Semantic Feature Selection Using WordNet

Stephanie Chua
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak.
hlchua@calm.unimas.my

Narayanan Kulathuramaiyer
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak.
nara@fit.unimas.my

Abstract

The web has caused an explosion of documents, requiring the need for an automated text categorization system. This paper explores the notion of semantic feature selection by employing WordNet [1], a lexical database. The proposed semantic approach employs noun synonyms and word senses for feature selection to select terms that are semantically representative of a category of documents. The categorical sense disambiguation extends the use of WordNet, which has been typically used for text retrieval and word sense disambiguation [2]. Our experiments on the Reuters-21578 dataset have shown that automated semantic feature selection is able to perform better than well known statistical feature selection methods, Information Gain and Chi-Square as a feature selection method.

1. Introduction

Text categorization is defined as assigning new documents to a set of pre-defined categories based on the classification patterns suggested by a training set of categorized documents [3]. One major difficulty in text categorization is the high dimensionality of the feature space. There are thousands of terms occurring in documents and it has been noted that learning algorithms cannot handle such a large number of terms. Feature selection is performed to reduce the dimensionality of the feature space. There are a number of feature selection methods, which over the years, are used in a wide range of text categorization tasks. Among the more popular statistics based ones are Document Frequency (DF), InfoGain (IG), Chi-Square (Chi2) and Mutual Information [4], [5].

The works by [6] has shown that approaches considering semantics are promising. Their work on cascaded feature selection extracts two sets of documents, one with all parts of speech (POS) and another with only nouns from the Reuters-21578 dataset. These terms are then looked up in WordNet and the synonyms associated with those terms are used as features for representing documents. Our work, however, explores the use of synsets and word senses to capture the actual context of a word's usage.

2. WordNet

WordNet is an online lexical reference system developed by the Cognitive Science Laboratory at Princeton University by a group led by Professor George A. Miller [1]. At present, WordNet contains more than 150,000 different unique terms. These are organized into some 115,000 word meanings or set of synonyms called synsets. The lexicon in WordNet is divided into nouns, verbs, adjectives and adverbs. Only nouns are considered in our research as preliminary experiments indicate that the use of other part-of-speech does not significantly enhance performance.

3. Using WordNet for feature selection

This research focuses on the feature selection process where WordNet is employed to discover synonymous terms based on cross-referencing. We will compare the WordNet synonyms approach with statistical methods such as Chi2 and IG. IG is a feature selection technique that makes use of the presence and absence of a term in a document to select its features. Chi2 measures the degree of independence between a term and a category. It selects features with high dependency on a particular category. The WordNet synonyms approach will explore the use of synonyms and word senses to derive a better set of features for category representation to achieve better categorization effectiveness.

In the WordNet synonyms approach, feature selection is based on terms with overlapping word senses co-occuring in a category. The co-occurrence of terms with the same synset signature is used as an indicator of significant terms to represent a category. For terms co-occuring in a category, the correct sense is determined based on the synset signature cross-referencing. Cross-referencing is done by checking the list of noun synsets

Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)
0-7695-2100-2/04 $ 20.00 IEEE

for all senses for similarity in the signatures. The senses of different terms with overlapping synsets aggregate the semantic context of a category. The original terms from the category that belongs to the similar synsets will then be identified and added as features for category representation. The WordNet synonyms approach is denoted as WN in all the figures.

The example below illustrates the approach used. Let us look at all the senses for five nouns; 'corn', 'maize', 'acquisition' and 'ship'. Each sense has a signature, which is referred to as a synset containing synonyms to reflect a sense.

Table 1. List of terms and their synsets for each sense

Terms	Synsets for all senses
Corn	Sense 1: {corn, maize, Indian corn, Zea mays} Sense 2: {corn} Sense 3: {corn, edible corn} Sense 4: {corn, clavus} Sense 5: {wheat, corn} Sense 6: {corn whiskey, corn whisky, corn}
Maize	Sense 1: {corn, maize, Indian corn, Zea mays} Sense 2: {gamboge, lemon, lemon yellow, maize}
Acquisition	Sense 1: {acquisition} Sense 2: {acquisition} Sense 3: {learning, acquisition} Sense 4: {skill, accomplishment, acquirement, acquisition, attainment}
Ship	Sense 1: {ship}

From Table 1, we can see that there are two identical synsets. Sense 1 of 'corn' and sense 1 of 'maize' have identical synsets with the same signatures. These synsets will be used in feature selection to select the original terms found in each category. Categorical sense disambiguation is performed here to automatically disambiguate semantically related terms. By finding identical synsets with the same signatures, the dominant senses can be determined for the synonymous terms in each category.

