A COMPARATIVE STUDY OF FEATURE SELECTION METHODS

Wanwan Zheng and Mingzhe Jin

Graduate School of Culture and Information Science,
Doshisha University, Kyoto, Japan

ABSTRACT

Text analysis has been attracting increasing attention in this data era. Selecting effective features from datasets is a particular important part in text classification studies. Feature selection excludes irrelevant features from the classification task, reduces the dimensionality of a dataset, and improves the accuracy and performance of identification. So far, so many feature selection methods have been proposed, however, it remains unclear which method is the most effective in practice. This article focuses on evaluating and comparing the available feature selection methods in general versatility regarding authorship attribution problems and tries to identify which method is the most effective. The discussions on general versatility of feature selection methods and its connection in selecting the appropriate features for varying data were done. In addition, different languages, different types of features, different systems for calculating the accuracy of SVM (support vector machine), and different criteria for determining the rank of feature selection methods were used to measure the general versatility of these methods together. The analysis results indicate the best feature selection method is different for each dataset; however, some methods can always extract useful information to discriminate the classes. The chi-square was proved to be a better method overall.

KEYWORDS

Feature Selection Methods, Effectiveness, General Versatility, Authorship Attribution

1. INTRODUCTION

A feature is a measurable property or characteristic of a phenomenon being observed. Using a set of features, classifications can be performed using machine learning algorithm. There are four main kinds of features that contain authorial impressions for authorship: lexical, character, syntactic, and semantic features (Nagaprasad et al., 2014). Among them lexical *n*-gram features are the most widely used. As far as authorship attribution approaches are concerned, previous studies have suggested plenty of other kinds of features, mean word length, mean sentence length, and vocabulary richness measures etc. However none of these features has been proven satisfactory in all cases. Mosteller and Wallace (1964) proposed a semiautomatic selection procedure to determine the most useful terms, and the most frequent ones composed mainly of various function words (determiners, preposi-tions, conjunctions, pronouns, some adverb and verbal forms). Burrows (2002) proposed considering the first 40～150 most frequent word types; and function words constituted a large proportion among these word types. In this article, word-unigrams, word-bigrams, tag-unigrams, and tag-bigrams are used as lexical features, and xuci from Chinese are used as function words. In the fields of computational linguistics and probability, an *n*-gram is a contiguous sequence of *n* items from a given sample of text or speech. An *n*-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram". In this article, word-unigrams data is referred to all the word-tokens, word-bigram data is a sequence of two adjacent

DOI: 10.5121/ijnlc.2018.7501

words shifting one word every time from the beginning. Tag is the abbreviation of part-of-speech (abbreviated as POS) tag.

Some features, such as character and lexical features can increase the dimensionality of the feature sets considerably. However, most features are irrelevant and lead to poor performance of classifiers. Therefore, the dimensionality reduction which attempts to reduce the size of feature space without sacrificing the performance of text classification, has been deemed as a critical step. In general, without informative features, it is difficult to train a model with low generalization error, however, it was reported that if relevant features can be extracted, even a simple method can show remarkable results. In such cases, feature selection is most commonly used for dimensionality reduction in the field of text classification. However, in applications, selecting the appropriate feature selection methods remains hard for a new application because so many feature selection methods are available. Yang and Pedersen (1997) evaluated document frequency thresholding (*df*), information gain (IG), mutual information (MI), chi-square (CHI), term strength (TS) five selection measures for the topical text classification in a comparative study. Their experiments indicated that IG, or CHI tends to achieve the best results, and *df* performed similarly. In Sebastiani (2002), DIA association factor (DIA), IG, MI, CHI, NGL coefficient (NGL), relevancy score (RS), odds ratio (OR) and GSS coefficient (GSS) eight methods were tested, however none of them was shown to be robust across different classification applications, and OR and CHI were the best selection functions. CHI, a distance-based FS method, has severed as a well-known feature selector in numerous studies (Moh'd Mesleh, 2011; Parlar and Ayşe özel, 2016; Liu *et al.*, 2018; Zareapoor and Seeja, 2015) and has been proved to be effective. Cui et al. (2006) conducted experiments on the sentiment classification for large-scale online product reviews to show that CHI significantly reduced the dimension of the feature vector instead of degrade the performance. Savoy (2013) argued that term frequency (*tf*) and *df* tend to have good overall performance. By far, the above experiments have only been performed in English corpora, and their effectiveness has not been evaluated in other languages. In this article, Japanese, Chinese, and English corpora are used to evaluate feature selection methods.

