Effect of Pruning on Feature Ranking Metrics in Highly Skewed Datasets in Text Classification

Muhammad Nabeel Asim[1], Abdur Rehman[2], Muhammad Idrees[3]

Al-Khwarizmi Institute of Computer Science, University of Engineering and Technology, Lahore, Pakistan[1]
Department of Computer Science and Engineering University of Gujrat, Gujrat Pakistan[2]
Department of Computer Science and Engineering, University of Engineering & Technology, Narowal, Pakistan[3]

Abstract
A variety of feature ranking algorithms are available for text data to select appropriate features for a classification task. To improve the feature selection process, data is preprocessed to remove too frequent and too rare terms, called pruning. Although not required for non-text data, pruning has become and essential step to simplify the feature selection of text data, which results in boosting the overall classification performance. In this paper we have studied the effect of pruning on eight well known feature selection metrics, namely NDM, IG, ODDS, CHI, DFS, POIS, GINI and ACC2. while evaluation of FR metrics is done using featured micro and macro F1 measure by using SVM classifier. Experimental results on five bench mark datasets, including WAP, RE0, RE1, K1a and K1b, show that pruning adversely affect three feature ranking algorithms IG, DFS and ACC2, for which pruning reduces the overall efficiency of the classification. While pruning improves the classification performance for the rest five FR metrics.

Key words:
Text Classification, ranking algorithms

1. Introduction

An immense amount of data is being generated on Internet every minute [1]; as email users send 204,000,000 messages, Google receive 4,000,000 search queries, twitter user 277,000 tweets per minutes. It's a huge challenge to search information in this giant data in short span of time. High dimensionality of data is the main provoking element of this research. In fact it is impossible to search relative information from such huge raw data without classifying it [9]. To retrieve and search data into small number of documents which belong to our query class is more efficient and less time consuming instead of searching in whole repository.

In automatic text classification, we assign categories or classes to documents in a collection of N documents having a set of M categories [14]. A set containing all the documents under consideration is called corpus. Documents can be in hard or soft form. Documents in hard form are categorized manually by human experts while documents in soft form are either categorized by human experts or by using some sort of classification algorithms. Text classification is an example of content classification in which a document belongs to a class, if particular amount of data/contents present in a document is same as that of the class. In library science, a document assumed to be a part of class if at least 20% content discussed in the document belongs to that class [15]. Auto-matic text classification mostly follows the ”Bag of Words” representation which considers the occurrence of a word in documents regardless of its order is called term count(tc) or term frequency (tf).

Text classification has a lot of applications in several domains such as text mining or searching for a specific information [5]; separation of legitimate emails from Spam emails and finding customers interest from their comments in social media [15].

The process of classification is divided into three steps [17]: first feature extraction, in which dimensionality of data is reduced by generating new features from already present features, second step is feature selection where from a set of large features only highly invidious feature among the classes are selected, third step is classification in which a highly discriminative set of features are given to classifier which assigns them labels from a set of known categories. Before feature selection metrics applied, text data needs to be pre-processed [4] i.e removal of stop words and stemming of data. Stop words are grammatical structuring words like ”is”, ”am”, ”the” etc. and do not convey any meaningful information are removed using a dictionary/vector of stop words; while stemming is to convert the inflected form of words to their base form. Text data contains fewer rare terms/features and a

number of those features which frequently present in documents. Pruning is a step prior to feature selection as frequently adopted by practitioners to remove outliers and too rare terms by applying specific threshold criterion [5]; while feature selection is used to remove non-informative, non-relevant features and to select top ranked features. Features are not independent, they provide clear information to classifier when combined with other features and may provide ambiguous or no-information to classifier alone.

Manuscript received October 5, 2017
Manuscript revised October 20, 2017

Pruning is necessary pre-step to feature selection [4]. But in highly skewed dataset, classes which occur very few times would have relatively fewer features than frequently occurring classes, such classes may be unable to pass the given threshold test offered by pruning; which in case of pruning will receive no allocation in training phase, will produce errors in testing phase of a classifier [6]. In pruning Upper and lower threshold values are selected for document frequency. Lower threshold value is absolute in which we remove words which occur in three or less documents while in upper threshold those words are discarded which present in 25% or more of documents [7].

Training and prediction phase are two processes of text clas-sification. First phase trains the classifier on already present data so that incoming data be assigned to their respective labels. In other words it determines the decision boundary of the classifier. We showed by experimentations on five bench mark datasets the role of pruning by using eight feature ranking methods and evaluate results using featured macro and micro F1 measures using SVM classifier. We take the difference of non-pruned and pruned empirical values and evaluate them using micro and macro F1 measures and show their illustration by using both graphical and tabular forms. If difference is non-negative, then before applying FR metric there is no need for pruning and vice versa.

Organization of this paper into different sections is as follows: related work is discussed in section II, experimental setup is presented in section III; finally, section IV and V is about conclusion of the paper.

2. RELATED WORK

The process of feature selection can be done by using three techniques. One of them is Filter method. In filter approach, FR methods are applied on datasets for selection of highly invidious features having high discriminative power without the involvement of any classification technique [16]. In filters method absence of classifier in the process of feature selection reduce the efficiency of classification process. Wrapper approach selects a subset of features, trains the classifier on given subset; test the error on subset of features other than training subset then selects a subset whose error is minimum[15]. Third approach in feature selection is embedded approach which selects features based on classification model during learning phase of classifier.

Mostly algorithms use document frequency to rank the features such as odds ratio, information gain and chi squared [7]. Document frequency measures can be represented in the form of confusion matrix as shown in table I.

TABLE I: Confusion matrix

	t_j	$\bar{t}_j$
Positive Class	t_p	f_n
Negative Class	f_p	t_n

Definitions of document frequency measures are given as.

True Positives (t_p)
Positive documents containing the term
False Positives (f_p)
Negative documents containing the term
True Negatives (t_n)
Negative documents not containing the term
False Negatives (f_n)
Positive documents not containing the term

This paper deals with filter based FR metrics and we present in this section all the measures that we used in our experimental evaluation.

A. Balanced Accuracy Measure (ACC2)

Accuracy measure (ACC) is a well known feature selection technique widely used in single label text classification. It is simply the difference of true positives and false positives of a term. It works well in balanced dataset but perform poorly on unbalanced dataset because this algorithm is biased toward tp.

