Evaluation of Feature Combination Approaches for Text Categorisation

Robert Neumayer and Kjetil Nørvåg

Department of Computer and Information Science,
Norwegian University of Science and Technology, Trondheim, Norway

Abstract. Text categorisation relies heavily on feature selection. Both the possible reduction in dimensionality as well as improvements in classification performance are highly desirable. To the end of feature selection for text, a range of different methods have been developed, each having unique properties and selecting different features. However, it remains unclear which of them can be combined and what benefits this brings with it. In this paper we present correlation methods for the analysis of feature rankings and evaluate the combination of features according to these metrics. We further show results of an extensive study of feature selection approaches using a wide range of combination methods. We performed experiments on 19 test collections and report our findings.

1 Introduction

The automatic assignment of text documents in predefined categories is denoted to as text categorisation (TC) and has been subject to intense research for decades. However, driven by the ongoing growth of online sources and widespread availability of text documents of all kinds in electronic form, text categorisation has not lost attraction as a research area. Besides, a lot of research output has successfully been turned into applications by industry or is followed up in other research projects, e.g. News or e-mail articles are automatically sorted and delivered to end users. Spam detection techniques have reached a high level of accuracy and in many cases keep inboxes useful. Together with the growth of user generated content on the Web, this generates a strong strong demand for highly effective solutions to the text categorisation problem.

Using all the terms in the collection as feature set leads to the problem of high-dimensionality shared with all other research areas dealing with text. This dimensionality is often prohibitively high for many learning algorithms which are later used to decide which category a document is assigned to. For this reason it is required to limit the space complexity of the text categorisation problem. Feature selection is vital to facilitate such a reduction in dimensionality and most machine learning algorithms could not be applied at all without it.

A range of methods have been suggested and evaluated to this end – with varying performance. Computational resources have become easier available and make the computation of multiple feature rankings possible or applicable. For this reason it has become feasible to use more than one feature selection technique and combine their impact on classification performance. However, not

M. Kryszkiewicz et al. (Eds.): ISMIS 2011, LNAI 6804, pp. 438–448, 2011.

enough research has been done on the possible benefits of combining the results of more than one method. Some methods are more promising to combine than others, which ones to choose from is one of the central questions we try to resolve.

The main four contributions of this paper are: a) we compare a range of feature selection and introduce new ranking merging methods which we compare to existing work; b) we further examine ways of combining them and provide a thorough analysis of which methods to combine based on both the correlations of individual rankings and performance considerations; c) additionally, we document possible performance increases and provide hints as to when the different combination methods work best; d) we show improvements by means of feature ranking merging in an extensive study based on 19 different text test collections, focusing on possible generalisations of our findings.

We continue with giving an overview of related work in the area of text categorisation in Sec. 2. This is followed by an overview of the 15 feature selection methods to be used in Sec. 3. After that, we give an analysis of the combination of these methods in Sec. 4. In Sec. 4 we provide several combination methods based on both ranks and individual values. We further describe experimental results in Sec. 5. Then, we conclude and give an outlook on future work.

2 Related Work

A good overview and a comprehensive survey of the whole area of text categorisation is given by Sebastiani in [11]. Feature selection for text categorisation is surveyed in [12]. An comparison of feature selection using linear classifier weights is given in [5]. Unsupervised feature selection has been used in the context of P2P systems in [7]. The results of a more recent and extensive empirical study of a wide range of single feature selection measures is presented in [2]. Here, the author compares a list of 11 feature selection methods. The evaluation is done on 19 test collections of different size and difficulty. The author uses one-against-all classification and as such averages all results over 229 binary classification problems. A possible combination of methods is not considered.

Social choice voting models have successfully been applied to improve meta search in information retrieval in [6]. The authors show that the Condorcet ranking merging method outperforms the Borda method with respect to precision achieved on the TREC collection. Recent research shows the superiority of reciprocal rank merging for the information retrieval problem of similarity ranking merging. This is shown by several TREC experiments in [1].

Combination experiments for text categorisation are reported in [9]. Experiments are done with four different feature selection methods and a test collection sampled from RCV1-v2. It is shown that certain combination methods improve peak R-precision and F_1. Feature selection combination was, for example, suggested in [10]. The authors selected feature selection methods based on 'uncorrelatedness' and presented results for two document collections. Both studies only partly work with benchmark collections and the results are therefore difficult to compare also due to the impact of different preprocessing techniques applied.

