Mühendislik Bilimleri ve Tasarım Dergisi
9(4), 1398 – 1047, 2021
e-ISSN: 1308-6693

Journal of Engineering Sciences and Design
DOI: 10.21923/jesd.906211

Araştırma Makalesi Research Article

CLASSIFICATION OF USER COMMENTS IN A MOBILE APPLICATION USING DATA AUGMENTATION WITH MACHINE LEARNING TECHNIQUES

Özer ÇELİK, Gürkan KAPLAN[*]

Eskişehir Osmangazi University, Faculty of Science and Letter, Department of Mathematics and Computer Science, Eskişehir, Turkey

Keywords

Text Classification, Machine Learning, Artificial Intelligence, Natural Language Processing.

Abstract

With the increasing use of social media in recent years, there are too many comments to be followed on almost every issue. These comments contain both important and unimportant information. But, it is almost impossible to follow of so many comments nowadays. In this study, text classification of user comments made to the Anadolu University mobile application was made. It was estimated whether the comments made on the application were related to the content or the application. In addition, the effect of oversampling and undersampling on text classification performance was investigated. For this purpose, synthetic minority oversampling technique (Smote), condensed nearest neighbor undersampling technique (CNN) and random undersampling (RUS) technique were applied to the data set. 1008 user comments received from mobile application were classified by these techniques. In the Smote oversampling classification, ANN algorithm was found to have the best classification with 93.57% accuracy. In the undersampling classification, Random Forest algorithm was found to have the best classifications with 72.22% accuracy. In the random sampling classification, Extreme Gradient Boosting algorithm was found to have the best classification with 84.44% accuracy.

MAKİNE ÖĞRENMESİ TEKNİKLERİ İLE VERİ ÇOĞALTMA KULLANARAK BİR MOBİL UYGULAMADA KULLANICI YORUMLARININ SINIFLANDIRILMASI

Anahtar Kelimeler

Metin Sınıflandırma, Makine Öğrenmesi, Yapay Zeka, Doğal Dil İşleme.

Öz

Son yıllarda sosyal medya kullanımının artması ile beraber neredeyse her konuda takip edilemeyecek kadar çok yorum bulunmaktadır. Bu yorumlar hem olumlu hem de olumsuz yorumlar içermektedir. Fakat günümüzde çok sayıda yorumu takip etmek neredeyse imkansızdır. Bu çalışmada açık erişimli Anadolu Üniversitesi'nin mobil uygulamasına yapılan kullanıcı yorumlarının çeşitli makine öğrenmesi teknikleri ile metin sınıflandırması yapıldı. Uygulamaya yapılan yorumların içerikle mi yoksa uygulama ile mi ilgili olduğu tahmin edilmeye çalışıldı. Buna ek olarak aşırı örnekleme ve az örneklemenin metin sınıflandırma performansına etkisi incelendi. Bu amaçla sentetik azınlık aşırı örnekleme tekniği (Smote), yoğun en yakın komşu az örnekleme tekniği (CNN) ve rasgele az örnekleme tekniği (RUS) veri setine uygulandı. Mobil uygulamadan alınan 1008 kullanıcı yorumu içerik ve uygulama açısından süreçlerden geçirilerek sınıflandırıldı. Smote aşırı örnekleme sınıflandırmasında ANN algoritması %93.57 doğrulukla en iyi sınıflandırma olarak bulundu. CNN algoritmasında Rassal Orman algoritması %72.22 doğrulukla en iyi sınıflandırmalar olarak bulundu. RUS tekniğinde ise Aşırı Gradient artırma %84.44 doğrulukla en iyi sınıflandırma olarak bulundu.

Alıntı / Cite

Çelik, Ö., Kaplan, G., (2021). Classification of User Comments in A Mobile Application Using Data Augmentation with Machine Learning Techniques, Journal of Engineering Sciences and Design, 9(4), 1398-1407.

[*] Corresponding author: gkaplan@ogu.edu.tr, +90-222-239-3750

Yazar Kimliği / Author ID (ORCID Number)	Makale Süreci / Article Process	
Ö. Çelik, 0000-0002-4409-3101	Başvuru Tarihi / Submission Date	30.03.2021
G. Kaplan, 0000-0002-6393-5546	Revizyon Tarihi / Revision Date	05.08.2021
	Kabul Tarihi / Accepted Date	10.08.2021
	Yayım Tarihi / Published Date	20.12.2021

1. Introduction

Text classification has many different application areas. For example, determining the subject of a new book in the library and determining the appropriate place among the books with a similar theme is a text classification problem. If this is done by computer instead of human labor, the process is called computerized text classification. Many applications such as filtering spam, determining the author or language of a text, document indexing, determining word meaning are examples of text classification applications. On the other hand, other types of classification applications can be realized with text classification solution methods. In the application of the classification of speeches, it can be ensured that speech is assigned to the appropriate class by performing text classification after speech recognition process.