Balance accuracy measure (ACC2) is an enhanced version of accuracy (ACC) measure [15]; it is the absolute difference of true positive rate (tpr) and false positive rate (fpr) of a term. As tpr and fpr are normalized terms, obtained after division of tp and fp with their class size respectively, it solves the problem of biasing toward more frequent features. Formulas for these equations given:

$$Accuracy\ Measure = ACC = tp - fp \tag{1}$$

$$Balanced\ Accuracy\ Measure = ACC2 = |\,tpr - fpr\,| \tag{2}$$

In equation 2 values of t_{pr} and f_{pr} are described in equation 3 and 4.

$$tpr = \frac{tp}{tp + fn} \tag{3}$$

$$ftpr = \frac{tn}{tn + fp} \tag{4}$$

B. Normalized Difference Measure (NDM)

Balanced accuracy measure assigns score to a term on the basis of |tpr -fpr|. ACC2 assigns equal rank to different terms, which has same value of |tpr -fpr| but different values of tpr and fpr. According to NDM [15], features at top left and bottom right are more important as compared to features on the diagonal axis.

$$NDM = \frac{|tpr - fpr|}{\min(tpr, fpr)} \quad (5)$$

C. Information Gain (IG)

Information gain (IG) is widely used algorithm for feature selection in text classification. This technique counts the amount of information about classification problem weather it is increased or decreased by addition or removal of a term from the feature sub set. Information of a feature f can be measured as

$$IG_f = e(p;n)[P_w e(tp; fp) + P_{\overline{w}} e(fn; tn)] \quad (6)$$

Where p and n represents the number of positive and negative instances, further e (p, n) can be calculated as

$$-p\frac{p}{p+n}\log 2\frac{p}{p+n} - \frac{n}{p+n}\log 2\frac{n}{p+n}$$

p_w and $p_{\overline{w}}$ can be calculated as

$$p_w = \frac{(tp+fp)}{N}, P_{\overline{w}} = 1 - P_{term}$$

D. Chi-Squared (CHI)

Widespread use of CHI metric in data mining applications make it favored method as it depicts, features which are present or absent are independent of class labels or not[18]. Chi square do not perform well when there exist infrequent terms in data sets but its performance can be improved by applying pruning on data sets having a certain threshold level [19]. Performance of chi square decrees in document or text classification when they have less term count. Score of ith feature of kth class can be calculated as :

$$CHI = \frac{(tp \times tn - fn \times fp)^2}{(tp+fp)(fn+tn)(tp+fn)(fp+tn)} \quad (7)$$

E. Gini index (GINI)

Gini Index is a distribution estimation criterion of a term over different classes given as:

$$GI(t) = \sum_{j=i}^{M} P(t \mid Cj)^2 \; P(Cj \mid t)^2 \quad (8)$$

F. Odds Ratio (OR)

OR is the fraction of true positive and negative to false positive and negative. It assigns highest score to rare terms which are present in negative class[20]. In order to attain non zero value of false positive and negative this algorithm need to retain a large number of features in the vector. Mathematical formulation of OR is given below:

$$OR = \frac{t_p \times t_n}{f_p \times f_n} \quad (9)$$

G. Distinguishing feature selector (DFS)

DFS is a probabilistic based feature ranking metric proposed by Uysal and Gunal [21]. It assign high rank to features which occur more time in one class and less time in other class. DFS metric assigns score values between 0.5 and 1.0[21].

$$DFS = \sum_{c=i}^{n} \frac{P(C_i \mid f)}{P(\bar{f} \mid C_i) + P(f \mid \overline{C}_i) + 1} \quad (10)$$

Where n is the number of classes, P (Ci) is probability of ith class and $P(\bar{f} \mid C_i)$ is probability of absence of feature f when class Ci is given while $P(f \mid \overline{C}_i)$ is feature likelihood when classes other than Cj are given.

H. Poisson ratio (POIS)

This algorithm is mostly used for feature selection in information retrieval to expand user query[22]. It calculates the deviation of a term from the distribution. A term which fits into the distribution is being marked independent of the given class. Mathematical formulation is given as

$$POIS = \frac{(ap - \hat{ap})^2}{\hat{ap}} + \frac{(bnp - \hat{bnp})^2}{\hat{bnp}} + \frac{(cp - \hat{cfp})^2}{\hat{cfp}} + \frac{(dtn - \hat{dtn})^2}{\hat{dtn}}$$

$$\hat{ap} = N(C)(1 - e^{(-\lambda)}), \hat{ap} = N(C)e^{(-\lambda)},$$
$$\hat{cfp} = N(\overline{C})(1 - e^{(-\lambda)}, \hat{d}tn = N(\overline{C})e^{(-\lambda)},$$
$$\lambda = F / N \quad (11)$$

Where ap and bnp represents the presence or absence of a term or features in a particular class respectively. If a term is present but not belonging to class C is represented by quantity cf p; dtn represents if t and C are both absent from the documents. While hat values are predicted values of non-hat quantities.

3. EXPERIMENTAL SETUP

This section briefly explains the characteristics of five skewed datasets (Wap, RE0, RE1, K1a, and K1b) which are used in experimental evaluation of eight featured feature ranking metrics and results. Evaluation of FR metrics is done using micro and macro F1 measures and results are shown in tabular forms. Quality of features which are selected by FR algorithms are being assessed by SVM classifier.

A. Datasets used

We used five data sets which includes two highly skewed subsets of Reuters datasets RE0 and RE1 which are used by Forman [7], given by University of Minnesota. Three highly unbalanced subset of WebACE project (WAP, K1a and K1b) are used. A detailed summary of five data sets such as total number of documents, number of terms, class skew and number of classes is presented in table II. A pre-processing step is already applied on datasets we obtained from the Internet data repository i.e. removal of stop words and stemming. A pre-processing step before applying any FR metrics, is excessively used in data mining and machine learning applications, is pruning which removes too frequent and rare terms. In pruning lower threshold is a fix bound in which those features are removed which belong to less than three documents [7], while in upper bound too frequent features are removed which present in 25% or more of documents [7].