In addition, many sophisticated machine learning algorithms, such as support vector machine (SVM), Naïve Bayes (NB), and k-nearest neighbour (K-NN), have been extensively applied to text classification in recent years. An experimental study by [4] involving SVM, K-NN, decision trees, Rocchio, and NB, showed that SVM and K-NN performed best, while all these classifiers had similar effectiveness on categories with over 300 positive training examples each. According to Sebastiani (2002), SVM provided a sharp contrast with relatively unsophisticated and weak methods such as Rocchio. Chandrashek and Sahin (2014) used the classifier accuracy and the number of reduced features to compare the stability of feature selection techniques. And SVM was proved to be a more reliable feature selection algorithm. Further, as SVM performs classifications without feature selection, it is used to avoid double selection in this article. In addition, the leave-one-out cross-validation (LOOCV) method is chosen to obtain a higher accuracy.

This article focuses on evaluating and comparing the effectiveness and general versatility of feature selection methods in authorship attribution. The primary overview of feature selection methods is provided. Next, the effectiveness and general versatility are discussed to select useful features. Moreover, several different languages, including Japanese, Chinese, and English, several types of features (Japanese: word-unigrams, tag-unigrams, tag-bigrams; Chinese: the function words xuci, word-bigrams; English: the spam dataset) are used to measure the general versatility of these methods. Finally, SVM is used to measure the effectiveness of feature selection methods. Moreover, several systems are used to calculate the accuracy of SVM and determine the rank of feature selection methods based on four criteria. Consequently, the effectiveness and general versatility of feature selection methods are evaluated by the integrated analysis of four rankings. If one feature selection method places in the first half of the top-ranking more than three times, it can be considered that the feature selection method is effective and universally valid.

This article is organized as follows: section 2 outlines the main characteristics of the corpora used in this work; section 3 presents the selected feature selection methods; section 4 explains how to measure the effectiveness and general versatility of feature selection methods using SVM; and section 5 draws the main conclusion.

2. EVALUATION CORPORA

To measure the general versatility of feature selection methods, three corpora, which are written in Japanese, Chinese, and English and six kinds of feature sets are used. For Japanese and Chinese corpora, three authors are chosen respectively, and only novels are used. In this article, the spam dataset in R is used for the English corpus. The Japanese, Chinese, and English corpora are shown in Table 1, Table 2, and Table 3, respectively. The lexical features used in this article are word-unigrams, word-bigrams, tag-unigrams, and tag-bigrams. The xuci of Chinese are used as function words. The feature sets of word-unigrams, tag-unigrams, and tag-bigrams from Japanese corpus are extracted. Xuci and word-bigrams are extracted from Chinese corpus. For the spam dataset, fifty-eight variables (words) are used directly.

For Japanese and Chinese, all of texts are separated into words by MeCab and NLPIR with the default a dictionary installed, which are the most widely used morphological analyzer and syntactic analyzer in Japanese and Chinese. Morphological analysis and syntactic analysis by machines may involve errors, however, since all texts are processed on the same basis, the errors also equally distributed.

Table 1. Japanese corpus.

Author	Type	Number
Souseki Natume	Novel	25
Riiti Yokomitu	Novel	20
Kyouka Izumi	Novel	24

Table 2. Chinese corpus.

Author	Type	Number
Yan Mo	Novel	10
Congwen Shen	Novel	10
Ailing Zhang	Novel	10

Table 3. English corpus.

