Analytical Evaluation of Term Weighting Schemes for Text Categorization

Hakan Altınçay* and Zafer Erenel

Department of Computer Engineering, Eastern Mediterranean University

Famagusta, Northern Cyprus

E-mail: {hakan.altincay, zafer.erenel}@emu.edu.tr

Abstract

An analytical evaluation of six widely used term weighting techniques for text categorization is presented. The analysis depends on expressing the term weights using term occurrence probabilities in positive and negative categories. The weighting behaviors of the schemes considered are firstly clarified by analyzing the relation between the occurrence probabilities of terms which receive equal weights. Then, the weights are expressed in terms of ratio and difference of term occurrence probabilities where the similarities and differences among different schemes are revealed. Simulations show that the relative performance of different schemes can be explained by the ways they use ratio and difference of term occurrence probabilities in generating the term weights.

Keywords: Contour lines, term occurrence probability, term weighting, relative weights, text categorization

*Corresponding author: E-mail: hakan.altincay@emu.edu.tr, Tel: +90 392 6302842, Fax: +90 392 3650711
1 Introduction

Growing computational power, cheaper storage mediums and increased demand for sharing information in the last decade led to an exponential increase in the volume of information on the web. Since textual documents generally constitute most of the information on the web, researchers have been endeavoring to define enhanced vector representations to increase the success of text categorization systems. The Bag of Words (BOW) approach is the most common way to represent documents where each word (or, term) is a unique feature (Sebastiani, 2002). Taking into account the fact that all terms may not be useful for the classification of documents, feature selection schemes such as mutual information and chi-square are generally used to select a subset of features (Chen et al., 2009; Yang and Pedersen, 1997; Zheng et al., 2004; Liu and Yu, 2005; Liu et al., 2009). Simulations have shown that most text categorization schemes achieve their best scores when a subset of the available terms are used (Rogati and Yang, 2002).

Term weighting which aims to quantify the relative importance of different terms in the selected set, is the next step applied after feature selection where discriminative terms are assigned larger weights (Debole and Sebastiani, 2003; Liu et al., 2009). Term weighting has been traditionally formulated as the product of the term frequency and inverse document frequency, \( tf \times idf \) (Debole and Sebastiani, 2004). Alternative forms of weights are generally formulated by replacing the idf term and \( tf \) is generally normalized to eliminate the document length effect. The use of available feature selection metrics such as chi-square \( (tf \times \chi^2) \) for this purpose is extensively studied (Debole and Sebastiani, 2003). Recently, various alternative schemes are proposed for term weighting (Lan et al., 2009; Liu et al., 2009; Tsai et al., 2008). For instance, relevance frequency based weighting \( (tf \times rf) \) is one of them which delivers the best results on several benchmark datasets (Lan et al., 2009).

The available feature selection metrics and newly proposed schemes considered for term weighting are based on different methodologies having inherent assumptions. In order to develop
better term weighting schemes, a deeper analysis for their weighting behaviors is essential. If a common analysis framework can be developed, analytical comparison will be possible by using common parameters which may help to clarify their strengths and weaknesses and hence provide increased understanding for their inconsistent performances on different categories. In fact, the decision boundaries induced by several feature selection schemes computed by Forman (2003, 2007) provide some intuition about their relative behaviors. The analysis presented in this paper also includes the derivation of exact analytical expressions of the decision boundaries presented in (Forman, 2003). In the following context, “the weights generated by a scheme” will correspond to the right multiplicand of the actual term weights that is specific to the scheme under concern in analyzing different approaches.

This paper firstly emphasizes that the term weighting schemes can be expressed as a function of two term specific parameters namely, term occurrence probability in the category under concern (positive category) and its complement (negative category) (Erenel et al., 2009). Then, the analytic expressions of the patterns formed by the terms receiving equal weights in the two dimensional space of term occurrence probabilities are derived. These equations are equated to a constant and the relation between term occurrence probabilities is computed. These relations are also presented in the form of contour lines which are curves along which the term weighting schemes (functions of term occurrence probabilities) provide constant values to illustrate the distribution of equally weighted terms. It is shown that the contour lines of relevance frequency and mutual information are linear lines that pass through the origin. In other words, the weights computed by these two schemes for terms on such lines are constant. These lines are considered as a reference for the investigation of the characteristics of the weights of the other schemes. The analysis is done by deriving the analytic expressions of their weights in terms of the ratio and difference of term occurrence probabilities. Since the terms on a linear line through the origin have the same term occurrence probabilities ratio, the position of each term on such a line is represented using only the difference of term occurrence probabilities. This helps to simplify the functional form of weights and express them using only the term occurrence
probability difference parameter. Consequently, categorization of different schemes to reveal
their similarities and differences becomes possible. The analytical findings are verified through
simulations which show that the relative performance of different schemes can be explained by
the ways they use the ratio and the difference of term occurrence probabilities in generating the
weights. The observations from the analysis and the following simulations are used to modify \( rf \)
and \( MI \) for further investigation of the corresponding practical implications. It is shown that,
when appropriate expressions of the term occurrence probability difference are considered, the
performances of these schemes can be further improved.

A brief review about major components of text categorization including the classification
schemes and feature extraction methods is presented in Section 2. The notation used in the
current study is presented in Section 3. Formulations of six available weighting schemes in terms
of term occurrence probabilities are presented in Section 4. A comparison of these schemes in
terms of their weighting characteristics on linear lines passing through the origin is presented
in Section 5. In order to verify the importance of the observations from performance point of
view, the results of the experiments conducted on Reuters-21578 ModApte Top10 (Debole and
Sebastiani, 2004) and WebKB (Craven et al., 1998) datasets are presented in Section 6. The
last part, Section 7 presents the conclusions drawn from this study.

2 Related Work

Several classification techniques have been explored for text categorization, including Roc-
chio (Joachims, 1997), naive Bayes (Chen et al., 2009), \( k \)-nearest neighbor (Tan, 2005), multi-
layer neural networks (Selamat and Omatu, 2004) and support vector machines (SVM) (Joachims,
1998). Rocchio and \( k \)-nearest neighbor are similarity-based classifiers. In \( k \)-nearest neighbor ap-
proach, estimation of the best-fitting value of \( k \) during training to maximize the generalization
ability is also studied (Guo et al., 2004). Rocchio achieves learning by combining document vec-
tors into a prototype vector for each category. The computation of the cosine similarity scores
between the prototype vectors and the unseen document vector constitutes the categorization phase. Naive Bayes is another popular scheme which assumes that the features (words) occur independently of each other in the documents. Two common models considered are multi-variate Bernoulli and multinomial event. Multi-variate Bernoulli model represents documents with binary vectors whereas the latter utilizes the frequency of each word in computing the probability of each category (Chen et al., 2009). Multilayer neural networks implement linear discriminants in a space where the inputs have been mapped nonlinearly. The complexity of the network (the number of hidden layers and hidden nodes and parameter values such as learning rate, momentum rate, number of iterations and mean squares error) needs to be adjusted to maximize the success of categorization (Duda et al., 2001). The number of features used is generally kept smaller compared to the other schemes due to performance reasons (Selamat and Omatu, 2004). On the other hand, the robustness of SVM in very high dimensional feature space sets it as the strongest classification scheme in text categorization domain (Leopold and Kindermann, 2002).

Feature selection followed by term weighting are the two primary areas of term scoring. The main aim in feature selection is to reduce the dimensionality of the feature space by ranking the words in the text documents and selecting those with the highest scores. In local selection approach, a set of features is selected using relevant and irrelevant documents of each category (Zheng and Srihari, 2003; Liu et al., 2009). Hence, a different feature set may be used for each category. On the other hand, global feature selection exploits the relevant documents of all categories to form a common feature set (Özgür et al., 2005). Term weighting which aims to take into account the relative importance of different terms during categorization is generally implemented as the product of term frequency of each term and a positive weight that is proportional with its ability to discriminate documents in different categories. Some of the well known feature selection metrics that are also exploited for feature weighting are mutual information (MI), Chi-square ($\chi^2$) and odds ratio ($OR$) (Liu et al., 2009; Mladenic and Grobelnik, 2003; Sebastiani, 2002; Yang and Pedersen, 1997). As it was also stated in Section 1,
rf delivers the best results among all on several benchmark datasets (Lan et al., 2009). When used together with the term frequency, recent studies have shown that the probability-based weighting scheme \((\text{prob})\) which exploits two relevance indicators (Liu et al., 2009) and balanced relative frequency \((\text{brf})\) approach (Tsai et al., 2008) provide competitive or better performances when compared to more popular methods.

