Comparative Performance Analysis of Techniques for Automatic Drug Review Classification

Alper Kürşat Uysal[*]

Eskisehir Technical University, Faculty of Engineering, Department of Computer Engineering, Eskisehir, Turkey.
*akuysal@anadolu.edu.tr

Abstract

This study analyses the effectiveness of six text feature selection methods for automatic classification of drug reviews written in English using two different widely-known classifiers namely Support Vector Machines (SVM) and naïve Bayes (NB). In the study, a recently published public dataset namely Druglib including drug reviews in English was utilized in the experiments. For evaluation, Micro-F1 and Macro-F1 success measures were used. Also, 3-fold cross-validation is preferred to perform a fair evaluation. The feature selection methods used in the study are Distinguishing Feature Selector (DFS), Information Gain (IG), chi-square (CHI2), Discriminative Features Selection (DFSS), Improved Comprehensive Measurement Feature Selection (ICMFS), and Relative Discrimination Criterion (RDC). However, experiments were performed using two settings in which stemming was applied and not applied. Experiments indicated that ICMFS feature selection method is generally superior to the other feature selection methods according to the overall highest Micro-F1 and Macro-F1 scores achieved on drug reviews. While the highest Micro-F1 score was achieved with the combination of NB classifier and ICMFS feature selection method, the highest Macro-F1 score was achieved with the combination of NB classifier and DFSS feature selection method. The highest Micro-F1 and Macro-F1 scores were achieved for the cases that stemming algorithm was not applied.

Keywords: Pattern recognition, text classification, drug reviews, feature selection.

1. Introduction

Sentiment analysis has become popular due to the increase in online social platforms which people interact with. Sentiment analysis aims to automatically assign opinions of people into predefined categories [1]. It is possible to state that these categories are positive and negative in most cases. Opinions of people are generally stored as electronic text documents and it is necessary to apply text processing methods in order to construct automatic classification frameworks. Sentiment analysis has some application domains such as classification of movie reviews [2], restaurant reviews [3], product reviews [4] and drug reviews [5]. This study specifically focuses on classification of drug reviews and a literature review about only this topic is given in the following paragraph.

Na et al. proposed a linguistic rule-based approach for classification of drug reviews [6]. In this approach, a general lexicon including 9630 terms and a domain-specific lexicon including 10 terms about medical keywords are constructed. A score indicating sentiment level is assigned to each term and a rule-based approach using these scores is constituted. The authors stated that the proposed approach is an effective solution. Cavalcanti and Prudencio proposed a method focused on target based sentiment analysis on drug reviews [7]. In this method, a supervised learning approach depending on linguistic resources and application domain is used. They aimed to automatically detect different targets such as dosage and to assign these targets into pre-defined classes. Gopalakrishnan and Ramaswamy proposed a neural-network based method for classification of data obtained from social platforms related to healthcare [8]. In order to perform the study, they compiled reviews belonging to two different drugs from the Internet. According to the findings of the study, they stated that neural network based approach gives better performance than the one based on Support Vector Machines. Gräßer et al. performed a study which automatically detects different targets in drug reviews [5]. The study aims to determine the satisfaction level of the patients using the reviews obtained from the websites containing pharmacology reviews. Logistic regression was used to determine the sentiment levels in the comments. They performed the experiments with two different settings where training data is compiled using data from the same domain and from different domains. The performance of logistic regression in automatic classification of pharmacology reviews was investigated for these two different settings.

Feature selection is an active research field in the text classification. The impact of feature selection methods on classifier performance for drug review classification has not been investigated yet according to the literature.

The contributions of the paper can be summarized as follows. This study aims to examine the effects of text feature selection methods on the performance of the classifiers using a recently published public dataset containing drug reviews and detect the most successful frameworks for drug review classification. Therefore, in this study, the effectiveness of 6 feature selection methods in the literature has been investigated.

This paper is organized as follows. In Section 2, feature selection methods and classification algorithms are described. Then, the experimental results are presented in Section 3. In Section 4, some concluding remarks are given.

