Dewy index based Arabic Document classification with Synonyms Merge Feature Reduction

E. M Saad, M H Awadalla and A F Alajmi

**Communication & Electronics Dept., Faculty of Engineering, Helwan University
Egypt**

Abstract

Feature reduction is an important process before documents classification. The classification performance is impact by the quality of the selected. A new semantic approach is presented using synonym merge to preserve features semantic and prevent important terms from being excluded. The resulting feature space were then processed with five feature selection methods, ID, TFIDF, CHI, IG and MI. experiment show that classification performance is increased after merging terms and yielding best performance for CHI and IG selection method. A promising classification technique is presented based on Dewey decimal classification system, which uses filtered indexes and three levels of classes from Dewey system to classify and label Arabic documents. The technique shows along with synonyms merge a promising result.

Keywords: *Dimension reduction, Arabic text Classification, synonyms.*

1. Introduction

Over the past decade there was a rapid growth of data, which raised the question of the ability to process accurate information from text. Text mining and, natural language processing aim at information extraction and organizing. The tasks of text mining such as classification assigns labels to texts (documents) based on their similarities to the assigned classes, Whereas, clustering groups texts which have similar conceptual contents. Text summarizing, summarizes the text according to its base idea, and question answering searches a precise answer for a given question from a collection of texts. Whichever the task in question, the collection of text (Data set) must be processed first in order to extract useful information. One of the main preliminary requirements of text processing is feature extraction. In the context of text, features are words that appear in the text, but words at it appears in text are not a very good representation. Further processing is needed to produce meaningful and unified form of features, such as stemming, and stop word removing, which are grammatical words that by removing the efficiency of classification increases. Features are then represented. vector space model (VSM) [1] is a is an algebraic model for representing text documents, which uses category as the class label and the word or phrase as features of vector space. There are different methods to generate the values of features such as statistics based, grammar based, semantic based, etc.

There are two problems associated with features, the large scale of features, and the importance of a particular feature for understanding the text in question. For many applications dimension reduction is needed before further processing. Dimension reduction (DR) is done to reduce the large feature space extracted from the text. Another reason for reducing feature space is to eliminate noisy components to better represent [2], because it could handle other common words that were not eliminated by the stop word removal process. Dimension reduction (DR) also have the advantage of reducing the computing time, and cost, thus makes it more suitable requirement wise [3] for further processing to improves the text classification (task) efficiency and performance and, provides better data understanding that improves the clustering and classification result [4] by removing redundant and irrelevant terms from the corpus.

There are two classes of dimension reduction techniques, feature selection (FS), and feature extraction (FE) [3][5][6][7][8][9]. FE reduce data dimension by projecting the high dimensional data into lower space through transformations. FS searches a subset of the most representative features according to some criterion.

The goal of this paper is to present a semantic approach for dimension reduction. The approach is based on the fact that different words may have similar meaning (synonyms), and by looking at the count of the words in a text (TF, DF, etc.) alone we may exclude important features from the list of feature which might affect the quality of the classification process. Thus a synonym merge approach is presented to reduce the dimension of the feature space and maintain the semantic relation between the features. Furthermore a Dewey based categorization technique is presented, which uses a derived list of indexes from Dewey Decimal indexing system and three level of classification labels. Where a document may belong to more than one class and the classes are labeled using Dewey main category and two subcategories.

In the next section a review of related work to dimension reduction is presented. Section 3 shows different reduction techniques. Section 4 discusses the synonym merge reduction technique. In section 5 a Dewey index based classification method is explained. Experiment

result is shown in section 6 and, work is concluded in section 7.

2. Related Work

There are two classes of dimension reduction techniques, feature selection (FS), and feature extraction (FE). Feature selection selects a representative subset of the input feature set, based on some criterion. An estimate function is used to rank original features according to the calculated score value for each feature. This value represents the quality or importance of a word in the collection. The features then ordered in descending or ascending order to the values, and then select a suitable number of words of highest orders.
Feature selection methods are classified into four categories [11]: filter, wrapper, hybrid and embedded, the hybrid method has better efficiency in time complexity and granularity.
The two main feature selection methods are wrapper and filtering [2], [10]:

1- Wrapper (unsupervised): select a subset of features by evaluation function based on learning algorithms that will take these selected features [2]. Also called subset selection [4], it searches the set of possible features for the optimal subset. Wrapped method applied to high dimension feature space is NP hard optimization problem, which can hardly run effectively due to the time complexity
2- Filtering (supervised): The base in a filtering method is to set up a criterion (evaluation function, score function, feature evaluation index, filtering function) for measuring the feature so that whether it should be remained can be then decided. The most important features, whose evaluation function values are the largest or the smallest, are kept, and the others are filtered. [2]. Feature filtering ranks the features by a metric and eliminates all features that do not achieve an adequate score [4] defined by the threshold. The speed of filter method is fast because it does not consider the combination of different feature, but it generally catches coarse results only and that is why it is simple but efficient [7].

