INVESTIGATION OF TERM WEIGHTING SCHEMES IN CLASSIFICATION OF IMBALANCED TEXTS

Behzad Naderalvojoud, Ahmet Selman Bozkir and Ebru Akcapinar Sezer
Hacettepe University, Computer Engineering Department, Ankara, Turkey

ABSTRACT

Class imbalance problem in data, plays a critical role in use of machine learning methods for text classification since feature selection methods expect homogeneous distribution as well as machine learning methods. This study investigates two different kinds of feature selection metrics (one-sided and two-sided) as a global component of term weighting schemes (called as *tffs*) in scenarios where different complexities and imbalance ratios are available. Traditional term weighting approach (*tfidf*) is employed as a base line to evaluate the effects of *tffs* weighting. In fact, this study aims to present which kind of weighting schemes are suitable for which machine learning algorithms on different imbalanced cases. Four classification algorithms are used to indicate the effects of term weighting schemes on the imbalanced datasets. According to our findings, regardless of *tfidf*, term weighting methods based on one-sided feature selection metrics are better approaches for SVM and k-NN algorithms while two-sided based term weighting methods are the best choice for MultiNB and C4.5 on the imbalanced texts. As a result, the use of term weighting methods based on one-sided feature selection metrics is recommended for SVM and *tfidf* is suitable weighting method for k-NN algorithm in text classification tasks.

KEYWORDS

Class imbalance problem, machine learning, text classification, term weighting, feature selection

1. INTRODUCTION

In machine learning, text classification is a supervised learning task which can predict the categories of unlabeled documents based on an inductive model learned from labeled documents. The common machine learning algorithms which have been used for this purpose include support vector machine (SVM), k-nearest neighbor (k-NN), naïve Bayesian (NB), neural networks (NN), decision trees (C4.5) and Rocchio (Ogura et al, 2011; Liu et al, 2009; Chawla et al, 2011). Binary classification by machine learning algorithms is usually performed based on a fundamental assumption that the distributions of two classes should be close to each other. In other words, there should be as many positive examples as negative ones (Chawla et al, 2004). This mentioned fundamental requirement cannot be always met since there are many imbalanced datasets relying on real world examples, (e.g. cancer detection, network intrusion detection, credit card fraud detection, oil-spill detection). At this point, classifiers generally present weak performance while the dominant class is well classified; the examples belonging to the minor class tend to be misclassified. Nonetheless, the aim of these classifiers is to generate a model that best fits the training data with minimum error rate. Furthermore, they consider the global quantities in generating the model.

Class imbalance problem occurs in text classification tasks when the numbers of positive samples are significantly lower than negative ones. There are other domain characteristics that aggravate the problem such as (1) class complexity (2) size of training set and (3) subclusters (Japkowicz and Stephen, 2002). In typical binary imbalanced text classification, the positive class consists of the documents that belong to one subject and negative class consists of all other remaining items. Thus, increment in the number of negative class samples leads to growth of class complexity. In this case, the positive class can be formed as a cluster while the negative class cannot. Therefore, raising the degree of imbalance by incrementing the negative documents with different subjects causes aggravation of class distribution and growing the number of subclusters. In order to generate a classification model with low generalization error for minor class, existence of adequate number of samples in the training data set is crucial. Therefore, the datasets which have insufficient number of positive samples tend to be misclassified since the classification algorithms aim to build models which have generalization capability.

Class imbalance problem also exists in the multi classification schemas when one class is assumed as a target category (positive class or minor class) and the union of the other classes are considered as negative class (majority class) (Ogura et al, 2011). In this case, most of the machine learning methods are often biased to the majority class and ignore the minor class since they attempt to minimize the global parameters such as total error rate and do not take the class distribution into consideration (Japkowicz and Stephen, 2002).

An inevitable stage in the text classification task is representing the textual documents in a realizable form for any classifier. As a well-known method, vector space model (VSM) is known as a text representation model which makes a transformation from content of the natural language texts into a vector of term space (Salton and Buckley, 1988). In this model, assigning a weight for each term is effective to represent data, since the importance of each term in different documents can vary. This issue can be taken into consideration in the imbalanced cases. Thus, *tfidf* as a basic term weighting scheme is used in text classification tasks. This method belongs to information retrieval field and does not need any prior information about the categories; hence it is called as unsupervised term weighting approach (Lan et al, 2009). In the text classification, since the labeled documents are available, this information can be used as a global parameter in the term weighting scheme. Thus, the term weighting approaches which use the prior known information, are called supervised approaches in the literature (Debole and Sebastiani, 2004).