### 3 Notation

The information elements used in text categorization are presented in Table 1. Let \(c_i\) denote the \(i\)th positive category and \(\overline{c_i}\) denote the complement (negative) category which includes the documents from all categories other than \(c_i\). \(A\) represents the number of documents that belong to \(c_i\) and contain term \(t_k\), \(B\) is the number of documents that belong to \(c_i\) but do not contain term \(t_k\), \(C\) denotes the number of documents that belong to \(\overline{c_i}\) and contain term \(t_k\) and \(D\) is the number of documents that belong to \(\overline{c_i}\) but do not contain term \(t_k\) (Lan et al., 2009).

Using the contingency table, the a priori probabilities of the categories can be computed as \(P(c_i) = \frac{N^+}{N}\) and \(P(\overline{c_i}) = \frac{N^-}{N}\) where \(N^+ = (A + B)\) denotes the the number of documents in the positive category, \(N^- = (C + D)\) denotes the the number of documents in the negative category and \(N\) is the total number of training documents. The probability of the documents which belong to \(c_i\) and contain at least one occurrence of \(t_k\) equals to \(P(t_k, c_i) = A/N\) which can also be expressed as \(P(t_k, c_i) = P(t_k|c_i) \times P(c_i)\), where \(P(t_k|c_i)\) represents the probability of documents from \(c_i\) in which \(t_k\) occurs at least once. Similarly, \(P(t_k, \overline{c_i}) = P(t_k|\overline{c_i}) \times P(\overline{c_i}) = C/N\) denotes the probability of documents which belong to the negative category and contain at least one occurrence of \(t_k\). \(P(t_k|\overline{c_i})\) represents the probability of documents from \(\overline{c_i}\) in which \(t_k\) occurs.
at least once. Let $p^+$ and $p^-$ denote $P(t_k|c_i)$ and $P(t_k|\bar{c}_i)$ respectively which can be computed as
\[ p^+ = \frac{A}{A + B}, \]
\[ p^- = \frac{C}{C + D}. \] 
For a term weighting problem, $P(c_i)$ and $P(\bar{c}_i)$ are specific to the category and independent of the terms. For the simplification of formulations, $P(c_i)$ is denoted by $p$ and hence $P(\bar{c}_i) = (1-p)$ in the remaining part of this paper.

4 Weight patterns as contour lines

In this section, the weighting patterns of six widely used term weighting schemes mentioned above are studied. The expressions for term weights in terms of $p^+$ and $p^-$ are firstly derived. These expressions are then equated to a constant and the relationship between these parameters is found for a constant weight. The relations are also plotted as contour lines on the $(p^+, p^-)$ plane. The terms of “Earn” category in Reuters-21578 dataset are ranked using $\chi^2$ method for this purpose and the weights computed from the top 5000 are considered.

4.1 Relevance frequency ($rf$)

$rf$ is originally defined as (Lan et al., 2009)
\[ rf = \log\left(2 + \frac{A}{C}\right) \] 
Since $p^+ = \frac{A}{Nxp}$ and $p^- = \frac{C}{Nx(1-p)}$, $rf$ can be expressed as
\[ rf = \log\left(2 + \frac{p^+}{p^-} \frac{p}{1-p}\right) \] 
where logarithm to the base 2 is considered. In the other schemes described below, the same base value is used. In order to compute contour lines, let $rf = \tau$ where $\tau$ is an arbitrary constant. Then,
\[ 2 + \frac{p^+}{p^-} \frac{p}{1-p} = 2^\tau = \xi \]
The equation given above describes the relation between \( p^+ \) and \( p^- \) that corresponds to the \( rf \) weight value denoted by \( \tau \). After some manipulations, we obtain

\[
p^- = \frac{p}{(1-p)(\xi - 2)} p^+ \tag{5}
\]

Consequently, the terms on the line given in Eq. (5) are assigned the same weight, \( \tau \) by \( rf \).

In other words, \( p^+ \) and \( p^- \) are linearly proportional for a given weight. This means that the weight of a term having \( p^+ = x \) and \( p^- = y \) is the same as the weight of a term having \( p^+ = kx \) and \( p^- = ky \), where \( k \) is an arbitrary constant. This is also obvious from Eq. (2). Replacing \( A \) and \( C \) by \( kA \) and \( kC \) respectively does not change the weight value. Using “Earn” category where \( p = 2877/6491 = 0.4432 \), the average \( rf \) value of the top 500 terms is set as the \( \tau \) value of the first contour. The \( \tau \) value for the second contour is computed using the terms ranked after the first group having ranks in between 501 and 1000. This procedure is repeated to compute the values of \( \tau \) for plotting ten different contours. The reason for computing averages is to make sure that the contours correspond to weights that approximately characterize the whole set. Alternatively, the weights of ten randomly selected terms can also be used. Ten contour lines obtained for \( rf \) are presented in Figure 1 together with the corresponding values of \( \tau \). For example, when \( p^+ = 1.00 \) (i.e. the term occurs in all positive documents) and \( p^- = 0.63 \) (i.e. the term occurs in 63 percent of negative documents) or, when \( p^+ = 0.50 \) and \( p^- = 0.31 \), the value of weight is 1.71.

4.2 Mutual information (\( MI \))

\( MI \) is defined as (Sebastiani, 2002)

\[
MI = \log \frac{P(t_k, c_i)}{P(t_k)P(c_i)} \tag{6}
\]

Since \( P(t_k) = P(t_k|c_i)P(c_i) + P(t_k|\neg c_i)P(\neg c_i) \), the corresponding expression becomes

\[
MI = \log \left( \frac{p^+}{p^+ + (1-p)p^-} \right) \tag{7}
\]
As before, let $MI = \tau$. Then

$$\frac{p^+}{p^+ (1 - p)p^-} = 2^\tau = \xi$$  \hspace{1cm} (8)

After some manipulations, we get

$$p^- = \frac{1 - \xi p}{\xi (1 - p)p^+}$$  \hspace{1cm} (9)

As seen in Eq (9), for the terms receiving the same weight, there is a linear relation between $p^+$ and $p^-$ as in $rf$. The contour lines obtained for $MI$ are presented in Figure 2. Due to very small dynamic range of the weights, we considered only the top 500 terms for this plot where $\tau$ values are computed as the average weights of 50 terms.

4.3 Odds ratio ($OR$)

$OR$ is defined as (Sebastiani, 2002)

$$OR = \frac{P(t_k|c_i)(1 - P(t_k|\bar{c}_i))}{(1 - P(t_k|c_i))P(t_k|\bar{c}_i)}$$  \hspace{1cm} (10)

$$= \frac{p^+ (1 - p^-)}{p^- (1 - p^+)}$$

Let $OR = \xi$. Then, we obtain

$$p^- = \frac{p^+}{\xi + (1 - \xi)p^+}$$  \hspace{1cm} (11)
As seen in Eq. (11), for the terms receiving the same weight, relation between $p^+$ and $p^-$ is not linear in the case of OR. This means that the weight of a term having $p^+ = x$ and $p^- = y$ is not the same as the term having $p^+ = kx$ and $p^- = ky$ in this case. The contour lines obtained for OR are presented in Figure 3. In general, $\log(OR)$ or $\log(2 + OR)$ are used in practical implementations due to performance reasons (Lan et al., 2009). However, this does not effect the generality of the current analysis. The type of the contour lines would be the same. Only the position of contours would change due to applying $\log(.)$ on the currently used weights. In particular,

$$\log(2 + OR) = \log\left(2 + \frac{p^+ (1 - p^-)}{p^- (1 - p^+)}\right)$$

Letting $\log(2 + OR) = \tau$ and performing some manipulations, the relation becomes

$$\frac{p^+ (1 - p^-)}{p^- (1 - p^+)} = 2^\tau - 2 = \hat{\xi}$$

from which an expression similar to the one given in Eq. (11) can be obtained.
4.4 Balanced relative frequency (brf)

brf is originally defined as (Tsai et al., 2008)

\[ brf = \log \left( \frac{AB}{CD} \right) \]  

