

ANALYSIS OF THE ROLE OF ENTROPIES AND ASSIGNMENT OF RANKS TO THE FEATURES IN GENRE DISCRIMINATION

HEMLATA PANDE^a AND H. S. DHAMI^b

Abstract

Genre or style, an important property of text, and automatic text genre discrimination is becoming important for classification and retrieval purposes as well as for many natural language processing tasks. Various methods with feature cue vectors have been used for genre discrimination, which utilize different statistical measures corresponding to a range of linguistic features. Since all features of a text are not equally useful so for better results and to reduce the dimensionality of used spaces, it becomes necessary to select those features which have the capacity to discriminate between categories in a better manner. In order to achieve this aim, a technique known as feature selection is commonly applied in text classification. These techniques attempt to select the subset of features in text categorization that are most useful for the categorization task. We propose to use a ranking technique for feature selection with the help of a completely randomized design method for features and square deviations of property weights with their applications to familiar measures such as syntactic features, vocabulary features and 'most frequent words'. It has been established that besides part of speech probabilities, their entropies play an important role in the discrimination process.

Keywords: Categorization, Entropy, Features, Rank, Completely randomized design, Square deviations.

1. Introduction

Text classification research has been mostly focused on subject or prepositional content of text. As text-based applications become more varied and the amount of information increases tremendously, different aspects of text can be useful for various purposes including classification. We focus on text genre or the style of text that often characterizes the purpose for which the text has been written. Examples for genre are: research article, novel, poem, news article, editorial, homepage, advertisement, manual, court decision etc.

Text categorization is the task of automatically sorting a set of documents into categories from a predefined set. This task, that falls at the crossroads of information retrieval and machine learning, has witnessed a booming interest in last few years due to its applications in automated indexing of scientific articles according to predefined thesauri terms, filing patents into patent directories, selective dissemination of information to information consumers, automated population of hierarchical catalogues of web resources, spam filtering, identification of document genre, authorship attribution, survey coding, and even automated essay grading. Classificatory tasks in computational

^a Dept. of Mathematics, University of Kumaun, S. S. J. Campus Almora, INDIA
E-mail: h.lata.pande@gmail.com, pande_hemlata1011@yahoo.com

^b Dept. of Mathematics, University of Kumaun, S. S. J. Campus Almora, INDIA
E-mail: profdhami@rediffmail.com

linguistics are performed on the basis of text genre detection and authorship attribution. Genre determination of text usually refers to identification of the kind of the text and is very useful in many natural language processing tasks owing to the fact that the approaches to text genre detection are focused on the use of simple measures that can be easily detected and reliably counted, as discussed in the works of Kessler et al (1997). Parsing accuracy and POS tagging accuracy can be increased by taking genre into account. Word senses are also related to the genre of the text, and in information retrieval genre could help to give better search results. A lot of work has been done in text discrimination with the help of different statistical measures. In a comparative analysis of written and spoken genres of English, Biber (1988) has listed an array of text features. He has also outlined five dimensions which can be used to characterize genre (Biber 1989). Karlgren and Cutting (1994) have taken a set of properties of these distinguishing dimensions and has used them to compute a score to categorize texts in pre defined text genre categories. Lee and Myaeng (2004) have used a deviation formula and a discrimination formula of statistical features for feature extraction from training documents. They have also tested the viability of using genre classification information in the automatic classification of documents into subject areas and have concluded that the knowledge of the genre of a given document helps in correct classification. Kelih et al (2005) have tested impact of word length for genres and authors by taking Russian letters and poems written by three different authors and have found that word length is characteristic of genre rather than of authorship. They (Kelih et al, 2006) have also tested the impact of sentence length in text typology. Stamatatos et al (2000a) have used the frequencies of occurrence of the most frequent words as discriminators of text genre and have tested the role of frequencies of occurrence of the most common punctuation marks for accurate text categorization. Stamatatos et al in another work (2000b) have used multiple regression analysis and discrimination analysis for 22 vectors of style marker of token level, syntax level and analysis level for text categorization in terms of genres and also in terms of authors. They have observed that the size of the training set plays a significant role in the results and have also calculated the contribution of each level style markers to the classification procedure. Wolters and Kirsten (1999) have explored the impact of the inclusion of function word frequencies and POS frequencies in genre classification. Klavans and Kan (1998) have focused on the determination of the role of verbs in document analysis by taking verb profile as one of many factors in the determination of text type. Borja and Chappas (2006) have compared the style of various forms of media texts using measures of vocabulary richness and have also used three parameters of Sichel distribution for this purpose. Antic et al (2006) have studied the suitability of parameters of two parametric Poisson models in discriminating text sorts. Stadlober and Djuzelic(2006) have applied multivariate statistical methods for text classification by text length and by moments of the distribution of word length.

In the paper, we have determined the results for the classification process for the four kinds of articles, by taking three types of features separately. The three types of features have been combined in order to obtain better results for the process. Ranking technique has been selected for inferring about the order of the features, that is, which feature has to be given preference over others? The process has been to prefer the features with the lower rank over the features with higher rank so that better results for the classification process could be obtained. The section 3 of the paper introduces the concept of entropies of POS (Parts of speech) and tests their role in discrimination of texts in various categories.

