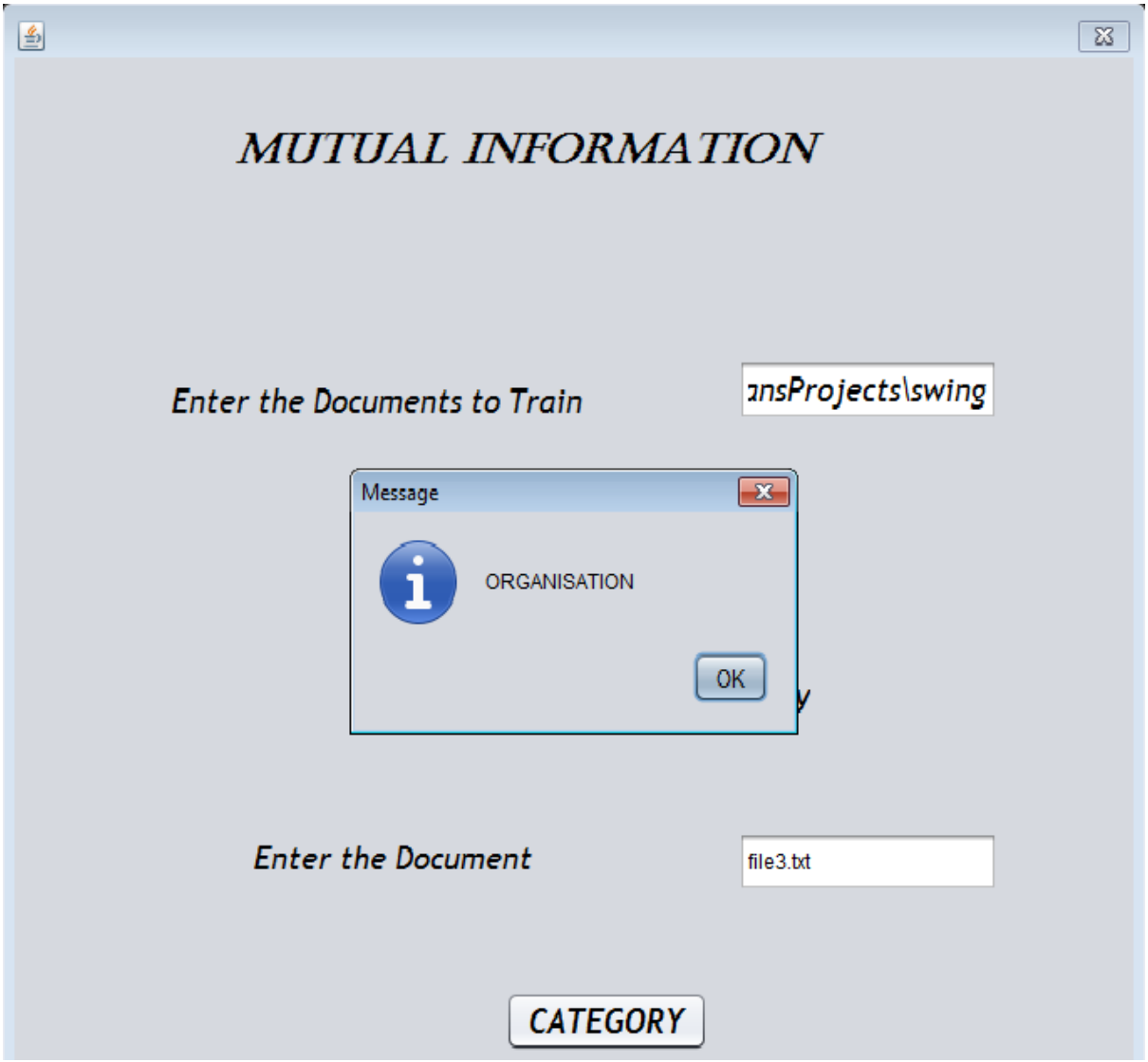


Figure 4.1.3 Output

CHAPTER 5
SYSTEM DESIGN

Systems design is the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering. If the broader topic of product development "blends the perspective of marketing, design, and manufacturing into a single approach to product development, then design is the act of taking the marketing information and creating the design of the product to be manufactured. Systems design is therefore the process of defining and developing systems to satisfy specified requirements of the user.

The logical design of a system pertains to an abstract representation of the data flows, inputs and outputs of the system. This is often conducted via modelling, using an over-abstract model of the actual system. In the context of systems design are included. Logical design includes ER Diagrams. The physical design relates to the actual input and output processes of the system. This is laid down in terms of how data is input into a system, how it is verified/authenticated, how it is processed, and how it is displayed as in physical design. Figure 5.1 demonstrates the system architecture in detail. The main steps of the system are the following:

5.1 ACTIVITY DIAGRAM

Activity diagram is an important diagram in UML to describe dynamic aspects of the system. Activity diagram is basically a flow chart to represent the flow form one activity to another activity. The activity can be described as an operation of the system. So the control flow is drawn from one operation to another. This flow can be sequential, branched or concurrent. Activity diagrams deals with all type of flow control by using different elements like fork, join etc.

