Turkish Journal of Engineering

Turkish Journal of Engineering (TUJE)
Vol. 3, Issue 4, pp. 168-178, October 2019
ISSN 2587-1366, Turkey
DOI: 10.31127/tuje.554417
Research Article

AUTOMATIC DETECTION OF CYBERBULLYING IN FORMSPRING.ME, MYSPACE AND YOUTUBE SOCIAL NETWORKS

Çiğdem İnan Acı *[1], Eren Çürük [2] and Esra Saraç Eşsiz [3]

[1] Department of Computer Engineering, Mersin University, 33340, Mersin, Turkey
ORCID ID 0000-0002-0028-9890
caci@mersin.edu.tr

[2] Department of Electrical-Electronics Engineering, Mersin University, 33340, Mersin, Turkey
ORCID ID 0000-0002-4631-7834
erencuruk@gmail.com

[3] Department of Computer Engineering, Adana Alparslan Türkeş Science and Technology University, 01250, Adana, Turkey
ORCID ID 0000-0002-2503-0084
esarac@adanabtu.edu.tr

* Corresponding Author
Received: 16/04/2019 Accepted: 03/05/2019

ABSTRACT
Cyberbullying has become a major problem along with the increase of communication technologies and social media become part of daily life. Cyberbullying is the use of communication tools to harass or harm a person or group. Especially for the adolescent age group, cyberbullying causes damage that is thought to be suicidal and poses a great risk. In this study, a model is developed to identify the cyberbullying actions that took place in social networks. The model investigates the effects of some text mining methods such as pre-processing, feature extraction, feature selection and classification on automatic detection of cyberbullying using datasets obtained from Formspring.me, Myspace and YouTube social network platforms. Different classifiers (i.e. multilayer perceptron (MLP), stochastic gradient descent (SGD), logistic regression and radial basis function) have been developed and the effects of feature selection algorithms (i.e. Chi2, support vector machine-recursive feature elimination (SVM-RFE), minimum redundancy maximum relevance and ReliefF) for cyberbullying detection have also been investigated. The experimental results of the study proved that SGD and MLP classifiers with 500 selected features using SVM-RFE algorithm showed the best results (F_measure value is more than 0.930) by means of classification time and accuracy.

Keywords: *Cyberbullying, Automatic Detection, Social Networks, Feature Selection, Classification.*

1. INTRODUCTION

Traditional bullying is defined as aggressive acts that repeatedly occur between individuals with power imbalances that cause harm or distress. Cyberbullying is defined as deliberate and continuous actions that are aggressive towards vulnerable people by using many electronic methods such as internet, e-mail, blog, text and social media message (Snakenborg *et al.*, 2011). Traditional bullying and cyberbullying have similarities, such as power imbalance among individuals, aggressiveness, and the introduction of negative actions. In a study conducted in 2012, the researcher has identified two different types of cyberbullying, as indirect and direct (Langos, 2012). While direct cyberbullying is the only act between the victim and attacker who performs the action, in the indirect cyberbullying, the attacker uses many electronic media to carry action into platforms that can be accessed by more and more people. In addition to these types of actions, cyberbullying actions that may be of different types are still being investigated by researchers (Kowalski *et al.*, 2019). Cyberbullying actions are different in content and contain many features. However, in general, the most common cyberbullying activities include sexuality, gender differences, disability, racism, terrorism, personal character, belief, behavior, external appearance, and weight. The realization of cyberbullying acts, to be able to hide the identity in cyberspace, can be expressed as the key to cyberbullying. Thus, the individuals who cannot make any bullying in real life can turn into cyberbullies (Poland, 2010).

Cyberbullying is carried out in different species depending on the aggressors of attackers and the gender and age of the victim. It has been found that the social-emotional consequences of cyber victimization are comparable to the victimization of traditional bullying (Diamanduros *et al.*, 2008). Initial research has shown that exposure to cyberbullying can negatively affect physical and social development. It can also lead to psychological, emotional and academic problems (Hinduja *et al.*, 2008; Li, 2005; Wolak *et al.*, 2007; Ybarra *et al.*, 2007). In addition, it has been observed that victims of cyberbullying were adversely affected by violence, loneliness, suicidal tendencies (Andreou, 2004). The work done so far, they show that cyberbullying is increasingly occurring and that it causes many negative effects. Therefore, it is necessary to take some precautions to detect and prevent cyberbullying. So the first and most important part of the fight against cyberbullying is the reporting of cyber action. Informing should be done after reporting and detection. In this process, the information of both the victim and the attacker should be shared by informing the official units and internet service providers first. However, reporting gives us information about only bullying acts, and we cannot prevent action. This means that the situation requires online software to automatically detect and prevent cyberbullying.