Classification results for three different types of features - Our analysis of 95 categorized test articles (defined later) by discriminant analysis with 3 different sets of vectors, 11 parts of speech frequencies, 5 vocabulary richness measures and probabilities of 48 most frequent words, revealed that the average of calculated F-values (

$$F = \frac{1}{\frac{\alpha}{P} + \frac{(1-\alpha)}{R}} \quad \text{for } \alpha = 0.5, \text{ where P and R are precision and recall respectively)}$$

for different categories are only 0.59, 0.58, 0.67 respectively, whereas the desired values are closer to 1. In order to improve the results for better categorization of texts, we have selected a combination of features from parts of speech properties, from vocabulary measures and from the most frequent words.

2. Method

For the current study, 188 articles, 96 press news reportages, 44 press editorials, 23 creative writings (novels and short stories) and 25 recipes have been taken from the ICE India Corpus^c compiled by Shivaji University, Kolhapur, India, and by Freie Universität Berlin, Germany. 49% of all the articles (48 reportages, 22 editorials, 11 creative writings and 12 recipes) were taken as training articles and the remaining (48 reportages, 22 editorials, 12 creative writings and 13 recipes) have been used for the testing purposes. For categorizing the 95 test articles we have used the linear discriminant analysis (with the help of statistiXL^d) along the lines of Karlgren and Cutting (1994). The results of this process depend on selected features because all features are not equally effective in the discrimination process. Three types of features were used: POS features, vocabulary features and most frequent words in the following manner-

i) Eleven part of speech features: namely frequencies of occurrence of noun, pronoun, adjective, adverb, verb, determiner, conjunction/preposition, modal, *to*, cardinal number and interjection have been selected by tagging the texts with the Go tagger^e.

ii) Five vocabulary richness measures propounded by Yule, Honore, Brunet, Sichel, and Simpson as cited in Stamatatos et al. (2000b) have been used, where V is vocabulary of a text N is total word tokens in the text and V_i is number of words i times

$$\left. \begin{aligned} K &= \frac{10^4 (\sum_{i=1}^{\infty} i^2 V_i - N)}{N^2}, \quad R = \frac{(100 \log N)}{\left(1 - \left(\frac{V_1}{V}\right)\right)}, \quad W = N^{V^{-\alpha}}, \\ S &= \frac{V_2}{V}, \quad D = \sum_{i=1}^V V_i \frac{i(i-1)}{N(N-1)} \end{aligned} \right\} \dots\dots\dots(2.1)$$

iii) Forty eight most frequent words of English from the 50-words list of Stamatatos et al (2000a) have been selected except *n't* and *'s*.

The calculated values of precision and recall, defined in equation (2.2), for the above three cases have been enumerated in the table1.

^c <http://www.ucl.ac.uk/english-usage/ice/index.htm>.

^d <http://www.statistixl.com>.

^e <http://uluru.lang.osaka-u.ac.jp/~k-goto/download/download.cgi?name=GoTagger07.zip>

Table 1. Values of precision and recall for categorization of text articles by POS frequencies, vocabulary measures and 48 most frequent words

| Category | 11 POS frequencies | | Vocabulary measures | | 48 most frequent words | |
|--------------------|--------------------|--------|---------------------|--------|------------------------|--------|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| 1.Recipes | 0.455 | 0.385 | 0.529 | 0.692 | 0.765 | 1.0 |
| 2.Creative-writing | 0.889 | 0.667 | 0.5 | 1.0 | 0.778 | 0.583 |
| 3.Editorials | 0.55 | 0.5 | 0.5 | 0.773 | 0.444 | 0.545 |
| 4.Reportages | 0.618 | 0.708 | 0.75 | 0.3125 | 0.691 | 0.604 |
| Average Value: | 0.628 | 0.565 | 0.569 | 0.694 | 0.669 | 0.683 |

$$\left. \begin{aligned}
 \text{precision} &= \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\text{retrieved documents}|} \\
 \text{recall} &= \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\text{relevant documents}|}
 \end{aligned} \right\} \dots\dots\dots (2.2)$$

Table 1 depicts that by using the three types of features, the values of precision and recall are not very fine. We have made an attempt in this paper to apply the concept of entropy, and of ranks for features in order to improve text categorization results. Ranking Technique is a way of gathering information on a specific subject, and ranking the importance of the information, in order to meet some decision-making objective. e.g., by gathering the information for the data of the feature values corresponding to various features for the considered training articles (the set of articles, each labeled with their kind and used for extraction of some statistical measures corresponding to some features of texts) to determine - what feature has the highest priority for the categorization of these (training) articles as well as the remaining set of articles called the testing articles.

3. Selection of Features and Role of Entropies of POS

We have chosen a combination of three different types of features by applying filters of completely randomized design and ‘square deviations from the mean of feature weights’, since the number of features was large.