The common strategies proposed in the class imbalance problem literature are addressed at data and algorithmic level. At algorithmic level, the employed strategies include determining the decision threshold (Chen et al, 2006), adjusting the probabilistic estimate at the information gain and Bayesian based methods such as decision tree and naïve Bayes respectively (Kibriya et al, 2005). At data level, the proposed approaches include the different forms of resampling methods (Chawla et al, 2004) and instance weighting schemes (Liu et al, 2009). In this study we focus on the data level approaches. The first approach is resampling data in via under sampling the majority class and over sampling the minority class. Moreover, Liu investigated several resampling techniques in the realm of imbalanced text classification (2004). Chawla et al. proposed a synthetic technique for over sampling the minority class samples named SMOTE (2011).

Another approach at the data level is using instance weighting methods in representation of data. In their study, Debole and Sebastiani, replaced the *idf* by category-based feature selection metrics (i.e. chi square, information gain and gain ratio) that had been used in the term selection phase (2004). They employed SVM as learning method with Reuters-21578 and showed supervised term weighting cannot be consistently superior to *tfidf*. In another study, (Lan et al, 2009) proposed a supervised term weighting method, *tf.rf*, based on distribution of relevant documents in the collection. Their proposed method was providing better performance than the other weighting schemes based on information theory and statistical metrics in combination with SVM and k-NN algorithms. On the other hand, a simple probability based term weighting scheme was proposed to better distinguish documents in minor categories (Liu et al, 2009). Moreover, Sun et al. provided a comparative study on the effectiveness of resampling and instance weighting strategies using SVM (2009).

To best of our knowledge, in most of the studies the proposed solutions for dispelling the class imbalance problem were evaluated by using one or two classifiers (especially by SVM). In this study, we try to survey the instance weighting strategy in combination with four algorithms which work based on four different approaches. Thus, the following objectives will be addressed in this study:

- Investigation of the supervised and unsupervised weighting approaches on imbalanced datasets as well as compatibility of each weighting method with machine learning algorithms.
- Comparing the effect of two-sided feature selection metrics (metrics that consider the negative, non-relevant, features as well as the positive, relevant, ones) with one-sided metrics (metrics that take only the positive features into consideration) at the term weighting perspective.

In fact, we try to discuss which kind of feature selection metrics (as a component of term weighting scheme) can be beneficial to represent imbalanced data and which term weighting schemes are suitable for which machine learning algorithms. For this purpose, four different classifiers (SVM, k-NN, NB and C4.5) are employed in the experiments. The main reason of this selection is that they are based on different approaches (i.e. perceptron based, instance based, probabilistic based and information gain based).

2. FEATURE SELECTION AND TERM WEIGHTING

Feature selection is often employed in text classification tasks in order to reduce dimensionality when documents are represented as a set of words without considering the grammar and order of the words. On the other hand, it has positive effects on improving the classification accuracy by reducing over fitting problem (Liu et al, 2009). In this study information gain with local policy is used as feature selection metric since it has introduced better performance on the imbalanced text classification (Tasci and Gungor, 2013).

Feature selection metrics can be used as a global factor of term weighting function since they evaluate the importance of a term for a specific category. In this study, two approaches are used in the formula of different feature selection metrics; (1) one-sided and (2) two-sided metrics. One-sided metrics take only positive features (i.e. relevant terms) into consideration since they compute the relevancy power of terms for a category. We test two common one-sided metrics i.e. *RF* and *Odds Ratio* (Lan et al, 2009) in the experiments. Two-sided metrics consider both positive and negative features implicitly. In other words, they can take into account either the relevancy or non-relevancy power of terms for a category. We also investigate the effect of two well-known two-sided feature selection metrics i.e. *Information Gain* and *Chi Square* which are based on probabilistic and information theories (Debole and Sebastiani, 2004). The mentioned feature selection metrics in the experiments and their formulas have been summarized in Table 1.

In text classification, term weighting is usually realized by methods taken from information retrieval and text search fields. There are three assumptions behind these traditional methods. They consider following points (1) multiple appearances of a term in a document are no less important than single appearance (*tf* assumption); (2) rare terms are no less important than frequent terms (*idf* assumption); (3) for the same quantity of term matching, long documents are no more important than short documents (*normalization* assumption) (Debole and Sebastiani, 2004).