TABLE II: Summary of the five datasets used for experiments

Dataset	Total Docs	Number of Terms	Number of Classes	Min Class size	Max Class size
Wap	1560	6852	20	5	341
Categories	Culture, Media, Multimedia, Business, Politics, Cable, Online, Review, Health, Sports, Art, Variety, Television, Music, Entertainment, Stage, Film, People, Industry, Technology				
K1a	2340	8589	20	9	449
Categories	E, Ec, B, Ea, H, Ev, Ecu, Er, T, Et, Es, P, Em, S, Ep, Emu, Eo, Ei, Ef, Emm				
K1b	2340	8589	6	60	1389
Categories	Politics, Sports, Health, Tech, Business, Entertainment				
RE0	1504	2886	13	11	608
Categories	lei, housing, bop, wpi, retail, ipi, jobs, reserves cpi, gnp, interest, trade, money-fx				
RE1	1657	3037	25	10	371
Categories	cotton, zinc copper, ship, carcass, alum, tin, iron oilseed gold, meal, wheat, orange, rubber, cofee, livestock, gas, veg, flr, cocoa, pet, grain, crude, nat, sugar				

In our experimentations we show the role of pruning on FR metrics which present in 25% or more of documents [7] using SVM classifier. We make two groups of datasets, on one group before applying any FR algorithm we applied pruning and on other group we do not apply pruning and then compare the results. Results are discussed in Results section III-D. For cross validation of results we use split of datasets, although there is no hard and fast rule for splitting we use 70% of data in training phase and 30% in testing phase.

B. Classification and Feature Ranking Algorithms used

Classification is done using SVM classifier [10]. In experimental setup, LibSVM library [11] for SVM classifier is used with Weka 3. We explore the effect of pruning and non-pruning on eight well known feature ranking algorithms (NDM, ACC2, IG, POIS, CHI, DFS, GINI, ODDS). After feature selection we evaluate characteristics of features on subsets of different sizes of top ranked features(10, 20, 50, 100, 200, 500,1000, 1500).

C. Evaluation Measures

Performance of classifiers is evaluated using macro and micro averaged F1 measure.

$$F1=\frac{1}{\frac{\alpha}{P_{recision}}+\frac{(1+\alpha)}{R_{ecall}}}=\frac{(\beta^2+1)\times P_{recision}\times R_{ecall}}{\beta^2\times P_{recision}+R_{ecall}} \qquad (12)$$

A combined measure obtained by joining precision and recall is F1 measure which is a weighted harmonic mean.

$$F1=\frac{2\times P_{recision}\times R_{ecall}}{P_{recision}+R_{ecall}} \qquad (13)$$

In macro average precision and recall are computed locally for each class then average is taken globally over each category. Mathematical formulation is given by Sebastiani [12]. Putting Eq. 14 into Eq. 13 gives the desired macro-averaged F1 measure. Macro-averaged assigns equal rank/weight to each class despite of class frequency [13]. Superscript denotes macro averaging.

$$R^m=\frac{\sum_{j=1}^{C} Rj}{C} \qquad P^m=\frac{\sum_{j=1}^{C} Pj}{C} \qquad (14)$$

In micro average F1 is measured globally for each class, where each class recall and precision are considered separately [13]. Micro average precision and recall are given as:

$$p^\mu=\frac{\sum_{j=1}^{C} t_{pj}}{\sum_{j=1}^{C}(tp_j+fp_j)} \quad R^\mu=\frac{\sum_{j=1}^{C} t_{pj}}{\sum_{j=1}^{C}(tp_j+fn_j)} \qquad (15)$$

D. Results

In this section Tables are shown which contain difference of pruned and unpruned F1 measure for eight feature ranking algorithms on all bench mark test points.

1) Wap Dataset: Performance of eight feature ranking met-rics on pruned and unpruned versions of WAP dataset using macro and micro F1 evaluation measure is shown in Figure 1 and 2. Classification results for macro and micro F1 measure on unpruned data is shown in Figure 1a and 2a respectively; results by applying pruning as a pre-processing step are shown in Figure 1b for macro F1 measure and in Figure 2b for micro F1 measure. We conclude that performance of chi square, gini index and poisson ratio is very low on unpruned data as compared to performance of these algorithms on pruned data. ACC2 has outperformed other seven metrics in case of micro F1 measure on unpruned data but on pruned data performance of ACC2 is decreased while NDM performance is enhanced by applying pruning. In case of macro F1 measure for subsets of 1000 to 1500 top ranked features DFS metric is the highest scorer on unpruned data whereas its performance considerably deteriorates on pruned data.

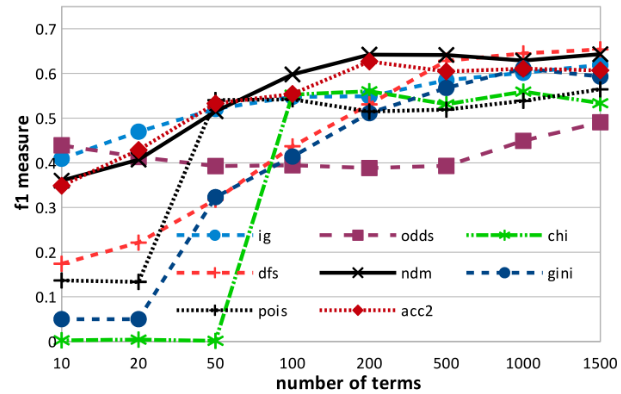
Table III and IV illustrate the percentage difference of the performance of eight feature ranking metrics on pruned and unpruned data for macro and micro F1 measure respectively on WAP dataset. As the difference table shows IG metric is better performer in case of unpruned data than pruned data. It is obvious from the difference table and we can deduce that an overall trend for WAP dataset is such that in which performance of ACC2, DFS and IG is high on unpruned data as compared to on pruned data. The performance of other five metrics is high on pruned data as compared to their performance on unpruned data.