Examples of score function used are document frequency (DF), Chi-square (CHI) (X2 statistics), information gain (IG), mutual information gain (MI), term strength, term contribution (TC)[2]. IG is one [12] of the most effective techniques. term frequency (TF), document frequency (DF), information gain (IG), mutual information (MI), odd-ratio, Chi-square [13][11].

On the other hand, feature extraction methods transforms the original features into new, lower dimension space and creates new features, by computing new features as some function of the old ones. They are categorized into linear and non-linear algorithms [6]. The new space sometimes called "concept Space" [3] or latent sub-space [7]. Feature Extraction is helpful in solving the problems related to synonymy and polysemy [9], were the loss of great part of [8] useful information of original feature set are avoided. Transformed features generated by feature extraction may provide a better discriminative ability than the best subset of given features, but these new features may not have a clear physical meaning [8]. The complexity of feature extraction algorithms are often too high to be applied on large scale text processing tasks [6]. Furthermore, it is difficult to provide a direct semantic interpretation to the new features. Examples of linear feature reduction are fisher discriminant method, principal component analysis PCA, Latent semantic analysis LSA, linear discriminant analysis, maximum margin criterion (MMC) and orthogonal centroid algorithm (OCA). Non-linear feature extraction transformation algorithms are Locally Linear Embedded (LLE), ISOMAP and Laplacian Eigenmaps.

In the next two section a review of previous work done on feature reduction.

2.1 Feature Selection Methods:

Feature selection algorithms are widely used in the area of text processing due to their efficiency [6].
Fouzi .etl [20] compares five reduction techniques, root-based and, light stemming, document frequency DF, tfidf, and latent semantic indexing LSI. Then it shows that df , tfidf, and lsi methods were superior to the other techniques in term of classification problem. Furthermore, Savio [21] discusses four reduction methods DF, category frequency – Document frequency CF-DF, TFIDF, and principal component analysis PCA. The experiment shows that reduction of DF=15.2%, CF-DF=36.4%, TFIDF=78.8%, and PCA=98.9%. Which conclude that PCA is the most effective method in term of reduction with some decrease in classification effeciency.
Yang et al. [12] experimented with the first five score functions on Reuters21578 collection, and concluded that DF, CHI, and IG [14] are more effective than the others. The functions were Document frequency DF, Information gain IG, mutual information MI, X2-test CHI, and term strength.
Wang [2] Applies feature reduction method called variance-mean based feature filtering that aims at keeping the best features and at the same time improves performance. Features are represented as terms-documents matrix, and the values are the probability that term t will occur in document i. Then two vectors mean E and variance D are computed. The variance of E is computed to show degree of dispersion among classes and, the mean of D are computed, which shows average level of the degree of variability within every class term t can show. The bigger D(E) the more distinguishable among classes using that term w. and the smaller E(D) the more cohesive within each single class averagely using that term w is. So the more distinguishable among classes and cohesive within each single class using that

term w is the more possible the term t should remain. D(E) and E(D) based criterion can be used to evaluate the importance of the candidate term t. with the evaluation function F= β* D(E)/E(D) where β is a tuning parameter. If F>f (threshold) then the term is selected. Comparison with DF and CHI results in similar performance. Performance reaches 0.92 of macro-f1 value with order 400.

Wangi [16] investigates the use of Hill Climbing (HC), Simulated annealing (SA), and threshold accepting (TA) optimization techniques as feature selection techniques to reduce dimension of an e-mail, and improve the classification filter performance. It was found that that simulated annealing has the best performance. The approach starts with transferring the email to vectors of TF-IDF. Then apply the feature selection techniques to choose best discriminating feature sets. Performances were compared with Linear Discriminant Analysis and the accuracy was 90% for LDA, 93.6 for HC, 94.6 for TA and 95.5 for SA.

Zifeng [5] propose an approach of feature selection based on Constrained LDA (CLDA) to overcome the problem of un-interpable features resulting from LDA transformation. Like LDA, the proposed approach will find a subset of features which maximize the discriminant capability between classes with a linear. The work is based on selecting but not transforming features by LDA to preserve structure information between-class and within-class for text categorization. Constrained LDA (CLDA) models feature selection as a search problem in subspace and finds optimal solution subject to some restrictions. The CLDA is transformed into a process of scoring and sorting of features. Experiments on 20 Newsgroups and Reuters-21578 show that CLDA is consistently better than information gain IG and CHI with lower computational complexity.