2. Materials and Methods
2.1 Feature Selection Methods

Filter-based feature selection methods are mostly used for text classification due to their speed [9]. They work faster because they do not need to interact with classifiers during feature selection. In this study, 6 filter-based feature selection methods were used. These methods are Distinguishing Feature Selector (DFS), Information Gain (IG), Chi-square (CHI2), Discriminative Features Selection (DFSS), Improved Comprehensive Measurement Feature Selection (ICMFS), and Relative Discrimination Criterion (RDC).

2.1.1 Distinguishing Feature Selector (DFS)

The formula of DFS was constructed with 4 pre-determined constraints [9]. The DFS score for a feature is calculated using the following equation.

$$DFS(t) = \sum_{i=1}^{M} \frac{P(C_i | t)}{P(\bar{t} | C_i) + P(t | \bar{C_i}) + 1} \qquad (2.1)$$

In (2.1), the conditional probability $P(C_i | t)$ refers to the probability of the existence of the class C_i in the presence of the term t. M refers to the number of classes in the dataset. The conditional probability $P(\bar{t} | C_i)$ refers to the probability of the absence of the term t in the case where the class C_i exists. The conditional probability $P(t | \bar{C_i})$ refers to the existence of the term t in the presence of other classes except class C_i.

2.1.2 Information Gain (IG)

IG method is used to analyze the contribution of the presence and absence of a term to the correct classification [10]. This method can be expressed using the following equation. A high score produced by IG method for a term means that the term is very discriminative for classification.

$$IG(t) = -\sum_{i=1}^{M} \frac{P(C_i) \log P(C_i)}{P(t)} + \sum_{i=1}^{M} P(C_i | t) \log P(C_i | t) + P(\bar{t}) \sum_{i=1}^{M} P(C_i | \bar{t}) \log P(C_i | \bar{t}), \qquad (2.2)$$

In (2.2), M represents the number of classes in the dataset. $P(C_i)$ refers to the probability of class C_i in entire dataset. $P(t)$ and $P(\bar{t})$ denote the probability of the existence and absence of the term t in the dataset. The expressions $P(C_i | t)$ and $P(C_i | \bar{t})$ denote the probability of the existence of class C_i for the cases where term t is present and not present, respectively.

2.1.3 Chi-square (CHI2)

CHI2 is a statistical method which is used to investigate whether different events are independent of each other or not. When Chi-square method is adapted as a feature selection approach, two different events are the existence of features and the existence of classes. CHI2 score for a feature can be calculated using the following equation.

$$CHI2(t, C) = \sum_{t \in \{0,1\}} \sum_{C \in \{0,1\}} \frac{(N_{t,c} - E_{t,c})^2}{E_{t,C}}, \qquad (2.3)$$

In (2.3), the value N indicates the observed frequencies and the value E indicates the expected frequencies for features and classes, respectively [9]. With this equation, a local class-based score is obtained for each feature. The maximum function is used to transform the class-based scores produced by this method into a single value.

2.1.4 Discriminative Features Selection (DFSS)

DFSS is a recent feature selection method that aims to construct a feature set consisting of features with high document frequency and high average term frequency within the documents of a class [11]. DFSS score for a feature can be calculated using the following equation.

$$DFSS(t, C) = \frac{\text{tf}(t, C) / \text{df}(t, C)}{\text{tf}(t, \bar{C}) / \text{df}(t, \bar{C})} \times \frac{a}{(a + b)} \times \frac{a_i}{(a + c)} \times \left| \frac{a}{(a + b)} - \frac{c}{(c + d)} \right| \qquad (2.4)$$

In (2.4), a represents the amount of documents in class C including the term t. b represents the amount of documents in class C not including the term t. c represents the amount of documents in classes except the class C including the term t. d represents the amount of documents in classes except the class C not including the term t. The values $\text{tf}(t, C)$ and $\text{tf}(t, \bar{C})$ refer to the frequency of the term in the class C and in the other classes, respectively. $\text{df}(t, \bar{C})$ represents the number of documents in classes except class C containing the term

t. df(t,C) indicates the number of documents in the class C containing the term t. The maximum function is used to transform the class-based scores produced by this method into a single value.