Data	Number
The spam dataset	spam:1813, nospam:2788, variables:58

3. SELECTION FUNCTION

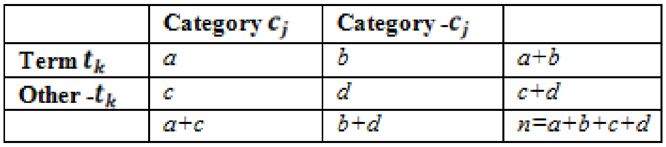
There are different independent feature-scoring functions to rank the features based on their discriminative power. In this article, twenty-two types of feature selection methods were gathered based on the previous research, which are shown in Table 4. Among these feature selection methods, eleven types are computed by R packages and eleven types are computed by numerical calculation methods.

Table 4. Twenty-two types of feature selection methods.

R packages	**Numerical Calculation Methods**
Boruta (from Boruta)	df (Document Frequency)
Relief (Relief from Fselector)	idf (Inverse Document Frequency)
CORELearn (from CORELearn)	tf (Term Frequency)
VIF (Unconditional Variable-Importance Meaturefrom party)	TF-IDF (Term Frequency-inverse Document Frequency)
VIT (Conditional Variable-Importance Meature from party)	MD (Mahalanobis's Distance)
IG (Information Gain from FSelector)	PMI (Pointwise Mutual Information)
GR (Gain radio from FSelector)	OR (Odds Radio)
SU (Symmetrical Uncertainty from FSelector)	DIA (Darmstadt Indexing Approach)
RF (MeanDecreaseGini from Randomforest)	CC (Correlation Coefficient)
oneR (oneR from FSelector)	CHI (Chi-square)
Xgboost (from Xgboost)	GSS (GSS coefficient)

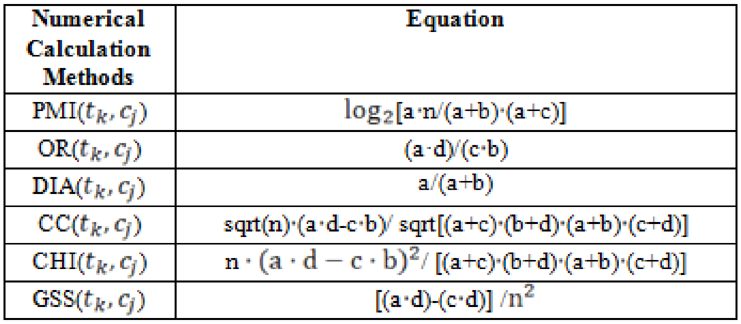
According to Savoy (2015), the capability of a term t_k can be measured based on a given category (or author) c_j, with $j = 1,2,...,C$, by using a contingency table for each pair ($t_{k,,}$ c_j) as shown in Table 5. In Table 5, the value indicates the number of texts belonging to the category c_j, in which the term t_k occurs. When considering all other classes (denoted by $- c_j$), the term t_k appears in other b texts. Thus, for the whole corpus, term t_k occurs in $a+b$ texts, while $a+c$ texts labelled with the category c_j can be counted. To measure the association between term t_k and category (or author) c_j, the numerical calculation methods are used. Some equations are shown in Table 6.

Table 5. Example of a contingency table for a term t_k and a category c_j.



	Category c_j	**Category -c_j**	
Term t_k	a	b	$a+b$
Other -t_k	c	d	$c+d$
	$a+c$	$b+d$	$n=a+b+c+d$

Table 6. Equations for the selected methods.



Numerical Calculation Methods	**Equation**
PMI(t_k, c_j)	$\log_2$[a·n/(a+b)·(a+c)]
OR(t_k, c_j)	(a·d)/(c·b)
DIA(t_k, c_j)	a/(a+b)
CC(t_k, c_j)	sqrt(n)·(a·d-c·b)/ sqrt[(a+c)·(b+d)·(a+b)·(c+d)]
CHI(t_k, c_j)	$n \cdot (a \cdot d - c \cdot b)^2$/ [(a+c)·(b+d)·(a+b)·(c+d)]
GSS(t_k, c_j)	[(a·d)-(c·d)] /n^2