Semantically related terms that are found in each category are listed out from the list of similar synsets. Their frequencies are then obtained from the index file. Therefore, in the example shown, there are two features selected to represent the category. They are 'corn' and 'maize'. The rationale behind choosing synonymous terms in representing a category is that these terms have a stronger discriminative power in representing a category, as they are conceptually close.

Word sense disambiguation (WSD) refers to the process of disambiguating words by telling which sense an ambiguous word belongs to. WordNet has also been used in the process of WSD [2]. WSD is a complicated process where the sense of every term in every sentence is important. Our focus in this research is not on WSD. However, we simplify the WSD process in this research by identifying significant synsets for a category by finding the overlapping synset sense signatures. We applied a simplified version of WSD because past works have reported the importance of WSD in improving categorization performance [7], [8].

4. Experiments

The experiments were conducted using the multinomial naïve Bayes scheme from the Waikato Environment for Knowledge Analysis (WEKA) [9] machine learning tool. We chose this machine learning scheme because of its simplicity and fast processing. Although support vector machines (SVM) has been shown to outperform other schemes [10], our objective has not been to show which machine learning scheme is able to perform best. On the other hand, we are interested in the relative performance of the feature selection methods regardless of what machine learning schemes used.

The effectiveness of each feature selection method is evaluated using f-measure value, which gives the same weight to both precision and recall. Statistical methods like Chi2 and IG were compared with the WordNet synonyms approach to see how well a semantic approach performs as compared to a statistical approach.

Experiments were also carried out to see the significance of using the nouns from WordNet that occur in each category as features to contrast with the performance of the WordNet synonyms approach. Both experiments rank features according to their frequencies. WordNet nouns approach is denoted as WN1 in Figure 4 and 5.

5. Results and analysis

As a benchmark, we perform experiments on the Reuters-21578 dataset using the ModApte split [11]. This split assigns documents from April 7, 1987 and before to the training set and documents from April 8, 1987 and after to the test set. Only documents with an occurrence of at least one "Topics" category are used. Figure 1 shows the macro-averaged f-measure comparison between Chi2, IG and WN for the Reuters-21578 dataset. It shows that the WordNet synonyms approach is able to outperform the statistical feature selection methods, Chi2 and IG. However, this paper explores the performance of the WordNet synonyms approach on the top ten

categories of the Reuters-21578 "Topics" dataset to provide an insight as to why this approach is able to outperform the statistical approaches.

The macro-averaged and micro-averaged f-measure is reported for the experiments carried out on the top ten categories of the Reuters-21578 "Topics" dataset. The WordNet synonyms approach is able to outperform Chi2 and IG in all the experiments performed except at 10 and 20 terms. We will analyse why this is so in the following sections. The results for the macro-averaged f-measure and the micro-averaged f-measure for the Reuters-21578 top ten categories are shown graphically in Figure 2 and Figure 3 respectively.

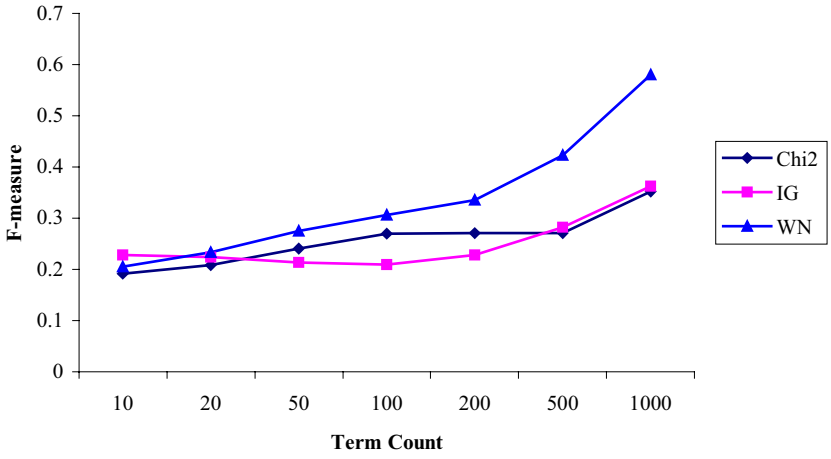

Figure 1. Macro-averaged f-measure comparison between Chi2, IG and WN for the Reuters-21578 dataset