Ren [22] propose an improved LAM feature selection algorithm (ILAMFS). The algorithm is based on combining the gold segmentation and the LAM algorithm based on the characteristics and the category of the correlation analysis, filtering the original feature set, and retaining the feature selection with strong correlation and weak category. Then, weighted average and Jaccard coefficient of feature subsets make redundancy filtering. Finally, obtains an approximate optimal feature subset. LAM algorithm has been improved, an improved LAM proposed feature selection algorithm (ILAMFS). the accuracy of selection in the threshold, feature selection and, efficiency of the running time are improved.

Xi [9] propose a two-stage feature selection method based on the Regularized Least Squares-Multi Angle Regression and Shrinkage (RLS-MARS) model. First, measure the features, and select the important features, by applying a new weighting method, the Term Frequency Inverse Document and Category Frequency Collection normalization (TF-IDCFC), and using the category information as a factor. Next, the RLS-MARS model is used to select the relevant information, while the Regularized Least Squares (RLS) with the Least Angle Regression and Shrinkage (LARS) can be viewed as an efficient approach. The experiments demonstrate the effectiveness of the new feature selection method for text classification in several classical algorithms: KNN and SVMLight. The performance of the new algorithm is similar to the feature selection by χ 2 CHI statistics less number of features was chosen and, outperforms χ 2 methods when the dimensionality grows higher.

Furthermore, features are selected using hybrid FS method based on improved particle swarm optimization PSO and, support vector machine SVM in JIN.[11], named FS_PSO_TC. It integrates the advantages of term frequency-inverse document frequency DF-IDF as inner-class measure and Chi-square as inter-class, and introduce a feature selection method based on swarm intelligence. The improved particle swarm optimization used to select fine features on the results of coarse grain filtering, and utilizing support vector machine to evaluate feature subsets and taking the evaluations as the fitness of particles. Experiments show the method reducing effectively the high dimension while catching better categorization efficiency.

A collaborative filtering method in [15] is used to reduce feature space, which utilizes traditional feature reduction techniques along with a collaborative filtering method for to predict the value of missing features for each class. Information Gain (IG) used to identify non-trivial noun phrases with semantic meanings in the documents, noun phrases (NP) chunking is adopted for this purpose. Chunking groups together semantically related words into constituents. Experiment indicates an improvement in classification accuracy over the traditional methods for both Support Vector Machines and AdaBoost classifiers.

2.2 Feature Extraction Methods

Feature extraction methods project the high dimension feature space in to a lower one. Furthermore, rough set theory has been applied to feature reduction area. [14] Rough set theory can discover hidden patterns and dependency relationships among large number of different feature terms in text datasets. No additional information about the data is required such as thresholds or domain knowledge. Essential part of information can be identified through generating reducts (i.e., the minimal sets of attributes which has the same distinguish capability as the original set of attributes), thereby reducing the irrelevant /redundant attributes as well as maintain the discemibility power with respect to the task of pattern recognition or classification [14].

Jensen [13] reviews techniques that preserve the underlying semantics of the data, using crisp and fuzzy rough set-based methodologies. This paper reviews techniques, which employs rough set based methodology which belongs to rough sets, fuzzy rough sets, and rough set-based feature grouping. It was shown how fuzzi-fying a particular evaluation function, the rough set dependency degree, can lead to group and individual

selection based on linguistic labels—more closely resembling human reasoning.

A set-based case-based reasoning (CBR) approach is proposed [14] to tackle the task of text categorization (TC). The initial work of integrating both feature and document reduction/selection in TC using rough sets and CBR properties is presented. Rough set theory is incorporated to reduce the number of feature terms through generating reducts. Two concepts of case coverage and case reachability in CBR are used in selecting representative documents. The main idea is that both the number of features and the documents are reduced with minimal loss of useful information. Experiments on the text datasets of Reuters21578 show that, although the number of feature terms and documents are reduced greatly, the problem-solving quality in tem of classification accuracy is still preserved. In average, 43.6% documents can be reduced and the classification accuracy is still preserved.

Chouchoulas [23] proposed a Rough Set-based approach (RSAR) and test it using Email messages; based on their work, Bao developed a rough set-based hybrid method using Latent Semantic Indexing (LSI) and Rough Set theory to TC

Cheng and Zhang [18] propose a method TFERS based on rough set theory and correlation analysis, in which a new formulation for attribute importance is proposed based on the classification capability of attributes. This formulation also avoids the recalculation of attribute importance. In text preprocessing phase, the term vector space representation of text is extended to concept ('synset') level based on Wordnet. As a result, dimension of the feature vector is reduced. A complete text feature extraction method TFERS is proposed, which includes text preprocessing, construction of the text feature vector, calculation of attribute significance, and attributes reduction. In the process of attributes reduction, correlation analysis is incorporated in order to get satisfactory feature reduction. The results of the simulation experiment and text classification show the validity of TFERS.