2. Literature Survey

The problem of classification of multiple media such as video was solved by reducing the problem of classification of texts related to multiple media in the document (Tantuğ, 2016). With the advances in technology, text classification has become increasingly important with computer aided communication tools such as e-mail, forum and chat rooms. Since applications such as blog areas, which are constantly updated, are used by millions of people, tracking is very difficult. Although studies focused on content analysis in the past, the number of those focusing on classification of the contents were limited. The reason for this is that it is difficult to classify the mood of a text. With the advances in artificial intelligence studies, this difficulty has been eliminated. Natural language processing models require the use of prior knowledge for text classification. Machine learning approaches use supervised learning algorithms to create annotated models. For text classification, machine learning techniques tend to achieve better results than natural language processing techniques, as they can better adapt to different areas and conditions (Chaffar and Inkpen, 2011).

In literature, Tufekci et al., used the reduced feature vector to classify web-based news texts using Turkish grammar features. Results obtained from Naive Bayes, SVM, C4.5 and RF classification methods were generally higher, but the highest success was obtained from Naive Bayes algorithm with 92.73% accuracy (Tüfekci et al., 2012). Unlike other languages, when the literature is searched, text classification has not been studied very much on Turkish texts. It is noteworthy that the number of Turkish text classification studies is small. Amasyalı and Yıldırım developed a system for text classification and achieved a success rate of 76% (Amasyalı et al., 2004). Amasyalı and Diri used the N-grams character and examined some classification algorithms to determine the author of the text, the type of text and the gender of the author. Success in these problems was 83%, 93% and 96%, respectively [Amasyalı et al., 2006]. Yildiz et al. achieved a 96.25% success rate with Naive Bayes algorithm by proposing a new feature extraction method for text classification (Yildiz et al., 2007). Güven et al. Applied Latent Semantic Analysis method in N-gram word documents (Güven et al., 2006). Özgür et al. developed anti-spam filtering methods for Turkish and agglutin languages in general and achieved 90% success with the help of ANN and Bayesian Networks algorithms (Özgür et al., 2004). Güran et al. applied a variety of classification methods to a Turkish dataset in which the words unigram, bigram and trigram were represented and generally achieved high classification rates. The best results were 95.83%, 93.17% and 52.83%, respectively (Güran et al., 2009). Twitter text classifications were made in some articles (Sriram et al., 2010; Jordan et al., 2019). Short text classification was made and semantic correlations were examined (Shi et al., 2018). Troll account detection was made and 93.93% success was achieved (Bengisu et al, 2021).

3. Material and Method

The purpose of machine learning is to identify complex problems with the help of computers and to present rational solutions to them. This shows that machine learning is closely related to areas such as statistics, artificial intelligence, data mining and computer science, and requires a interdisciplinary study. Regression methods are integral components of data analysis to explain the relationship between one or more explanatory variables and a result variable. One of the first studies with machine learning techniques was done by Vapnik (Vapnik, 1995). In this study, Vapnik used the Support Vector Machine for the solution of the regression problems. The method used in this study showed high success rates in many regression and time series prediction problems (Müller et al., 1997). Today, with the developments in technology, the use of machine learning techniques is increasing. The following part deals with machine learning techniques used in this study.

In the case of a large number of socio economic and medical research results consisting of two or multi-level categorical data, Logistic Regression (LR) Analysis is preferred to investigate cause-effect relationship between the dependent variable and the independent variables.

In the K-Nearest Neighborhood (KNN) classification method, data is assigned to the class most frequently represented among the closest data in the sample. The closest data is determined by calculating the Euclidean distance function. K, in this classification, is the nearest number of data to be considered (Schlögl et al., 2005). Parameters used in this study: n_neighbors: 2, algorithm: ball tree, leaf_szie: 30, metric: minkowski.

Support vector machine (SVM) is a machine learning technique used in various text classification problems. SVMs follow the principle of structural risk minimization. The aim in viewing the data as points in a high-dimensional feature space is to fit a hyperplane between the positive and negative samples to maximize the distance between the data points and the plane. (Schwarm and Ostendorf, 2015). Parameters used in this study: leaf_szie: 30, metric: minkowski.

Naive Bayes (NB) is a simple model that works well on text. This method, commonly used for classification, is based on Bayes theorem, which is a fundamental theorem of probability. Naive Bayes classifiers suppose that an attribute value is independent of classes. This situation is called class conditional independence. Bayesian formula is as (1).