Tfidf as a standard term weighing scheme is used in information retrieval and text classification tasks. It is formulated in form of multiplying term frequency (*tf*) by inverse document frequency (*idf*). The common and normalized form of that are shown in Equations 1 and 3 respectively (Salton and Buckley, 1988):

$$tfidf(t_i, d_j) = tf(t_i, d_j) \times idf(t_i) \tag{1}$$

$$idf(t_i) = log\left(\frac{N}{N_{t_i}}\right) \tag{2}$$

$$W_{i,j} = \frac{tfidf(t_i, d_j)}{\sqrt{\sum_{k=1}^{|T|} tfidf(t_k, d_j)^2}} \tag{3}$$

where $tf(t_i, d_j)$ denotes the number of times that term t_i occurs in document d_j, N is the number of all documents in the training set, N_{t_i} denotes the number of documents in the training set in which term t_i occurs at least once and $|T|$ denotes the number of unique terms which have been extracted from the training set. In this study, *tfidf* is used as a standard term weighting scheme throughout the experiments. At supervised term weighting, feature selection metrics are replaced instead of *idf* in the Equations 1 and 3. We named that as *tffs* in this study.

3. EXPERIMENTS

In this study, the effect of each feature selection metric is investigated over the imbalanced text classification by considering as a global component of the term weighting function. At the experiment stage, we have used R8 dataset which was extracted from Reuters-21578 and 20Newsgroups datasets which are publicly available at (Dataset for single-label text categorization, 2014) for single label text categorization. These two datasets have been widely used in text classification researches (i.e. Debole and Sebastiani, 2004; Sun et al, 2009; Ogura et al, 2011).Pre-processing steps have been applied on the datasets such as removing the 524 SMART stop words and applying Porter's Stemmer algorithm. We conducted two types of experiments for balanced and imbalanced cases. In order to control the state of imbalance and degree of complexity, we selected one category as the positive class and the remaining portion as the negative one as (Ogura et al, 2011) had done. The R8 dataset has eight categories with imbalanced number of documents for categories and consequently it has lower complexity than the 20Newsgroups dataset. In the 20Newsgroups dataset, there exist 20 categories with almost equal number of documents. Thus, with one vs. all configuration, we can make an imbalanced case with high complexity due to the abundance of different categories in the negative class. First, we tried to make 1:1 configuration for R8 dataset by selecting the largest category among the others (i.e. *earn* category) as positive class and the sum of the other categories were considered as negative class. For 20Newsgroups dataset, *sci.space* was selected as positive class and *sci.electronics* was chosen as negative class. In the second stage, the imbalance situation was constituted on the R8 and 20Newsgroups datasets by selecting the *trade* and *sci.space* categories as positive class respectively with the consideration of the union of the other categories as the negative class. Thus, 1:20 imbalance ratio was approximately obtained for each dataset with different degree of complexities. The experiments were performed on the original training and test sets for the both datasets as shown in Table 2. By using information gain metric, the top 25 features were selected from each category for both datasets.

Table 1. All metrics used in the experiments as the global factor of term weighting schemes

Metric name	Type	Formula
Chi square	Two-sided	$X^2 = N\dfrac{(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$
Information gain	Two-sided	$IG = \frac{a}{N}log\frac{aN}{(a+c)(a+b)} + \frac{b}{N}log\frac{bN}{(b+d)(a+b)} + \frac{c}{N}log\frac{cN}{(a+c)(c+d)} + \frac{d}{N}log\frac{dN}{(b+d)(c+d)}$
Odds ratio	One-sided	$OR = log\frac{ad}{bc}$
Relevance frequen	One-sided	$RF = \log\left(2 + \frac{a}{\max(1,c)}\right)$

Notation:
a denotes the number of documents belongs to positive class which contains term t_i
b denotes the number of documents belongs to positive class which does not contain term t_i
c denotes the number of documents belongs to negative class which contains term t_i
d denotes the number of documents belongs to negative class which does not contain term t_i
N denotes the number of all documents in the data training set

Four popular classification algorithms i.e. libSVM (Chang and Lin, 2011), Multinomial Naïve Bayes (MultiNB) (Kibriya et al., 2005), decision tree (C4.5) (Chawla et al., 2011) and k-Nearest Neighbors (k-NN) (Ogura et al., 2011) were used to evaluate the weighting methods. In fact, we evaluate the compatibility of each classifier with each of the term weighting functions. Furthermore, we used linear kernel with default parameters for libSVM and selected k=5, 15, 25 and 35 for k-NN algorithm. For k-NN, we computed the average of the results which are obtained from different values of k in the experiments. To evaluate the results, F_1-score metric obtained from Precision (P) and Recall (R) values is used via following formulas: (1) F_1 = 2PR/(P+R), (2) P = TP/ (TP+FP) and (3) R = TP/(TP+FN) where TP, FP and FN are true positives, false positives and false negatives, respectively.