Shaiei and wang [17] conduct a study on three different document representation methods for text used together with three Dimension Reduction Techniques (DRT). The three Document representation methods considered are based on the vector space model, and they include word, multi-word term, and character N-gram representations. The dimension reduction methods are independent component analysis (ICA), latent semantic indexing (LSI), and a feature selection technique based on Document Frequency (DF). Results are compared in terms of clustering performance, using the k-means clustering algorithm. Experiments show that ICA and LSI are clearly better than DF on all datasets. For word and N-gram representation, ICA generally gives better results compared with LSI. Experiments also show that the word representation gives better clustering results compared to term and N-gram representation. The results show that for all datasets, clustering quality using ICA is better than using LSI in the whole range of dimensionalities investigated. For low dimensionalities, especially lower than 50, for all datasets, the DF based method has the worst performance among the dimension reduction methods used.

Ren[19] performs dimension reduction using Linear discriminant analysis (LDA) to maximize class separability in the reduced dimensional space.

Haifeng [8] present a weighted method based on the sample distribution, which will make the between-class and within-class scatter matrixes with poor scatter be weighted, to enhance the categorization ability after dimensional reduction and to improve the dimensional reduction effect of linear feature extraction method based on scatter difference. experiment show this method is superior to the original maximum scatter difference method in precision rate and recall rate

Thang and, cheng [7] propose a feature reduction method based on probabilistic mixture model of directional distributions. The main idea states that if documents can be viewed as directional data, so can words in the current context. Attributes of a word data point will be its frequencies of appearance in the documents. A mixture model of von Mises-Fisher distributions is applied for clustering the word space, Which results in a set of mean vectors, each of which potentially represents a group of words of the same topic. A projection matrix is then created based on word-to-mean cosine measure, by calculating the cosine distance of each word-mean vector pair. Hence, after the linear transformation, documents in the reduced-dimension space have the number of attributes equal to the number of potential topics in the document corpus [7]. A mixture of distributions is utilized to decompose the word space into a set of sub-topics, which are represented by their mean vectors. Through this matrix, the document corpus is transformed into a new feature space of much lower dimension. Experiments on various benchmark datasets shows that proposed method performs comparably with Latent Semantic Analysis (LSA), and much better than standard methods such as Document Frequency (DF) and Term Contribution (TC).

There was some work done to produce a universal dimension reduction methods [3].were both feature selection and feature extraction techniques are applied.

Zhu [3] applied feature selection and feature extraction to SVMs. In the feature selection case, experimental results show that when the linear kernel is used for SVMs, the performance is close to the baseline system, and when nonlinear kernel is employed, feature selection methods get the performance decrease sharply. On the contrary, principal component analysis (PCA), one of feature extraction methods, gets excellent performance with both linear and nonlinear kernel functions. It examines the ability of feature selection methods to remove irrelevant features, and combine PCA, a feature extraction method, with SVMs as a solution to the problem of synonym, and study the influence of different kernel functions on the performance of dimension reduction methods. Experimental results over

two different datasets show that when the linear kernel is employed for SVMs, feature selection methods achieve better performance compared to the baseline system. However, when the polynomial kernel is combined with feature selection methods, the performance decreases dramatically, and is much worse than the baseline system. On the contrary, PCA perform well no matter which kernel is employed. Employ dimension reduction methods, feature selection and feature extraction, for SVMs as the preprocessing of text categorization.
Ning and Yang [6], propose feature selection algorithm called Trace Oriented Feature Analysis (TOFA). The main function of TOFA is a unified framework that integrates feature extraction algorithms such as unsupervised Principal Component Analysis and supervised Maximum Margin Criterion. TOFA can process supervised problem and, unsupervised and semi-supervised problems. Experimental results on real text datasets demonstrate the effectiveness and efficiency of TOFA. The main contributions of this paper are: (1) by formulating the feature extraction algorithms as optimization problem in continuous solution space a unified feature extraction objective function, where many commonly used previous work are special cases of this unified objective function; (2) by formulating the feature selection algorithms as the optimization problem in a discrete solution space, we propose a novel feature selection algorithm by optimizing our proposed unified objective function in the discrete solution space; and (3) through integrating the objective function of feature extraction and solution space of feature selection, our proposed feature selection algorithm can find the optimal solution according to the unified objective function. Experimental results on real text datasets show the effectiveness and efficiency of TOFA for text categorization.

2.3. Classification

Text classification is an important task of text processing. A typical text classification process consists of the following steps: preprocessing, indexing, dimensionality reduction, and classification [24]. A number of statistical classification and machine learning techniques has been applied to text classification, including regression models like linear least square fit mapping LLSF [12], K-Nearest Neighbor classifiers [5][9][12], Decision Tress, Bayesian classifiers, Support Vector Machines [2][3][15][19][9][11][6-6], Neural Networks [20][21] and, AdaBoost [15]. SVM has been applied to text classification [11] and achieved remarkable success, [17] proposed a hybrid method used transductive support vector machine (TSVM) and simulated annealing (SA), which selected top 2 thousands high CHI-square value features to form dataset and gained better classification results compared to standard SVM and TSVM.