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \tag{1}$$

A decision tree (DT) is defined as a classification that repeatedly divides a dataset into smaller subdivisions relative to a set of tests defined in each node of the tree. Decision trees are not parameterized and do not require a hypothesis about the input data. In addition, they handle nonlinear relationships between data, allow for missing data, and can handle both numerical and categorical inputs (Friedl and Brodley, 1997). Parameters used in this study: criterion: entropy, random_state: 0, splitter: best, max_depth: none, min_samples_split: 2, min_samples_leaf: 1, min_weight_fraction_leaf: 0.0, max_features: none, max_leaf_nodes: none, class_weight: none.

Random Forest (RF) is a statistical learning algorithm that uses a large set of decision tress for both regression and classification tasks. Due to its high accuracy, robustness and ability to deliver information according to the order of its properties, RF is effectively applied to various machine learning applications including bioinformatics and medical imaging. The basic Random Forest classification describes: a descriptive evaluation of training data, the most accurate Random Forest values, and a set of predictive accuracy measures with evaluation of results (Petkovic et al., 2018). Parameters used in this study: criterion: entropy, n_estimaters: 100, random_state: 0, max_depth: none, min_samples_split: 2, min_samples_leaf: 1, max_features: auto, bootstrap: true, oob_score: false, verbosa: 0, ccp_alpha: 0.

The main purpose of Adaboost (ADB) classification is to combine the outputs of a series of weak learners. Each weak learner represents a decision tree and an artificial network. In each round, the weights of misclassified samples are increased, and the weights of the correctly classified samples are decreased. This process continues until changes in weights become insignificant. Parameters used in this study: base_estimator: none, n_estimaters: 50, learning_rate: 1, algorithm: Samme.r, random_state: 0.

Gradient Boosting (GB) is a decision tress-based method and it is based on a gradient increase that differs from the random forest based on the boot stack. The approach is generally used as a basic learner with decision trees in a fixed size and in this context it is called gradient tree strengthening (Hu and Min, 2018). This machine learning method is widely used to provide the most advanced results in some challenging datasets. The main idea of this model is to create a strong classifier by improving weak classifiers through multiple iterations to achieve the desired final combination. Each iteration reduces the remains of the previous model. It is designed to improve the previous result by creating a new combination model in the direction of the remaining (Yang et al., 2018). Parameters used in this study: n_estimaters: 500, max_depth: 4, min_samples_split: 5, learning rate: 0.01, loss: ls.

Extreme Gradient Boosting (XGBoost) is a supervised learning algorithm that helps to predict the result by combining the predictions of a simpler and weaker model. It is faster than Gradient boosting algorithm. It tries to minimize the error by learning iteratively from previously created weak models (Monisha et al., 2018). XGBoost has always been faster than other apps and really faster compared to other algorithms. XGBoost is well-versed in datasets in classification and regression predictive modelling.

Artificial neural networks (ANN) have been developed based on the human brain's biological neural networks and are an information processing system designed to perform the functions of these networks (Celik and Osmanoglu, 2019). An ANN has hundreds of neurons interconnected. Inputs take various forms and structures of information based on a weighting and attempt to learn about the data to produce accurate output. ANNs use a learning rule called back propagation.

3. 1. Resampling Techniques

In resampling techniques, the imbalance is that the sample size of one class is much higher or lower than the other class or classes. Data samples belonging to small classes are misclassified more often than those belonging to common classes (Sun et al., 2007). Oversampling and inadequate sampling in data analysis are techniques used to adjust the class distribution of a data set.

Synthetic Minority Oversampling Technique(SMOTE) is a technique used to appropriately increase the number of samples in your data set. This technique creates new samples from the samples in the dataset. The technique does not change the number of majority samples. Newly created samples are not copies of existing ones. The technique takes the properties in that area for the treated class and its nearest neighbors and creates new samples by combining the characteristics with the neighbors. This approach enhances the properties of each class and generalizes the samples. Smote only changes the number of minorities. Synthetic samples generated from large and relatively few specific decision zones. The working algorithm of the technique is shown in Table 1 (Chawla et al., 2002):

Table 1. Algorithm of Smote

```
ALGORITHM 1: Smote's code block
Input: T ← Number of minority class samples
Input: N ← Amount of Smote
Input: k ← Number of nearest neighbour
If  N<100 then
    T ← (N/100) x T
    N ← 100
end
N ← (int) (N/100)
Numattrs ← 0
Sample[ ] ← [ ]
Newindex ← 0
Synthetic[ ] ← [ ]
for i=1 to T do
    nnarray ← Compute k nearest neighbours for I
    Populate(N, i, nnarray)
end
Populate(N, i, nnarray)
while N<>0 do
    nn ← rand(1)
    for attr=1 to matters do
        dif ← Sample[ nnarray[nn]][attr] – Sample[i][attr]
        gap ← rand(1)
        Synthetic[newindex][attr] ← Sample[i][attr] + gap x dif
    end
newindex ← newindex +1
N ← N – 1
end
```

The Condensed Nearest Neighbor uses the nearest 1 neighbor to decide whether to use a sample. The algorithm is running as followed:

- Get all minority samples into a C set.
- Add a sample from the desired class to C and also add the other sample of that class to the S set.
- Go through the S set and classify each sample with the nearest 1 neighbour rule.
- If the selected sample is not correctly classified, add it to the C set.
- Repeat until there are no misclassified samples in the S set.