TABLE III: Performance % difference Table of eight FR metrics on pruned and unpruned data for wap dataset using macro F1 measure; here Fx = F1 score of x metric on unpruned data - F1 score of x metric on pruned data

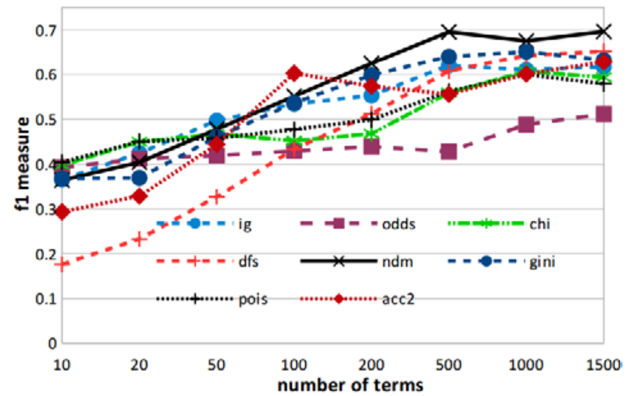
Features	Feature Ranking Algorithms							
	ΔF_{ig}	ΔF_{odds}	ΔF_{chi}	ΔF_{dfs}	ΔF_{ndm}	ΔF_{gini}	ΔF_{pois}	ΔF_{acc2}
10	4.356	4.641	-39.531	-0.221	-0.427	-31.830	-26.758	5.577
20	4.521	0.095	-44.581	-1.120	0.287	-31.971	-31.716	10.023
50	2.379	-2.641	-46.545	-0.960	3.743	-13.726	8.257	8.839
100	1.351	-3.494	10.073	0.319	4.689	-12.225	6.474	-4.831
200	-0.394	-5.069	9.250	1.908	1.750	-8.697	1.554	5.341
500	-3.498	-3.483	-2.769	2.076	-5.398	-7.239	-4.274	4.900
1000	-0.838	-3.906	-4.728	0.263	-4.641	-3.973	-6.085	0.996
1500	0.213	-2.099	-5.978	0.182	-5.321	-3.920	-1.443	-2.114

TABLE IV: Performance % difference Table of eight FR metrics on pruned and unpruned data for wap dataset using micro F1 measure; here Fx = F1 score of x metric on unpruned data - F1 score of x metric on pruned data

Features	Feature Ranking Algorithms							
	ΔF_{ig}	ΔF_{odds}	ΔF_{chi}	ΔF_{dfs}	ΔF_{ndm}	ΔF_{gini}	ΔF_{pois}	ΔF_{acc2}
10	2.886	0.483	-58.558	0.021	0.509	-50.746	-23.685	2.642
20	3.165	-0.441	-62.868	0.373	0.657	-53.442	-29.679	6.348
50	1.019	-1.206	-66.260	0.214	-0.411	-11.809	3.685	4.476
100	1.883	-3.311	1.678	3.104	2.129	-8.306	1.631	2.438
200	0.932	-5.495	-0.845	2.158	0.993	-4.538	0.166	0.976
500	0.282	-3.880	-0.839	0.357	-2.367	-2.841	-1.772	0.881
1000	1.001	-3.824	-1.880	0.816	-1.767	-2.024	-1.361	-0.082
1500	0.049	-4.064	-1.488	0.370	-1.930	-0.978	-2.391	-0.359

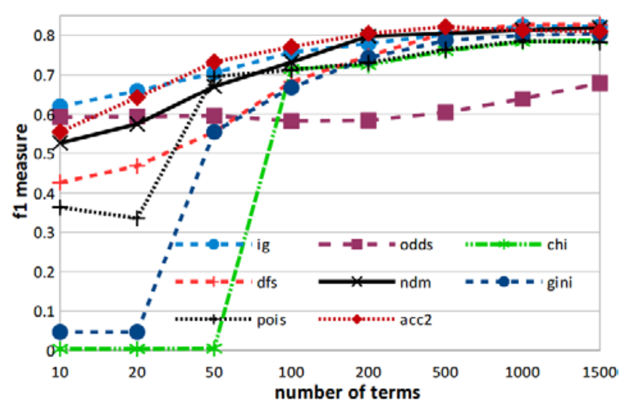


(a) Macro F1 measure evaluation using SVM

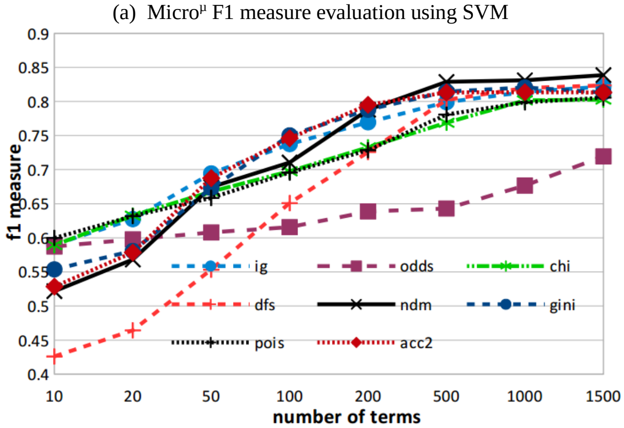


(b) Macro F1 measure evaluation using SVM classifier using pruning as pre processing step

Fig. 1 Graphical illustration of outcomes for classification on WAP dataset



(a) Micro$^{\mu}$ F1 measure evaluation using SVM



(b) Micro$^{\mu}$ F1 measure evaluation using SVM classifier using pruning as pre processing step

Fig. 2 Graphical illustration of outcomes for classification on WAP dataset

2) K1b Dataset: Results of difference between unpruned and pruned datasets after applying FR metrics for macro and micro F1 measures are shown in Tables V and VI. Each one of the three feature ranking metrics DFS, and IG have 18.75% performance on pruned data while 81.25% performance on unpruned data collectively using both micro and macro F1 measures. Gini index performed 0% on unpruned data for macro F1 measure and chi square also showed 0% performance for both micro and macro F1 measure. Performance of NDM, poisson and odds ratio is low on unpruned data.

TABLE V: Performance % difference Table of eight FR metrics on pruned and unpruned data for K1b dataset using macro F1 measure

TABLE VI: Performance % difference Table of eight FR metrics on pruned and unpruned data for K1b dataset using micro F1 measure; here Fx = F1 score of x metric on unpruned data - F1 score of x metric on pruned data

Features	Feature Ranking Algorithms							
	ΔF_{ig}	ΔF_{odds}	ΔF_{chi}	ΔF_{dfs}	ΔF_{ndm}	ΔF_{gini}	ΔF_{pois}	ΔF_{acc2}
10	6.994	3.487	-26.310	-0.049	-2.042	-21.011	-10.730	5.408
20	3.762	-0.157	-29.548	3.413	2.412	-23.797	-14.593	4.948
50	2.426	0.258	-32.678	1.472	0.808	-1.642	-1.773	2.528
100	0.939	0.155	-35.007	0.784	-0.011	-0.164	1.109	1.583
200	0.936	-0.351	-35.550	-0.067	0.374	-0.267	1.035	0.318
500	0.145	-0.075	-36.503	0.005	-1.121	-0.010	-0.163	-0.117
1000	-0.522	-1.440	-37.156	0.259	-0.327	0.106	-0.309	0.063
1500	0.144	-1.615	-36.943	0.070	-0.896	0.137	-0.669	-0.135