Table 1. Notation for feature selection

Variable	Explanation
N	total #documents in the collection.
N_{C_k}	#documents in category C_k.
$N_{\overline{C_k}}$	#documents not in category C_k.
N_F	#documents containing feature F.
$N_{\overline{F}}$	#documents not containing feature F.
N_{F,C_k}	#documents containing feature F in category C_k.
$N_{\overline{F},C_k}$	#documents not containing feature F in category C_k.
$N_{F,\overline{C_k}}$	#documents containing feature F not in category C_k.
$N_{\overline{F},\overline{C_k}}$	#documents not containing feature F not in category C_k.

Initial results of an evaluation study on a large set of categorisation problems were presented in [8]. However, in this paper we present a more in-depth analysis of feature correlation and provide experimental results for this analysis.

3 Feature Selection Methods

We list the used notation and the different feature selection methods we use in this paper in Tab. 1. We chose to use a generalised notation since we consider it easier to follow compared to the wide range of different notations. The individual feature selection methods are given by name and abbreviation in Tab. 2. A method is called unsupervised if it does not rely on previously assigned labels (methods belonging there are shown in the first part of the table). A method is called supervised if it does rely on previously assigned labels to compute the discriminative power of a feature (shown in the second part of the table). This represents a good overview of methods used in various recent studies.

4 Combination of Feature Selection Methods

The question of which feature selection techniques to combine is the most important decision. We present important considerations in the following.

4.1 Compatibility between Methods

The range of values provided by the different methods might be inhibitive in their combination based on feature value. This becomes apparent in Fig. 1. We present the value distribution of the top 100 measure for two selected methods in the training set of the 20newsgroups collection and the distribution of a random ranking for the purpose of comparison (random numbers are generated in experiments). The values are normalised between zero and one. Nevertheless the distribution is important since only measures with similar distributions can be combined in a straight-forward way. In the worst case this will lead to single techniques having little or even no effect on the final ranking (e.g. when

Table 2. Notation, unsupervised, and supervised feature selection methods used

Method	Explanation
Document Freq. (DF)	The number of documents a term occurs in
Inverse Document Freq. (IDF)	$\frac{N}{DF(F)}$
Collection Freq. (CF)	The number of occurrences of a term in a collection
Inverse Collection Freq. (ICF)	$CF(F)\log_2 \frac{N}{DF(F)}$
Term Freq. Doc. Freq. (TFD)	$(n_1 \times n_2 + c(n_1 \times n_3 + n_2 \times n_3))$, where c is a constant, $c \geq 1$, n_1 is the number of documents without the feature, n_2 is the number of documents where the feature occurs exactly once, n_3 is the number of documents where the feature occurs twice or more.
Information Gain (IG)	$-\sum_{k=1}^{C} \frac{N_{C_k}}{N} \ln \frac{N_{C_k}}{N} + \frac{N_F}{N} \sum_{k=1}^{C} \frac{N_{F,C_k}}{N_F} \ln \frac{N_{F,C_k}}{N_F}$ $+\frac{N_{\overline{F}}}{N} \sum_{k=1}^{C} \frac{N_{\overline{F},C_k}}{N_{\overline{F}}} \ln \frac{N_{\overline{F},C_k}}{N_{\overline{F}}}$
Mutual Information (MI)	$\sum_{v_f \epsilon \{1,0\}} \sum_{v_{C_k} \epsilon \{1,0\}} P(F = v_f, C_k = v_{C_k})$ $\ln \frac{P(F=v_f, C_k=v_{C_k})}{P(F=v_f)P(C_k=v_{C_k})}$
Odds Ratio (OR)	$\ln \frac{P(F\|C_k)(1-P(F\|\overline{C_k}))}{P(F\|\overline{C_k})(1-P(F\|C_k))} = \ln \frac{\left(\frac{N_{F,C_k}}{N_{C_k}}\right)\left(1-\frac{N_{F,\overline{C_k}}}{N_{\overline{C_k}}}\right)}{\left(\frac{N_{F,\overline{C_k}}}{N_{\overline{C_k}}}\right)\left(1-\frac{N_{F,C_k}}{N_{C_k}}\right)}$
Class Discrimination Meas. (CDM)	$\sum_{k=1}^{\|C\|} \left\| \log \frac{\left(\frac{N_{F,C_k}}{N_{C_k}}\right)}{\left(\frac{N_{F,\overline{C_k}}}{N_{\overline{C_k}}}\right)} \right\|$
Word Freq. (WF)	N_{F,C_k}
χ^2 statistic (χ^2)	$\frac{N \times \left((N_{F,C_k} \times N_{\overline{F},\overline{C_k}}) - (N_{F,\overline{C_k}} \times N_{\overline{F},C_k})\right)^2}{N_F \times N_{\overline{F}} \times N_{C_k} \times N_{\overline{C_k}}}$
NGL-Coefficient (NGL)	$\frac{\sqrt{N}(N_{F,C_k} N_{\overline{F},\overline{C_k}} - N_{F,\overline{C_k}} N_{\overline{F},C_k})}{\sqrt{N_F N_{\overline{F}} N_{C_k} N_{\overline{C_k}}}}$
Categorical Proportional Difference (CPD)	$\frac{N_{F,C_k} - N_{F,\overline{C_k}}}{N_F}$
GSS-Coefficient	$N_{F,C_k} N_{\overline{F},\overline{C_k}} - N_{F,\overline{C_k}} N_{\overline{F},C_k}$
Bi-Normal Separation (BNS)	$\left\| F^{-1}\left(\frac{N_{F,C_k}}{N_{C_k}}\right) - F^{-1}\left(\frac{N_{F,\overline{C_k}}}{N_{\overline{C_k}}}\right) \right\|$