In this study, automatic detection of cyberbullying was performed on datasets obtained from YouTube, Formspring.me and Myspace social network platforms. A method consisting of four main steps has been identified and implemented on the datasets. First, upper/lower-case conversion, stemming and stopwords removal were used for pre-processing. In the second step, feature extraction was performed and the performances of Multilayer Perceptron (MLP), Stochastic Gradient Descent (SGD), Logistic Regression (LR) and Radial Basis Function (RBF) classifiers were measured using the obtained features. Then, the performances of classifiers are tested using Chi2, Support Vector Machine-Recursive Feature Elimination (SVM-RFE), Minimum Redundancy Maximum Relevance (MRMR) and ReliefF feature selection algorithms to reduce classification time. The best results are achieved by an SVM-RFE algorithm using the selected 500 features. The performance criteria (i.e. F_measure) for test data classification is 0.953, 0.911 and 0.988 for YouTube, Formspring.me and Myspace datasets respectively. In addition, classification durations after applying feature selection algorithms were reduced by 20 times for YouTube, 2.5 times for Formspring.me and 10 times for Myspace datasets.

The remainder of this article follows as Section 2 presents related work on cyberbullying detection. Section 3 gives the details of materials and methods used to detect cyberbullying. Section 4 describes the comparison of the results obtained from experimental studies. Section 5 concludes the study.

2. RELATED WORK

In recent years, a number of studies have been performed for the analysis of cyberbullying detection. When some of these studies are examined, it is seen that classifiers like SVM, k Nearest Neighbours (kNN), Naïve Bayes (NB), J48 decision tree were used frequently (Dadvar and Jong, 2012a; Dinakar *et al.*, 2011; Eşsiz, 2016; Kontostathis, 2009; Ozel *et al.*, 2017).

Noviantho and Ashianti (2017) compared the performances of SVM and NB classifiers as a classification method for cyberbullying detection. Average accuracy of 92.81% was measured for the NB classifier while the SVM had 97.11% accuracy on average in the study conducted. In other steps of the study, the performances of the kNN and J48 decision tree methods were also measured and the best values were obtained as a result of the classification with SVM classifier.

Different labeling and weighting methods are used in addition to cyberbullying detection and text mining classification methods. N-gram method was used for labeling and term frequency * inverse document frequency (tf * idf) method was used for weighting by Yin *et al.* (2009). A supervised machine learning approach was applied, YouTube comments were collected, manually tagged, and binaries and multi-class classifications were applied on three different topics: sexuality, physical appearance, intelligence, and perception by Dinakar *et al.* (2011). In the test results, 80.2% accuracy was achieved in binary classifications and 66.7% accuracy in multi-class tests. In a study conducted at the Massachusetts Institute of Technology, a system has been developed that performs cyberbullying detection in textual contexts from YouTube video comments. The system defines the interpretation as a sequence of classes in sensitive subjects such as sexuality, culture, intelligence, and physical characteristics, and identifies which interpretation belongs to which class (Dadvar *et al.*, 2012b). In a study of the basic text mining system using the bad-of-word approach, an accuracy of 61.9% was achieved in a model that was developed by developing emotional and contextual features (Yin *et al.*,

2009). Bullying traces were defined using a variety of natural language processing techniques by Dadvar and Jong (2012a). Online and offline cyberbullying patterns were examined and emotion analysis system and secret Dirichlet discrimination methods were used to determine the type of bullying. In this method, bullying patterns were not correctly detected.

Cyberbullying has recently been recognized as a serious health problem among online social network users and has an enormous influence in developing an effective perception model. In order to increase the accuracy of the cyberbullying detection studies and to obtain faster methods, feature selection methods are used in the detection of cyberbullying. By Algaradi *et al.* (2016), a set of attributes based on the content of the network, activity, user, and tweet produced from Twitter data is proposed for feature selection. A supervised machine learning method has been developed to detect cybercrime on Twitter and three different feature selection algorithms, Chi2 test, information gain, and Pearson correlation, have been used. As a result of the experiments, a F_measure value of 0.936 was reached in the model developed based on the proposed features. These results show that the proposed model provides a suitable solution for detecting cyberbullying in online communication environments. An effective approach to identify cyberbullying messages from social media by weighting the feature selection was proposed by Nahar *et al.* (2012). Datasets contain data collected from three different social networks: Kongregate, Slashdot, and MySpace. The weighted tf * idf scheme is used on bullying contained attributes. The number of bad words is scaled by two factors. LibSVM has been applied to the classification problem of two classes. For the MySpace dataset, 0.31 and 0.92 F_measure values were obtained for the basic and weighted tf * idf approximations. In another study conducted in this area, the Ant Colony Optimization algorithm was used in selecting subsets of attributes. Formspring.me, YouTube, MySpace and Twitter datasets by combining a novel method is proposed by Eşsiz (2016). In the experimental results, it is observed that the proposed method gives better results than the classical feature selection methods such as information gain and Chi2.