The basic purpose of activity diagrams is to captures the dynamic behaviour of the system. Activity is a particular operation of the system. Activity diagrams are not only used for visualizing dynamic nature of a system but they are also used to construct the executable system by using forward and reverse engineering techniques.

5.1.1 PREPROCESSING

The first step is the pre-processing of the datasets. In this step each document is parsed, non-alphabetic characters and mark-up tags are discarded, case-folding is performed. At the end of these processes the category list, term list, term matrix and category matrix of the training documents and term matrix and category matrix of the test documents are created. Finally the feature weighting is calculated for each remaining word in the documents.

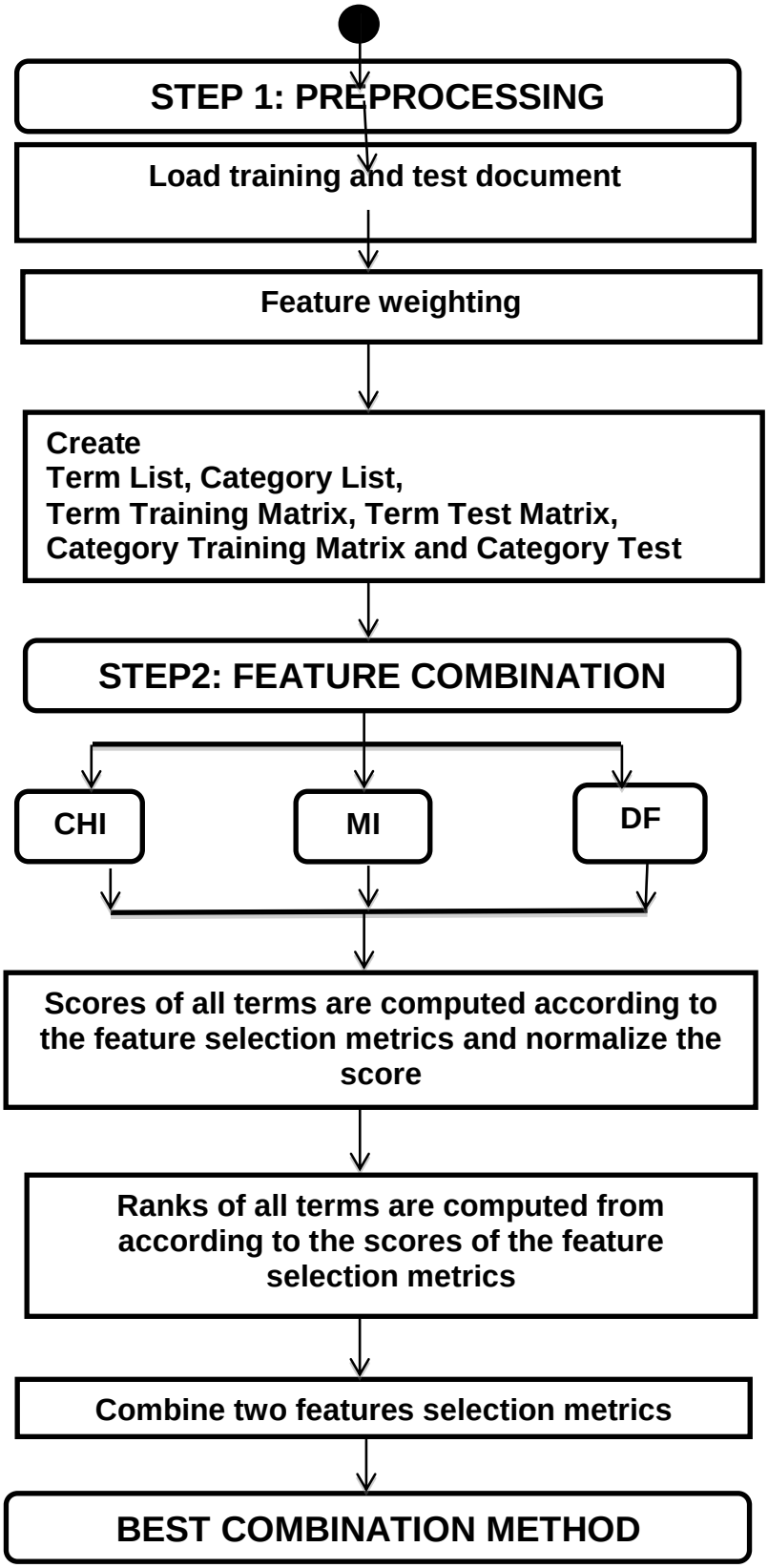
5.1.2 FEATURE COMBINATION

The second step combines the feature selection metrics with the varied combination approaches. Each feature selection metric computes the scores of the all terms and gives the higher scores to terms considered more informative for classification. In order to combine the outputs of different feature selection methods, the scores of each term from the varied feature selection methods are normalized. Then the ranks of the all terms are computed by ranking the scores of the terms with standard competition ranking. In this step all possible binary-combinations of different feature selection methods are generated.