Assuming recipes, creative-writings, editorials and reportages as first, second, third and fourth categories respectively, we have calculated the differences in means of feature values for each category with respect to the other categories by applying Completely Randomised Design with hypothesis -

H_0 : the means of different treatments do not differ significantly and

H_1 : there is a significant difference in means,

For the purpose we have assumed 4 different treatments corresponding to determination of the weights of features for training articles of 4 categories, that is, assuming category 1 for recipes, 2 for creative writings, 3 for press editorials and 4 for press news reportages, treatment 1 has been taken for the determination of weights for the training articles of recipe and similarly for the other categories. If a selected feature f has significant difference in means for any one of the four categories with respect to the other categories, we have denoted denote this as $\langle(i,j)\rangle_f$ (where i and j represent any two categories having significance difference in their mean values for the property f , i and j both varying from 1 to 4).

For better results in categorization of test texts besides significant difference in means, it is also necessary that various texts of a particular category take feature values which do not differ very much from the mean value of the feature for that category. It shall enable the comparison of means of different categories as comparison of the texts of the different categories and thus the sum of square deviations of feature values of objects from a particular category from their respective means should be lower. But as for different features, feature values are in different ranges so for comparison of their square deviations we have taken weight of a feature in such a way that it is independent of scale.

Defining the weight of a feature as the feature value divided by its maximum value in 93 training texts, we have

$$w_{pq} = \frac{|f_{pq}|}{\max_{1 \leq p \leq 4} |f_{pq}|} \quad q=1,2,3,\dots,n_p \dots\dots\dots(3.1) \text{ where } f_{pq} \text{ represents feature value of } q^{\text{th}} \text{ object}$$

of p^{th} category n_p is no. of training texts in p^{th} category.

Since the square deviation increases with the number of objects, we have computed the values for the resultant variance S_{ij}^2 corresponding to each pair (i, j) of categories with significance difference in their means of the feature by the following formula:

$$s_{ij}^2 = \frac{\sum_{k=1}^{n_i} (w_{ik} - \bar{w}_i)^2 + \sum_{k=1}^{n_j} (w_{jk} - \bar{w}_j)^2}{n_i + n_j - 2} \dots\dots\dots (3.2) ,$$

n_i is the number of objects of the i^{th} category.

The average value of S_{ij}^2 has been determined for all discriminating pairs, e.g. if the means of a feature f differ significantly for the first and second category and for second and fourth category only, then we have denoted this as $\langle(1,2),(2,4)\rangle_f$ and have determined the values of s_{12}^2 and s_{24}^2 from equation (3.2) and then have calculated their average value as : $\text{Average}(s_{ij}^2) = \text{average of } s_{12}^2 \text{ and } s_{24}^2$. The process has been thrashed out in detail for the property ‘occurrence of noun in text’ as-

The frequency of occurrence of noun in a text has been determined for all the considered texts and feature has been taken 'probabilities of occurrence of noun in the text'. We have obtained the weights for this feature, as defined by the equation (3.1) for the 4 categories and the determined values have been shown in Table2.

Table2

| Treat-ments | | Weights.(w _{pq}) |
|-------------|-----------|--|
| 1 | Category1 | 0.685, 0.637, 0.706, 0.685, 0.667, 0.733, 0.640, 0.692, 0.767, 0.851, 0.827, 0.833 |
| 2 | Category2 | 0.479, 0.416, 0.385, 0.377, 0.363, 0.390, 0.402, 0.316, 0.457, 0.413, 0.295 |
| 3 | Category3 | 0.513, 0.534, 0.517, 0.491, 0.513, 0.508, 0.395, 0.554, 0.481, 0.465, 0.564, 0.500, 0.554, 0.539, 0.431, 0.464, 0.463, 0.489, 0.468, 0.494, 0.526, 0.472 |
| 4 | Category4 | 0.543, 0.637, 0.607, 0.833, 0.679, 0.733, 0.691, 0.627, 0.710, 0.680, 0.427, 0.687, 0.743, 0.668, 0.619, 0.535, 0.594, 0.574, 0.688, 0.558, 0.679, 0.489, 0.454, 0.975, 0.690, 0.845, 0.693, 0.685, 0.871, 0.877, 0.748, 0.573, 0.556, 0.808, 0.717, 0.520, 0.741, 0.577, 0.769, 0.635, 0.616, 0.718, 0.507, 0.717, 0.964, 0.643, 0.648, 1.000 |

If $\sum_j y_{ij} = T_i$, be the observation total of i^{th} treatment which is replicated r_i times ($i = 1, 2, 3, 4$), then we have

$T_1 = 8.724, T_2 = 4.293, T_3 = 10.937, T_4 = 32.546, r_1 = 12, r_2 = 11, r_3 = 22, r_4 = 48$.
Therefore

$$\text{Treatment sum of squares} = \sum_{i=1}^4 \frac{T_i^2}{r_i} - \frac{(\sum_{i=1}^4 T_i)^2}{\sum_{i=1}^4 r_i} = 1.197,$$

$$\text{Total sum of squares} = \sum_{i,j} y_{ij}^2 - \frac{(\sum_{i=1}^4 T_i)^2}{\sum_{i=1}^4 r_i} = 2.106,$$

Thus the analysis of variance table can be designed in the form of Table 3.