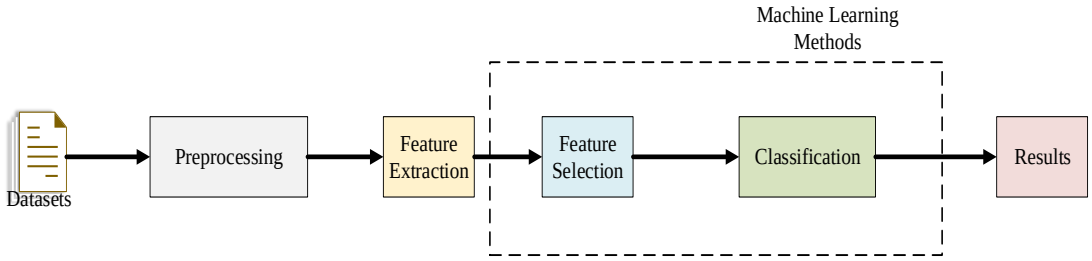

Fig. 1. The architecture of the proposed model

3. MATERIALS AND METHODS

Three different datasets obtained from YouTube, Formspring.me and Myspace social networks were used in the study. Experimental results were obtained by applying four main steps; pre-processing, feature extraction, feature selection, and classification. The architecture of the proposed model consisting of these steps is shown in Fig. 1. Different methods have been tested and applied to determine the optimal method for each step. Classification performances were compared by using SVM, RBF, MLP, SGD and LR classifiers. In addition, MRMR, ReliefF, Chi2, and SVM-RFE feature selection algorithms have been used to demonstrate the effect of selection algorithms on classification time and performance.

3.1. Datasets

The Formspring.me dataset is an XML file consisting of 13158 messages from the Formspring.me website published by 50 different users. This dataset was prepared for a study conducted in 2009 (Yin *et al.*, 2009). The dataset is divided into two classes as "Cyberbullying Positive" and "Cyberbullying Negative". While negative messages represent messages that do not contain cyberbullying, Positive messages represent messages that contain cyberbullying. The Cyberbullying Positive class has 892 messages and the Cyberbullying Negative class has 12266 messages. To separate the dataset into training and test sets, the "holdout" method used in datasets with similar dimensions was applied (Bing, 2011). Some well-known data clusters, such as Reuters and 20NewGoups (20NG), have similar size and number of samples as specified in (Chakrabarti, 2003). So that have been used the same methods as in these examples.

Myspace dataset consists of messages collected from Myspace group chats. Group chats contained in the dataset are labeled and grouped into 10 message groups. For example, if a group conversation contains 100 messages, the first group includes 1-10 messages, the second group includes 2-11 messages, and the last message group includes 91-100 messages. Labeling is done once for every group of 10 and it is labeled whether there are messages containing bullying in those 10 messages. There are a total of 1753 message groups in this dataset, divided into 10 groups with 357 positive and 1396 negative labels.

The last dataset is made up of comments collected from YouTube, the world's most popular video site. YouTube has a large audience, it is becoming a platform where some bad behaviors such as cyberbullying are frequently seen. YouTube dataset used in this study is labeled as positive (hosting cyberbullying) or negative (not hosting cyberbullying) in a study conducted in 2013 (Dadvar *et al.*, 2013). This corpus is a collection of 3464 messages written by different users. Approximately 75% of the samples in all dataset were randomly selected as the training set and the rest were taken as the test set. For all datasets, the number of comments in the training and test clusters for both positive and negative classes is

presented in Table 1.

Table 1. The number of training and test messages of datasets

Dataset	Label	Training	Test
Formspring.me	Positive	669	223
	Negative	9199	3067
Myspace	Positive	267	90
	Negative	1047	349
YouTube	Positive	312	105
	Negative	2285	762

3.2. Pre-processing

In the pre-processing step, combinations of three different pre-processing steps: conversion to lower/upper case, stemming and with or without stop words were used. It is desirable to observe the effects on the classification performance of these pre-processing combinations in the processes carried out in this stage. The first pre-processing criterion is the lowercase/uppercase conversion. Normally, writing a word with a lowercase or uppercase does not cause any change in the meaning of that word, so all the words used are usually converted to lower case letters.

Sometimes, in blogs, forums, and other electronic communication platforms, a word can be written in capital letters to emphasize its significance or to mean loud. In this case, two different cases were examined in the pre-processing step of the case conversion of this study:

1. For the first case, all words are converted to lower case.
2. In the second case, all the words except the words with all the letters written in upper case are converted to lower case, and all the words written in upper case are left with uppercase letters. For example, if a word is written in ABCD format, the word is preserved in upper case format, but if it is written as Abcd or aBCd, that word is converted to abcd format.