However, if a, b, c, d is the number of texts, then at least two drawbacks can be considered: first, missing values are likely to be obtained, as PMI and OR, because some of a, b, c, d tend to be 0 easily in real applications. To explain the second drawback in detail, an example is given as follows. Both term t_1 and t_2 appear in two texts (a = 2); the appearance frequency of t_1 is 100, and the appearance frequency of t_2 is 1. In this case, c is surely the same. If *b* and *d* are also the same, t_1 and t_2 will obtain the same value. Thus, the importance of t_1 and t_2 will be seen as the same, and this is not optimal. In this article, the number of appearance frequency is used instead of the number of texts.

So far, a local utility value for each term t_k and category c_j have been obtained. The value can be used to make a feature ranking when associated with a binary classification problem. The ranking provides a measurement of the feature's effectiveness in discriminating the different classes. In the case of multiclass classification, more than two authors are involved. Each class is one group and the rest of the classes are one group. In this way, it becomes a binary classification problem. For example, there are A, B, and C three classes, by making two classes into one group, A (BC), B (AC), C (AB) can be obtained. According to Table 5, each term t_k can obtain three values (x, y, z). Further, the weighted mean (Equation 1 and Equation 2) is used to compute the value of term $t_{k.}$

$$\mathrm{U}_{wmean}(t_k) = \sum_{j=1}^{|c|} prob[c_j] \cdot f(t_k, c_j) \tag{1}$$

$$prob[c_j] = (a + c) / n \tag{2}$$

On the other hand, according to the definition of *df*, *tf*, *idf*, TF –IDF, it is inappropriate to use the weighted mean. The *df* is the document frequency, indicating the number of texts indexed by term $t_{k.}$ The larger the *df* result, the better the corresponding term. Further, the estimation for *tf*, *idf*, TF –IDF and MD are shown in Equations (3)～(6). In Equation (3), *tf* ($t_{k,,}$ *d*) is in article *d*, term t_k' s *tf* value. $n_{t_k,d}$ is the appearance frequency of term t_k in article *d*. $\sum_{s\in d} n_{s,d}$ is sum of the appearance frequency of all terms.

$$tf(t_k, d) = \frac{n_{t_k,}d}{\sum_{s\in d} n_{s,d}} \tag{3}$$

$$idf(t_k) = \log \frac{N}{df(t_k)} + 1 \tag{4}$$

$$TF - IDF = tf \cdot idf \tag{5}$$

$$MD(t_k) = \frac{(\overline{c_{k,j}} - \overline{-c_{k,j}})^2}{\mathrm{cov}(c_{k,j}, -c_{k,j})} \tag{6}$$

Here is an example of *tf*: the appearance frequency of term t_1 in c_1 and c_2 is 5, 1 respectively, the appearance frequency of term t_2 in c_1 and c_2 is 5, 5 respectively. Obviously, t_1 is more useful to discriminate c_1 and c_2. Therefore, for *df*, *tf*, *idf*, TF – IDF, the addition should not be used and the subtraction is optimal. In this article, based on the above analysis, Equation (7) as the weighting for *df*, *tf*, *idf*, TF –IDF is used. The weighting for MD is shown in Equation (8).

$$\mathrm{U}(t_k) = |A(t_k)-B(t_k)|+|A(t_k)-C(t_k)|+|B(t_k)-C(t_k)| \tag{7}$$

$$MD(t_k) = MD_{A(BC)}(t_k) + MD_{B(AC)}(t_k) + MD_{C(AB)}(t_k) \tag{8}$$

4. Measure Effectiveness and General Versatility of Feature Selection Methods Using SVM

SVM is a marginal classifier that maximizes the margin between the data samples of two classes. Three steps are performed in this work:

(1) Apply the feature selection methods shown in Table 4 to make feature rankings. Based on the rankings, all the features are rearranged. The most useful feature is shown in the first column and the second useful feature is in the second column. By using this method, new datasets are made.

(2) Based on the new datasets, increasing one by one from two features to compute each of the accuracy of SVM.