Since the tabulated value of F at 0.05 level of significance is less than the calculated value, we can reject H_0 at 5% level of significance which implies that the means of at least weights of two categories differ significantly at 5% level.

Table3. Analysis of variance table.

| Source of variation | Degrees of freedom (d.f.) | Sum of squares | Mean sum of squares | F |
|---------------------------|---------------------------|----------------|---------------------|-------|
| Treatments | 3 | 1.197 | 0.399 | 39.06 |
| Error (within treatments) | 89 | 0.909 | 0.0102 | |
| Total | 92 | | | |

For pairs of treatments (i, j) with significant differences in their means, we should have:

$$|\mu_i - \mu_j| > t_{0.05} \text{ for error d.f.} \times \sqrt{S_E^2 \left(\frac{1}{r_i} + \frac{1}{r_j} \right)},$$

μ_i be the mean of i^{th} treatment .

Thus probability of 'noun' can be differentiated among 5 pairs out of 6 pairs of categories $\langle (1,2), (1,3), (2,3), (2,4), (3,4) \rangle$ *Noun probability.*

Calculated values of S_{ij}^2 for weights of nouns for above categories are:

$$s_{12}^2 = 0.00437, s_{13}^2 = 0.00309, s_{13}^2 = 0.00211, s_{24}^2 = 0.01422, s_{34}^2 = 0.0120,$$

the average value of these, $\text{Average}(S_{ij}^2)$ is 0.00716. Similarly the average value of

S_{ij}^2 for all features that have significance difference in means for at least two categories has been obtained.

We know that any feature can be considered as a better discriminator among all other comparing features if it differs significantly in all the 6 pairs and also has lower value of average standard deviation as compared to other features. To meet this requirement, we have ranked features in two ways:

- by number of pairs which the feature can discriminate and
- by value of $\text{Average}(S_{ij}^2)$

We determined the average value of ranks assigned by the above-mentioned two ways and arranged features first by ascending order of their average ranks. If the averages ranks were the same then we arranged the features in ascending order on the basis of their ranks obtained by first way. Where these values were also the same, we formed the order on the basis of their average (S_{ij}^2) . This process was applied to different part of speech features, vocabulary features and most frequent words in the following manner:

(1) Part of Speech features: For these features we tabulated 28 parts of speech tags (S. No. 1-28) out of 36 parts of speech tags (excluding /FW, /LS, /POS, /SYM, /NN, /NNPS, /NNP, /NNS) as assigned by the Go tagger (In order to reduce the effect of capitalized words we considered only nouns and not its different sub-categories) while the other seven parts of speech (S. No. 29-35) were calculated by the assigned tags. The probability of each part of speech was calculated by assuming 29 POS in all (S. No. 1-29) as all words must be assigned to one of these 29 parts of speech (assigned as a noun or any POS from S. no. 1-28).

Table 4. Part of speeches and used tags.

| S. No | Part of Speech | S. No. | Part of Speech | S. No. | Part of Speech |
|-------|--|--------|--|--------|------------------------------------|
| 1 | Coordinating conjunction (/CC) | 13 | wh-determiner (/WDT) | 25 | Pre-determiner (/PDT) |
| 2 | "Adjective, comparative" (/JJR) | 14 | wh-adverb (/WRB) | 26 | Particle (/RP) |
| 3 | Possessive pronoun (/PRP\$) | 15 | Non wh-determiner (/DT) | 27 | Modal (/MD) |
| 4 | "Adverb, superlative" (/RBS) | 16 | Adjective, non superlative & non comparative (/JJ) | 28 | To (/TO) |
| 5 | "Verb, gerund/present participle" (/VBG) | 17 | Personal pronoun (/PRP) | 29 | Noun (/NNPS+/NNP+/NNS+/NN) |
| 6 | "Verb, non 3rd ps. sing. Present" (/VBZ) | 18 | "Adverb, comparative" (/RBR) | 30 | Pronoun (/PRP\$+PRP+WP\$+WP) |
| 7 | Possessive wh-pronoun (/WP\$) | 19 | "Verb, past tense" (/VBD) | 31 | Adjective (/JJR+/JJS+/JJ) |
| 8 | Preposition/subord. Conjunction (/IN) | 20 | "Verb, non-3rd ps. sing." (/VBP) | 32 | Adverb (/RBR+/RBS+/WRB+/RB) |
| 9 | "Adjective, superlative" (/JJS) | 21 | wh-pronoun (/WP) | 33 | Verb (/VB+/VBD+/VBG+/VBN+/VBP+VBZ) |
| 10 | Adverb [non wh, non superlative & non comparative] (/RB) | 22 | Existential there (/EX) | 34 | Determiner (/DT+/WDT) |
| 11 | "Verb, base form" (/VB) | 23 | Interjection (/UH) | 35 | Preposition/conjunction (/CC+/IN) |
| 12 | "Verb, past participle" (/VBN) | 24 | Cardinal number (/CD) | | |