(14)

Since,

\[ \begin{align*}
A &= \frac{p^+}{p - 1 - p} \\
C &= \frac{p^+}{p - 1 - p} \\
B &= \frac{(1 - p^+)}{(1 - p^-) (1 - p)} \\
D &= \frac{p}{(1 - p^-) (1 - p)} 
\end{align*} \]  

(15)

brf can be expressed as

\[ brf = \log \left( \frac{p^+ (1 - p^+)}{p^- (1 - p^-) (1 - p)} \right). \]  

(16)

After setting \( brf = \tau \) and performing some manipulations presented in Sec. A of the Appendix, the following equation is obtained.

\[ \frac{p^+ (1 - p^+)}{p^- (1 - p^-)} = \xi. \]  

(17)

Rearranging the terms, the relation becomes

\[ (p^+)^2 - \xi(p^-)^2 - p^+ + \xi p^- = 0. \]  

(18)
This means that there is a quadratic relation between $p^+$ and $p^-$ for the terms receiving the same weight.

Since $brf$ may provide negative weights, the expression should be slightly modified as in OR.

In this study, we modified $brf$ as $\hat{brf} = \log(2 + \frac{AB}{CD})$ which does not make any change on contour types. The corresponding expression becomes

$$\hat{brf} = \log \left( 2 + \frac{p^+}{p^-} \frac{p^+ (1 - p^-)}{p^- (1 - p^+)} \right).$$

After setting $\hat{brf} = \tau$ and performing some manipulations, the same equation as the one given in Eq. (17) can be obtained as

$$\frac{p^+ (1 - p^-)}{p^- (1 - p^+)} = \hat{\xi}.$$  

The contour lines obtained for $\hat{brf}$ are presented in Figure 4.

### 4.5 Probability based approach ($prob$)

$prob$ is originally defined as (Liu et al., 2009)

$$prob = \log \left( 1 + \frac{A^2}{B^2} \right)$$  

\[21\]
It can be expressed in terms of \(p^+\) and \(p^-\) as

\[
prob = \log \left( 1 + \frac{p^+}{(1-p^+)} \frac{p^+ p}{p^- (1-p^-)} \right)
\]  \hspace{1cm} (22)

Letting \(prob = \tau\), we obtain

\[
\frac{p^+}{(1-p^+)} \frac{p^+ p}{p^- (1-p^-)} = 2\tau - 1 = \hat{\xi}
\]  \hspace{1cm} (23)

which can be written as

\[
p(p^+)^2 + \hat{\xi}(1-p)p^+p^- - \xi(1-p)p^- = 0
\]  \hspace{1cm} (24)

As seen in Eq. (24), the relation between \(p^+\) and \(p^-\) is quadratic. The contour lines obtained for \(prob\) are presented in Figure 5.

**4.6 Chi-square (\(\chi^2\))**

\(\chi^2\) is defined as (Sebastiani, 2002)

\[
\chi^2 = \frac{N[P(t_k,c_i)P(\bar{t}_k,\bar{c}_i) - P(t_k,\bar{c}_i)P(\bar{t}_k,c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}
\]  \hspace{1cm} (25)
Figure 6: Contour lines for $\chi^2$ on the $(p^+, p^-)$ plane.

Note that $P(\bar{t}_k|c_i) = 1 - P(t_k|c_i)$, $P(\bar{t}_k|\bar{c}_i) = 1 - P(t_k|\bar{c}_i)$ and $P(\bar{t}_k) = 1 - P(t_k)$. After some manipulations, we obtain

$$\chi^2 = \frac{Np^2(1-p)^2[p^+(1-p^-) - p^-(1-p^+)]^2}{[p.p^+ + p^-(1-p)][(1-p^+)+p+(1-p^-)(1-p)]p(1-p)}$$

(26)

Since $p$ and $N$ are fixed for all terms, setting $\chi^2 = \tau$, we get

$$\frac{(p^+ - p^-)^2}{p.p^+ + p^-(1-p) - 2p(1-p)p^+.p^- - p^2(p^+)^2 - (1-p)^2(p^-)^2} = \frac{\tau}{Np(1-p)} = \xi$$

(27)

where $\xi$ is another constant in terms of $\tau$. Rearranging the terms, the following relation can be obtained.

$$(1 + \xi p^2)(p^+)^2 + (1 + \xi(1-p)^2)(p^-)^2 + [2\xi p(1-p) - 2]p^+.p^- - \xi p.p^+ - \xi(1-p)p^- = 0$$

(28)

As seen in Eq. (28), there is a quadratic relation between $p^+$ and $p^-$ as in $brf$. The contour lines obtained for $\chi^2$ are presented in Figure 6.

5 Weight patterns on $p^- = \beta p^+$

The weighting schemes can be categorized into two groups as providing either linear or nonlinear contour lines on the $(p^+, p^-)$ plane. $rf$ and $MI$ provide linear contour lines whereas $OR$, $prob$, $\chi^2$
Figure 7: The positions of the top 5000 terms of “Earn” category.

and bref generate nonlinear contour lines. The expressions for nonlinear contours are complex and hence comparison among different weighting schemes is not straightforward. However, visual inspection of the contour lines provide some hints about meaningful similarities. For instance, for small values of \( p^+ \) and \( p^- \), it can be seen that the contour lines generated by OR and bref are very close to linear lines through the origin. As a matter of fact, a further analysis is needed to achieve a deeper understanding of their relative weighting behaviors. The analysis carried out for this purpose is based on computing the weights generated on linear lines passing through the origin. Consider the line \( p^- = \beta p^+ \). Each \( \Delta = |p^+ - p^-| \) value specifies a unique term located on this line. In other words, two different terms on \( p^- = \beta p^+ \) can not have the same \( \Delta \) value. Without any loss of generality, assume that \( \beta < 1 \). Then, \( \Delta = (p^+ - p^-) \). Using \( p^- = \beta p^+ \), we can obtain

\[
\begin{align*}
p^+ &= \frac{1}{1-\beta} \Delta \\
p^- &= \frac{\beta}{1-\beta} \Delta
\end{align*}
\]  

Replacing \( p^+ \) and \( p^- \) terms in the weight formulas, we can compute the weights of terms as a function of \( \Delta \). This will simplify the functional form of weights and express them using only the difference parameter. Consequently, categorization of different schemes to reveal their
similarities and differences will be possible. The analysis is restricted to small values of $\Delta$. The reason of this is twofold. Firstly, the majority of terms are very close to the origin (Forman, 2003). This can also be seen in Figure 7 where the positions of the top 5000 terms of “Earn” category are presented. As it can also be seen in Eq. (29), both $p^+$ and $p^-$ are directly proportional to $\Delta$ which means that terms located close to the origin on $(p^+, p^-)$ plane have small $\Delta$ values. Consequently, the analysis will cover vast majority of terms. Secondly, as it will be seen in the following sections, the weight expressions will have simpler forms when $\Delta$ is kept small and hence comparison between different schemes will be possible. This analysis will also reveal the exact form of almost linear contour lines provided by OR and $brf$ around the origin.

For $rf$, the expression for weights can be obtained using Eq. (3) as

$$rf = \log \left( 2 + \frac{\Delta}{1-\beta} \left( \frac{p}{1-p} \right) \right) = \log \left( 2 + \frac{1}{\beta} \left( \frac{p}{1-p} \right) \right)$$

which is constant in terms of $\Delta$. Similarly, the expression can be computed for $MI$ using Eq. (7) as

$$MI = \log \left( \frac{\Delta}{p^{1-\beta} + (1-p)(1-\beta)} \right) = \log \left( \frac{1}{p + \beta(1-p)} \right)$$

Figure 8: The weights of terms lying on $p^- = \beta p^+$ as a function of $\Delta = (p^+ - p^-)$ computed by $rf$. 
which is also constant in terms of $\Delta$. In fact, since the weights are equal for all terms on $p^- = \beta p^+$, the weight is constant on this line. This can be seen in Figures 8 and 9 respectively for $rf$ and $MI$ where the weights are presented for four values of $\beta$.