one method provides consistently higher values than the other). These findings should be taken into account when deciding what techniques to combine. We therefore stress the importance of normalisation for ranking combination.

4.2 Fitness of Individual Methods

The performance of the individual methods is definitely an important criterion when choosing which combinations to combine. We assembled a list of methods with varying individual performance and different numbers of features.

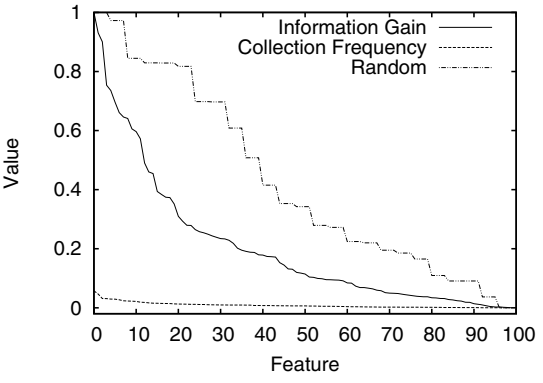

Fig. 1. Normalised feature values for top 100 features in the 20-news collection for two selected feature selection methods compared to the random distribution

4.3 Correlations of Individual Methods

In this section we analyse different possibilities of which rankings or feature selection techniques to combine. The more correlation there is between two rankings the less benefit can be expected from their combination (i.e. if two methods provide equally good results but have low correlation, it can be assumed that different features are responsible). The main goal here is to find groups of non-correlating rankings produced by different feature selection methods. We use the Spearman rank coefficient to find pairwise correlations between rankings. It is a measure for correlation between two rankings, operating on the rank on their elements rather than their numeric values; the coefficient is given in the following (d denotes the difference in ranks and n the length of the rankings, i.e. the number of features selected per ranking).

$$R = 1 - \frac{6 \sum d^2}{n^3 - n}$$

4.4 Actual Combination Techniques

A range of methods have been suggested for ranking, albeit often in different settings like Condorcet merging and Borda merging which initially are originally used for defining winners of elections.

Two or several rankings can be combined by using the ranks of the terms in the individual rankings. When dealing with two rankings of a term t_i, the ranks of this term $r_j(t_i)$ are used rather than the plain values. If term t_x is ranked first in r_1 and second in ranking r_2, these rank values are $r_1(t_x) = 1$ and $r_2(t_x) = 2$ respectively. If ranks from several methods are combined the final list is sorted according to the newly computed rank values.