3) K1a Dataset: K1a dataset having percentage difference of values on pruned and unpruned data for eight feature ranking metrics at different test points is shown in Tables VII and VIII using macro and micro F1 measures respectively. DFS and IG have 25% performance on pruned data while 75% performance on unpruned data collectively using both micro and macro F1 measures. Performance of ACC2 metric on unpruned data is 81.25% and only 18.75 % on pruned data for both macro and micro F1 measures. Chi square and Gini Index on average attain highest values of F measure in 0% of cases using unpruned dataset. Performance of NDM, poisson and odds ratio is relatively high on pruned data.

TABLE VII: Performance % difference Table of eight FR metrics on pruned and unpruned data for K1a dataset using macro F1 measure; here Fx = F1 score of x metric on unpruned data - F1 score of x metric on pruned data

Features	Feature Ranking Algorithms							
	ΔF_{ig}	ΔF_{odds}	ΔF_{chi}	ΔF_{dfs}	ΔF_{ndm}	ΔF_{gini}	ΔF_{pois}	ΔF_{acc2}
10	4.656	1.477	-39.482	-0.052	2.454	-23.552	-25.637	3.335
20	5.359	2.382	-43.662	0.912	3.618	-28.328	-31.092	8.746
50	4.333	0.518	-47.957	0.804	2.212	-10.624	7.755	0.484
100	0.358	-2.573	-48.749	5.259	-2.710	-4.212	6.570	4.659
200	0.610	-2.191	-50.528	1.516	-2.888	-1.927	4.009	2.329
500	-4.522	-4.576	-59.696	-0.083	-2.536	-2.745	-5.283	8.236
1000	-2.168	-2.405	-63.744	0.269	-6.812	-5.625	-2.188	-0.454
1500	-1.237	-4.659	-66.006	0.591	-6.896	-2.505	-7.201	1.852

TABLE VIII: Performance % difference Table of eight FR metrics on pruned and unpruned data for K1a dataset using micro F1 measure; here Fx = F1 score of x metric on unpruned data - F1 score of x metric on pruned data

Features	Feature Ranking Algorithms							
	ΔF_{ig}	ΔF_{odds}	ΔF_{chi}	ΔF_{dfs}	ΔF_{ndm}	ΔF_{gini}	ΔF_{pois}	ΔF_{acc2}
10	3.151	-0.350	-61.665	-0.291	-1.523	-41.608	-23.512	2.888
20	4.762	-0.723	-64.890	0.258	-0.932	-46.047	-28.092	7.917
50	2.957	-1.684	-69.154	0.435	0.545	-11.383	1.935	3.034
100	1.658	-3.976	-70.888	2.635	0.893	-6.402	3.339	1.443
200	1.321	-5.775	-73.919	1.868	0.184	-2.867	1.740	1.043
500	-0.812	-3.373	-75.747	1.434	-2.151	-1.310	-0.997	1.103
1000	-0.261	-5.560	-79.611	0.142	-3.887	-1.464	-2.483	1.573
1500	-0.590	-3.619	-81.607	-0.090	-3.594	-0.742	-3.061	0.194

4) RE1 Dataset: Table IX shows that each of four feature

Features	Feature Ranking Algorithms							
	ΔF_{ig}	ΔF_{odds}	ΔF_{chi}	ΔF_{dfs}	ΔF_{ndm}	ΔF_{gini}	ΔF_{pois}	ΔF_{acc2}
10	4.366	0.600	-72.080	0.214	0.701	-42.287	-26.523	6.479
20	2.219	0.290	-73.895	4.138	1.324	-44.377	-27.076	2.951
50	2.468	-1.223	-75.119	1.805	1.133	-3.363	1.919	3.124
100	0.379	-0.614	-73.969	2.607	-0.036	-1.252	3.550	3.041
200	0.302	-0.498	-72.257	0.378	-1.610	-0.572	2.424	0.689
500	-0.321	-1.394	-75.043	0.105	0.436	-0.512	-2.345	1.924
1000	0.684	-0.969	-76.187	1.119	-0.664	-1.511	-2.847	1.373
1500	-0.442	-2.429	-73.843	-1.528	-0.682	-1.098	-1.892	-3.721

ranking metric IG, DFS, GINI and CHI perform better on unpruned data for five top ranked subset of features out of eight in case of macro F measure. Odds ratio, NDM and poisson ratio show 50% of performance for each of pruned and unpruned data and vice versa. Chi square performed better for unpruned data at one test point only. From Table X it can be seen that IG and DFS attain 100% performance on pruned data while ACC2 just performed 12.5% on unpruned data. NDM is the worst scorer for unpruned data showing 0% performance.

TABLE IX: Performance % difference Table of eight FR metrics on pruned and unpruned data for RE1 dataset using macro F1 measure; here Fx = F1 score of x metric on unpruned data - F1 score of x metric on pruned data

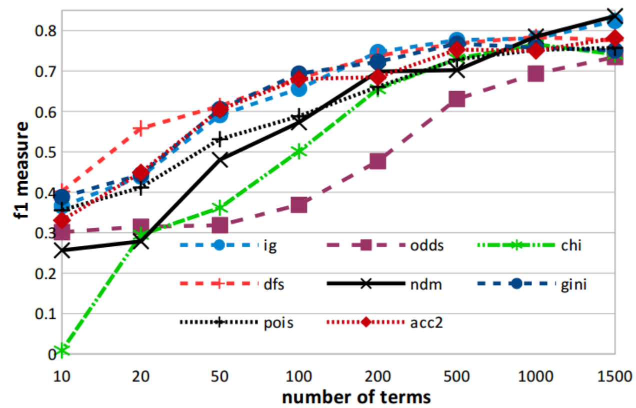
Features	Feature Ranking Algorithm							
	ΔFig	ΔFodds	ΔFchi	ΔFdfs	ΔFndm	ΔFgini	ΔFpois	ΔFacc2
10	1.555	1.904	-58.343	-0.532	3.687	3.490	0.740	6.048
20	1.533	-0.709	-3.762	1.733	2.723	3.936	-0.299	6.296
50	1.350	1.871	0.575	4.376	2.255	1.095	-1.433	-1.458
100	-0.686	-0.782	-1.212	3.377	-1.817	3.943	3.864	1.211
200	1.177	-0.944	-2.197	0.330	-0.963	-1.266	1.924	-1.980
500	-2.975	0.241	-2.130	-0.046	-0.575	4.895	-2.399	0.302
1000	-0.877	-0.370	-0.404	-0.354	-4.480	-0.280	0.746	-1.330
1500	0.465	1.720	-3.748	2.178	1.441	-0.117	-0.671	0.994