Table 2. Properties of datasets

Dataset	# of training documents	# of test documents	# of classes
R8	5485	2189	8
20 Newsgroups	11293	7528	20

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Balanced Case

In the first stage of experiments, we took the 1:1 balanced situation into consideration combined with different complexity. Fig. 1 shows the results of the supervised (*tffs*) and unsupervised (*tfidf*) term weighting schemes over the R8 dataset using the four different classifiers. It is observed that the SVM performs significantly better than the other classifiers. It also shows the compatibility of SVM with two-sided feature selection metrics when they are used in the term weighting scheme. According to obtained results, *tfidf* weighting gives better results than the supervised ones for k-NN, C4.5 and MultiNB. Among these classifiers, C4.5 and MultiNB are more sensitive to weighting schemes. Nonetheless, term weighting based on one-sided metrics are better approach for them in comparison with two-sided ones.

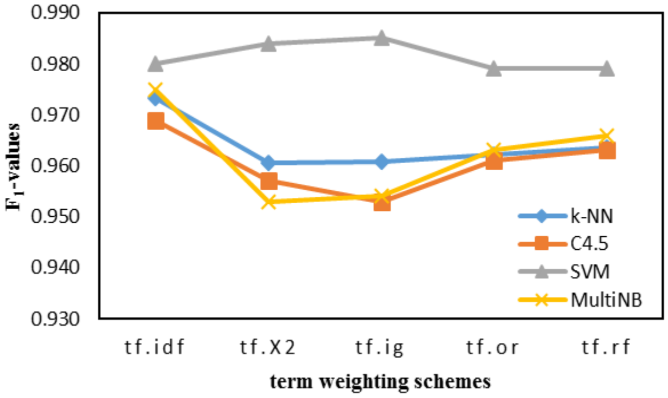

Figure 1. The F_1-values of five weighting schemes tested over R8 dataset with balanced setting using four different classifies.

We compared the previous observation with the results obtained from 20Newsgroups dataset. Fig. 2 indicates the performance of weighting schemes over the 20Newsgroups dataset using the same classifiers. As shown in Fig. 2, both C4.5 and MultiNB methods perform better than the k-NN and SVM.

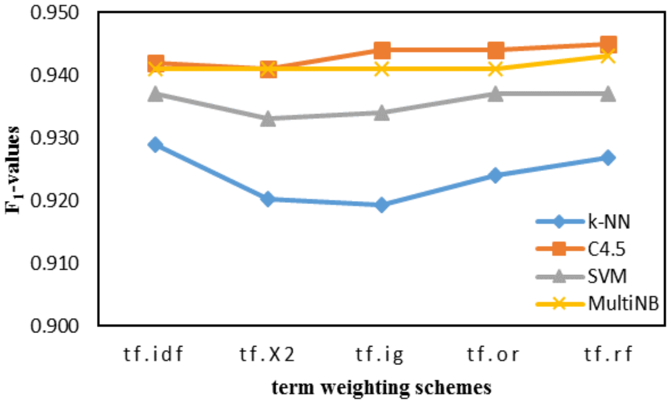

Figure 2. The F_1-values of five weighting schemes tested over 20Newsgroups dataset with balanced setting using four different classifies.

It is noted that the observation is different than the R8 dataset since its complexity is different from the 20Newsgroups. Also we selected two similar categories for 20Newsgroups dataset while the positive class in R8 dataset is less similar to negative class. This leads to increase in the error region between positive and negative classes in the training set and consequently raises the generalization error for the model obtained from SVM. Hence the performance of SVM degrades in the 20Newsgroups dataset. According to both observations, we can conclude that the performance of one-sided metrics is better than the two-sided ones excluding SVM which can work well with two-sided based metrics, shown in Fig. 1.