Garcia et al (2006) described a new approach to text categorization where the focus is on the amount of information (entropy) in the text. We have utilized the concept of entropy by considering all the sub-categories of parts of speech mentioned in above table from numbers 1-29 and by considering only 14 categories of parts of speech (numbers 22-35) and have represented them by en_1 and en_2 respectively. Thus

$$en_1 = -\sum_i p_i \log_2 p_i, i = 1, 2, \dots, 29, \quad p_i \text{ be the probability of } i^{\text{th}} \text{ POS in}$$

text by assuming 29 categories of POS, (S. No. 1-29 in Table 4.)

$$en_2 = -\sum_i p_i \log_2 p_i, i = 1, 2, \dots, 14, \quad p_i \text{ be the probability of } i^{\text{th}} \text{ POS in}$$

text by assuming 14 categories of POS, (S. No. 22-35 in Table 4.)

Entropies of pronouns were calculated by considering four sub-categories of pronouns, namely possessive pronoun, personal pronoun, wh-pronoun and possessive wh-pronoun, of adjective from its three sub-categories, adverb from 4 sub-categories, verb from 6 sub-categories, determiner from 2 sub-categories and preposition/conjunction from 2 categories. The entropy of that part of speech has been taken equal to zero for which no sub-categories have been considered as is the case of noun.

If a POS with non-zero entropy has n subcategories then its maximum entropy will be

$$n \log_2 n \text{ and therefore relative entropy} = \frac{\text{entropy}}{n \log_2 n}.$$

For each POS with non-zero entropy, relative entropy was calculated and a new feature 'sum of entropies of POS' denoted by 'sum' is defined as:

$$\text{sum} = \sum_i (\text{entropy})_i \dots \dots \dots (3.3), \text{ where } (\text{entropy})_i \text{ is relative entropy of } i^{\text{th}}$$

POS.

The suitability of these entropic measures can be visualized in the discrimination of different categories. We can demonstrate this fact with the help of following figure in which the plotted values of en_1 for 93 training texts depict significant differences in the three categories of recipes, editorials and creative-writings. The figure also demonstrates the applicability of this measure in the measurement of the discrimination between creative-writings and reportages.

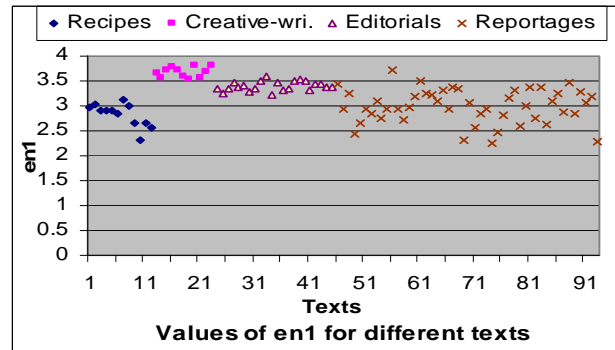


Figure1. Calculated values for training texts.

On the basis of training articles of three types (recipes, creative-writings, editorials), we can consider the numerical values in the following manner:

For creative-writings $en_1 > 3.54$, in case of editorials $3.12 < en_1 < 3.54$, and for recipes $en_1 < 3.12$. We have then classified texts of these three categories on the base of the values of en_1 , as shown in the following classification table for the test articles of the three types:

Table5. Classification of testing articles among three categories based on en_1 value.

| | Recipes | Creative-writings | Editorials | Total |
|-------------------|---------|-------------------|------------|-------|
| Recipes | 13 | 0 | 0 | 13 |
| Creative-writings | 0 | 10 | 2 | 12 |
| Editorials | 1 | 1 | 20 | 22 |

The table exhibits that the classification exercise is 100% for recipes, 91% for Editorials and 83.3% for creative writings.

In a similar manner the separation process between reportages and creative-writings can be classified by taking: $en_1 < 3.5146$ for reportages, which results in correct classification of 11 creative writings out of 12 and 44 reportages out of 48.

Application of this approach in the selection of ranges of en_2 among recipes, creative-writings and editorials under the conditions- for creative-writings $en_2 > 2.861$, for editorials $2.63 < en_2 < 2.861$, and for recipes $en_2 < 2.63$ and also the condition $en_2 < 2.85$ for creative-writings between reportages and creative writings yielded more than 92% average accuracy rate for the three categories of creative-writings, recipes, editorials and also for discrimination of reportages from creative writings in training texts. And respectively had a discrimination of 100% for recipes, 83.3% for creative-writings, 72.7% for editorials correctly among three categories and an average 88.5% between creative writings and reportages for the test texts. The feature 'sum' described by equation (3.3) was used in separation of sets of categories in case of editorials from recipes and creative-writings from recipes.

(2) Vocabulary measures: The five measures under this category were obtained with the help of 5 vocabulary richness statistics defined by equation 2.1.

(3) Most frequent words: We selected 47 words from the list of 50 most frequent words (excluding *n't*, *'s* on account of their meager occurrence and "to", which has been already considered as /TO in POS features) mentioned by Stamatatos et al (2000a).