Another possibility is to use the values given by the individual methods. Term t_x might for example have different values in different rankings. To get a final value for a term across multiple rankings these individual values might be combined by, e.g. building the sum or average over the values. These final values are then sorted and the top k features selected as input to the classifier.

Table 3. Ranking Merging methods used

Method	Explanation
Highest Rank (HR)	A feature's highest rank in all single rankings.
Lowest Rank (LR)	The lowest of all rankings is used as final score.
Average Rank (AR)	The average over all single ranks is used.
Borda Ranking Merging (BRM)	Gives scores according to the length of the single rankings. If the size of a ranking is n and an element is ranked at the ith position the score $\frac{i}{n}$. This technique is also applicable for individual rankings with different lengths. The final scores are the sum of the individual scores.
Condorcet Ranking Merging (CRM)	A majoritarian method favouring the candidate beating every other candidate in pair-wise comparisons. If, e.g., feature a is higher ranked than b in any of the methods, it a clearly beats b. For aggregation the number of pair-wise wins or ties is summed for each candidate and the one with the highest score is the overall winner.
Reciprocal Ranking Merging (RRM)	In this setting, the score for a feature is the sum of 1 divided by the rank in the single rankings.
Divide by Max. then OR (DMOR)	The average over all single feature values in this setting we normalise by the maximum.
Divide by Length then OR (DLOR)	Normalisation by the length of the vector.
Pure Round Robin (RR)	One feature from each ranking is added to the final ranking in turn until the desired number of features is reached.
Top N Ranking Merging (Top$N$)	The top n features from each ranking in turn are added until enough features are collected.
Weighted N Ranking Merging (WN)	The first n % are taken from the first ranking, the remaining $1 - n$ % are composed of the other rankings in equal parts.

We introduce a third group of methods based on round robin algorithms and weighted combinations. The rationale behind the weighted methods is that the whole set of features selected by one method is more than the sum of its parts. This means that it's well possible that the performance of a method is influenced not by the single features but that there is an underlying dependence on the features. With weighted ranking merging the majority of the features is selected from one method and only a smaller fraction from additional methods.

5 Experiments

The following experiments were performed using the 20newsgroups data set[1], which has become very popular for text experiments in the field of machine

[1] http://people.csail.mit.edu/jrennie/20Newsgroups

Table 4. Spearman rank coefficient measure for the full set of documents for each feature selection method. We list statistics of correlation values for each method. The evaluation considers all features in the training set of the 20news collection.

	IG	OR	WF	MI	CS	DIA	NGL	CPD	BNS	CDM	GSS	DF	TFD	W	CF	ICFI
IG	1.0	.33	.91	.95	.81	**-.28**	**.10**	**-.32**	**-.03**	**-.10**	.90	.92	.80	.21	.88	.86
OR	.33	1.0	.41	.38	.60	**.12**	**.14**	.51	**-.03**	.52	.53	**.10**	**.14**	**.14**	**.13**	**.13**
WF	.91	.41	1.0	.98	.85	**-.00**	.27	**-.16**	**.02**	**-.03**	.92	.83	.75	.24	.81	.79
MI	.95	.38	.98	1.0	.84	**-.07**	.23	**-.22**	**-.02**	**-.05**	.91	.88	.78	.24	.85	.83
CS	.81	.60	.85	.84	1.0	**-.05**	.19	.17	**.02**	.31	.95	.59	.57	.30	.60	.59
DIA	**-.28**	**.12**	**-.00**	**-.07**	**-.05**	1.0	.80	.44	**-.10**	.34	**-.15**	**-.28**	**-.22**	.18	**-.25**	**-.25**
NGL	**.10**	**.14**	.27	.23	.19	.80	1.0	.20	**-.12**	.22	**.14**	**.09**	**.10**	.29	**.10**	**.10**
CPD	**-.32**	.51	**-.16**	**-.22**	.17	.44	.20	1.0	**-.23**	.85	**-.03**	**-.52**	**-.43**	.17	**-.48**	**-.47**
BNS	**-.03**	**-.03**	**.02**	**-.02**	**.02**	**-.10**	**-.12**	**-.23**	1.0	**-.38**	**.08**	**.01**	**.09**	**-.11**	**.05**	**.06**
CDM	**-.10**	.52	**-.03**	**-.05**	.31	.34	.22	.85	**-.38**	1.0	**.11**	**-.36**	**-.30**	.26	**-.32**	**-.32**
GSS	.90	.53	.92	.91	.95	**-.15**	**.14**	**-.03**	**.08**	**.11**	1.0	.74	.69	.24	.73	.72
DF	.92	**.10**	.83	.88	.59	**-.28**	**.09**	**-.52**	**.01**	**-.36**	.74	1.0	.84	.16	.93	.90
TFD	.80	**.14**	.75	.78	.57	**-.22**	**.10**	**-.43**	**.09**	**-.30**	.69	.84	1.0	.16	.96	.97
W	.21	**.14**	.24	.24	.30	.18	.29	.17	**-.11**	.26	.24	.16	.16	1.0	.17	.16
CF	.88	**.13**	.81	.85	.60	**-.25**	**.10**	**-.48**	**.05**	**-.32**	.73	.93	.96	.17	1.0	1.0
ICF	.86	**.13**	.79	.83	.59	**-.25**	**.10**	**-.47**	**.06**	**-.32**	.72	.90	.97	.16	1.0	1.0