The weight expressions for $rf$ and $MI$ can be easily computed for $\beta > 1$ as well. In this case, $\Delta = |p^+ - p^-| = (p^+ - p^-)$ and hence

\[
p^+ = \frac{1}{\beta - 1} \Delta \]
\[
p^- = \frac{\beta}{\beta - 1} \Delta
\] (32)

Using these modified forms of $p^+$ and $p^-$, it can be easily shown that the weights for $rf$ and $MI$ are also constant for $\beta > 1$. However, for the schemes having nonlinear contour lines, the weights are not expected to be constant in terms of $\Delta$. The following sections present the expressions for the remaining term weighting schemes.

5.1 **OR**

Using Eq. (29), the weight expression for $OR$ that is given in Eq. (10) can be rewritten as

\[
OR = \frac{1}{\beta} \times \frac{(1 - \beta) - \beta \Delta}{(1 - \beta) - \Delta}
\] (33)
Figure 10: The weights of terms lying on $p^- = \beta p^+$ as a function of $\Delta = |p^+ - p^-|$ computed by $OR$: (a) $\beta < 1$, (b) $\beta > 1$.

Let us denote $OR$ which is a function of $\Delta$ using $f(\Delta)$. The Taylor series expansion of $f(\Delta)$ at point $a$ can be computed as

$$ f(\Delta) = f(a) + f'(a)(\Delta - a) + O(\Delta^2) $$

As mentioned above, we are interested in small values of $\Delta$ near $a = 0$. Because of this, $O(\Delta^2)$ can be neglected. Hence the expression in Eq. (33) can be approximated at $a = 0$ as (see Sec. B of the Appendix)

$$ OR \approx \frac{1}{\beta}(1 + \Delta). $$

The slope of the corresponding function is $1/\beta$ which is inversely proportional to $\beta$. Figure 10 (a) presents the weights computed using the original expression given in Eq. (33) that are assigned to the terms lying on the line $p^- = \beta p^+$ for $\Delta \in [0.005, 0.1]$ and $\beta \in \{1/5, 1/4, 1/3, 1/2\}$.

The assumption that the term $O(\Delta^2)$ can be neglected is verified by the almost linear lines obtained in the figure where it can be concluded that the weights assigned by $OR$ can be assumed as linearly proportional to $\Delta$ for small values.

In the case of $\beta > 1$, using Eq. (32) the weight expression for $OR$ scheme can be computed
Figure 11: For $\Delta = 0.05$, the positions of terms having angles in $\{10, 20, \ldots, 80\}$ in degrees.

as

$$OR = \frac{1}{\beta} \times \frac{(\beta - 1) - \beta \Delta}{(\beta - 1) - \Delta}.$$  \hspace{1cm} (36)

Using Taylor series expansion, the expression for $OR$ is approximately equal to

$$OR \approx \frac{1}{\beta}(1 - \Delta).$$  \hspace{1cm} (37)

The slope is $-1/\beta$ for $\beta > 1$. This can also be seen in Figure 10 (b) which corresponds to Eq. (36) for $\Delta \in [0.005, 0.1]$ and $\beta \in \{2, 3, 4, 5\}$. Using the expressions obtained for $OR$, it is observed that

- For $\beta < 1$, the weights increase as $\Delta$ increases
- For $\beta > 1$, the weights decrease as $\Delta$ increases

Let us summarize this information as $\{\uparrow, \downarrow\}$ to be used later for comparison. The first arrow represents the increasing behavior of weights for $\beta < 1$ and similarly the second arrow represents the decreasing behavior for $\beta > 1$.

In order to investigate the effect of $\beta$ on the weighting characteristics of $OR$, Eqs. (33) and (36) should be evaluated for a fixed $\Delta$ value. This is achieved for $\Delta = 0.05$ by considering eight different values, $\beta \in \{0.18, 0.36, 0.58, 0.84, 1.19, 1.73, 2.75, 5.67\}$ which correspond to the
Figure 12: The weights generated using $OR$ for $\Delta = 0.05$ and various $\beta$ values. The corresponding angle is also presented in degrees.

angles $\theta \in \{10, 20, 30, \ldots, 80\}$ in degrees. The positions of the corresponding terms on the $(p^+, p^-)$ plane are presented in Figure 11. The weights generated accordingly are illustrated as a function of $\beta$ and the corresponding angle in Figure 12. It can be seen in the figure that the weights corresponding to a fixed $\Delta$ value decrease as $\beta$ increases which can also be seen in Eqs. (35) and (37). This means that this scheme favors positive terms.

The weight expressions can be similarly computed for $\log(2 + OR)$. Using Taylor series, the corresponding expression can be approximately derived as (see Sec. C of the Appendix)

$$\log(2 + OR) = \log\left(2 + \frac{1}{\beta} \times \frac{(1 - \beta) - \beta \Delta}{(1 - \beta) - \Delta}\right) \approx \log(2 + \frac{1}{\beta}) + \frac{\log(e)}{2\beta + 1} \Delta \quad (38)$$

The expression obtained is similar to that of $OR$ in the sense that both are increasing functions of $\Delta$ with nonzero intercept values on $p^-$ axis. For $\beta > 1$, the weight expression can be computed as (see Sec. D of the Appendix)

$$\log(2 + OR) = \log\left(2 + \frac{1}{\beta} \times \frac{(\beta - 1) - \beta \Delta}{(\beta - 1) - \Delta}\right) \approx \log(2 + \frac{1}{\beta}) - \frac{\log(e)}{2\beta + 1} \Delta \quad (39)$$

As in the case of $OR$, this is a decreasing function of $\Delta$ with nonzero intercept value on $p^-$ axis.
5.2 \textit{brf}

The weights provided by \textit{brf} can be computed using Eqs. (16) and (29) as

$$\text{brf} = \log\left(\frac{p^2}{\beta(1-p)^2} \times \frac{(1-\beta) - \Delta}{(1-\beta) - \beta\Delta}\right)$$  (40)

Using Taylor series expansion, the expression for \textit{brf} at \(\alpha = 0\) is computed as

$$\text{brf} \approx \log\left(\frac{p^2}{\beta(1-p)^2}\right) - \log(e)\Delta$$  (41)

If \(\beta > 1\) is considered, using Eq. (32) we can rewrite Eq. (16) as

$$\text{brf} = \log\left(\frac{p^2}{\beta(1-p)^2} \times \frac{(\beta - 1) - \Delta}{(\beta - 1) - \beta\Delta}\right)$$  (42)

and hence \textit{brf} can be approximately computed as

$$\text{brf} \approx \log\left(\frac{p^2}{\beta(1-p)^2}\right) + \log(e)\Delta$$  (43)

A similar analysis can be done for \(\hat{\text{brf}}\) which can be written as

$$\hat{\text{brf}} = \log\left(2 + \frac{p^2}{\beta(1-p)^2} \times \frac{(1-\beta) - \Delta}{(1-\beta) - \beta\Delta}\right)$$  (44)

Using Taylor series expansion, the expression for \(\hat{\text{brf}}\) can be approximately derived for \(\beta < 1\) as

$$\hat{\text{brf}} \approx \log\left(2 + \frac{p^2}{\beta(1-p)^2}\right) - \frac{p^2 \log(e)}{p^2 + 2\beta(1-p)^2}\Delta$$  (45)

If \(\beta > 1\) is considered, Eq. (16) becomes

$$\hat{\text{brf}} = \log\left(2 + \frac{p^2}{\beta(1-p)^2} \times \frac{(\beta - 1) - \Delta}{(\beta - 1) - \beta\Delta}\right)$$  (46)

Then, the weight function can be obtained as

$$\hat{\text{brf}} \approx \log\left(2 + \frac{p^2}{\beta(1-p)^2}\right) + \frac{p^2 \log(e)}{p^2 + 2\beta(1-p)^2}\Delta$$  (47)

As seen in the equations above, the weights are of similar forms for both \textit{brf} and \(\hat{\text{brf}}\). More specifically, the weights are linearly proportional to \(\Delta\) where the slope is negative for \(\beta < 1\) and positive for \(\beta > 1\). Due to the negative weights generated by \textit{brf}, \(\hat{\text{brf}}\) is considered in the remaining part of this study.
Figure 13: The weights of terms lying on $p^{-} = \beta p^{+}$ as a function of $\Delta = |p^{+} - p^{-}|$ computed by $\hat{b}r$ : (a) $\beta < 1$, (b) $\beta > 1$.