We observed that out of all these 90 features, 69 features differ in at least 2 categories and the values of $Average(S_{ij}^2)$ for these features lie in the range 0.003-0.057. The process of assignment of ranks was conducted in the following two manners:

a) For categories: Ranks 1, 2, 3, 4, 5, 6 were assigned to features which have significant difference in means for 6, 5, 4, 3, 2, 1 pairs of categories respectively.

b) For $Average(S_{ij}^2)$: Ranks 1, 2, 3, 4, 5, 6 were assigned if the value of $Average(S_{ij}^2)$ for the feature lay in ranges 0-.01, .01-.015, .015-.02, .02-.025, .025-.035, 0.35-0.45 respectively while the rank 7 was assigned for higher values of $Average(S_{ij}^2)$. The motive behind this ranking was to assign lower ranks to features with a lower value of $Average(S_{ij}^2)$. The justification behind assigning the highest value of the rank equal to 7 is that ranks depending on categories are only up to 6 and if we consider much higher values of ranks for $Average(S_{ij}^2)$ then the average of the two ranks will be in the same order as that of II rank.

The discussed features generate following table for assignment of ranks.

Table6. Features arranged by their ranks.

| | Feature | $Average(S_{ij}^2)$ | Rank by I manner | Rank by II manner | Average Rank |
|----|--|---------------------|------------------|-------------------|--------------|
| 1 | Vocabulary richness function of Simpson, D | 0.010997 | 1 | 2 | 1.5 |
| 2 | Vocabulary richness function of Yule, K | 0.011012 | 1 | 2 | 1.5 |
| 3 | Pronouns | 0.011607 | 1 | 2 | 1.5 |
| 4 | Verbs | 0.012158 | 1 | 2 | 1.5 |
| 5 | en ₁ | 0.003395 | 2 | 1 | 1.5 |
| 6 | Nouns | 0.007161 | 2 | 1 | 1.5 |
| 7 | Subordinate conjunctions or prepositions | 0.009721 | 2 | 1 | 1.5 |
| 8 | Of* | 0.019679 | 1 | 3 | 2 |
| 9 | en ₂ | 0.003486 | 3 | 1 | 2 |
| 10 | Prepositions or conjunctions | 0.009981 | 3 | 1 | 2 |

| | | | | | |
|----|--|----------|---|---|-----|
| 11 | Adverbs | 0.020742 | 1 | 4 | 2.5 |
| 12 | Cardinal numbers | 0.015923 | 2 | 3 | 2.5 |
| 13 | Determiners | 0.017864 | 2 | 3 | 2.5 |
| 14 | Non wh-determiners | 0.0189 | 2 | 3 | 2.5 |
| 15 | Personal pronoun | 0.012377 | 3 | 2 | 2.5 |
| 16 | Vocabulary richness function of Honore, R | 0.013688 | 3 | 2 | 2.5 |
| 17 | Vocabulary richness function of Sichel, S | 0.014355 | 3 | 2 | 2.5 |
| 18 | Vocabulary richness function of Brunet, W | 0.006452 | 4 | 1 | 2.5 |
| 19 | Adverbs which are not either wh-adverbs or superlative or comparative | 0.02273 | 2 | 4 | 3 |
| 20 | sum | 0.018079 | 3 | 3 | 3 |
| 21 | Were* | 0.002379 | 5 | 1 | 3 |
| 22 | The* | 0.026484 | 2 | 5 | 3.5 |
| 23 | Verb, past tense | 0.029389 | 2 | 5 | 3.5 |
| 24 | Possessive pronoun | 0.030108 | 2 | 5 | 3.5 |
| 25 | Verb, gerund/present participle | 0.018569 | 4 | 3 | 3.5 |
| 26 | Would* | 0.012267 | 5 | 2 | 3.5 |
| 27 | Wh-pronouns | 0.048334 | 1 | 7 | 4 |
| 28 | Have* | 0.035895 | 2 | 6 | 4 |
| 29 | That* | 0.037309 | 2 | 6 | 4 |
| 30 | Was* | 0.04017 | 2 | 6 | 4 |
| 31 | Had* | 0.044692 | 2 | 6 | 4 |
| 32 | Coordinating conjunction | 0.02787 | 3 | 5 | 4 |
| 33 | But* | 0.029569 | 3 | 5 | 4 |
| 34 | Non superlative non comparative adjectives | 0.030964 | 3 | 5 | 4 |
| 35 | It* | 0.031005 | 3 | 5 | 4 |
| 36 | What* | 0.032422 | 3 | 5 | 4 |
| 37 | To | 0.020219 | 4 | 4 | 4 |
| 38 | And* | 0.020315 | 4 | 4 | 4 |
| 39 | As* | 0.022386 | 4 | 4 | 4 |
| 40 | All* | 0.024072 | 4 | 4 | 4 |
| 41 | There* | 0.017028 | 5 | 3 | 4 |
| 42 | Verb, base form | 0.046489 | 2 | 7 | 4.5 |
| 43 | Has* | 0.036117 | 3 | 6 | 4.5 |
| 44 | Models | 0.037161 | 3 | 6 | 4.5 |
| 45 | Not* | 0.039373 | 3 | 6 | 4.5 |