(3) To determine the effectiveness and general versatility of feature selection methods, four tables are made. The first is the ranking of the accuracy of SVM, when the same number of features for each selection method is taken (Table 7); the second is the ranking of the accuracy of SVM, when the best number of features is taken (Table 8). The best number of features is determined by where the highest and the most stable value of the accuracy of SVM is taken; the third is the mean accuracy of SVM (Table 9); and the last is the maximum value of the accuracy of SVM (Table 10).

Table 7. Ranking of the accuracies of SVM when taking the same number of features.
(tag-unigram of Japanese)

Boruta	**VIF**	**COREIearn**	**MD**	**Relief**	**VIT**	**IG**	**RF**	**SU**	**oneR**	**GR**
0.94	0.88	0.87	0.87	0.86	0.84	0.81	0.81	0.80	0.80	0.80
CHI	**Xgboost**	**CC**	***tf***	**PMI**	**OR**	***tf-idf***	**DIA**	**GSS**	***idf***	***df***
0.78	0.78	0.78	0.75	0.70	0.68	0.68	0.64	0.62	0.57	0.55

Table 8. Ranking of the accuracies of SVM when taking the best number of features.
(tag-unigram of Japanese)

Boruta	**Xgboost**	***df***	**VIF**	**GR**	**IG**	**SU**	**RF**	**Relief**	**COREIearn**	**VIT**
0.96	0.94	0.93	0.88	0.87	0.86	0.86	0.86	0.84	0.84	0.84
GSS	**MD**	**oneR**	**OR**	**CHI**	**CC**	***tf***	**PMI**	**DIA**	***idf***	***tf-idf***
0.84	0.84	0.83	0.83	0.81	0.81	0.81	0.81	0.80	0.80	0.80

Table 9. Ranking of the mean accuracies of SVM (tag-unigram of Japanese).

Boruta	**Xgboost**	**RF**	**VIF**	**IG**	**MD**	**VIT**	**SU**	**Relief**	**COREIearn**	**CHI**
0.93	0.92	0.82	0.81	0.81	0.81	0.81	0.80	0.80	0.80	0.80
GR	***tf***	**CC**	**oneR**	***df***	**OR**	**PMI**	**GSS**	**DIA**	***tf-idf***	***idf***
0.80	0.79	0.79	0.78	0.78	0.76	0.75	0.74	0.72	0.70	0.69

Table 10. Ranking of the maximum accuracies of SVM (tag-unigram of Japanese).

Xgboost	**Boruta**	***df***	**COREIearn**	**VIF**	**CHI**	**RF**	**GR**	**MD**	**Relief**	**IG**
0.97	0.96	0.93	0.90	0.88	0.87	0.87	0.87	0.87	0.86	0.86
SU	**VIT**	**OR**	**PMI**	**GSS**	**CC**	***tf***	**oneR**	**DIA**	***idf***	***tf-idf***
0.86	0.86	0.86	0.86	0.84	0.84	0.84	0.83	0.80	0.80	0.80

How is the best number of features decided? Using step (2), Figure 1 is obtained. Figure 1 is the result of Boruta using Japanese tag-unigrams and is shown as an example. Refer to Figure 1, the highest and the most stable value of the accuracies of SVM is taken as the best number. Thus, the best number for Boruta is 20. About Table 8, the minimum value in Table 11 (VIF: 9) is defined as the same number of features. While Table 9 and Table 10 can be built by step (2).

According to Tables 7~10, for tag-unigrams of Japanese, Boruta, CORELearn, GR, IG, MD, Relief, RF, SU, VIF, VIT, CHI, and Xgboost have placed in the first half of top-ranking more than three times, thus they are considered to be good feature selection methods. The results are shown in Table 12.