TABLE X: Performance % difference Table of eight FR metrics on pruned and unpruned data for RE1 dataset using micro F1 measure; here Fx = F1 score of x metric on unpruned data - F1 score of x metric on pruned data

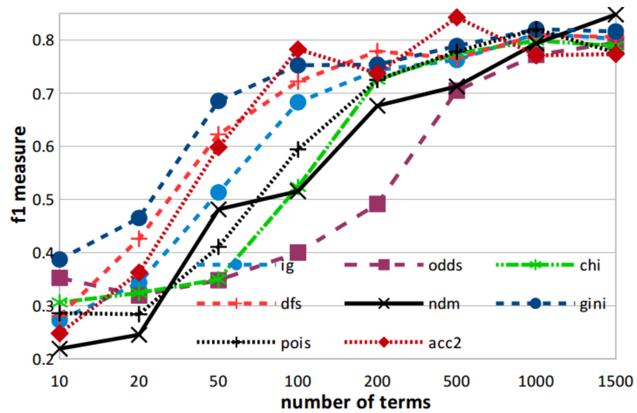
Features	Feature Ranking Algorithms							
	ΔF_{ig}	ΔF_{odds}	ΔF_{chi}	ΔF_{dfs}	ΔF_{ndm}	ΔF_{gini}	ΔF_{pois}	ΔF_{acc2}
10	1.345	0.000	-76.715	0.188	-1.987	0.022	-1.799	2.409
20	2.382	-0.740	-4.964	8.557	-0.506	-1.097	1.760	1.038
50	2.108	-0.308	1.059	2.025	-0.343	1.868	-0.181	2.558
100	1.831	-0.922	-0.109	1.876	-0.128	-0.441	-0.321	0.251
200	0.919	-0.666	-0.829	0.698	-0.180	-0.383	1.047	0.683
500	0.874	0.781	1.304	1.613	-0.072	0.607	-0.256	-0.122
1000	0.724	0.260	-0.934	1.610	-0.454	0.371	-0.281	0.750
1500	0.589	-0.629	-0.165	1.674	-0.182	-1.133	-0.058	0.952

5) RE0 Dataset: Figure 3 and 4 represents the result of micro and macro F1 measure for both pruned and unpruned version of data on RE0 dataset. Results on pruned dataset are shown in figure 3b and 4b for macro and micro F1 measure respectively. Figure 3 and 4 shows performance of chi square is very low on unpruned data as compared to its performance on pruned data. Conversely performance of DFS on unpruned data is relatively high as compared to its performance on pruned data.

Eight feature ranking metrics having percentage difference of performance for micro and macro F1 measure on RE0 dataset is shown in Table XI and XII. In case of micro F1 measure performance of IG and DFS is high on all test points for unpruned data and performance of ACC2 is high only at one point for pruned data, other five metrics show mixed performance. In macro F1 measure odds ratio performed better for pruned data on all test points, CHI and GINI show good performance on seven test points while DFS performed poor on six test points.

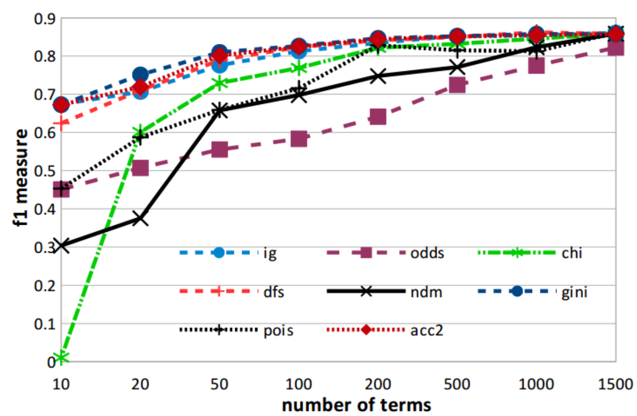


(a) Macro F1 measure evaluation using SVM with pruning

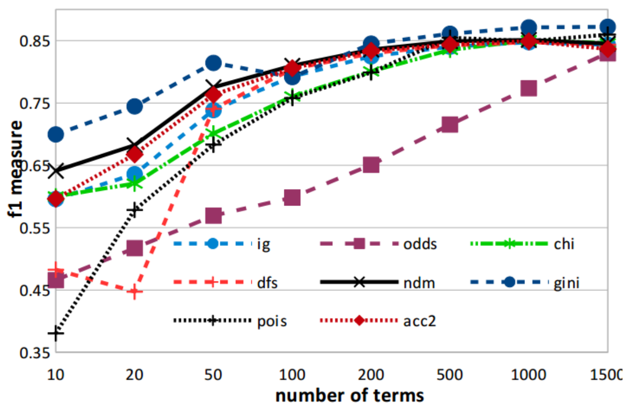


(b) Macro F1 measure evaluation using SVM

Fig. 3 Graphical illustration of outcomes for classification on RE0 dataset



(a) Micro$^{\mu}$ F1 measure evaluation using SVM with pruning



(b) Micro$^{\mu}$ F1 measure evaluation using SVM

Fig. 4 Graphical illustration of outcomes for classification on RE0 dataset

TABLE XI: Performance % difference Table of eight FR metrics on pruned and unpruned data for RE0 dataset using macro F1 measure; here Fx = F1 score of x metric on unpruned data - F1 score of x metric on pruned data