Table 5. Overlap within feature rankings for the 20news collection. 1000 features are selected and we count the number of features occurring in both rankings.

	IG	OR	WF	MI	CS	DIA	NGL	CPD	BNS	CDM	GSS	DF	TFD	W	CF	ICF
IG	1.0	.31	.76	.95	.69	**.01**	.35	**.00**	.60	**.09**	.86	.51	.59	**.07**	.55	.56
OR	.31	1.0	.26	.31	.34	**.08**	**.21**	**.01**	**.25**	.63	.27	.29	.28	**.13**	.28	.27
WF	.76	.26	1.0	.74	.56	**.00**	.31	**.00**	.41	**.06**	.85	.70	.76	**.05**	.74	.75
MI	.95	.31	.74	1.0	.74	**.01**	.38	**.00**	.64	**.12**	.85	.47	.55	**.08**	.52	.53
CS	.69	.34	.56	.74	1.0	**.02**	.44	**.00**	.66	**.23**	.67	.28	.36	**.10**	.33	.34
DIA	**.01**	**.08**	**.00**	**.01**	**.02**	1.0	**.04**	**.11**	**.01**	**.12**	**.01**	**.00**	**.00**	**.04**	**.00**	**.00**
NGL	.35	**.21**	.31	.38	.44	**.04**	1.0	**.00**	.40	**.15**	.35	**.19**	**.23**	**.10**	**.22**	**.23**
CPD	**.00**	**.01**	**.00**	**.00**	**.00**	**.11**	**.00**	1.0	**.00**	**.02**	**.00**	**.00**	**.00**	**.02**	**.00**	**.00**
BNS	.60	**.25**	.41	.64	.66	**.01**	.40	**.00**	1.0	**.15**	.52	**.16**	**.25**	**.09**	**.22**	**.24**
CDM	**.09**	.63	**.06**	**.12**	**.23**	**.12**	**.15**	**.02**	**.15**	1.0	**.09**	**.00**	**.01**	**.16**	**.02**	**.03**
GSS	.86	.27	.85	.85	.67	**.01**	.35	**.00**	.52	**.09**	1.0	.56	.63	**.06**	.60	.61
DF	.51	.29	.70	.47	.28	**.00**	**.19**	**.00**	**.16**	**.00**	.56	1.0	.87	**.02**	.89	.85
TFD	.59	.28	.76	.55	.36	**.00**	**.23**	**.00**	**.25**	**.01**	.63	.87	1.0	**.03**	.93	.92
W	**.07**	**.13**	**.05**	**.08**	**.10**	**.04**	**.10**	**.02**	**.09**	**.16**	**.06**	**.02**	**.03**	1.0	**.03**	**.04**
CF	.55	.28	.74	.52	.33	**.00**	**.22**	**.00**	**.22**	**.02**	.60	.89	.93	**.03**	1.0	.96
ICF	.56	.27	.75	.53	.34	**.00**	**.23**	**.00**	**.24**	**.03**	.61	.85	.92	**.04**	.96	1.0

learning and has been used for example in [4]. The data set consists of newsgroup postings from the 20 newsgroups. From each newsgroup, 1.000 articles from the year 1993 have been selected; after removing duplicate articles (mostly cross-postings to several newsgroups), 18.846 unique messages remain. Each text consists of the message body and in addition the 'Subject' and the 'From' header lines which we discarded before analysis. We use the predefined 'bydate' split, which is divided into training (60%) and testing (40%).