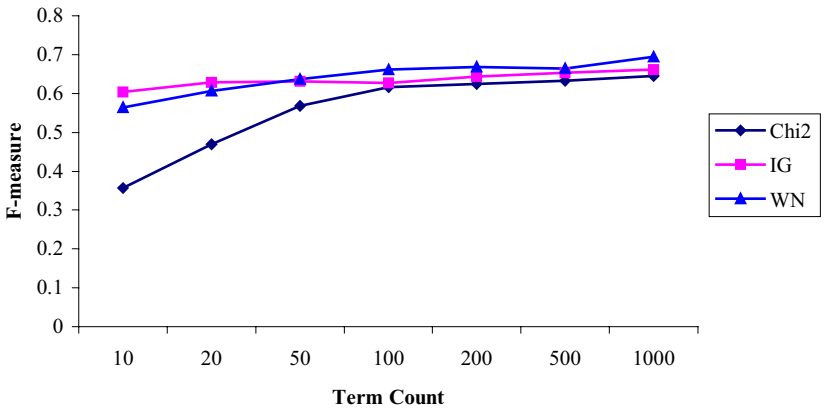

Figure 2. Macro-averaged f-measure comparison between Chi2, IG and WN for Reuters-21578 top ten categories

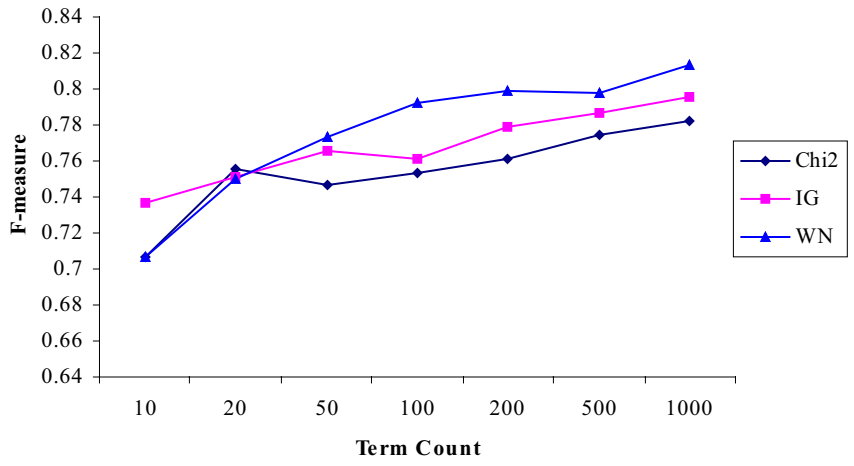

Figure 3. Micro-averaged f-measure comparison between Chi2, IG and WN for Reuters-21578 top ten categories

By choosing semantically related features, closer concepts can be identified within the category. This point will be illustrated in detail when we evaluate the performance of the categories individually. With the use of WordNet in feature selection, insignificant words can be filtered. These consist of non-English words, wrongly spelt words, insignificant abbreviations and names. Therefore, insignificant terms found in the dataset, such as, "twa", "shr", "lme", "natl" and so on will be filtered and omitted. Terms like "lt", "govodi" and "pik" are actually meaningless in representing a category. If human experts were asked to choose a set of words to represent a category, they would not have chosen those words that are mentioned. However, statistical feature selection techniques will include such words if the statistical data show that the words are significant statistically. These statistical techniques do not take into consideration whether a term is misspelt or is reflective of a category. By incorporating WordNet in feature selection, we can actually tackle this problem by filtering these insignificant terms and at the same time, make use of the available synonym relationship in WordNet to identify semantic features in a category.

From Figure 2 and 3, we can see that the WordNet synonyms approach generally performs better than the statistical feature selection methods, Chi2 and IG with the exception of results at term count 10 and 20. At term count 50 and above, the WordNet synonyms approach outperforms both Chi2 and IG.