Random Undersampling Technique randomly removes samples that belong to the Majority class. You can modify the samples while performing these operations. This is to reduce the unbalanced of the data set. However, there are some drawbacks of these operations, such as increasing variance and removing important instances from the class (Fernández et al., 2018).

3.2. Proposed Method

In this study, text classification was performed with various machine learning techniques of user comments made to the mobile application of Anadolu University. it was estimated whether the comments made on the application were related to the content or the application. In addition, the effect of oversampling and undersampling on text classification performance was investigated. 1008 user comments received from mobile application were classified following examination in terms of content or application. Before this classification process, dataset was pre-processed and the required scaling was made. After this stage, comments were passed through various machine learning techniques and it was determined which statistical classification gave better results in different situations. Python programming language was used for the machine learning techniques applied to the dataset used in the study.

In data Pre-Processing stage; before the machine learning techniques were applied, dataset was analyzed using the following steps. The flowchart of the process is as follows.

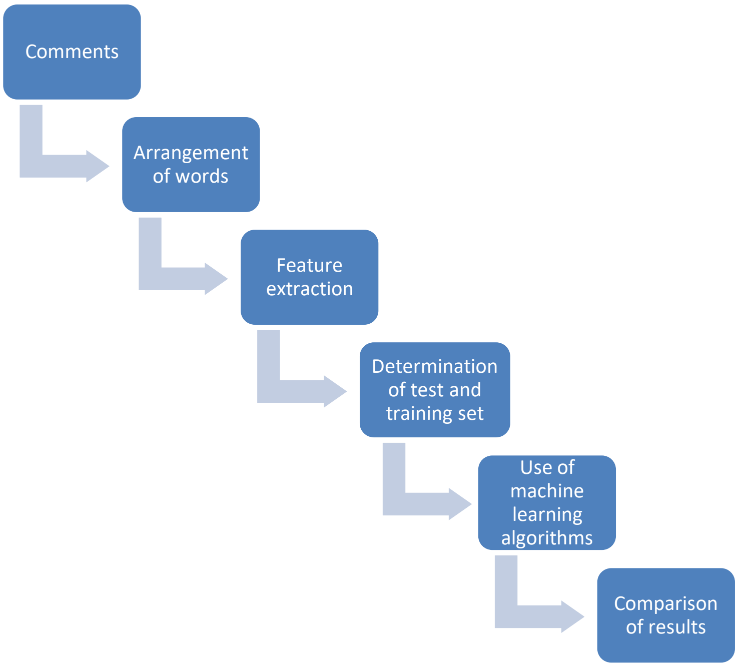
Comments
Arrangement of words
Feature extraction
Determination of test and training set
Use of machine learning algorithms
Comparison of results

Figure 1. Data Pre-Processing State

1- user name, gender and playstore scoring sections were not included in the training because text classification will be made. Only people's comments were included.
2- By applying filtering according to Turkish alphabet, punctuation marks, numbers, etc. non-alphabet characters were deleted.
3- By getting token each word in a sentence:
- All uppercase letters converted to lowercase.
- The words called stopwords in the Turkish language were deleted by applying "stopwords" filter.
- The root of each word was determined using TurkishStemmer and the sentences were re-created with the root form.

4- The most used 1000 words for the processed matrix were determined by "CountVectorizer Max Features" method, and 1008x1000 dimensional matrix was formed (Application: 933, Content: 75).

5- The comments made to the dataset were divided into 933 and 77 comments as application and content respectively and an unbalanced dataset was obtained in the first phase. To get rid of this problem, the dataset was resampled. The unbalanced dataset creates difficulties for typical classifiers such as decision tree induction systems or multilayer sensors designed to optimize overall accuracy without taking into account the relative distribution of each class (Estabrooks, 2000). Resampling methods are commonly used to cope with the problem of class imbalance. Although it is very easy to implement these approaches, setting them in the most effective way involves several challenges. In particular, it should be well established whether over-sampling is more effective than inadequate sampling and which sampling rate should be used (Estabrooks et al., 2004).