3. Feature Selection techniques:

3.1 Stemming:

Root based stemming and light stemming results in a considerable reduction in feature dimension. Root-based depends on pattern matching to extract the root of the word after prefix, suffix, and infix removal. This technique reduces over 40% of the feature but it does not preserve the semantic of the features, because the root could generate many words with different meaning. on the other hand, light stemming which strips off prefixes and suffixes without removing the infixes has more ability of preserving the meaning and reduces the feature space 30%, but it still have the problem of dealing with two similar words (in-term of semantic) as different feature because of their difference in infixes.
We developed a technique based on morphological word weights using HMM [25]. A hidden markov model is used to match a word with a pattern and thus remove prefixes and suffixes. The pattern then transformed to a unified pattern called Masdar (مصدر). This technique reduces the feature space by 40% and at the same time preserves the semantic of the features.

3.2 Document frequency (DF)

Document frequency refers to the number of documents that a feature appears in. The selection of features is based on the high value of DF. By experiment 60% reduction achieved by removing terms which occurs only in one document. DF can be used as a criterion for selecting good terms [17]. The main idea behind using document frequency is that rare terms either do not capture much information about one category, or they do not affect global performance. DF is simple and, as effective as more advanced feature selection methods [10-24].

3.3 Term Frequency- Inverse Term Frequency (TFIDF):

Term-Frequency-Inverse-Term-Frequency is formulated as:

$$\text{tf-idf}(t,d) = \text{tf}(t,d) \times \text{idf}(t)$$

Where the term frequency refers to the number of occurrences of term *i* in document and, the inverse document frequency is a measure of the general importance of the term (obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient).
The *inverse document frequency*

$$\text{idf}(t) = \log \frac{|D|}{|\{d : t \in d\}|}$$

with

- $|D|$: the total number of documents in the document set.

- $|\{d : t \in d\}|$: number of documents where the term t appears.

A high weight in tf–idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. The tf-idf value for a term will be greater than zero if and only if the ratio inside the idf's log function is greater than 1. Depending on whether a 1 is added to the denominator, a term in all documents will have either a zero or negative idf, and if the 1 is added to the denominator a term that occurs in all but one document will have an idf equal to zero.

3.4 X^2 statistic Chi-square

X2 statistic (CHI) measures the lack of independence between term and category. Then it compares with x2 distribution to measure the degree of freedom to judge extremeness.
Term goodness measure is expressed by:

$$X^2(t,c) = \frac{N \times (AD - CB)}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$

Where,
T: term
C: category
A: number of times t and c co-occure
B: number of time the t occur without c
C: number of times c occure without t
D: number of times neither C nor t occur
N: the total number of documents.

If t and c are independent then X2 statistic equal 0. The X2 statistic is computed for each category between each unique term and a training corpus and that category.
Then combine the category-specific scores of each term into two scores:

$$X^2_{avg}(t) = \sum_{i=1}^{m} P_r(C_i) X^2(t, C_i)$$
$$X^2_{max}(t) = \max_{i=1}\{X^2(t, C_i)\}$$

3.5 Information Gain:

IG measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document [5]. Giving a corpus of training text, we compute the information gain of each term, and then remove those features whose information gain was less than some pre-determined threshold.
$\{C_i\}_{i=1}^{m}$ is the set of categories in the target space.
IG of term t is:

$$G(t) = -\sum_{i=1}^{m} P_r(C_i) \log P_r(C_i) + P_r(t) \sum_{i=1}^{m} P_r(C_i|t) \log P_r(C_i|t) + P_r(t) \sum_{i=1}^{m} P_r(C_i|t) \log P_r(C_i|t)$$

After computing IG, remove terms whose information gain less than predefines threshold.

3.6 Mutual Information (MI)

MI is commonly used in statistical language modeling of word associations and related applications [12].
MI between c and t is defined as:

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

Then,

$$I(t,c) = log \frac{A \times N}{(A + C) \times (A + B)}$$

Where,
A: number of times t and c co-occur
B: number of times t occurs without c
C: number of times c occur without t.
N: total number of documents.
If t and c are independent then I=0.
To measure thegoodness of a term two measures are defined:

$$I_{avg}(t) = \sum_{i=1}^{m} P_r(C_i) I(t, C_i)$$
$$I_{max}(t) = \max_{i=1}\{I(t, C_i)\}$$

4. Synonym Merge Reduction

The main idea behind synonym merge is to preserve important terms from being excluded. Term-document matrix is constructed to apply merge over all training set. A dictionary of synonyms [25] is used, were terms with similar meaning are giving one group code. Checking for terms synonym then merging terms with same group id into one feature. Word weights are extracted After a text is being processed and, tokenized [26].