Figure 13 (a) presents the weights assigned using Eq. (44) for $\beta < 1$ and (b) Eq. (46) for $\beta > 1$. It can be seen that the weights are close to linear and,

- For $\beta < 1$, the weights decrease as $\Delta$ increases
- For $\beta > 1$, the weights increase as $\Delta$ increases

This information is represented as $\{↓, ↑\}$. For $\Delta = 0.05$, the weights generated are illustrated as a function of $\beta$ and the corresponding angle in Figure 14. It can be seen in the figure that the weights decrease as $\beta$ increases. This means that this scheme favors positive terms as OR.

5.3 prob

The weights computed using the prob approach can be expressed for $\beta < 1$ as a function of $\Delta$ using Eqs. (22) and (29) as follows:

\[
prob = \log \left( 1 + \frac{p}{\beta(1-p)} \times \frac{\Delta}{[(1-\beta) - \Delta]} \right)
\]  (48)
Figure 14: The weights generated using $\hat{brf}$ for $\Delta = 0.05$ and various $\beta$ values. The corresponding angle is also presented in degrees.

Using Taylor series representation, Eq. (48) can be approximated as (see Sec. E of the Appendix)

$$prob \approx \frac{p \times \log(e) \Delta}{\beta(1-p)(1-\beta)}$$  \hspace{1cm} (49)$$

which shows that the weights are linearly proportional to $\Delta$.

For $\beta > 1$, the expression for $prob$ can be derived using Eqs. (22) and (32) as

$$prob = \log \left(1 + \frac{p}{\beta(1-p)} \times \frac{\Delta}{[(\beta-1)-\Delta]}\right)$$  \hspace{1cm} (50)$$

and hence,

$$prob \approx \frac{p \times \log(e) \Delta}{\beta(1-p)(\beta-1)}$$  \hspace{1cm} (51)$$

for small $\Delta$.

The slopes of the weight functions are $\frac{p \times \log(e)}{(1-p)} \times m$ where,

$$m = \begin{cases} 
\frac{1}{\beta} \times \frac{1}{(1-\beta)} & \beta < 1, \\
\frac{1}{\beta} \times \frac{1}{(\beta-1)} & \beta > 1 
\end{cases}$$  \hspace{1cm} (52)$$

Figure 15 (a) presents the weights computed for $\beta < 1$ and (b) illustrates the weights computed for $\beta > 1$ using Eqs. (48) and (50) respectively. The validity of the linear proportionality assumption can also be seen in the figures. Using the expressions obtained for $prob$, it can
Figure 15: The weights of terms lying on $p^+ = \beta p^+$ as a function of $\Delta = |p^+ - p^-|$ computed by prob: (a) $\beta < 1$, (b) $\beta > 1$.

It should be noted that the expressions computed using first order approximation are naturally linear for OR, brf and prob. However, absence of an intercept in prob differentiates it from OR and brf. Since the term frequencies scaled by the weights are considered in a single composite feature vector during categorization, the relative values of these weights are more important than their absolute values. More specifically, consider two different terms having the same term frequency value. The relative influence of these terms in categorization is specified by the relative values of the weights computed by the scheme under concern but not the absolute weight values. It can be seen in Eq. (51) that doubling $\Delta$ leads to a doubled weight for small values of $\Delta$ whereas the increase is much less in the case of OR and brf. This point will be further investigated in Section 5.5.

For $\Delta = 0.05$, the weights generated are illustrated as a function of $\beta$ and the corresponding angle in Figure 16. It can be seen in the figure that, when $\Delta = 0.05$,
Figure 16: The weights generated using prob for $\Delta = 0.05$ and various $\beta$ values. The corresponding angle is also presented in degrees.

- For $\beta < \frac{1}{2}$, the weights decrease as $\beta$ increases
- For $\frac{1}{2} < \beta < 1$, the weights increase as $\beta$ increases
- For $\beta > 1$, the weights decrease as $\beta$ increases

5.4 $\chi^2$

The weights provided by $\chi^2$ can be written in terms of $\Delta$ for $\beta < 1$ using Eq. (26) as

$$
\chi^2 = \frac{Np(1-p)\Delta}{p+\beta(1-p)}
$$

Using Taylor series expansion, Eq. (53) is approximately equal to

$$
\chi^2 \approx \frac{Np(1-p)(1-\beta)}{p+\beta(1-p)} \Delta
$$

For $\beta > 1$ skipping the intermediate steps, Eq. (26) can be written as

$$
\chi^2 \approx \frac{Np(1-p)(\beta-1)}{p+\beta(1-p)} \Delta
$$

Figure 17 (a) presents the weights for $\beta < 1$ which are computed using Eq. (53). Part (b) presents the weights assigned for $\beta > 1$ for which the term $(1-\beta)$ is replaced by $(\beta-1)$ in
Figure 17: The weights of terms lying on $p^- = \beta p^+$ as a function of $\Delta = |p^+ - p^-|$ computed by $\chi^2$: (a) $\beta < 1$, (b) $\beta > 1$.

Eq. (53). The linear dependence of weights on $\Delta$ values can be easily seen in the figures. The slope is positive for all $\beta$ values. This information is represented as $\{↑, 1\}$.

For $\Delta = 0.05$, the weights generated are illustrated as a function of $\beta$ and the corresponding angle in Figure 18. It can be seen in the figure that,

- For $\beta < 1$, the weights decrease as $\beta$ increases
- For $\beta > 1$, the weights increase as $\beta$ increases

5.5 Discussions

A summary of the findings through the analysis carried out is presented in Table 2. The second column presents the characteristics of contour lines. $rf$ and $MI$ provide linear contour lines whereas the others provide nonlinear lines. For a given line having slope $\beta$, the third column provides the values of weights as a function of $\Delta$. For instance, it can be seen in Eq. (49) that

$$f_1(\beta) = \frac{p \times \log(e)}{\beta (1-p)(1-\beta)}.$$

As seen in the table, for $rf$ and $MI$, the weights are constant in terms of $\Delta$. In other words, whatever the difference between $p^+$ and $p^-$ is, the weight is the same. Consider
Table 2: A summary of the main properties of six weighting schemes obtained from the analytical evaluations

<table>
<thead>
<tr>
<th>Weighting scheme</th>
<th>Contour lines</th>
<th>Weights on $p^- = \beta p^+$ as a function of $\Delta$</th>
<th>Weights on $p^- = \beta p^+$ as a function of $\Delta$ for ${\beta &lt; 1, \beta &gt; 1}$</th>
<th>Relation between weights and $\beta$ for fixed $\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rf</td>
<td>Linear</td>
<td>$f_{rf}(\beta)$</td>
<td>constant</td>
<td>${\downarrow}$</td>
</tr>
<tr>
<td>$MI$</td>
<td>Linear</td>
<td>$f_{MI}(\beta)$</td>
<td>constant</td>
<td>${\downarrow}$</td>
</tr>
<tr>
<td>prob</td>
<td>Nonlinear</td>
<td>$f_1(\beta)\Delta$, for $\beta &lt; 1$</td>
<td>${\uparrow, \downarrow}$</td>
<td>${\uparrow, \downarrow}$</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>Nonlinear</td>
<td>$f_3(\beta)\Delta$, for $\beta &lt; 1$</td>
<td>${\uparrow, \downarrow}$</td>
<td>${\downarrow, \uparrow}$</td>
</tr>
<tr>
<td>$OR$</td>
<td>Nonlinear</td>
<td>$f_5(\beta)(1 + \Delta)$, for $\beta &lt; 1$</td>
<td>${\uparrow, \downarrow}$</td>
<td>${\downarrow}$</td>
</tr>
<tr>
<td>$log(2 + OR)$</td>
<td>Nonlinear</td>
<td>$f_6(\beta) + f_{7}(\beta)\Delta$, for $\beta &lt; 1$</td>
<td>${\uparrow, \downarrow}$</td>
<td>${\downarrow}$</td>
</tr>
<tr>
<td>brf</td>
<td>Nonlinear</td>
<td>$f_8(\beta) - log(e)\Delta$, for $\beta &lt; 1$</td>
<td>${\uparrow, \downarrow}$</td>
<td>${\downarrow}$</td>
</tr>
<tr>
<td>$\overline{brf}$</td>
<td>Nonlinear</td>
<td>$f_9(\beta) - f_{10}(\beta)\Delta$, for $\beta &lt; 1$</td>
<td>${\uparrow, \downarrow}$</td>
<td>${\downarrow}$</td>
</tr>
</tbody>
</table>
Figure 18: The weights generated using $\chi^2$ for $\Delta = 0.05$ and various $\beta$ values. The corresponding angle is also presented in degrees.