Features	Feature Ranking Algorithms							
	ΔF_{ig}	ΔF_{odds}	ΔF_{chi}	ΔF_{dfs}	ΔF_{ndm}	ΔF_{gini}	ΔF_{pois}	ΔF_{acc2}
10	9.308	-5.127	-29.764	12.132	3.724	0.106	6.961	8.246
20	9.563	-0.462	-2.943	13.204	3.330	-2.144	12.788	8.614
50	7.759	-2.959	1.210	-0.901	-0.093	-7.949	12.058	0.629
100	-2.676	-3.119	-2.237	-3.965	5.731	-5.891	-0.608	-10.223
200	0.362	-1.462	-7.013	-4.100	2.245	-3.093	-6.400	-5.283
500	1.411	-7.407	-4.383	0.280	-1.051	-2.115	-4.966	-8.974
1000	-3.144	-7.835	-3.296	-2.507	-0.844	-6.268	-6.775	-2.082
1500	2.128	-6.075	-4.762	-3.010	-1.217	-6.524	-1.908	0.795

TABLE XII: Performance % difference Table of eight FR metrics on pruned and unpruned data for RE0 dataset using micro F1 measure; here Fx = F1 score of x metric on unpruned data - F1 score of x metric on pruned data

Features	Feature Ranking Algorithms							
	ΔF_{ig}	ΔF_{odds}	ΔF_{chi}	ΔF_{dfs}	ΔF_{ndm}	ΔF_{gini}	ΔF_{pois}	ΔF_{acc2}
10	7.759	-1.511	-58.957	14.142	-33.734	-2.747	7.240	7.597
20	7.094	-0.987	-1.985	26.279	-30.687	0.645	0.878	5.191
50	3.709	-1.392	2.825	4.844	-11.767	-0.436	-2.435	3.843
100	2.014	-1.497	0.828	1.911	-11.210	3.509	-4.221	1.843
200	0.940	-0.973	2.107	1.127	-8.821	0.155	2.906	1.033
500	1.241	0.901	-0.319	0.730	-7.796	-0.915	-3.958	0.730
1000	0.682	0.136	-0.537	1.440	-2.816	-1.347	-3.671	0.536
1500	1.759	-0.772	1.245	1.447	1.345	-1.503	0.087	2.056

4. Discussion

Dimensionality reduction is an emerging area of research, which attempts to improve the accuracy and execution time of classification by choosing relevant features. Pruning is a preprocessing step used to remove noisy and out lier terms from training corpus. Too rare and too frequent terms are removed from the training corpus during pruning. In this paper, our focus is to study behavior of eight well known feature ranking metrics on pruned and unpruned datasets. Our experiments show some interesting results.

TABLE XIII: Datasets containing number of terms before and after pruning

Datasets	Wap	RE0	RE1	K1a	K1b
Number of Terms before pruning	8460	2886	3758	16383	16372
Number of Terms after pruning	6852	2327	3037	8589	8589

Table XIII represents the number of terms in the original dataset and number of terms after pruning. As we mention in the text that Fx = F1 score of x metric on unpruned data -F1 score of x metric on pruned data, So at a particular test point if the score of Fx is positive, its mean algorithm performed well on unpruned data as compared to its performance on pruned data. Conversely if the value of Fx is negative its mean performance of feature ranking metric at pruned data is low as compared to its performance on unpruned data. We calculate the percentage of number of cases when a FR metric shows positive

F1 score for macro and micro F1 evaluation measure on five benchmark datasets. Table XIV and XV show the percentage performance of eight feature ranking metrics for unpruned cases on five bench mark datasets.

In case of macro F1measure Table XIV show that on unpruned five datasets average performance of three feature ranking metrics ACC2, DFS and IG is 72.5%, 65% and 67.5% respectively (higher than 50%), conversely performance of these three FR metrics on pruned data is 27.5%, 35% and 32.5%, which shows that these three metrics performed better on unpruned data as compared to pruned data. It can also be seen that both micro and macro average performance of other five FR metrics (NDM, CHI, Odds, POIS, GINI) on unpruned data is poor than their performance on pruned data

TABLE XIV: FR metrics containing % of highest macro F1 values using unpruned data

FR metrics	RE0	Wap	RE1	K1b	K1a	average
ig	75	62.5	62.5	75	62.5	67.5
odds	0	25	50	25	37.5	27.5
chi	12.5	25	12.5	0	0	10
dfs	37.5	62.5	62.5	87.5	75	65
ndm	50	50	50	50	37.5	47.5
gini	12.5	0	62.5	0	0	15
pois	37.5	37.5	50	37.5	37.5	40
acc2	50	75	62.5	87.5	87.5	72.5

TABLE XV: FR metrics containing % of highest micro F1 values using unpruned data

FR metrics	RE0	Wap	RE1	K1b	K1a	average
ig	100	100	100	87.5	62.5	90
odds	25	12.5	37.5	37.5	0	22.5

chi	50	12.5	25	0	0	17.5
dfs	100	100	100	75	75	90
ndm	12.5	50	0	37.5	37.5	27.5
gini	37.5	0	50	25	0	22.5
pois	50	37.5	25	25	37.5	35
acc2	100	75	87.5	75	100	87.5

5. Conclusion

High dimensionality is an intrinsic property of text data. Filtering appropriate features to reduce dimensionality in order to improve classification performance becomes essential for text data. Feature ranking metrics are confused by the presence of too rare or too frequent terms and may select such features in the feature set. To study the effect of pruning, we performed feature selection using eight feature ranking metrics on pruned and un-pruned datasets. Our experimental results showed that ACC2, DFS and IG have in-built strength to deal with rare and frequent features, as their performance is degraded by applying pruning. Performance of other five feature ranking metrics (NDM, CHI, Odds, POIS, GINI) is degraded if pruning is not applied. Better performance of five feature ranking metrics on pruned data show that these feature ranking metrics include some too rare terms in the selected features by ranking them higher. It is also observed that terms which are more concentrated in one class than other classes are highly discriminative, as compared to the terms which are uniformly distributed in all classes.

References

[1] J.James “Data never sleeps 2.0”2014[Online].

[2] Liu, Bing, et al. ”Text classification by labeling words.” AAAI. Vol. 4.2004.

[3] Ozgur,¨ Levent, and Tunga Gung¨or¨. ”Two-Stage Feature Selection for Text Classification.” Information Sciences and Systems 2015. Springer International Publishing, 2016. 329-337.

[4] V. Srividhya and R. Anitha, ”Evaluating preprocessing techniques in text categorization,” International Journal of Computer Science and Applica-tion, vol. 47, no. 11, 2011.