Algorithm (construct synonym tree) Figure 1.
A synonyms tree is constructed from list of synonyms in such a way:
- Null node having 28 branches constituting the Arabic alphabet.
- Each of the 28 node have 28 branches and so on until a word is constructed from root node to the n-1 node (n is the number of letters in a word)
- A leaf node contains the group ID.

Algorithm (synonym check)

Do for all terms
- From root select branch which contains the next word letter.

Continue until last term letter.
- If word does not belong to a synonym group the return null.
- If group found then return group ID.
- Merge all features with the same synonyms group id.

The result of the synonym check algorithm is a semantically reduced feature space. The resulting features are processed with a feature selection method to produce the final feature space that are used for later processing such as classification and clustering.

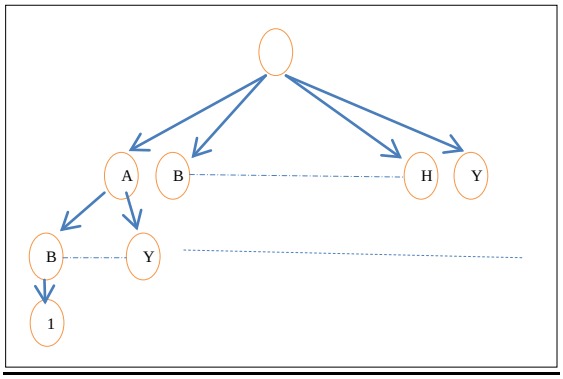

Fig.1 Synonym Tree

5. Dewy based Categorization

The Dewey Decimal Classification (DDC) is a proprietary system of library classification. The DDC attempts to organize all knowledge into 10 main classes [27] (table.1). The ten main classes are each further subdivided into ten divisions, and each division into ten sections, giving ten main classes, 100 divisions and 1000 sections. The system is made up of seven tables and ten main classes, each of which is divided into ten secondary classes or subcategories, each of which contain ten subdivisions.

Table.1: The main 10 classes

Class ID	***Class Name***
000	Computer science, information and general works
100	Philosophy and psychology
200	Religion
300	Social sciences
400	Language
500	Science (including mathematics)
600	Technology and applied Science
700	Arts and recreation
800	Literature
900	History, geography, and biography

Three levels of classes were used in this research and, the indexes were filtered to match the feature format, which is the morphological weight of the terms.

Algorithm (Dewey Classification)
For all terms in feature space do
- Search the term in the index list.
- If found set class ID to term.

End
Check for terms class ID:
- Choose class(s) which has the majority of terms belonging to.
- A document may belong to more than one class.
- Label document with class label.

An example of the classifying financial documents in Figure.2.

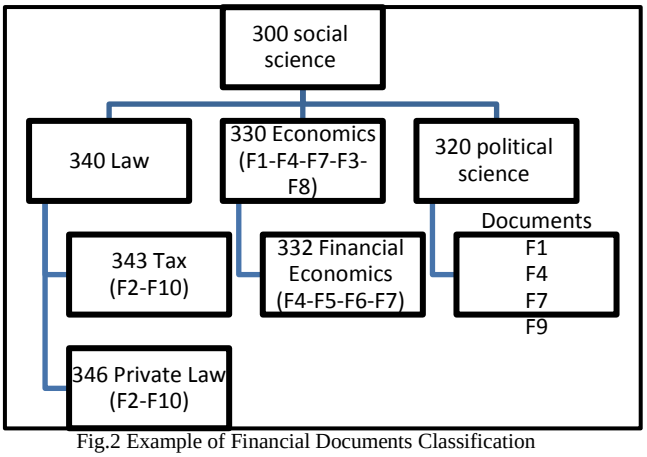

Fig.2 Example of Financial Documents Classification

6. Experiment and results

6.1 Data set:

An in house Arabic documents data set was used in this experiment. Documents represent various categories, news, finance, sport, culture, and science. The documents were tokenized to produce bag of words by eliminating unwanted characters. Then stop words were removed from the resulting documents. Removing stop words reduced the size of the features by 25% in average. Table 1. Shows the number of documents divided into the categories.

Table.2: The Date Set

Category name	***Number of documents***
News	50
Finance	30
Sport	30
Culture	60
Science	40
Total	210

Several reduction techniques were applied to the data which are shown in table 2.