6- The test and training clusters used in the dataset were determined as 0.3 and 0.7, respectively.

7- Oversampling was performed using Smote technique.
- Before: Application: 933, Content: 75
- After: Application: 933, Content: 933
- CountVectorizer (1000 feature selected with Max Features parameter)
- 500 of 1000 features selected with Principal Component Analysis(PCA) technique (n=500)

8- Undersampling was performed using CNN technique.
- Before: Application: 933, Content: 75
- After: Application: 156, Content: 75

9- Random Resampling(RUS) was performed.
- Before: Application: 933, Content: 75
- After: Application: 75, Content: 75

10- Comparing to results of three technique.

In statistical statistical analysis, Accuracy ratios were calculated using confusion matrix after the machine learning techniques were applied to the dataset used in the study. The confusion matrix is the matrix that gives the numbers of correctly and incorrectly classified data groups in a dataset.

The Accuracy Rate (ACC), a commonly used success evaluation method, was used in our study. Accuracy rate is the ratio of samples determined by the systems as the correct result(True Positive (TP) and True Negative (TN)) to all sample number. And the error rate is the rate of the sample number calculated false (False Positive (FP) and False Negative (FN)) to all sample number. It is expected to have the accuracy rate is higher than the false rate at the end of the study [Celik and Osmanoglu, 2019]. Confusion matrix is shown in Table 2.

Table 2. Confusion matrix

Data Set	Actual(1)	Actual(0)	Accuracy(%)
Predict (1)	TP	FP	Precision Score
Predict (0)	FN	TN	Negative Predictive Value (NPV)
Accuracy (%)	Recall Score, Sensitivity	Specificity	ACC

Success scores are calculated with the help of the confusion matrix. The success measures and formulas used in our study, which were calculated with the help of Confusion Matrix;

$$ACC = (T P + T N) / (T P + T N + F P + F N) \quad (2)$$
$$Precision = TP / (TP + FP) \quad (3)$$
$$NPV = TN / (TN + FN) \quad (4)$$
$$Recall = TP / (TP + FN) \quad (5)$$
$$Specificity = TN / (FP + TN) \quad (6)$$

Ther are several more accuracy scores calculated with the help of confusion matrix. In addition to, power of the study, type II error, type I error are calculated respectively via TP value, FN value and FP value. For all analysis and processing, a computer with Windows 10 64-bit operating system, quad-core Intel Skylake Core i5-6500 CPU with 3.2 GHz 6MB Cache and 8GB 2400MHz DDR4 Ram memory were used.

4. Experimental Results

In this study, classification of the user comments made to the mobile application of Anadolu University was carried out by using machine learning techniques. Full results of the study is shown in Table 3 and ROC graphics of the study are shown in Figure 2, Figure 3, Figure 4. The confusion matrixes of the methods that give the best results from the techniques are shown in Table4, Table5 and Table 6.

Table 3. Results of the study

	(%)	LR	KNN	SVM	BNB	GNB	DT	RF	ADB	GB	XGB	ANN
Specificitiy	**RUS**	90,48	81,82	93,33	77,78	66,67	78,57	92,86	77,78	90,91	95,24	84,21
	CNN	50,00	39,29	0,00	25,00	46,43	44,00	58,33	50,00	53,33	50,00	40,00
	SMOTE	86,91	58,26	87,76	75,07	90,15	83,44	86,09	77,85	82,33	84,36	91,43
Sensitivity	**RUS**	70,83	50,00	60,00	47,22	50,00	76,47	58,06	72,22	73,91	75,00	61,54
	CNN	72,58	75,00	69,44	69,12	79,55	76,60	75,00	76,92	75,44	75,93	74,47
	SMOTE	96,56	100,0	93,80	95,81	92,66	95,63	96,90	93,62	97,12	96,44	95,71
NPV	**RUS**	73,08	34,62	53,85	26,92	53,85	84,62	50,00	80,77	76,92	76,92	61,54
	CNN	22,73	50,00	0,00	4,55	59,09	50,00	31,82	45,45	36,36	40,91	45,45
	SMOTE	96,64	100,0	93,66	96,64	92,16	95,90	97,01	94,40	97,39	96,64	95,52
Precision	**RUS**	89,47	89,47	94,74	89,47	63,16	68,42	94,74	68,42	89,47	94,74	84,21
	CNN	90,00	66,00	100,0	94,00	70,00	72,00	90,00	80,00	86,00	82,00	70,00
	SMOTE	86,64	34,25	88,01	70,55	90,75	82,53	85,62	75,34	80,82	83,56	91,78
ROC AUC	**RUS**	71	**78**	76	67	60	67	75	73	75	**78**	64
	CNN	60	50	54	50	**65**	47	56	51	64	63	59
	SMOTE	91	65	91	83	92	87	91	85	88	90	**93**
ACC	**RUS**	80,00	57,78	71,11	53,33	57,78	77,78	68,89	75,56	82,22	**84,44**	71,11
	CNN	69,44	61,11	69,44	66,67	66,67	65,28	**72,22**	69,44	70,83	69,44	62,50
	SMOTE	91,42	65,71	90,71	83,04	91,43	88,93	91,07	84,46	88,75	89,82	**93,57**