In the individual category evaluation, we found that both category "Acq" and "Earn" have the highest number of training examples, thus producing the two highest values of f-measure. On the other hand, categories with a low number of training examples perform less effectively. This was proven in the correlation test that was carried out to see whether the number of training examples actually affects the performance of the classifier. We observe a strong correlation with a positive value of 0.81. Furthermore, another factor is observed, where categories with more synonyms found in WordNet, performed better. The correlation for the number of synonyms found and the performance of the classifier has the positive value of 0.88. Thus, both categories "Acq" and "Earn" again outperformed the other categories with "Corn" having the lowest performance.

To look into why WN performs less effectively at 10 and 20 terms, the results for the individual categories are analyzed. Table 2 shows the results for each category at term count of 10 and 20. The results printed in bold shows the highest of three f-measures.

Table 2. F-measure for each category for Chi2, IG and WN

Category/Term Count		Chi2	IG	WN
Acq	10	0.760	**0.790**	0.581
	20	0.847	**0.859**	0.731
Corn	10	0.034	**0.400**	0.385
	20	0.094	0.346	**0.408**
Crude	10	0.342	**0.723**	0.718
	20	0.620	**0.716**	0.715
Earn	10	0.910	0.908	**0.938**
	20	0.915	0.922	**0.942**
Grain	10	0.013	**0.581**	0.529
	20	0.603	**0.620**	0.556
Interest	10	**0.410**	0.404	0.402
	20	0.444	**0.561**	0.410
Money-fx	10	0.306	**0.579**	0.517

	20	0.474	**0.608**	0.595
Ship	10	0.022	**0.649**	0.442
	20	0.063	**0.661**	0.590
Trade	10	0.646	0.576	**0.688**
	20	0.562	0.593	**0.702**
Wheat	10	0.120	0.429	**0.439**
	20	0.075	0.396	**0.421**

From Table 2, it is noted that IG generally performs better in categories "Acq", "Crude", "Grain", "Money-fx" and "Ship" while WN performs better in categories "Earn", "Trade" and "Wheat". For category "Corn", IG is better with 10 terms while WN performs better for 20 terms. IG only performs better in category "Interest" at 20 terms while Chi2 did better for 10 terms. Therefore, we will take a look at the top 20 terms for category "Acq" where IG performed better, category "Earn" where WN performed better and category "Interest" where Chi2 performed best at 10 terms. Listed in Table 3, Table 4 and Table 5 respectively, are the top 20 terms for category "Acq" , "Earn" and "Interest" for Chi2, IG and WN.

Table 3. Top 20 terms for category "Acq"

Feature selection scheme	Top 20 terms for category "Acq"
Chi2	shares, offer, lt, stake, merger, cts, acquisition, company, inc, acquire, net, loss, corp, usair, common, mln, unit, shr, stock, sell
IG	cts, shares, net, loss, lt, shr, offer, stake, company, merger, tonnes, wheat, acquisition, trade, inc, mln, profit, qtr, corp, acquire
WN	title, company, shares, pct, corp, offer, share, stock, stake, acquisition, merger, common, unit, buy, agreement, board, shareholders, sell, american, investment

Table 4. Top 20 terms for category "Earn"

Feature selection scheme	Top 20 terms for category "Earn"
Chi2	vs, cts, net, loss, mln, shr, profit, revs, qtr, oper, dlrs, note, dividend, th, shrs, avg, div, earnings, prior, lt
IG	vs, cts, net, loss, mln, shr, profit, revs, qtr, dlrs, oper, note, dividend, trade, th, shrs, avg, tonnes, div, said
WN	vs, text, title, cts, net, loss, profit, revs, share, company, pct, note, corp, quarter, dividend, record, th, earning, stock, tax

Table 5. Top 20 terms for category "Interest"

Feature selection scheme	Top 20 terms for category "Interest"
Chi2	rate, rates, bank, money, fed, prime, pct, banks, stg, lending, base, funds, market, discounts, cut, bills, sterling, bundesbank, england, band
IG	rate, rates, bank, pct, money, fed, prime, market, stg, banks, loss, net, lt, cut, company, funds, dlrs, base, lending, discount
WN	pct, rate, market, fed, cut, prime, government, week, funds, federal, term, base, dealers, central, month, bills, reserve, february, growth, dollar

The similar terms between Chi2, IG and WN for category "Acq" are removed to see the combination of words that are unique to each feature selection method. From the 13 unique terms obtained, both Chi2 and IG have one term that strongly represents the "Acq" category, which is "acquire". With the WN approach however, "acquire" was not chosen because it is a verb and our approach only considers nouns. Together with other terms chosen by Chi2 and IG, their performance are better than WN, as abbreviations such as "lt" and "shr" are not chosen by WN although they are statistically strong terms.