6.2 Evaluation Criteria:

Text classifiers performance was evaluated using the F1 measure. This measure combines recall and precision in the following way:

Precision:

$$P = \frac{number\ of\ correct\ classes\ found}{number\ of\ classes\ found}$$

Recall:

$$P = \frac{number\ of\ correct\ classes\ found}{number\ of\ correct\ classes}$$

F-measure:

$$F = \frac{2.R.P}{R+P}$$

6.3 Results:

Figure.1 displays the performance curve for classification of the training set after term selection using IG, DF, MI, TFIDF, CHI respectively. Figure.2 shows the performance curve for classification of the training set for term selection using IG, DF, MI, TFIDF, CHI respectively after applying the synonym reduction processes discussed earlier.

Table.2: The Date Set

Dimension Reduction Technique	*Average Percentage of reduction*	*Average Percentage of reduction w/ syn merge*
DF	83%	87%
Tfidf	78%	82%
CHI	90%	93%
IG	89%	91%
MI	60%	72%

Comparison between fig.1, and fig.2 result shows improvement in both number of reduced features and classification performance measure F1-measure.

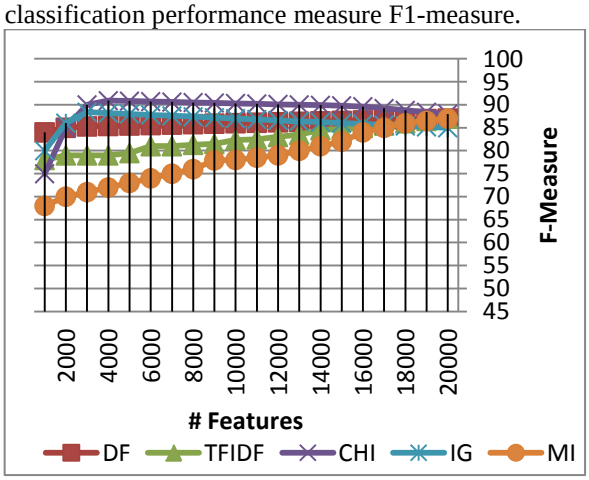

Fig.3 Classification performance without synonym merge

The computation of CHI, DF, and MI are similar to that of IG.The differences are the approaches to rank features. However, MI is not comparable with IG, DF, and CHI on text categorization. CHI and IG gives the best performance overall discussed selection methods.

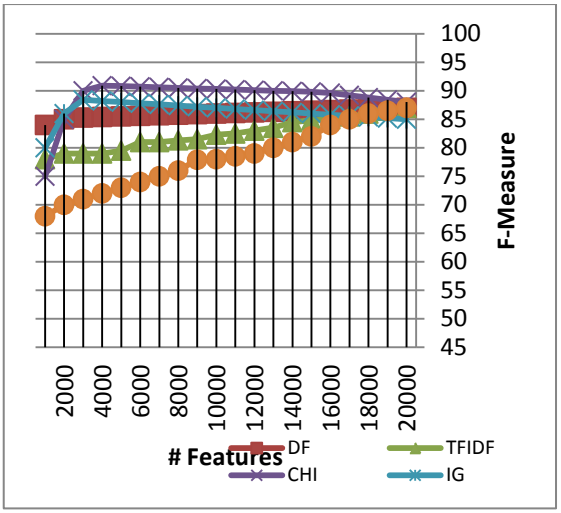

Fig.2 Classification performance after synonym merge

7. Conclusion:

In this paper a semantic feature reduction approach was presented with a Dewey based classification algorithm. The reduction is based on synonyms merge to overcome the problem of feature synonyms excluded during feature selection process. Terms with similar meaning are merged into one group then the resulting groups are used as the new features which results in two advantages, one is reducing the feature space, and the other is preserving the sematic of the feature without relying into complex methods. Five feature selection methods were applied after synonym merges, DF, TFIDF, CHI, IG, and MI to produce a more compact feature space. Experiment on using those methods with and without the synonym merge results in improvement of the feature reduction and the classification performance presented by the F-measure. Furthermore, a new approach based on Dewey indexing is presented. It classifies documents based on filtered version of Dewey indexes, and uses a hierarchy structure of three levels to produce labeled overlapped classes. Over the five features selection methods used CHI, and IG shows the best performance.

As a future word feature extraction methods will be experimented on with the new classification technique to decide the best performance yielding method.

8. References:

[1] [1] G. Salton, and M.l McGill, "An Introduction to Modem Information Retrieval", McGraw-Hill, 1983.