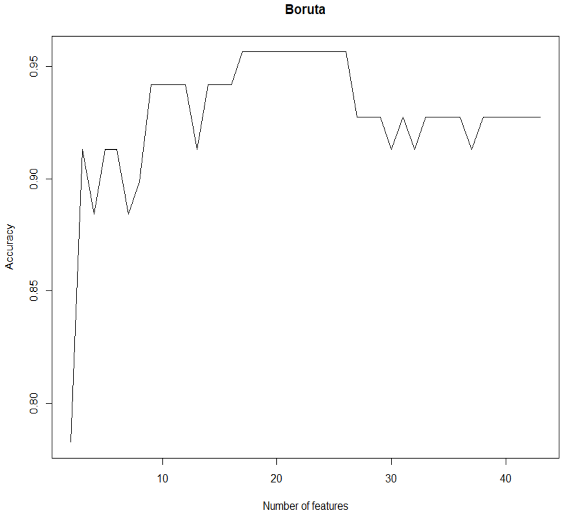


Figure 1. Result of Boruta using Japanese tag-unigrams.

Table 11. The best number of features for tag-unigrams of Japanese.

Xgboost	**Boruta**	**df**	**COREleam**	**VIF**	**CHI**	**RF**	**GR**	**MD**	**Relief**	**IG**
20	20	40	14	9	18	17	13	13	23	14
SU	**VIT**	**OR**	**PMI**	**GSS**	**CC**	**tf**	**oneR**	**DIA**	**idf**	**tf-idf**
12	10	33	27	30	20	27	34	42	40	39

Table 12. The best number of features for tag-unigrams of Japanese.

3 classes_tag-unigrams_Japanese : Boruta, CORELearn, GR, IG, MD, Relief, RF, SU, VIF, VIT, CHI, Xgboost
3 classes_word-unigrams_Japanese : Boruta, COREIearn, *df*, *idf*, PMI, Relief, *tf*, *tf-idf*, CHI
3 classes_tag-bigrams_Japanese : Boruta, CC, COREIearn, IG, MD, oneR, Relief, RF, SU, *tf*, VIF, VIT, CHI, Xgboost
3 classes_word-bigrams_Chinese : Boruta, CC, COREIearn, *df*, GR, IG, MD, oneR, RF, SU, *tf*, VIF, CHI
3 classes_Xuci_Chinese : Boruta, CC, COREIearn, DIA, GR, RF, SU, *tf,* VIF, CHI, Xgboost

For three classes, five types of features which were extracted from Japanese and Chinese are tested, and then five sets of the effective feature selection methods are obtained. Therefore, the results of the first half of the top rankings are summarized. Boruta, COREIearn, and CHI appear five times. RF, SU, *tf*, and VIF appear four times. CC, GR, IG, MD, Relief, and Xgboost appear three times.

English corpus spam in R is a two-class dataset. Because the spam dataset has 4,601 rows, 100-fold cross-validation is conducted instead of the time-consuming LOOCV. In this way, each has forty-six rows and is used for testing one time, all the rows can be predicted. One problem is how to extract the forty-six rows. There are two methods:

(1) Random extraction, in which the number of spam or no-spam that is used for testing cannot be predicted.

(2) Twenty-three are spam while twenty-three are no-spam.

In this article, (1) is used to build the ranking of twenty-two feature selection methods when taking the same number of features and the ranking when taking the best number of features; (2) is used to build the ranking by the mean accuracy of SVM and the ranking by the maximum accuracy of SVM. In this way, different methods are used to compute the accuracies of SVM. The result is shown in Table 13.

In Table 12 and Table 13, six kinds of features are tested, by using three languages. Based on an integrated consideration of the results in Table 12 and Table 13, in the first half of the top-ranking, CHI appears six times; Boruta, CORElearn, SU and VIF appear five times, however Boruta and CORElearn are not chosen by English; and RF, *tf*, GR, Relief, and Xgboost appear four times, however RF and *tf* fail to show their superiority in English. CHI tends to provide good overall performance. On the second-class performance level, SU and VIF are placed at a high rank. Although GR, Relief, and Xgboost are not as good as CHI, SU, and VIF, they are still worth considering to be used in authorship attribution.

In the same way, the ineffective feature selection methods are identified. DIA is considered as the worst function with a comparably large best number of features and has lower positive discrimination rate.