2.1.5 Improved Comprehensive Measurement Feature Selection (ICMFS)

ICMFS is an extended version of an existing method in the literature namely comprehensively measure feature selection [12]. The ICMFS score for a feature is calculated by the following equation.

$$\text{ICMFS}(t,\text{C}) = \frac{P(t\,|\,\text{C}) \times P(\text{C}\,|\,t)}{P(\text{C}) \times \max_{\text{C}_j \neq \text{C}} \left\{ \dfrac{P(t\,|\,\text{C}_j) + \Delta}{P_{avg} + \Delta} \right\}} \qquad (2.5)$$

In (2.5), Δ is a constant value, and the researchers suggesting this method used this value as 0.001. The P_{avg} value is the mean of $P(t\,|\,\text{C}_j)$ values for the classes except the class C_j. After calculating the ICMFS scores for each class, the maximum function is used to transform the class-based scores produced by this method into a single value.

2.1.6 Relative Discrimination Criterion (RDC)

RDC is a feature selection method that is implemented using an algorithm rather than a single formula [13]. The flow of the RDC feature selection method is as below.

POS = The amount of documents in positive class for two-class problem
NEG = The amount of documents in negative class for two-class problem
TCMAX = The maximum term count for term t
tp_tc = the number of documents that contain the term t with a term count tc from the documents of the positive class
fp_tc = the number of documents that contain the term t with a term count tc from the documents of the negative class
for tc = 1 to TCMAX do
$\qquad tpr_{tc} = tp_{tc}\,/\,POS$
$\qquad fpr_{tc} = fp_{tc}\,/\,NEG$
$\qquad D_{tc} = |\,tpr_{tc}\text{ - }fpr_{tc}\,|$
$\qquad RDC_{tc} = \dfrac{D_{tc}}{\min(tpr_{tc}, fpr_{tc})*tc}$
end
$AUC_{tc} = 0$
for tc = 1 to TCMAX do
$\qquad AUC_{tc} = AUC_{tc} + \dfrac{RDC_{tc} + RDC_{tc+1}}{2}$
end

The RDC algorithm produces class-based scores for each term. These class-based scores are globalized using weighted average of the class probabilities and a single score is obtained for each term.

2.2 Classification Algorithms

In this study, two classification algorithms which are known as successful in text classification problems are used. These algorithms are Support Vector Machines and naive Bayes. These classifiers are briefly described in the following subsections.

2.2.1 Support Vector Machines (SVM)

The philosophy of the SVM classifier is to maximize the margin on the classification hyperplane [14]. It aims to find a decision hyperplane whose distance is far away from data points which are members of two classes. In this study, linear Support Vector Machines, which is used frequently in text classification studies and known to be successful, has been used. In the experimental study, LIBSVM classification toolbox [15] was used. The linear Support Vector Machines was applied with the standard parameters of the LIBSVM toolbox.

2.2.2 Naïve Bayes (NB)

NB is a classification algorithm based on Bayes's theorem. The NB algorithm basically works by assuming that the features are independent of each other and that there is no correlation between them. Therefore, some probabilities are multiplied during the calculations. In text classification, multi-variate Bernoulli and multinomial event models are used rather than Gaussian event model during the implementation of naïve Bayes algorithm [16]. In this study, NB algorithm is implemented in Matlab environment and multivariate Bernoulli event model is used.

3. Results and Discussion

In this section, a comprehensive performance analysis of 6 feature selection methods was performed for drug review classification problem. In this analysis, experiments were also performed for the cases that stemming was used and not used. Two classifiers were used for performance analysis. In the experiments, term frequency - inverse document frequency (TF-IDF) method was applied as term weighting method.

3.1 Dataset

In the study, a publicly available dataset containing drug reviews, was utilized [5]. Experiments were conducted using positive and negative reviews in the dataset namely Druglib. The reviews in the study were scored by people on a scale of 1 to 10, and those with a general assessment score of 1-4 were labeled as negative comments, and those with a general assessment score of 7-10 were labeled as positive comments. In the experiments, the contents of three different fields including benefits, side effects and general comments, were combined. Information about the dataset is given in Table 1.

Table 1. Druglib dataset.

Class	Number of documents
positive	2800
negative	902

3.2 Success Measures

Micro-F1 and Macro-F1 success measures were used in this study. While the Micro-F1 score takes into account the classification performance in the entire dataset, the Macro-F1 score considers individual classification performances of the classes. In the case that the number of documents contained in the classes is not balanced and the success ratio of the large class is higher, considering only the Micro-F1 score may lead to neglect of the failure in the small class.