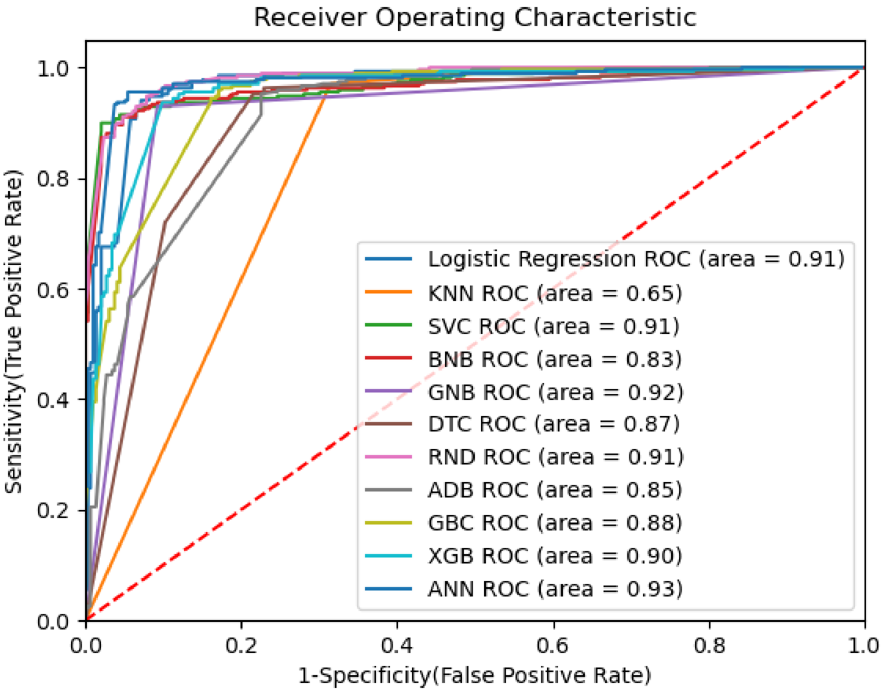

Figure 2. ROC graphs of oversampling technique (SMOTE)

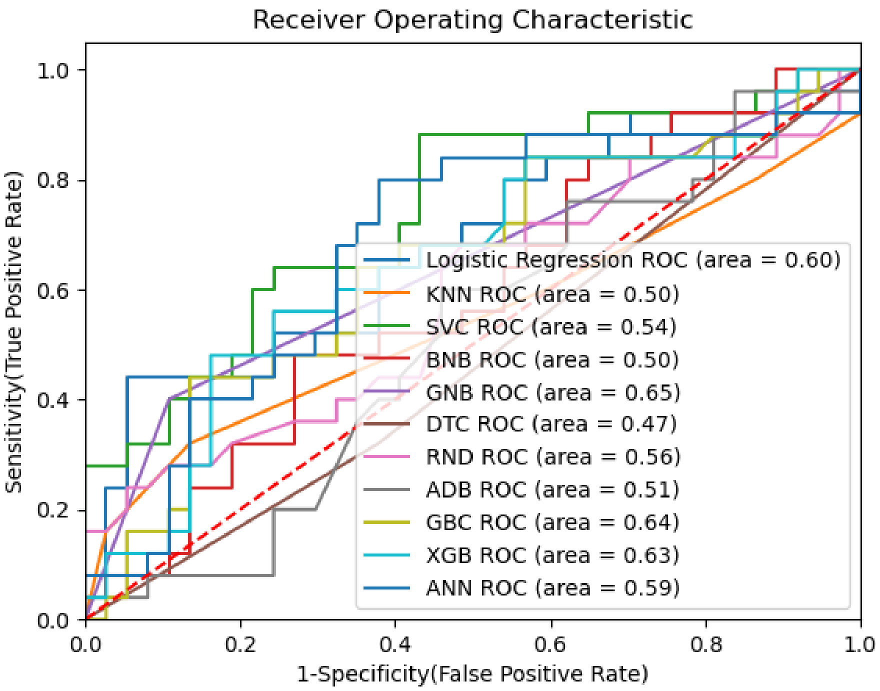

Figure 3. ROC graphs of undersampling technique (CNN)