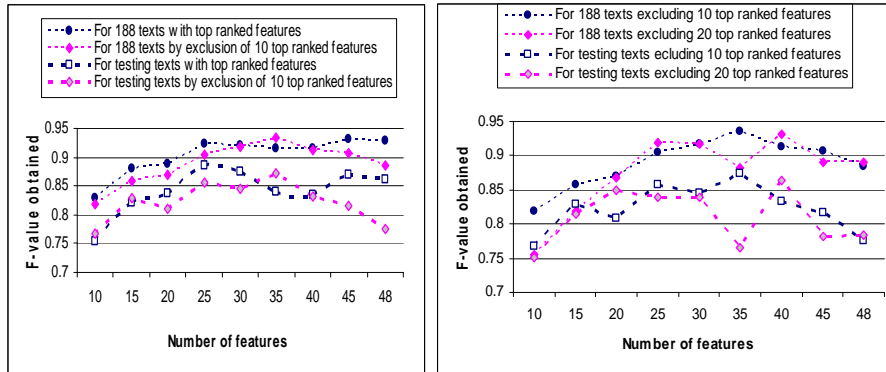
| | | | | | |
|----|---------------------------------------|----------|---|---|-----|
| 46 | He* | 0.044734 | 3 | 6 | 4.5 |
| 47 | With* | 0.026653 | 4 | 5 | 4.5 |
| 48 | Verb non 3rd ps. Singular, present | 0.027601 | 4 | 5 | 4.5 |
| 49 | Interjections | 0.027821 | 4 | 5 | 4.5 |
| 50 | Or* | 0.031601 | 4 | 5 | 4.5 |
| 51 | She* | 0.032086 | 4 | 5 | 4.5 |
| 52 | You* | 0.033083 | 4 | 5 | 4.5 |
| 53 | Her* | 0.037648 | 4 | 6 | 5 |
| 54 | Adjectives | 0.039904 | 4 | 6 | 5 |
| 55 | I* | 0.041086 | 4 | 6 | 5 |
| 56 | For* | 0.042158 | 4 | 6 | 5 |
| 57 | Is* | 0.044246 | 4 | 6 | 5 |
| 58 | Wh-determiner | 0.025662 | 5 | 5 | 5 |
| 59 | In* | 0.026881 | 5 | 5 | 5 |
| 60 | Wh-adverbs | 0.034098 | 5 | 5 | 5 |
| 61 | Verb non 3rd ps. Singular | 0.04863 | 4 | 7 | 5.5 |
| 62 | We* | 0.050809 | 4 | 7 | 5.5 |
| 63 | Be* | 0.051602 | 4 | 7 | 5.5 |
| 64 | They* | 0.052293 | 4 | 7 | 5.5 |
| 65 | Been* | 0.036836 | 5 | 6 | 5.5 |
| 66 | By* | 0.040053 | 5 | 6 | 5.5 |
| 67 | On* | 0.048059 | 5 | 7 | 6 |
| 68 | Said* | 0.050816 | 5 | 7 | 6 |
| 69 | Adjectives, comparative | 0.057112 | 5 | 7 | 6 |

In this case for each part of speech, the feature has been selected as its probability of occurrence and for frequent words (marked by * in table) value of 'frequency of word/total word tokens' has been used. This table envisages that all the three entropies, defined by us, have obtained positions in top 20 ranked features (two entropies, e_1 and e_2 takes position in top 10 ranked features) out of all the 69 features this implies that entropies of texts are good features (as getting lower ranks) for text categorization and therefore can be utilized as better measures for genre discrimination.

4. Observations and Results

In this section we will argue for the advantage of using the investigated method for the feature selection process. For this purpose we have applied discriminant analysis to all 188 (training as well as testing). We took the different features and observed that the average values of F when no ranks are excluded with first 10, 15, 20, 25, 30, 35, 40, 45, 48 features of Table5 are .828, .879, .888, .923, .921, .914, .915, .932, .928; when top 10 ranks are excluded .819, .858, .87, .905, .918, .935, .914, .907, .885 and when 20

ranks are excluded are .755, .818, .868, .918, .918, .882, .931, .889, .891 respectively. In a similar manner we have obtained the average values for the testing texts under the same conditions. The results are graphically represented in the following figures:



Figures 2 and 3. F values obtained for all articles and also for testing articles by taking different number of features.

From the above figures it is clear that the highest average value of F for the testing articles is obtained as 0.89 by using 25 top ranked features while it is 0.87 when 10 top ranked features are excluded and 35 features are used. In the case of the exclusion of 20 top ranked features, the maximum F value obtained was to be 0.86 in 40 features for the testing texts. These accentuate the inclusion of top ranked features for categorization of testing articles for getting better results.