4.2 Imbalanced Case

In the second stage of the experiments, we tested the behavior of term weighting schemes and classification algorithms over the 1:20 imbalanced case. First observation is that SVM performs well with one-sided term weighting methods and can even outperform *tf.idf*, while k-NN shows an adaptation with *tf.idf* and *tf.rf* term weighting schemes. On the contrary, MultiNB and C4.5 give better performance by two-sided methods and outperform *tfidf* (please see *tfidf* in the R8 dataset, shown in Fig. 3). In fact, Fig. 3 demonstrates the compatibility of one-sided methods with SVM, two-sided ones with MultiNB and C4.5, and both *tf.idf* and *tf.rf* with k-NN algorithm. It can be also observed that SVM and MultiNB effectively perform via supervised term weighting schemes on the imbalanced data. In order to expand the obtained results, we employed the same experiments on the 20Newsgroups dataset by using same imbalance ratio and more complexity configuration.

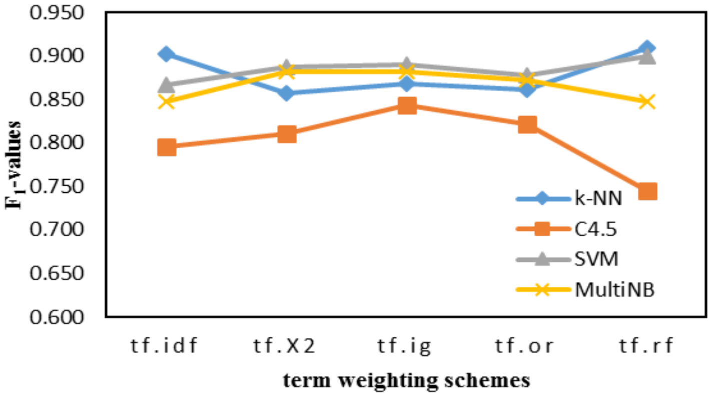

Figure 3. The F_1-values of five weighting schemes tested over R8 dataset with imbalanced setting using four different classifies.

Fig. 4 shows the classification performance of five term weighting schemes tested on the 20Newsgroups dataset using different classifiers. As shown in Fig. 4, *tfidf* outperforms the supervised term weighting schemes in the 20Newsgroups dataset which has more complexity than the R8. In the 20Newsgroups dataset, it is observed that as the degree of class complexity raises the number of subclusters increases. Therefore, it can be concluded that category based metrics cannot clearly make a contrast between documents of positive and negative classes. Nonetheless, *tfidf* which has no attention to category labels creates a good contrast in the imbalanced case with high complexity. Among the supervised weighting schemes, SVM and k-NN perform well with one-sided metrics, while C4.5 and MultiNB are compatible with two-sided metrics. This is similar to the previous observation which was obtained from R8 dataset. According to the both results in imbalanced cases, SVM with the term weighting schemes based on one-sided metrics usually performs well on the imbalanced datasets as shown in Figs. 3 and 4.

According to our findings, we can conclude that supervised term weighting schemes usually provide better representation of data for the classifiers on the imbalanced datasets with less complexity (as shown in Fig 3). Nonetheless, for high degree of complexity, *tfidf* seems a better term weighting scheme for the machine learning algorithms.

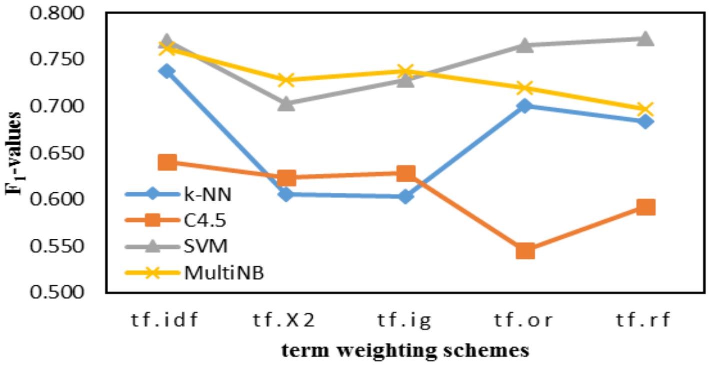

Figure 4. The F_1-values of five weighting schemes tested over 20Newsgroups dataset with imbalanced setting using four different classifies.

To determine the significance of the term weighting methods for each algorithm, we perform the ANOVA test on the F_1 values obtained from term weighting methods rather than t-test since it shows the significance of the results in more than 2 groups. As shown in Table 3, since the P-values of the tests are less than 0.05 for each case, there is a statistically significant difference between the mean F_1 values of levels at the 95.0% confidence level. Table 3 presents a multiple comparison of results to determine which algorithms differ significantly from others with respect to term weighting approaches. It can be observed that MultiNB and SVM significantly perform better than the others by using term weighting methods. At the Table 3, two and three homogenous groups are identified using columns of X's for R8 and 20Newsgroups datasets respectively. Within each column, the levels containing X's constitute groups which there are no statistically significant differences. To create a discrimination between means, Fisher's least significant difference (LSD) procedure is employed here.