Additionally, we use a set of categorisation problems also used for binary classification experiments in [2], which were initially used by Han and Karypis and

Table 6. Experimental results on 20news, single methods in (a), combinations in (b)

(a) Classification results for the 20news collection, 1000 features, individual methods

Method	Acc.
CF	66.76
TFD	67.75
DF	64.90
ICF	67.37
WF	71.14
IG	72.65
BNS	72.03
CPD	8.44
CHI	**73.49**
CDM	42.66
DIA	8.75
GSS	71.68
MI	72.94
NGL	60.62
OR	63.83

(b) Classification results for combinations for 1000 features. We list combinations and merging methods representing an improvement over the best single method, the best values are shown in bold font

Methods and combination type	Acc.
BNS-CHI-AvgMinMaxNorm	73.54
BNS-CHI-AvgRank	**74.03**
BNS-CHI-Borda	**74.03**
BNS-CHI-Condorcet	73.59
BNS-CHI-LowestRank	73.70
BNS-CHI-Reciprocal	**74.02**
BNS-DF-MI-CHI-WF-OR-AvgMinMaxNorm	73.54
BNS-IG-Condorcet	73.74
BNS-IG-HighestRank	73.77
BNS-IG-Reciprocal	73.77
BNS-IG-RoundRobin	73.79
BNS-IG-Top100RoundRobin	73.67
BNS-IG-Top50RoundRobin	73.70
BNS-MI-AvgRank	73.61
BNS-MI-Borda	73.61
BNS-MI-Condorcet	73.55
BNS-MI-Reciprocal	73.65
BNS-WF-AvgMinMaxNorm	73.67
IG-BNS-Condorcet	73.74
IG-BNS-HighestRank	73.77
IG-BNS-Reciprocal	73.77
IG-BNS-RoundRobin	73.73

originate from TREC, OHSUMED, Reuters. The collection sizes range from 204 to 31472 documents and the number of classes varies from six to 36 classes. All collections were already preprocessed by basic stemming and stop-word removal. Unless stated otherwise we always select the 1000 best features, this can either mean 1000 per method for the single runs or 1000 features as a combination of multiple rankings. We both assume 1000 features to be a reasonable dimensionality in terms of complexity and performance and fixed this parameter to limit the number of results.

5.1 Individual Pair-Wise Correlations

Spearman. The results of the computation of the Spearman coefficient for all terms in the corpus occurring more than once, i.e. 53000 terms is given in Tab. 4. It is shown that some methods have more un-correlated methods than others. In cases of methods which have rather low performance when used exclusively like DIA or NGL this is not surprising, in other cases like BNS (used, e.g., in [3]) it suggests that the combination of the method with others might be beneficial.

Table 7. Results on the 19 text collections, single methods in (a), combinations in (b)

(a) Averaged classification results over all 19 test collections, 1000 features, individual methods

Method	Acc.
TFD	85,24
DF	84,38
CF	84,90
WF	83,77
IG	**86,45**
BNS	84,04
CPD	71,02
CHI	86,36
CDM	73,85
DIA	52,98
GSS	85,88
MI	85,97
NGL	69,64
OR	83,68

(b) Classification results for combinations for 1000 features. We list combinations which represent an improvement over the best single method, the best values are shown in bold font

Methods and combination type	Acc.
IG-BNS-AverageMinMaxNorm	86,14
IG-BNS-Condorcet	86,38
IG-BNS-DLOR	86,82
IG-BNS-Main50	86,45
IG-BNS-Main60	86,45
IG-BNS-Main70	86,45
IG-BNS-Main80	86,45
IG-BNS-Main90	86,45
IG-BNS-RoundRobin	**86,65**
IG-BNS-Top100RoundRobin	**86,69**
IG-BNS-Top300RoundRobin	**86,66**
IG-BNS-Top50RoundRobin	**86,71**
BNS-DF-MI-CHI-WF-IG-OR-Condorcet	86,31
BNS-DF-MI-CHI-WF-OR-Main50	86,88
BNS-DF-MI-CHI-WF-OR-Main60	86,87

However, this only gives an overall view of the potential of combination of the methods. The correlation of all terms in the collection can only partly help to discriminate. If we look at the correlation only at the $top - n$ terms, the results might differ. It is for example possible that two methods have a low correlation overall, but a high correlation when only the top 1000 features are considered.