For category "Earn", 11 unique words are identified. Out of the 11 unique terms, Chi2 has 9 abbreviations and IG has 8 abbreviations, which are not proper English abbreviations. However, Chi2 and IG choose these terms because they give good statistical values. The only word that is reflective of the category "Earn" is "earnings" for Chi2 and "trade" for IG. In the case of WN, it has words like "share", "company", "earning", "stock" and "tax" that are reflective of the category "Earn". Therefore, in this category, WN is able to perform better and has a set of terms that better describe this category.

There are 12 unique terms for category "Interest". From the 12 unique terms, Chi2 has terms like "sterling", "bundesbank" and "england" that gives high statistical value. IG has 5 terms similar to Chi2 and also an abbreviation "lt" that is not chosen by WN. They also have one similar abbreviation that is not chosen by WN, "stg". This explains why they have better performance when compared to WN, as WN does not contain the abbreviations that gave good statistical values in machine learning.

Generally, the WordNet synonyms approach performs better from 50 terms onwards because it is capable of choosing terms that are more reflective of a category. With fewer terms at 10 and 20 terms, IG has a lot of terms

that has strong statistical values. Although those terms in general, are not representative of the category concerned, it is able to perform well in machine learning because of its statistics. In the case of WordNet synonyms approach, for term size 10 and 20, WordNet perform worse than the statistical methods because it ignores the terms like "lt" and "stg" that gives good statistical performance but are nonetheless meaningless in the language context. Therefore, the performance of the WordNet synonyms approach here is marginally reduced. However, the significance of these abbreviations becomes less significant as the number of terms increased, as we can see from both Figure 2 and 3 that WordNet synonyms approach performs better from 50 terms onwards.

To determine the significance of using WordNet synonyms approach for feature selection, we compare the use of WordNet nouns. From Figure 4, we can see that WN1 performs slightly better than WN at term count 10, 20 and 500. We attribute this to the existence of nouns with high frequencies that gives good statistical values for machine learning. However, in Figure 5, WN outperformed WN1 at all term counts. Generally, the WordNet synonyms approach can perform better than the WordNet nouns approach. Another plus point of WN over WN1 is that WN, which chooses only synonyms, reduces more of the features for each category. We have observed that WN has a feature space that is more than 40% smaller than that of WN1 but still gives an overall result that is better than WN1. This again helps in the dimensionality reduction problem.

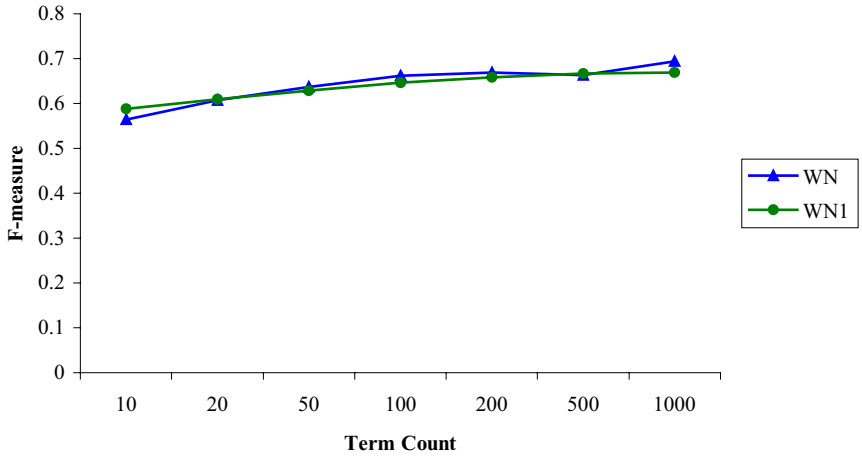

Figure 4. Macro-averaged f-measure comparison between WN and WN1 for Reuters-21578 top ten categories