Table 13. Effective feature selection methods for two classes.

2 classes_word_English :
Xgboost, VIF, SU, GR, Relief, TF-IDF, oneR, *idf*, CHI

5. CONCLUSIONS

This article attempted to find effective and universally valid feature selection methods. The analysis results indicate the best feature selection method is different for each dataset; however, there are some methods through which we can always extract useful features to discriminate different classes. CHI tends to provide good overall performance. And SU and VIF are placed on the second-class level of performance. Although GR, Relief and Xgboost are not as good as CHI, SU and VIF, they are still worthy of consideration in authorship attribution. Boruta also seems to be a good method. Further, it was found that VIF (mean decrease accuracy) has always been better than RF (mean decrease gini), although both of them are included in the RandomForest feature selection algorithm, and RF is the default setting.

The above conclusions are similar to Savoy (2013). However, in our analysis, *tf* and *df* also showed some promising results. *df* performed well in Japanese's word-unigrams and Chinese's word-bigrams, and *tf* performed well in word-unigrams and tag-bigrams of Japanese and word-bigrams and xuci of Chinese in our experiment. However, for Japanese and Chinese, even *df* and *tf* entered the top eleven, their rankings were not very high. For the spam dataset, *df* and *tf* were not good feature selection methods. Therefore it can be said that although *tf* and *df* sometimes perform well in the feature selection for texts, in a sense, they are not always effective. Finally the better effectiveness of CHI, SU, VIF are verified through this work.

REFERENCES

[1] Burrows, J. (2002) "Delta: a measure of stylistic difference and a guide to likely authorship", *Literary and Linguistic Computing*, 17(3), pp267-287.

[2] Chandrashek, G., Sahin, F. (2014) "A survey on feature selection methods", *Computers and Electrical Engineering* 40(1), pp16-28.

[3] Cui, H., Mittal, V., & Datar, M. (2006) "Comparative experiments on sentiment classifica-tion for online product reviews", *AAAI'06 Proceedings of the 21st National Conference on Artificial Intelligence*, pp1-11.

[4] Joachimas, T. (1998) "Text categorization with support vector machines: learning with many relevant features". *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pp137-142.

[5] Mosteller, F. & Wallace, D. L. (1964) "Applied Bayesian and Classical Inference: The Case of the Federalist Papers", Spring-Verlag New York, New York.

[6] Nagaprasad, S., Raghunadha, T., Vijayaral, P., Vinaya, A. B., &Vishnu, B. V. (2014) "Infuence of machine learning techniques on authorship attribution for Telugu text features", *Intionational Journal of Advanced Research in Computer Engineering & Technology* 3(11), pp3633-3640.

[7] Sebastiani, F. (2002) "Machine learning in automated text categorization", *ACM Computing Surveys* (CSUR) 34(1), pp1-47.

[8] Savoy, J. (2013) "Comparative evaluation of term selection functions authorship attribution", *Literary and Linguistic Computing* 30(2), pp246-261.

[9] Yang, Y. & Pedersen, J. O. (1997) "A comparative study on feature selection in text catego-rization", *Processing of the 14th International Conference on Machine Learning* (ICML), pp412-420.

[10] Moh'd Mesleh, A. (2011). "Feature sub-set selection metrics for Arabic text classification", *Pattern Recognition Letters*, 32, pp1922-1929.

[11] Parlar, T. & Ayşe özel, S. (2016). "A new feature selection method for sentiment analysis of Turkish reviews". 2016 International Symposium on Innovations in Intelligent Systems and Applications.

[12] Liu, H. Y., Zhou, M. C., Lu, X. S., & Yao, C. (2018). "Weighted Gini index feature selection method for imbalanced data". IEEE 15th International Conference on Networking, Sensing and Control.

[13] Zareapoor, M. & Seeja, K. R. (2015). "Feature extraction or feature selection for text classification: a case study on phishing Email detection". I. J. Information Engineering and Electronic Business, pp. 60-65.