Overlap Metric. The decision on which feature selection methods to compare also relies on the correlation for the respective top-n features. To this end we chose the overlap metric which simply calculates the percentage of features occurring in both rankings (the Spearman coefficient was undefined for some rankings due to a lack of co-occurring features).

Based on Tab. 5 we suggest the following combinations of methods: BNS OR WF CDM TFD. BNS has a low overlap ($< .25$) with nine other methods and therefore constitutes a good basis for combination. The other methods have reasonable overlap with each other and belong to different classes of methods (supervised/unsupervised).

We show the results for all single methods and the 20news collection in 6(a). The best method(s) in each column are printed in bold font. For the single methods we achieved the best results with the χ^2 method, the WF, IG, BNS, GSS and MI methods are not far behind (Tab. 6(a)). We then performed experiments with combinations and ranking merging, based on the analysis provided earlier. Out of the 364 experiments performed (all pairwise combinations plus the combination of all methods selected), 22 are improvements over the χ^2 method. The improvement is, however, limited with 74.03 over 73.49 with the best single

method. Reciprocal rank merging is included in four out of the five pairwise combinations and along with reciprocal rank merging is the most common method.

We show the results achieved on the collection of 19 collections in Tab. 7, the values listed are averages over all 19 results. The best single method in this context is IG with 86.36 per cent of the instances correctly classified, shortly followed by MI, GSS, and TFD. We found marginal improvements by merging shown in Tab. 7(b). The merging methods mainly relying on one method and taking in few features from the remaining methods perform more stable.

6 Conclusions and Future Work

We presented a range of methods for both feature selection and combination for text categorisation. In addition, we presented two classes of methods not previously used for categorisation methods (round robin based and weighted ranking merging). We further presented an extensive experimental evaluation of which feature selection methods to combine and performance evaluation on a diverse set of text categorisation benchmark collections. Future work will mainly deal with new feature ranking merging strategies in limited application domains and the development of strategies when to rely on which combination of methods.

Acknowledgements

We hereby express gratitude to Østein Løhre Garnes who helped with the initial implementation of some of the feature selection methods in his master's thesis.

References

1. Cormack, G.V., Clarke, C.L.A., Büttcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd ACM SIGIR, pp. 758–759 (2009)
2. Forman, G.: An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3, 1289–1305 (2003)
3. Forman, G.: BNS feature scaling: an improved representation over tf-idf for SVM text classification. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM), pp. 263–270 (2008)
4. Mitchell, T.: Machine Learning. McGraw Hill, New York (1997)
5. Mladenić, D., Brank, J., Grobelnik, M., Milic-Frayling, N.: Feature selection using linear classifier weights: interaction with classification models. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 25-29, pp. 234–241. ACM, New York (2004)
6. Montague, M., Aslam, J.A.: Condorcet fusion for improved retrieval. In: Proceedings of the 11th ACM International Conference on Information and Knowledge Management (CIKM), pp. 538–548 (2002)
7. Neumayer, R., Doulkeridis, C., Nørvåg, K.: A hybrid approach for estimating document frequencies in unstructured P2P networks. Information Systems 36(3), 579–595 (2011)

8. Neumayer, R., Mayer, R., Nørvåg, K.: Combination of feature selection methods for text categorisation. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 763–767. Springer, Heidelberg (2011)
9. Scott Olsson, J., Oard, D.W.: Combining feature selectors for text classification. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM), pp. 798–799 (2006)
10. Rogati, M., Yang, Y.: High-performing feature selection for text classification. In: Proceedings of the 11th ACM International Conference on Information and Knowledge Management (CIKM), pp. 659–661 (2002)
11. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
12. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference on Machine Learning (ICML), pp. 412–420 (1997)