[2] [2] YI WANG, and XIAO-JING WANG, "A NEW APPROACH TO FEATURE SELECTION IN TEXT CLASSIFICATION" , Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005

[3] [3] Muhua ZHU, Jingbo ZHU, and Wenliang CHEN, "Effect Analysis of Dimension Reduction on Support Vector Machines", Proceeding of NLP-KE'05
[4] [4] K. Mugunthadevi, S.C. Punitha, M. Punithavalli, and K. Mugunthadevi, "Survey on Feature Selection in Document Clustering", International Journal on Computer Science and Engineering (IJCSE), 2011
[5] [5] Cui Zifeng, Xu Baowen , Zhang Weifeng, Jiang Dawei, and Xu Junling, "CLDA: Feature Selection for Text Categorization Based on Constrained LDA", International Conference on Semantic Computing, 2007
[6] [6] Jun Yan, Ning Liu, Qiang Yang, Weiguo Fan, and Zheng Chen, "TOFA: Trace Oriented Feature Analysis in Text Categorization", Eighth IEEE International Conference on Data Mining, 2008.
[7] [7] Nguyen Duc Thang, Lihui Chen, and Chee Keong Chan, "Feature Reduction using Mixture Model of Directional Distributions", 10th Intl. Conf. on Control, Automation, Robotics and Vision, Hanoi, Vietnam, 17–20 December 2008
[8] [8] Liu Haifeng, Su Zhan, Yao Zeqing, and Zhang Xueren, "A Method of Text Feature Extraction Based on Weighted Scatter Difference", Second WRI Global Congress on Intelligent Systems, 2010.
[9] [9] LI Xi, DAI Hang, and WANG Mingwen, "Two-stage Feature Selection Method for Text Classification", International Conference on Multimedia Information Networking and Security, 2009.
[10] [10] Liu, T., Liu, S., Chen, Z. and Ma, W.Y. An Evaluation on Feature Selection for Text Clustering. In Proceedings of the Twentieth International Conference on Machine Learning (ICML'03), 2003
[11] [11] Yaohong JIN, Wen XIONG, and Cong WANG, "Feature Selection for Chinese Text Categorization Based on Improved Particle Swarm Optimization", IEEE, 2010.
[12] [12] Yiming Yang, and Jan O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization", Proceedings of 14th International Conference on Machine Learning, San Francisco, pp.412-420, 1997
[13] [13] Richard Jensen, and Qiang Shen "Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approaches", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 16, NO. 12, DECEMBER 2004
[14] [14] Yan Lps,I Mon Chi-Keung Shiu, Sankar Kumar Pal, And James Nga-Kwok Liu, "A Rough Set-Based Cbr Approach For Feature And Document Reduction In Text Categorization", Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, 26-29 August 2004
[15] [15] Yang Song1, Ding Zhou, Jian Huang, Isaac G. Councill, Hongyuan Zha, and C. Lee Giles "Boosting the Feature Space: Text Classification for Unstructured Data on the Web", Proceedings of the Sixth International Conference on Data Mining (ICDM'06)
[16] [16] Ren WangI, Amr M. Youssef , and Ahmed K. Elhakeem, "On Some Feature Selection Strategies for Spam Filter Design", IEEE CCECE/CCGEI, Ottawa, May 2006
[17] [17] Mahdi Shafiei, Singer Wang, Roger Zhang, Evangelos Milios, Bin Tang, Jane Tougas, and Ray Spiteri, "Document Representation and Dimension Reduction for Text Clustering", 2007 IEEE
[18] [18] Yiyuan Cheng, Ruiling Zhang, Xiufeng Wang, and Qiushuang Chen "Text Feature Extraction Based on Rough Set", Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008
[19] [19] Cheong Hee Park, "Dimension Reduction Using Least Squares Regression in Multi-labeled Text Categorization", IEEE, 2008
[20] [20] Comparing Dimension Reduction Techniques for arabic text classification using BPNN algorithm. First international conference on integrated intelligent computing, 2010
[21] [21] Savio L. Y. Lam, and Dik Lun Lee, "Feature Reduction for Neural Network Based Text Categorization" , 1999
[22] [22] Yong-gong Ren, Nan Lin, and Yu-qi Sun, "An Improved LAM Feature Selection Algorithm", Seventh Web Information Systems and Applications Conference, 2010.
[23] [23] A. Chouchoulas and Q. Shen, "Rough Set-Aided Keyword Reduction for Text Categorisation," Applied Artificial Intelligence, vol. 15, no. 9, pp. 843-873, 2001.
[24] [24] Guiying Wei, Xuedong Gao, and Sen Wu, "Study of text classification methods for data sets with huge features", 2nd International Conference on Industrial and Information Systems, 2010.
[25] [25] A. F. Alajmi, E. M. Saad and M. H. Awadalla, "Hidden markov model based Arabic morphological analyzer", International Journal of Computer Engineering Research Vol. 2(2), pp. 28-33, March 2011.
[26] [26] Mass'ad Abur-Rijaal, A pocket Dictionary of Synonyms and Antonyms, Librairie du Liban Publishers,2007.
[27] "The Alphabetic Dewey decimal index", Mohammed Sherief, Jomhoria Publishing, 1997.