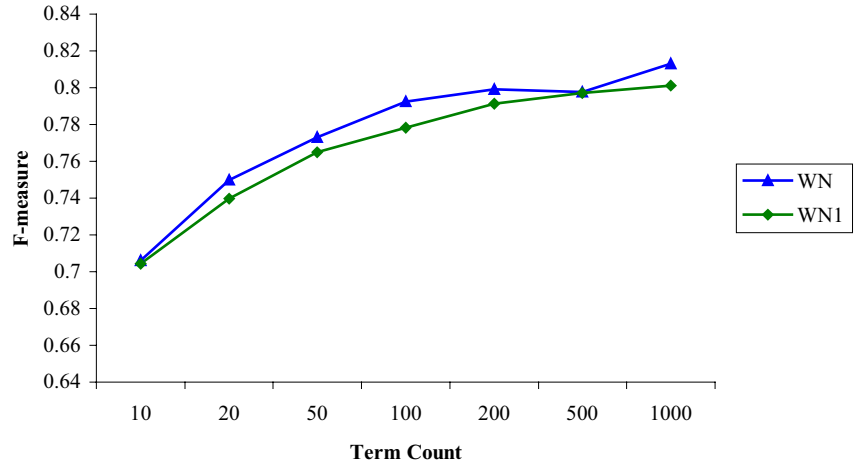

Figure 5. Micro-averaged f-measure comparison between WN and WN1 for Reuters-21578 top ten categories

6. Conclusion

This paper highlights that the WordNet based semantic feature selection is promising. From the experiments conducted, we can see that the WordNet based semantic approaches can improve the feature selection process in text categorization. It is able to provide a better set of features for category representation as the Reuters-21578 collection has many terms that are covered comprehensively by WordNet. The categorical sense disambiguation is promising as an effective approach to automatically disambiguate semantically related terms. Therefore, semantic feature selection is seen to be promising. Further works are being carried out to derive more meaningful features by employing hypernyms in WordNet.

However, for domain specific areas containing technical words, domain specific dictionaries need to be incorporated to complement the terms in WordNet.

7. Reference

[1] Miller G. A., Beckwidth, R., Fellbaum, C., Gross, D. and Miller, K. J., "Introduction to WordNet: An On-line Lexical Database", *International Journal of Lexicography*, Vol 3, No.4 (Winter 1990), pp. 235-244.

[2] Li, X., Szpakowicz, S., Matwin, S., "A WordNet-based Algorithm for Word Sense Disambiguation", in proceedings of the IJCAI-95, 1995, pp. 1368-1374.

[3] Sebastiani, F., "A Tutorial on Automated Text Categorisation", (Analia Amandi and Ricardo Zunino, editors), in proceedings of the ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, Buenos Aires, AR., 1999, pp. 7—35.

[4] Yang, Y. and Liu, X., "A re-examination of text categorization methods", in proceedings of the SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, Berkeley, CA, 1999, pp. 42—49.

[5] Yang, Y. and Pedersen, J. A., "Comparative Study on Feature Selection in Text Categorization", in proceedings of the 14th International Conference (ICML 97), Nashville, TE, USA, 1997, pp. 412—420.

[6] Masuyama, T. and Nakagawa, H., "Applying Cascaded Feature Selection to SVM Text Categorization", in the DEXA Workshops, 2002, pp. 241-245.

[7] Mihalcea, R. and Moldovan, D., "Semantic indexing using WordNet senses", in proceedings of the ACL Workshop on IR & NLP, Hong Kong, 2000.

[8] Gonzalo, J., Verdejo, F., Chugur, I. And Cigarran, J., "Indexing with WordNet Synsets can Improve Text Retrieval", in proceedings of the Coling-ACL'98 Workshop: Usage of WordNet in Natural Language Processing Systems, 1998, pp. 38-44.

[9] Witten, I. H. and Frank, E. *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 2000.

[10] Aas, K and Eikvil, L. "Text Categorization: a survey", Technical Report #941, Norweigan Computing Center, 1999.

[11] Apte, C., Damerau, F. J. and Weiss, S. M., "Automated learning of decision rules for text categorization", *ACM Transactions on Information Systems,* 12, 3, 1994, pp. 233 – 251.