We are now in a position to establish the superiority of the discussed methods over the prevalent methods. We have constructed the classification table for the testing articles corresponding to highest value of F, which is shown in the following table:

Table7. Classification table for testing texts with 25 top ranked features.

| Actual Category \ Predicted Category | Predicted Category | | | | | Total | Precision Recall |
|--------------------------------------|--------------------|------------------|-----------|-----------|----|-----------|---------------------|
| | Recipe | Creative-writing | Editorial | Reportage | | | |
| Recipe | 13 | 0 | 0 | 0 | 13 | 0.87/1 | |
| Creative-writing | 0 | 12 | 0 | 0 | 12 | 0.92/1 | |
| Editorial | 0 | 0 | 20 | 2 | 22 | 0.71/0.91 | |
| Reportage | 2 | 1 | 8 | 37 | 48 | 0.95/0.77 | |
| Total | 15 | 13 | 28 | 39 | 95 | | |

For this classification table the average values of precision and recall are 0.92 and 0.86 respectively, which are better than the values obtained with three types of features shown in table1. Repetition of the process with a slight change of assignment of

second rank as 1,2,3,4,5,6 for Average(S_{ij}^2) in ranges 0-.01,.01-.015,.015-.02,.02-.025,.025-.03 or higher reveals that the average value of F for the testing articles is maximum (which is 0.88) for top ranked features while the maximum of average F value with the exclusion of 10 top ranked and 20 top ranked features are 0.87 and 0.86 respectively.

The above discussion signifies the fact that the ranking technique assigns lower ranks to features, which in turn will give better results for the categorization of the testing articles. This technique enables better feature selection from training texts so that results for the testing texts can also be improved.

5. Conclusion

On the basis of the discussions and observations carried out in the present paper, it can be concluded that the concept of entropies of POS are good features for the discrimination of texts. The suitability of these entropic measures has been visualized in the discrimination of different categories.

We can also conclude that while applying the discrimination process, instead of taking some particular kinds of features separately as POS features, vocabulary measures and most frequent words-list we can take a set of features comprising all features by applying filters of completely randomized design and square deviations from the mean of feature weights which have a possibility to give better results as the results of the process depend on the features taken. This possibility has been expressed by us in the form of a ranking technique so that the features with lower ranks should be taken preferentially than the features with higher ranks. This technique shall enable better feature extractions from training texts so that results for the testing texts can also be improved.

Acknowledgement

The authors are grateful to the Council of Scientific & Industrial Research (CSIR), New Delhi, for providing financial assistance to carry out the research work on this interesting field of applied Mathematics in the form of a senior research fellowship to the first author (CSIR grant number : 09/428(0062)/2006-EMR-1).

References

- Antic G.; Stadlober E.; Grzybek P.; Kelih E. (2006). Word length and frequency distributions in different text genres. In M. Spiliopoulou et al. (Eds.): From Data and Information Analysis to Knowledge Engineering, pp. 310-317, Springer, Heidelberg.
- Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press.
- Biber, D. (1989). A typology of English texts. *Language*, **27**, pp. 3-43.
- Borja, M., C.; Chappas, C. (2006). A Stylometric Analysis of Newspapers, Periodicals and News Scripts. *Journal of Quantitative Linguistics*, **13**(2-3), pp. 285-312.
- Garcia, M.; Hidalgo, H.; Chavez, E. (2006). Contextual Entropy and Text Categorization. Fourth Latin American Web Congress (LA-WEB'06), pp. 147-153.
- Karlgren, J.; Cutting, D. (1994). Recognizing text genres with simple metric using discriminant analysis. *Proceedings of the 15th Conference on Computational Linguistics*, Vol. 2, pp. 1071-1075.

- Kelih, E.; Antic, G.; Grzybek, P.; Stadlober, E. (2005). Classification of Author and/or Genre? The Impact of Word Length. In: C. Weihs, and W. Gaul (Eds.), *Classification – The Ubiquitous Challenge*. Springer, Berlin, pp. 498–505.
- Kelih, E.; Grzybek, P.; Antic, G.; Stadlober, E. (2006). Quantitative Text Typology: The Impact of Sentence Length. In: M. Spiliopoulou et al. (Eds.): *From Data and Information Analysis to Knowledge Engineering*. Springer, Berlin, pp. 382–389.
- Kessler, B.; Nunberg, G.; Schuetze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting on Association for Computational Linguistics*, pp. 32–38.
- Klavans, J.; Kan, M., Y. (1998). Role of verbs in document analysis. *In proc. COLING/ACL Montreal*.
- Lee, Y.; Myaeng, S. (2004). Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization, *Proc. 37 HICSS*, 2004.
- Stadlober, E.; Djuzelic, M. (2006). Multivariate Statistic Methods of Quantitative Text Analysis. In: Grzybek, Peter (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht, NL: Springer; pp. 259-276.
- Stamatatos, E.; Fakotakis, N.; Kokkinakis, G. (2000a). Text genre detection using common word frequencies. In *18th International Conference on computational Linguistics*, 2000.
- Stamatatos, E.; Fakotakis, N.; Kokkinakis, G. (2000b). Automatic text categorization in terms of genre and author. *Computational Linguistics*, **26**(4), pp. 471–495.
- Wolters, M.; Kirsten, M. (1999). Exploring the use of linguistic features in domain and genre classification. In *Proceedings of EACL'99*, Morristown, NJ, USA. pp. 142–149.