A term $t_1$ for which $p^+ = 0.5$ and $p^- = 0.01$ which means that it appears in fifty percent of the positive documents and one percent of the negative documents. For $t_2$, assume that $p^+ = 1$ and $p^- = 0.02$. Both $t_1$ and $t_2$ are on the same line through the origin and hence the weights generated by $rf$ and $MI$ are the same for these terms where the $\Delta$ value of $t_1$, $\Delta_1 = 0.49$ and the $\Delta$ value of $t_2$, $\Delta_2 = 0.98$. In fact, it may be argued that $t_2$ is more informative about the category which is also reflected by a much bigger $\Delta$ value. However, this information is not considered by either $rf$ or $MI$. On the other hand, the weights provided by the other schemes having nonlinear contour lines are linearly proportional with $\Delta$.

Although there is a linear proportionality between the weights and $\Delta$ in all four term weighting schemes having nonlinear contour lines, the functional form of $\Delta$ and its influence on the overall weights are not the same. This results in differences in the relative values of the weights which is highly critical from categorization performance point of view since all terms are considered in a composite vector. Suppose that we have two different terms: $t_1$ for which $p^+ = 0.050$ and $p^- = 0.045$ and $t_2$ having $p^+ = 0.010$ and $p^- = 0.009$. For both terms $\beta = 0.9$ and the respective values of $\Delta$ are $\Delta_1 = 0.005$ and $\Delta_2 = 0.001$. The effect of the multiplicand $\Delta$ in prob is to generate weights $w_1 = 0.005 \times f_1(\beta)$ and $w_2 = 0.001 \times f_1(\beta)$ where
\[ w_1/w_2 = 5. \] On the other hand, the effect of the multiplicand \((1 + \Delta)\) in OR is to generate weights \(w_1 = 1.005 \times f_5(\beta)\) and \(w_2 = 1.001 \times f_5(\beta)\) where \(w_1/w_2 \approx 1.004\). It can be concluded the relative values of weights strongly depend on \(\Delta\) in the case of prob and \(\chi^2\) where there is a weak dependence in OR. For \(\log(2 + OR)\), brf and \(\widehat{brf}\), the relative values of weights also depend on \(f_i(\beta)\) values. For instance, in the case of brf the dependence of weights on \(\Delta\) becomes weaker as the value of \(f_k(\beta)\) increases.

The fourth column describes the relation between \(\Delta\) and corresponding weights for \(\beta < 1\) and \(\beta > 1\) using the expressions presented in the third column. As seen in this column, for a given \(\beta\), the weights generated by prob and \(\chi^2\) increase as \(\Delta\) increases for both \(\beta < 1\) and \(\beta > 1\). Consequently, these two schemes favor larger \(\Delta\) for both positive and negative terms. OR and brf provide weights that are also linearly proportional to \(\Delta\). However, OR favors larger \(\Delta\) values for positive terms as prob and \(\chi^2\) but it does not favor larger \(\Delta\) values for negative terms. On the contrary, brf favors larger \(\Delta\) values for negative terms and it does not favor larger \(\Delta\) values for positive terms. This is not intuitively reasonable since, for a fixed value of \(p^+ / p^-\), the weight of a positive term should not decrease as \(\Delta\) increases. This is in fact the case for all other weighting schemes considered in this study. These observations show that, although all weighting schemes aim at assigning larger weights to more important terms, considerable differences may exist between two schemes in ordering the terms according to their relative importance. This in fact explains the category dependent performance of the weighting schemes.

The last column presents a summary of the relations between the weights and \(\beta\) for a small but fixed \(\Delta\). The exact forms can also be easily obtained by replacing \(\Delta\) terms in weight expressions by a small constant. However, since the resultant behaviors will not be easily seen from the expressions, we preferred a simplified representation where it is only specified whether the corresponding function is increasing or decreasing in terms of \(\beta\). It can be seen in the table that, for rf, MI, OR and brf, the weights decrease as \(\beta\) increases which is shown by symbol \(\downarrow\). This means that these schemes favor positive terms more compared to negative terms for a
given $\Delta$. It was mentioned before that on $p^- = \beta p^+$ line, $brf$ favors larger $\Delta$ values for negative terms. In order to clarify this conflicting behavior, consider the contour lines of $\hat{brf}$ provided for a wider range of $\beta$ given in Figure 19. As seen in the figure, as $p^+$ and $p^-$ increase leading to a larger $\Delta$, the weights on the line $p^- = \beta p^+$ decrease when $\beta < 1$ and increase when $\beta > 1$.

Considering the terms presented in Figure 11 which have $p^+ < 0.35$ and $p^- < 0.35$, it can be seen that the weights decrease when $\beta$ increases.

$prob$ and $\chi^2$ behave differently from the rest when a particular $\Delta$ is considered. $\chi^2$ provides a symmetric behavior where minimum weights are generated for $\beta = 1$ and the weights increase as $\beta$ increases from one or decreases to zero as it can also be seen in Figure 18. This means that $\chi^2$ favors both negative and positive terms having the same $\Delta$ value. The behavior of $prob$ is more complicated. For negative terms, it is consistent with $rf$, $MI$, $OR$ and $brf$. However, the weights decrease as $\beta$ increases from zero to $1/2$ and then increases until $\beta = 1$. We believe that it is difficult to find a reasonable explanation for this behavior.

The expressions derived in terms of $\Delta$ and their characteristics summarized in Table 2 reveal the similarities and differences among different weighting schemes. In order to investigate the practical implications of these properties, experiments are conducted on Reuters-21578 dataset.
In particular, the effect of the multiplicand function of $\Delta$ in four weighting functions are studied.

6 Simulations

The ModApte split of the top ten classes of Reuters-21578 is one of the two datasets considered where the largest category “Earn” has 2877 training samples and 1087 test samples. The smallest category is “Corn” which has 182 training samples and 56 test samples. The negative categories are defined to include documents which belong to one or more of the remaining nine categories. The highly imbalanced category distribution of Reuters-21578 ModApte Top10 makes it significant among other datasets. WebKB is the other dataset that is a collection of web pages which belong to seven categories. They were collected by the Carnegie Mellon University Text Learning Group from several universities in 1997. Four of the categories namely, “Student”, “Faculty”, “Course” and “Project” which contain totally 4199 documents are generally used in text categorization experiments.

For each sample, stopwords are removed using SMART stoplist (Buckley, 1985). Subsequently, Porter stemming algorithm is applied (Porter, 1980). Term frequencies in each document are cosine-normalized. The SVM based classification toolbox, SVMlight with default parameters and linear kernel is employed as the classification scheme (Joachims, 1998, 1999). The commonly used evaluation metric for text classification known as $F_1$ measure is exploited. It is defined as the harmonic mean of precision ($P$) and recall ($R$) as

\[
F_1 = \frac{2 \times P \times R}{P + R},
\]

\[
P = \frac{TP}{TP + FP},
\]

\[
R = \frac{TP}{TP + FN}
\]

where $TP$, $FP$ and $FN$ denote true positives, false positives and false negatives respectively. For a given $K$-class problem, precision, recall and $F_1$ score are computed for each class separately. Macro $F_1$ score is then computed as the average of individual $F_1$ scores (Sebastiani, 2002).