3.3 Accuracy Analysis

In this section, classification performances were analyzed for 6 feature selection methods. Features selected by feature selection methods were sent as input to classifiers namely SVM and NB. The feature dimensions used in the experiments vary between 100 and 2000. Lowercase conversion and stemming are two different steps applied in the experiments besides TF-IDF term weighting method. A word list in English language was used while applying stop-word removal. The experiments were carried out for two different cases in which the Porter stemming algorithm [17] was used and not used. 3-fold cross validation method was applied in order to make a fair evaluation. 8119 and 11486 features were detected in the cases where Porter stemming algorithm was used and not used, respectively. Table 2 represents the results when the Porter stemming algorithm is used. However, Table 3 represents the results when the Porter stemming algorithm is not used. The highest scores in these tables for specific categories are shown in bold.

Table 2. The performance of SVM classifier when Porter stemming algorithm is used.

Feature size	Micro-F1 score						Macro-F1 score				
	100	300	500	1000	2000		100	300	500	1000	2000
DFS	75.90	75.71	73.63	72.41	71.63		44.76	62.60	63.09	62.92	61.12
IG	79.74	79.50	78.36	76.12	76.96		66.53	69.50	69.32	67.07	66.96
CHI2	79.82	**80.09**	78.63	75.87	76.12		66.70	70.45	69.67	67.10	67.57
DFSS	80.04	79.31	78.63	75.41	75.09		66.26	70.48	**70.52**	68.16	67.23
ICMFS	79.61	79.71	77.85	74.74	75.90		65.82	70.41	69.47	66.76	67.81
RDC	75.69	76.42	77.25	74.98	75.82		56.97	62.05	66.68	66.25	67.51

Table 3. The performance of NB classifier when Porter stemming algorithm is used.

Feature size	Micro-F1 score						Macro-F1 score				
	100	300	500	1000	2000		100	300	500	1000	2000
DFS	75.06	76.44	76.98	77.06	77.36		48.89	54.69	56.91	57.68	57.26
IG	78.69	79.90	80.04	80.04	80.20		63.80	66.81	67.50	67.87	67.02
CHI2	78.60	79.93	80.28	80.28	80.20		64.66	67.05	67.98	68.48	68.08
DFSS	77.96	79.63	79.36	79.55	80.17		59.46	65.09	65.37	66.77	68.93
ICMFS	77.50	79.50	**80.71**	80.15	80.63		59.58	65.77	68.32	68.48	**69.31**
RDC	75.44	76.14	78.69	80.15	80.31		60.05	64.31	67.23	69.22	69.01

In the case where SVM classifier was used in accordance with Table 2, the best Micro-F1 score was achieved by CHI2 feature selection method and the best Macro-F1 score was achieved by DFSS feature selection method. In the case where NB classifier was used in accordance with Table 3, both the best Micro-F1 score and the best Macro-F1 score were achieved by the ICMFS feature selection method. The best Micro-F1 score is achieved by the NB classifier and the best Macro-F1 score is achieved by SVM classifier.

In case where no stemming algorithm is used, the results achieved on the dataset are given in Table 4 and Table 5. The highest scores in these tables for specific categories are shown in bold.

Table 4. The performance of SVM classifier when a stemming algorithm is not used.

Feature size	Micro-F1 score						Macro-F1 score				
	100	300	500	1000	2000		100	300	500	1000	2000
DFS	76.06	74.71	73.17	71.20	70.74		45.96	61.96	63.06	61.65	62.04
IG	79.63	79.25	78.31	76.28	77.06		66.30	69.70	69.35	66.98	67.60
CHI2	**79.80**	78.61	77.55	76.09	76.58		66.86	68.18	68.32	67.26	67.76
DFSS	79.09	79.12	78.06	75.04	76.01		63.33	69.32	69.75	67.50	67.72
ICMFS	79.47	79.52	78.23	74.68	76.04		64.38	**69.91**	69.75	66.47	67.30
RDC	75.50	76.69	76.28	74.17	76.50		57.93	63.03	64.95	64.63	68.17

Table 5. The performance of NB classifier when a stemming algorithm is not used.