Table 3. ANOVA test for F_1values obtained from 5 weighting methods for each algorithm for imbalanced cases

	R8 with P-Value = 0.0003			20 Newsgroup with P-Value = 0.0002			
Algorithms	F means	Homogeneous Groups		F means	Homogeneous Groups		
C4.5	0.8034	X		0.6060	X		
KNN	0.8793		X	0.6661		X	
MultiNB	0.8662		X	0.7290			X
SVM	0.8842		X	0.7478			X

5. CONCLUSION

In this study, the effects of two kinds of supervised term weighting schemes (one-sided and two-sided term selection metrics) were investigated on the balanced and imbalanced texts with different degrees of complexity. *Tfidf* was used as a base line to evaluate the effect of supervised weighting methods on the imbalanced texts. We evaluated the performance of each weighting method by using four different machine learning algorithms (SVM, k-NN, MultiNB and C4.5). Actually, the appropriateness of weighting methods and machine learning algorithms were studied here and, to investigate this problem we generated datasets with two different complexity level such as balanced and imbalanced cases. According to our findings, in the balanced cases, almost all classifiers had a little impact on the weighting methods. Nonetheless, it can be seen that the supervised term weighting approach does not possess any effective superiority to *tfidf*. Furthermore, it was observed that one-sided based term weighting schemes outperform the two-sided based ones in the most balanced cases.

In the imbalanced cases, it is realized that all four classifiers were susceptible to the term weighting methods. Regardless of *tfidf*, one-sided term weighting methods are better approach for SVM and k-NN algorithms while two-sided methods are the best choice for MultiNB and C4.5. According to our results, it can be concluded that supervised term weighting methods based on one-sided term selection metrics are the best choice for SVM in the imbalanced datasets and k-NN algorithm usually perform well with *tfidf*. It should be also noted that MultiNB classifier presents interesting results on the imbalanced cases. As another finding, although supervised methods cannot constantly retain their superiority to *tfidf* on the more complex imbalanced datasets, they can provide effective results for classification algorithms.

REFERENCES

Chang, C. C. and Lin, C. J., 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST),* Vol. 2, No. 3, pp 27.

Chawla, N. V. et al, 2004. Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, Vol. 6, No. 1, pp 1-6.

Chawla, N. V. et al, 2011. SMOTE: synthetic minority over-sampling technique. *arXiv preprint arXiv:1106.1813*.

Chen, J. J. et al, 2006. Decision threshold adjustment in class prediction. *SAR and QSAR in Environmental Research*, Vol. 17, No. 3, pp 337-352.

Debole, F. and Sebastiani, F., 2004. Supervised term weighting for automated text categorization. In *Text mining and its applications*. Springer Berlin Heidelberg, pp. 81-97.

Japkowicz, N. and Stephen, S., 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, Vol. 6, No. 5, pp 429-449.

Kibriya, A. M. et al, 2005. Multinomial naive Bayes for text categorization revisited. In *AI 2004: Advances in Artificial Intelligence*. Springer Berlin Heidelberg, pp. 488-499.

Dataset for single-label text categorization, http://web.ist.utl.pt/acardoso/datasets/. (25.3.2014)

Lan, M. et al, 2009. Supervised and traditional term weighting methods for automatic text categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 31, No. 4, pp 721-735.

Liu, A. Y. C., 2004. *The effect of oversampling and undersampling on classifying imbalanced text datasets* (Doctoral dissertation, The University of Texas at Austin).

Liu, Y. et al, 2009. Imbalanced text classification: A term weighting approach. *Expert systems with Applications*, Vol. 36, No. 1, pp 690-701.

Ogura, H. et al, 2011. Comparison of metrics for feature selection in imbalanced text classification. *Expert Systems with Applications*, Vol. 38, No. 5, pp 4978-4989.

Salton, G. and Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, Vol. 24, No. 5, pp 513-523.

Sun, Y. et al, 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, Vol. 40, No. 12, pp 3358-3378.

Sun, A. et al, 2009. On strategies for imbalanced text classification using SVM: A comparative study. *Decision Support Systems*, Vol. 48, No. 1, pp 191-201.

Taşcı, Ş. and Güngör, T., 2013. Comparison of text feature selection policies and using an adaptive framework. *Expert Systems with Applications*, Vol. 40, No. 12, pp 4871-4886.