5.1.3 FEATURE SELECTION

The third step is feature selection that reduces the dimensionality by ranking all terms according to their importance estimated by combination and then selecting a given number of terms from the term list with the highest values.

STEPS INVOLVED IN PREPROCESSION AND COMBINATION:



CHAPTER 6

SUMMARY AND RESULT

The experiments show that substantial improvements can be achieved in text categorization by combining feature selection methods. Although many feature selection methods exist in text categorization, it is hard to state one is in general superior to others since the success of the methods depends on various variables. It is more likely that combining different feature selection methods obtains more effective performance in text categorization.

Comparing the performance of the individual methods with the performance of the combination methods demonstrate that combining two feature selection methods can significantly improve the performance of the individual methods in all dataset. In addition, in general rank combination achieves better performance than score combination in the case of global policy but score combination significantly achieves better performance in the case of local policy.

One of the important results of these experiments is that all combination methods improve the highest F-measure values of the individual metrics with almost all number of keywords from 10 to 2000. Especially, success of combining feature selection methods is more apparent when the keyword number is low. It is approved that more successful performances are achieved by less number of keywords compared to individual metrics. For individual metrics, the global policy is more successful than the local policy when the keyword number is high but it is outperformed by the local policy when the keyword number is low this assumption is still valid after combining feature selection methods.

As a feature work, we will test the combination of the multiple feature selection metrics. In this study we only focus on binary combinations since the previous studies conclude that the best results achieved by combining two feature selection metrics but their experimental setting is very limited. Thus, it is necessity to see the results of the combination of more than two methods in order to make a clear conclusion. In addition we plan to extend the experiments with new feature selection methods.

REFERENCES

[1] Sebastiani, F., “Machine Learning in Automated Text Categorization”, *ACM Computing, Surveys*, Vol. 34, No. 1, pp. 47, 2002.

[2] Yang, Y. and J. O. Pedersen, “A Comparative Study on Feature Selection in Text Categorization”, *Proceedings of the Fourteenth International Conference on Machine Learning*, pp.412-420, July 08-12, 1997.

[3] Forman, G., “An Extensive Empirical Study of Feature Selection Metrics for Text Classification”, *Journal of Machine Learning Research 3*, pp. 1289–1305, 2003.

[4]Joachims, T., “Text categorization with support vector machines: Learning with many relevant features”, *Proceedings 10th European Conference on Machine Learning (ECML), Springer Verlag*, 1998.

[5] Zhang, W., T. Yoshida, X. Tang, “Text classification based on multi-word with support vector machine”, *Knowledge-Based Systems 21*, pp.879–886, 2008.

[6]Eceozbilen,” Improving text categorization performance by combining feature selection methods”.

[7] Isabelle Guyon,” An Introduction to variable and feature selection”,2003.

[8]Yi-Wei Chen And Chih-Jen Lin,” Combining Svms With Various Feature Selection Strategies”.

[9]WlodzislawDuchAnd Karol GrudziNski,” *Weighting And Selection Of Features*”,
June 14-18 , 1999.

[10] Jana Novovi·cova.Antonin Malik, and PavelPudil,“*Feature Selection using Improved Mutual Information for Text Classification*”.

[11]Tasci, S., “An evaluation of existing and new feature selection metrics in text categorization”, Computer Engineering, Bogazici University, 2006.

[12]Li, Y., D. F. Hsu and S. M. Chung “Combining Multiple Feature Selection Methods for Text Categorization by Using Rank-Score Characteristics”, *International Conference on Tools with Artificial Intelligence - ICTAI* , pp. 508-517, 2009.

[13]Olsson, J. S. and D. W. Oard, “Combining Feature Selectors for Text Classification”, *CIKM’06*, 2006.

[14]Tax, D. M. J.; van Breukelen, M.; Duin, R. P. W. and Kittler, J., “Combining Multiple Classifiers by Averaging or by Multiplying”, *Pattern Recognition 33 (9)*, pp.1475 – 1485, 2000.

[15] Leopold, E. and J. Kindermann, “Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? ”, *Machine Learning*, 46, 423–444, 2002.