Feature size	Micro-F1 score						Macro-F1 score				
	100	300	500	1000	2000		100	300	500	1000	2000
DFS	75.28	76.96	77.20	77.09	77.55		49.98	56.92	58.15	58.24	59.74
IG	78.58	79.93	80.66	80.58	80.55		64.75	67.77	68.90	69.28	68.87
CHI2	78.52	79.96	80.85	80.96	80.88		65.55	67.97	69.45	69.98	70.07
DFSS	77.39	79.58	79.47	80.23	81.31		58.61	65.17	65.85	68.30	**71.04**
ICMFS	78.20	79.66	79.74	**81.44**	81.09		61.24	66.52	67.79	71.01	70.66
RDC	75.06	76.66	78.31	79.74	81.01		61.10	65.90	68.37	69.11	70.90

In the case where SVM classifier was applied in accordance with Table 4, the maximum Micro-F1 score was achieved by CHI2 feature selection method and the maximum Macro-F1 score was achieved by ICMFS feature selection method. In the case where NB classifier is used in accordance with Table 5, the maximum Micro-F1 score was achieved by ICMFS feature selection method and the maximum Macro-F1 score was achieved by DFSS feature selection method. Both the maximum Micro-F1 and the maximum Macro-F1 score were achieved by the NB classifier.

When two different classifier performances are considered, it can be said that the ICMFS feature selection method is more successful in automatic classification of drug reviews in general. While ICMFS feature selection method performed better in 4 out of 8 cases, CHI2 feature selection method performed better in 2 out of 8 cases and DFSS feature selection method performed better in 2 out of 8 cases in accordance with the highest performances. The maximum Micro-F1 score was achieved by the combination of the NB classifier and the ICMFS feature selection method in the case where stemming was not used. The maximum Macro-F1 score was achieved by the combination of the NB classifier and the DFSS feature selection method in the case where stemming was not used.

The most discriminative 10 features that feature selection methods selected are listed in Table 6 and Table 7. The features, put into the top-10 list by only one of the selection methods, are specified in bold. It can be stated that the features selected by RDC feature selection method differs from the ones selected by the other methods in general. In addition, the content of the identified features for the cases where stemming is used and not used are similar to each other.

Table 6. Top-10 features when Porter stemming algorithm is used.

Method	Features									
	1	2	3	4	5	6	7	8	9	10
DFS	effect	daili	morn	work	tablet	mg	**help**	year	**need**	side
IG	rash	sever	horribl	tablet	daili	told	discontinu	morn	vomit	pharmacist
CHI2	rash	sever	horribl	discontinu	told	vomit	pharmacist	daili	tablet	**suppos**
DFSS	effect	side	mg	work	year	day	daili	morn	drug	sever
ICMFS	effect	side	drug	day	sever	**take**	**time**	**pain**	**stop**	**doctor**
RDC	**greatli**	**strep**	**hive**	**terribl**	**cure**	**lump**	**poison**	**bc**	rash	**moistur**

Table 7. Top-10 features when a stemming algorithm is not used.

Method	Features									
	1	2	3	4	5	6	7	8	9	10
DFS	morning	daily	mg	effects	years	**worked**	**tablet**	**needed**	side	**reduced**
IG	severe	rash	reaction	horrible	morning	daily	told	vomiting	pharmacist	discontinued
CHI2	rash	severe	reaction	horrible	told	vomiting	pharmacist	discontinued	morning	daily
DFSS	effects	side	mg	day	daily	morning	years	severe	drug	**medication**
ICMFS	side	effects	day	drug	**taking**	severe	mg	**pain**	**doctor**	**time**
RDC	**greatly**	rash	**strep**	**hives**	pharmacist	**poison**	**bc**	**fda**	**lump**	**shock**