[5] Aggarwal, Charu C., and ChengXiang Zhai. ”A survey of text classifica-tion algorithms.” Mining text data. Springer US, 2012. 163-222.

[6] Forman, George. ”A pitfall and solution in multi-class feature selection for text classification.” Proceedings of the twenty-first international con-ference on Machine learning. ACM, 2004.

[7] Forman, George. ”An extensive empirical study of feature selection metrics for text classification.” Journal of machine learning research 3.Mar (2003): 1289-1305.

[8] Maimon, Oded, and Lior Rokach. ”Introduction to knowledge discovery and data mining.” Data Mining and Knowledge Discovery Handbook. Springer US, 2009. 1-15.

[9] Law, Martin HC, Mario AT Figueiredo, and Anil K. Jain. ”Simultaneous feature selection and clustering using mixture models.” IEEE transactions on pattern analysis and machine intelligence 26.9 (2004): 1154-1166.

[10] [Li, Tao, Shenghuo Zhu, and Mitsunori Ogihara. ”Using discriminant analysis for multi-class classification: an experimental investigation.” Knowledge and information systems 10.4 (2006): 453-472.

[11] Chang, Chih-Chung, and Chih-Jen Lin. ”LIBSVM: a library for support vector machines.” ACM Transactions on Intelligent Systems and Tech-nology (TIST) 2.3 (2011): 27.

[12] Sebastiani, Fabrizio. ”Machine learning in automated text categoriza-tion.” ACM computing surveys (CSUR) 34.1 (2002): 1-47.

[13] Uysal, Alper Kursat, and Serkan Gunal. ”A novel probabilistic feature selection method for text classification.” Knowledge-Based Systems 36 (2012): 226-235.

[14] Danso, Samuel, Eric Atwell, and Owen Johnson. ”Linguistic and statis-tically derived features for cause of death prediction from verbal autopsy text.” Language processing and knowledge in the web. Springer Berlin Heidelberg, 2013. 47-60.

[15] Rehman, Abdur, Kashif Javed, and Haroon A. Babri. ”Feature selec-tion based on a normalized difference measure for text classification.” Information Processing & Management 53.2 (2017): 473-489.

[16] Chandrashekar, Girish, and Ferat Sahin. ”A survey on feature selection methods.” Computers & Electrical Engineering 40.1 (2014): 16-28.

[17] Hira, Zena M., and Duncan F. Gillies. ”A review of feature selection and feature extraction methods applied on microarray data.” Advances in bioinformatics 2015 (2015).

[18] Forman, George. ”An extensive empirical study of feature selection metrics for text classification.” Journal of machine learning research 3.Mar (2003): 1289-1305

[19] H. Ogura, H. Amano, and M. Kondo, ”Comparison of metrics for feature selection in imbalanced text classification,” Expert Systems with Applications, vol. 38, no. 5, pp. 4978-4989, May 2011.

[20] Chen, Jingnian, et al. ”Feature selection for text classification with Na¨ıve Bayes.” Expert Systems with Applications 36.3 (2009): 5432-5435.

[21] Uysal, Alper Kursat, and Serkan Gunal. ”A novel probabilistic feature selection method for text classification.” Knowledge-Based Systems 36 (2012): 226-235.

[22] Ogura, Hiroshi, Hiromi Amano, and Masato Kondo. ”Feature selection with a measure of deviations from Poisson in text categorization.” Expert Systems with Applications 36.3 (2009): 6826-6832.

[23] Danso, Samuel, Eric Atwell, and Owen Johnson. ”Linguistic and statis-tically derived features for cause of death prediction from verbal autopsy text.” Language processing and knowledge in the web. Springer Berlin Heidelberg, 2013. 47-60.

[24] Rehman, Abdur, Kashif Javed, and Haroon A. Babri. ”Feature selec-tion based on a normalized difference measure for text classification.” Information Processing & Management 53.2 (2017): 473-489.

[25] Chandrashekar, Girish, and Ferat Sahin. ”A survey on feature selection methods.” Computers & Electrical Engineering 40.1 (2014): 16-28.

[26] Hira, Zena M., and Duncan F. Gillies. ”A review of feature selection and feature extraction methods applied on microarray data.” Advances in bioinformatics 2015 (2015).
[27] Forman, George. ”An extensive empirical study of feature selection metrics for text classification.” Journal of machine learning research 3.Mar (2003): 1289-1305
[28] H. Ogura, H. Amano, and M. Kondo, ”Comparison of metrics for feature selection in imbalanced text classification,” Expert Systems with Applications, vol. 38, no. 5, pp. 4978-4989, May 2011.
[29] Chen, Jingnian, et al. ”Feature selection for text classification with Na¨ıve Bayes.” Expert Systems with Applications 36.3 (2009): 5432-5435.
[30] Uysal, Alper Kursat, and Serkan Gunal. ”A novel probabilistic feature selection method for text classification.” Knowledge-Based Systems 36 (2012): 226-235.
[31] Ogura, Hiroshi, Hiromi Amano, and Masato Kondo. ”Feature selection with a measure of deviations from Poisson in text categorization.” Expert Systems with Applications 36.3 (2009): 6826-68

Authors Profile

Muhammad Nabeel Asim received his Bachelor degree from University of Management and Technology (UMT) and Masters degree in Electrical Engineering from University of Engineering and Technology, Lahore, Pakistan. Currently working as Research Officer at Al-Khawarizmi Institute of Computer Science (KICS)
University of Engineering and Technology (UET), Lahore Pakistan. His research interests are Bioinformatics, Artificial Intelligence, Machine Learning

Completed his PhD from UET Lahore in computer science. During his PhD, he has been part of Al-Khawarizmi Institute of Computer Science, UET Lahore for 8 year working as a researcher on different posts. His core expertise are in the field of machine learning and his focused area of research is “Text Classification”. He is currently working as Assistant Professor in Department of Computer Science, University of Gujrat.

Dr. Muhammad Idrees Completed his PhD from UET Lahore in computer science. He worked as an Assistant Professor and Head of Department of Computer Science and Information Technology, in BZU, Lahore and UoS, Lahore Campus for four years since 2011 to 2016. Now, he is working as an Assistant Professor and Head of Department in Computer science and Engineering, in University of Engineering and Technology Lahore, Narowal Campus since 2016 to date. His research interests include Bioinformatics, Databases, Software Engineering, Network Security, Artificial Intelligence and Embedded Systems.