In a recent study, it is observed that the $F_1$ scores of most weighting schemes plateau after
Table 3: $F_1$ scores obtained for the top ten categories of ModApte split for six weighting schemes.

<table>
<thead>
<tr>
<th>Category</th>
<th>rf</th>
<th>MI</th>
<th>prob</th>
<th>$\chi^2$</th>
<th>log($2+OR$)</th>
<th>$\hat{brf}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earn</td>
<td>98.58</td>
<td>98.53</td>
<td>97.57</td>
<td>97.61</td>
<td>98.53</td>
<td>98.72</td>
</tr>
<tr>
<td>Acquire</td>
<td><strong>97.47</strong></td>
<td>95.48</td>
<td>94.54</td>
<td>95.27</td>
<td>96.96</td>
<td>97.40</td>
</tr>
<tr>
<td>Money-fx</td>
<td>86.81</td>
<td>83.09</td>
<td>84.44</td>
<td>83.80</td>
<td>85.24</td>
<td><strong>87.91</strong></td>
</tr>
<tr>
<td>Grain</td>
<td><strong>96.58</strong></td>
<td>93.66</td>
<td>96.30</td>
<td>95.89</td>
<td>95.17</td>
<td>94.77</td>
</tr>
<tr>
<td>Crude</td>
<td><strong>92.35</strong></td>
<td>91.20</td>
<td>90.39</td>
<td>90.10</td>
<td>91.78</td>
<td>90.86</td>
</tr>
<tr>
<td>Trade</td>
<td>87.11</td>
<td>85.45</td>
<td>84.68</td>
<td>85.97</td>
<td>85.59</td>
<td><strong>87.39</strong></td>
</tr>
<tr>
<td>Interest</td>
<td>81.97</td>
<td>80.17</td>
<td>79.32</td>
<td>79.32</td>
<td><strong>82.30</strong></td>
<td>80.99</td>
</tr>
<tr>
<td>Wheat</td>
<td>84.67</td>
<td>80.60</td>
<td><strong>87.67</strong></td>
<td>86.11</td>
<td>82.96</td>
<td>83.72</td>
</tr>
<tr>
<td>Ship</td>
<td>82.28</td>
<td>80.25</td>
<td><strong>87.12</strong></td>
<td>86.42</td>
<td>81.01</td>
<td>81.53</td>
</tr>
<tr>
<td>Corn</td>
<td>86.79</td>
<td>85.71</td>
<td>85.98</td>
<td>84.91</td>
<td><strong>88.68</strong></td>
<td>86.00</td>
</tr>
<tr>
<td>Macro $F_1$</td>
<td>89.46</td>
<td>87.42</td>
<td>88.80</td>
<td>88.54</td>
<td>88.82</td>
<td>88.93</td>
</tr>
</tbody>
</table>

5000 features for SVM (Lan et al., 2009). Because of this, the top 5000 features ranked by $\chi^2$ are considered in the experiments. Instead of OR and $brf$, $log(2+OR)$ and $\hat{brf}$ are implemented due to practical reasons stated before. The schemes considered are used for weighting normalized term frequencies.

Table 3 presents the results achieved where the best result for each category is typed in boldface. As seen in the table, the relative performance of the schemes depend on the category. However, the effects of implicit similarities among the schemes can be seen in the table. For instance, “Wheat” and “Ship” are the two categories where $prob$ performs the best and the results provided by $\chi^2$ are ranked second. Moreover, both schemes perform much better compared to the others. It can be argued that a linear function of the form $k\Delta$ as a scale factor on the line $p^- = \beta p^+$ is useful for these two categories.
It is mentioned before that the behaviors of OR and brf are conflicting when their use of \(\Delta\) in weighting is considered. This is also evident from the results since there is not any category where both of them provide better results compared to the rest. In particular, on five categories, the best \(F_1\) score is achieved by either \(\log(2+OR)\) or \(\text{brf}\). However, in all five cases, the scheme providing worse performance than the other is ranked after \(rf\) which does not consider \(\Delta\) as a scale factor.

When the average performances, Macro \(F_1\) are considered, it is seen that \(rf\) is ranked first and the remaining are ranked in the order \(\text{brf}, \log(2+OR), \text{prob}, \chi^2, \text{and MI}\). However, when all ten categories are considered, each weighting scheme other than \(MI\) comes out to be the best on almost equal number of datasets. As a matter of fact, it can be argued that the use of \(\Delta\) as a factor in term weighting may be useful provided that the best-fitting function is selected.

For further investigation of \(\Delta\)’s importance, experiments are conducted on exploiting best-fitting function of \(\Delta\) explicitly for improving \(rf\). As seen in Table 3, the best results are obtained by the weighting schemes which implicitly considers \(\Delta\) as a multiplicative factor on categories “Wheat” and “Ship”. Taking this into account, the weight \(\Delta \times rf\) is applied on these categories. The experimental results are presented in Table 4. The \(F_1\) score is improved from 84.67 to 87.67 on “Wheat” and from 82.28 to 84.66 on “Ship” categories. Similarly, weight of the form

\[
w = \begin{cases} 
(1 + \Delta) \times rf & \beta < 1, \\
(1 - \Delta) \times rf & \beta > 1 
\end{cases}
\]  

(57)

is applied on “Interest” and “Corn” since \(\log(2+OR)\) is the best performing scheme on these categories. It can be seen in the table that the \(F_1\) score is improved to 82.78 on “Interest” category and to 87.85 on “Corn” category.

The experiments are also repeated for \(MI\). More specifically, the weight \(\Delta \times MI\) is applied on “Wheat” and “Ship” and weight of the form

\[
w = \begin{cases} 
(1 + \Delta) \times MI & \beta < 1, \\
(1 - \Delta) \times MI & \beta > 1 
\end{cases}
\]  

(58)

is applied on “Interest” and “Corn”. The \(F_1\) scores are improved on all four categories.
Table 4: $F_1$ scores obtained for investigating the practical implications of the analytical observations on Reuters-21578.

<table>
<thead>
<tr>
<th>Category</th>
<th>$\Delta \times rf$</th>
<th>$(1 + \Delta) \times rf, \beta &lt; 1$</th>
<th>$\Delta \times MI$</th>
<th>$(1 + \Delta) \times MI, \beta &lt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$(1 - \Delta) \times rf, \beta &gt; 1$</td>
<td>$(1 - \Delta) \times MI, \beta &gt; 1$</td>
<td></td>
</tr>
<tr>
<td>Interest</td>
<td>-</td>
<td>82.78</td>
<td>-</td>
<td>80.33</td>
</tr>
<tr>
<td>Wheat</td>
<td>87.67</td>
<td>-</td>
<td>85.31</td>
<td>-</td>
</tr>
<tr>
<td>Ship</td>
<td>84.66</td>
<td>-</td>
<td>86.75</td>
<td>-</td>
</tr>
<tr>
<td>Corn</td>
<td>-</td>
<td>87.85</td>
<td>-</td>
<td>86.54</td>
</tr>
</tbody>
</table>

Table 5: $F_1$ scores obtained for investigating the practical implications of the analytical observations on WebKB.

<table>
<thead>
<tr>
<th>Category</th>
<th>$rf$</th>
<th>$MI$</th>
<th>$\Delta \times rf$</th>
<th>$(1 + \Delta) \times rf, \beta &lt; 1$</th>
<th>$\Delta \times MI$</th>
<th>$(1 + \Delta) \times MI, \beta &lt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$(1 - \Delta) \times rf, \beta &gt; 1$</td>
<td>$(1 - \Delta) \times MI, \beta &gt; 1$</td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>91.24</td>
<td>89.14</td>
<td>92.35</td>
<td>-</td>
<td>90.76</td>
<td>-</td>
</tr>
<tr>
<td>Faculty</td>
<td>84.52</td>
<td>83.44</td>
<td>-</td>
<td>84.89</td>
<td>-</td>
<td>84.37</td>
</tr>
</tbody>
</table>

The experiments are also conducted on WebKB. Since the training and test sets are not fixed for this dataset, four-fold cross validation is performed and the average Macro $F_1$ scores are reported. On “Student” category, the highest Macro $F_1$ score (92.40) is provided by $\chi^2$ whereas the best Macro $F_1$ score (84.98) is achieved when $log(2 + OR)$ is used on “Faculty” category. Taking these into account, the weight $\Delta$ is applied on “Student” category and the weights presented in Eqs. (57) and (58) are applied on “Faculty”. The experimental results are presented on Table 5. On “Student” category, the Macro $F_1$ score is improved from 91.24 to 92.35 for $rf$ and from 89.14 to 90.76 for $MI$. On “Faculty” category, the score is improved from 84.52 to 84.89 for $rf$ and from 83.44 to 84.37 for $MI$.