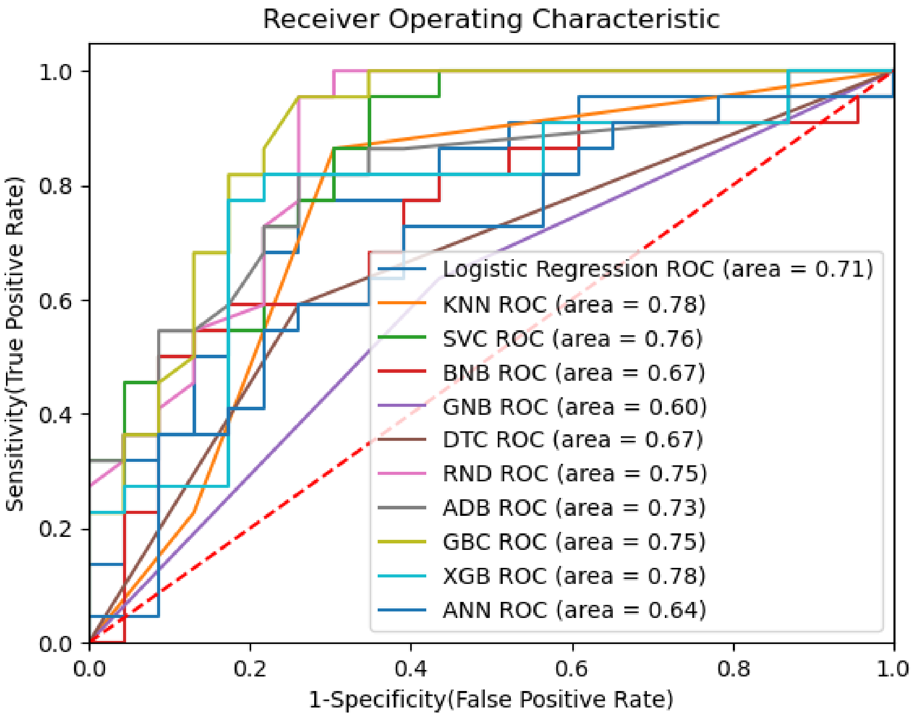

Figure 4. ROC graphs of random undersampling technique (RUS)

Table 4. Confusion matrix of artificial neural network algorithm for smote

Data Set	Actual (App)	Actual (Content)	Accuracy (%)
Predict (App)	268	24	91,78
Actual (Content)	12	256	95,52
Accuracy (%)	95,71	91,43	93,57

Table 5. Confusion matrix of logistic regression algorithm for CNN

Data Set	Actual (App)	Actual (Content)	Accuracy (%)
Predict (App)	35	15	70,00
Actual (Content)	9	13	59,09
Accuracy (%)	79,55	46,43	84,44

Table 6. Confusion matrix of gradient boosting algorithm for RUS

Data Set	Actual (App)	Actual (Content)	Accuracy (%)
Predict (App)	18	1	75,00
Actual (Content)	6	20	95,24
Accuracy (%)	94,74	76,92	84,44

As the number of units in the random resampling classification (n = 150) was low, the success rate was lower than the other two techniques. In addition, the low number of units in the data set applied to random resampling affected the decrease in the number of variables. For this reason, 980 variables (words) were included in this data set.

5. Results and Discussion

With this study, 1008 user comments received from mobile application were classified following examination in terms of content or application. 3 different techniques were used in the examination phase and the results were compared. In the oversampling classification, ANN algorithm was found to have the best classification with 93.57% accuracy. In the undersampling classification, Random Forest algorithm was found to have the best classification with 72.22% accuracy. In the random sampling classification, XGBoost algorithm was found to have the best classification with 84.44%accuracy.

We seen that Smote technique which is one of oversampling techniques makes better text classification than undersampling techniques. We think that this result is due to the following features of the Smote technique: This technique creates new samples from existing samples. It does not change the number of majority samples. Newly created samples are not the same as old samples. Therefore prevents overfitting. The technique takes the features in that field for the operand class and its nearest neighbors and creates new samples by combining the features with the neighbors. Due to similar situations, it develops the properties of each class and generalizes the samples. In addition, some algorithms were observed to be overfitting when applying undersampling and random undersampling techniques.

We realized how important data normalization is in this study process. So, we plan to work on dataset normalization in the future.

Conflict of Interest

No conflict of interest was declared by the authors.

References

Amasyalı, M.F., Yıldırım, T., 2004. Otomatik haber metinleri sınıflandırma, 224-226 pp, (in Turkish).

Amasyalı, M.F., Diri, B., 2006. Automatic Turkish text categorization in terms of author,genre and gender, In International Conference on Application of Natural Language to Information Systems, 221-226 pp.