4. Conclusion

In the study, the effectiveness of 6 text feature selection methods for automatic classification of drug reviews written in English were analyzed and the most successful classification schemes were determined. Two classification algorithms were used to investigate the success ratios of the feature selection methods. Experiments on the Druglib dataset were performed for the cases where the stemming was used and not used. Two different success measures were used in the experiments. According to the results of the experiments, the ICMFS feature selection method seems more successful than the others in automatic classification of drug reviews. The reason behind the success of ICMFS feature selection method may be the imbalance structure of the drug reviews. The implementation of ICMFS method specifically considers the effects of category sizes. The best Micro-F1 score was achieved by the combination of the NB classifier and the ICMFS feature selection method in the case where the stemming algorithm was not used. The best Macro-F1 score was achieved by the combination of the NB classifier and the DFSS feature selection method in the case where the stemming algorithm was not used. The stemming may not have a good effect on the classification performances as the drug reviews can contain some medical terms in addition to the terms used in daily language. As a future work, the performances of automatic classification schemes for drug reviews written in Turkish can be investigating by compiling and annotating reviews from web forums.

References

1. Uysal, A. K., Murphey, Y. L. Sentiment classification: Feature selection based approaches versus deep learning, proceedings of *17th IEEE International Conference on Computer and Information Technology (CIT)*, 2017, pp. 23-30.

2. Pang, B., Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, proceedings of *the 42nd annual meeting on Association for Computational Linguistics*, 2004, pp. 1-8: Association for Computational Linguistics.

3. Gan, Q., Ferns, B. H., Yu, Y., Jin, L., A Text Mining and Multidimensional Sentiment Analysis of Online Restaurant Reviews, *Journal of Quality Assurance in Hospitality & Tourism*, 2017, 18(4), 465-492.

4. Gui, L., Zhou, Y., Xu, R., He, Y., Lu, Q., Learning representations from heterogeneous network for sentiment classification of product reviews, *Knowledge-Based Systems*, 2017, 124, 34-45.

5. Gräßer, F., Kallumadi, S., Malberg, H., Zaunseder, S. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning, proceedings of *2018 International Conference on Digital Health*, 2018, pp. 121-125: ACM.

6. Na, J.-C., Kyaing, W. Y. M., Khoo, C. S. G., Foo, S., Chang, Y.-K., Theng, Y.-L. Sentiment Classification of Drug Reviews Using a Rule-Based Linguistic Approach, proceedings of *The Outreach of Digital Libraries: A Globalized Resource Network*, Berlin, Heidelberg, 2012, pp. 189-198: Springer Berlin Heidelberg.

7. Cavalcanti, D., Prudencio, R. Aspect-Based Opinion Mining in Drug Reviews, proceedings of *Portuguese Conference on Artificial Intelligence*, 2017, pp. 815-827: Springer.

8. Gopalakrishnan, V., Ramaswamy, C., Patient opinion mining to analyze drugs satisfaction using supervised learning, *Journal of Applied Research and Technology*, 2017, 15(4), 311-319.

9. Uysal, A. K., Gunal, S., A novel probabilistic feature selection method for text classification, *Knowledge-Based Systems*, 2012, 36, 226-235.

10. Forman, G., An extensive empirical study of feature selection metrics for text classification, *Journal of Machine Learning Research*, 2003, 3, 1289-1305.

11. Zong, W., Wu, F., Chu, L.-K., Sculli, D., A discriminative and semantic feature selection method for text categorization, *International Journal of Production Economics*, 2015, 165, 215-222.

12. Feng, L., Zuo, W., Wang, Y. Improved comprehensive measurement feature selection method for text categorization, proceedings of *Network and Information Systems for Computers (ICNISC), 2015 International Conference on*, 2015, pp. 125-128.

13. Rehman, A., Javed, K., Babri, H. A., Saeed, M., Relative discrimination criterion – A novel feature ranking method for text data, *Expert Systems with Applications*, 2015, 42(7), 3670-3681.

14. Joachims, T. Text categorization with support vector machines: Learning with many relevant features, proceedings of *10th European Conference on Machine Learning*, Chemnitz, Germany, 1998, vol. 1398, pp. 137-142.

15. Chang, C.-C., Lin, C.-J., LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, 2011, 2(3), 1-27.

16. Jiang, L., Cai, Z., Zhang, H., Wang, D., Naive Bayes text classifiers: A locally weighted learning approach, *Journal of Experimental & Theoretical Artificial Intelligence*, 2013, 25(2), 273-286.

17. Porter, M. F., An algorithm for suffix stripping, *Program*, 1980, 14(3), 130-137.