The experimental results have shown that considerable improvements can be achieved by
the explicit use of an appropriate function of $\Delta$ with $rf$ and $MI$ which have linear contour lines.

It can be argued that the best-fitting form can be computed for each category separately using the training data for achieving improved Macro $F_1$ scores.

7 Conclusions

In this study, analytic expressions of the patterns formed by the terms receiving equal weights on the two dimensional space of term occurrence probabilities are derived. The patterns are also presented in the form of contour lines which clarified the similarities and differences among the weighting schemes considered. Taking into account the fact that the contour lines of the scheme which provides the best overall performance are linear lines that pass through the origin, the characteristic of the weights of other schemes on these lines are investigated.

For the terms lying on the line $p^- = \beta p^+$, it is found that the weights generated can be expressed as a function of $\Delta$ and $\beta$. It is observed that the expressions computed accordingly for $prob$ and $\chi^2$ are similar whereas the behaviors of $OR$ and $brf$ are conflicting when their use of $\Delta$ in weighting is considered.

The analytical findings are verified through the simulations carried out. Both $rf$ and $MI$ are modified for further investigation of the corresponding practical implications. Experimental results have shown that $\Delta$ can be used as a scale factor for $rf$ and $MI$ which have linear contour lines when $\chi^2$ provides considerably better performance compared to the others. Similarly, the scale factor $(1 + \Delta)$ is shown to improve $rf$ and $MI$ when $log(2 + OR)$ performs the best. It can be concluded that the selection of category dependent form of weights in terms of $\beta$ and $\Delta$ is a promising research direction. Our following research will be on learning the best-fitting forms of weights for category dependent term weighting.

The functional forms computed for the schemes under concern are useful for describing their weighting behaviors. However, the reason for category dependent performances of the schemes is not clear. In particular, the main reason for the importance of the factor $\Delta$ on “Wheat” and
“Ship” categories of Reuters-21578 and on “Student” category of WebKB is not yet clarified. Similarly, the effectiveness of the factors $(1 + \Delta)$ and $(1 - \Delta)$ on “Interest”, “Corn” and “Faculty” should be questioned. We believe that these issues deserve further investigation.

Acknowledgements

We are grateful to the anonymous reviewers for their constructive suggestions. The numerical calculations reported in this paper were partly performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TR-Grid e-Infrastructure). This work was supported by the research grant MEKB-09-02 provided by the Ministry of Education and Culture of Northern Cyprus.
APPENDIX

A  Derivation of Eq. (17)

Set the value of \( b_{rf} \) given in Eq. (17) to \( \tau \) as follows:

\[
\begin{align*}
brf &= \log \left( \frac{p^+}{p^-} \frac{p}{(1-p^-)} \frac{1-p^+}{(1-p^-) (1-p)} \right) = \tau. \quad (59)
\end{align*}
\]

Then, we obtain

\[
\begin{align*}
p^+ \frac{p}{(1-p^-)} \frac{1-p^+}{(1-p^-) (1-p)} &= 2^{\tau}. \quad (60)
\end{align*}
\]

Since \( p \) is constant, moving it to the right hand side we obtain

\[
\begin{align*}
p^+ \frac{1-p^+}{p^- (1-p^-)} &= \frac{(1-p)^2}{p^2} \times 2^{\tau} = \xi \quad (61)
\end{align*}
\]

where \( \xi \) is another constant.

B  Derivation of Eq. (35)

Given that

\[
\begin{align*}
f(\Delta) &= \frac{1}{\beta} \times \frac{(1-\beta) - \beta \Delta}{(1-\beta) - \Delta}, \quad (62)
\end{align*}
\]

\( f'(\Delta) \) can be computed as

\[
\begin{align*}
f'(\Delta) &= \frac{1}{\beta} \times \frac{-\beta(1-\beta) + \beta \Delta + (1-\beta) - \beta \Delta}{(1-\beta - \Delta)^2} \\
&= \frac{1}{\beta} \times \frac{(1-\beta)^2}{(1-\beta - \Delta)^2}. \quad (63)
\end{align*}
\]

Hence, \( f'(0) = \frac{1}{\beta} \). Since \( f(0) = \frac{4}{\beta} \), using Eq. (34) we obtain \( OR \approx \frac{1}{\beta} (1 + \Delta) \).

C  Derivation of Eq. (38)

Let \( f(\Delta) \) denote the expression for \( \log(2 + OR) \) \( (\beta < 1) \) which can be written as

\[
\begin{align*}
f(\Delta) &= \log(e) \times \ln \left( 2 + \frac{1}{\beta} \times \frac{(1-\beta) - \beta \Delta}{(1-\beta) - \Delta} \right) \\
&= \log(e) \times \ln(u(\Delta)) \quad (64)
\end{align*}
\]
where \( \ln(.) \) denotes the natural logarithm and \( e \) denotes the base of natural logarithm. Note that \( \frac{\Delta}{\Delta \ln(u(\Delta))} = \frac{u'(\Delta)}{u(\Delta)} \) where \( u'(\Delta) \) denotes the derivative of \( u(\Delta) \) with respect to \( \Delta \). Then

\[
u'(\Delta) = \frac{1}{\beta} \frac{(1 - \beta)^2}{(1 - \beta - \Delta)^2}.
\]

(65)

Hence, \( f'(0) = \log(e) \times \frac{u'(0)}{u(0)} \) which can be computed as

\[
f'(0) = \log(e) \frac{\beta + 1}{2}\]

(66)

Since \( f(0) = \log(2 + \frac{1}{\beta}) \), using Eq. (34) we obtain Eq. (38).

**D Derivation of Eq. (39)**

Let \( f(\Delta) \) denote the expression for \( \log(2 + OR) \) \( (\beta > 1) \) which can be written as

\[
f(\Delta) = \log(e) \times \ln\left(2 + \frac{1}{\beta} \times \frac{(\beta - 1) - \beta \Delta}{(\beta - 1) - \Delta}\right)
\]

(67)

\[
= \log(e) \times \ln(u(\Delta))
\]

Then

\[
u'(\Delta) = \frac{1}{\beta} \frac{-(\beta - 1)^2}{(\beta - 1 - \Delta)^2}.
\]

(68)

Hence, \( f'(0) = \log(e) \times \frac{u'(0)}{u(0)} \) which can be computed as

\[
f'(0) = \frac{-\log(e)}{2\beta + 1}
\]

(69)

Since \( f(0) = \log(2 + \frac{1}{\beta}) \), using Eq. (34) we obtain Eq. (39).

**E Derivation of Eq. (49)**

Let \( f(\Delta) \) denote the expression for \( \text{prob} \) \( (\beta > 1) \) which can be written as

\[
f(\Delta) = \log(e) \times \ln\left(1 + \frac{p}{\beta(1 - p)} \times \frac{\Delta}{(1 - \beta - \Delta)}\right)
\]

(70)

\[
= \log(e) \times \ln(u(\Delta))
\]

Then

\[
u'(\Delta) = \frac{p}{\beta(1 - p)} \frac{(1 - \beta)}{(1 - \beta - \Delta)^2}.
\]

(71)
\[ u(0) = 1. \] Hence, \( f'(0) = \log(e) \times \frac{u'(0)}{u(0)} \) which can be computed as

\[ f'(0) = \frac{p \times \log(e)}{\beta(1 - p)(1 - \beta)} \]  \hspace{1cm} (72)

Since \( f(0) = \log(1) = 0 \), using Eq. (34) we obtain Eq. (49).