Bengisu, E. R. D. İ., Şahin, E. A., Toydemir, M. S., Dokeroglu, T. Makine Öğrenmesi Algoritmaları ile Trol Hesapların Tespiti. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 9(1), 430-442 pp, (in Turkish).

Celik, O., Osmanoglu, U.O., 2019. Comparing to Techniques Used in Customer Churn Analysis, Journal of Multidisciplinary Developments, 4(1):30-38 pp.

Chaffar S., Inkpen D., 2011. Using a heterogeneous dataset for emotion analysis in text, Butz C., Lingras P. (eds) Advances in Artificial Intelligence. Lecture Notes in Computer Science, vol 6657. Springer, Berlin, Heidelberg, 62-71 pp.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W. P., 2002. SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, 16, 321-357 pp.

Estabrooks A., 2000. A combination scheme for inductive learning from imbalanced data sets, Diss. DalTech.

Estabrooks A., Jo T., Japkowicz N., 2004. A multiple resampling method for learning from imbalanced datasets, Computational intelligence, 20(1): 18-36 pp.

Fernández, A., Garcia, S., Herrera, F., Chawla, N.V. , 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, Journal of artificial intelligence research, 61, 863-905 pp.

Friedl M.A., Brodley C.E., 1997. Decision tree classification of land cover from remotely sensed data, Remote sensing of environment, 61(3):399-409 pp.

Güran, A., Akyokuş, S., Bayazıt, N.G., Gürbüz, M.Z., 2009. Turkish text categorization using n-gram words, In Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, 369-373 pp.

Güven, A., Bozkurt, Ö.Ö., Kalıpsız, O., 2006. Advanced Information Extractionwith n-gram based LSI, In Proceedings of World Academy of Science, Engineering and Technology, 17:13-18 pp.

Hu J., Min J., 2018. Automated detection of driver fatigue based on EEG signals using gradient boosting decision tree model, Cognitive Neurodynamics, 431-440 pp.

Jordan, S. E., Hovet, S. E., Fung, I. C. H., Liang, H., Fu, K. W., & Tse, Z. T. H. (2019). Using Twitter for public health surveillance from monitoring and prediction to public response. Data, 4(1), 6.

Monisha A., Christina S.S., Santiago N., 2018. Decision Support System for a ChronicDisease- Diabetes, International Journal of Computer & Mathematical Sciences(IJCMS), 7(3):126-131 pp.

Müller K.R., Smola A., Ratsch G., Scholkopf B., Kohlmorgen J., Vapnik V., 1997.Predicting time series with support vector machines, International Conference on Artificial Neural Networks Springer, Berlin, Heidelberg, 999-1004 pp.

Özgür, L., Güngör, T., Gürgen, F., 2004. Adaptive anti-spam filtering for agglutinative languages: a special case for Turkish.", Pattern Recognition Letters, 25(16):1819-1831 pp.

Petkovic D., Altman R., Wong M., Vigil A., 2018. Improving the explainability of Random Forest classifier?user centered approach, Pacific Symposium on Biocomputing, 23:204-215 pp.

Schlögl A., Lee F., Bischof H., Pfurtscheller G., 2005. Characterization of four-class motor imagery EEG data for the BCI-competition, Journal of neural engineering, 2(4): L14.

Schwarm S.E., Ostendorf M., 2015. Reading level assessment using support vector machines and statistical language models, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 523-530 pp.

Shi, T., Kang, K., Choo, J., & Reddy, C. K. (2018, April). Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In Proceedings of the 2018 World Wide Web Conference (pp. 1105-1114).

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010, July). Short text classification in twitter to improve information filtering. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 841-842).

Sun, Y., Kamel, M.S., Wong, A.K., Wang, Y., 2007.Cost-sensitive boosting for classification of imbalanced data, Pattern Recognition, 40(12): 3358-3378 pp.

Tantuğ, A. C. , 2016. Metin Sınıflandırma, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5(2) (in Turkish).

Tüfekci, P., Uzun, E., Sevinç, B, 2012. Text classification of web based news articles byusing Turkish grammatical features, 20th Signal Processing and Communications Applications Conference, 1-4 pp.

Vapnik V., 1995The nature of statistical learning theory, Springer, 2nd edition, New York, USA,32-40 pp.

Yang L., Zhang X., Liang S., Yao Y., Jia K., Jia A., 2018. Estimating Surface Downward Shortwave Radiationover China Based on the Gradient Boosting Decision TreeMethod, Remote Sensing, 10(2): 185.

Yildiz, H.K., Gençtav, M., Usta, N., Diri, B., Amasyali, M. F., 2007. A new feature extraction method for text classification, IEEE 15th Signal Processing and Communications Applications, 1-4 pp.