The second pre-processing criterion is stemming. In this step, looking at the roots of words, words with the same root are removed from the feature subset. Porter's stemmer method was used to obtain the roots of the words in this study.

As the third and final pre-processing criterion, the meaningless words, called stopwords, have been examined. Meaningless words (prepositions, conjunctions, etc.) are defined as word groups that do not make sense when used alone. Generally, stopwords are not used as the feature in studies like subject classification, because they do not affect classification performance. However, the cyberbullying detection is slightly different from the subject classification, it was used as a pre-processing criterion in this study to investigate the effect of the stopwords on the classification performance.

All possible combinations of the three pre-processing methods mentioned above are considered. For a clearer understanding of these pre-processing steps, the eight different situations shown in Table 2 with code in the binary system.

3.3. Feature Extraction

In this study, only the comments in the dataset are used in the feature extraction step. Features such as username, age, gender are not included. The n-gram method is used for the feature extraction and n = 1 is taken. In addition, a document frequency (df) of 0.001 was used to discard misspelled words and very rarely used words from the subset of features.

Table 2. Binary representation of pre-processing methods

Pre-processing Methods	Lowercase(0) / Uppercase(1)	Stemming off(0) / Stemming on(1)	With Stopwords(0) / Without Stopwords(1)
000	0	0	0
001	0	0	1
111	1	1	1

The main problem in classification is the unbalanced distribution of the classes in the datasets used. In such problems, the main class to be used for analysis is represented by very few examples, relative to the other classes in the dataset. As shown in Section 3.1, the datasets used in this study are unbalanced. The number of positive messages in terms of cyberbullying in the datasets is very small compared to negative messages. When working on such datasets, it is ensured that the class distributions are equalized using methods such as oversampling or undersampling to remove the unbalanced distribution between classes. In oversampling, positive samples are replicated until the class distributions in the dataset are equalized. Since the number of positive samples in the dataset used in this study is not high enough, the undersampling method is not suitable for our datasets. Thus, in order to remove the imbalance of the class distribution, positive samples were replicated for each dataset and oversampling method is applied until a balanced dataset was formed. Table 3 shows the number of features obtained by oversampling and applying 0.001 df.

3.4. Feature Selection

After the feature extraction step, the feature selection methods were applied in order to select the best feature subsets. Chi2, ReliefF, MRMR, and SVM-RFE feature selection algorithms were used separately in this step. The results were compared with each other in order to determine the most suitable method of feature selection for cyberbullying detection by different methods used. The number of features is reduced to 10, 50, 100 and 500 using the feature selection algorithms to shorten the classification duration. Reducing the number of features can sometimes lead to lose some features that have low weight but has a positive effect on classification. This leads to a decrease in classification performance. It is very important to maintain the classification performance while reducing the classification duration.

Table 3. Number of extracted features

Pre-processing Methods	Datasets		
	Formspring.me	Myspace	YouTube
000	1865	9824	14496
001	1653	9547	14215
010	1802	8224	13128
011	1640	8017	12920
100	1980	10212	15961
101	1769	9936	15680
110	1931	8619	14481
111	1767	8413	14273

3.4.1. Chi2

The Chi2 test is (McHugh, 2013) a method used to statistically determine the independence between two variables. Chi2 is a statistical test that is often used to make a comparison between the expected data and the observed data according to a given hypothesis. When the Chi2 test is used as part of the feature selection, it is used to determine whether a particular term is associated with a particular class.

To illustrate the application of the Chi2 test in the selection of features with an example; we assume that our dataset has two classes (positive/negative) with N samples. When given a feature X, it can use the Chi2 test to assess the importance of distinguishing the class. By calculating Chi2 scores for all features, features can be sorted by Chi2 scores, then top-ranked features for model training are selected.

Table 4. Chi2 feature selection

	Positive Class	Negative Class	Total
Containing feature X	A	B	A+B=M
Not containing feature X	C	D	C+D=N-M
Total	A+C=P	B+D=N-P	N

A, B, C, D represent the observed value and E_A, E_B, E_C, E_D indicate the expected value. Based on the Null Hypothesis that the two events are independent, we can calculate the expected E_A value using Eq. (1)

$$E_A = (A + C)\frac{A + B}{N} \tag{1}$$

E_B, E_C, E_D values can be calculated with a similar formula. The basic idea is that if the two events are independent, the probability of X occurring in Positive class instances must equal the likelihood that X occurs in all instances of the two classes. Using the formula of the Chi2 test in Eq. (2). After a few simple steps of the process, a formula is obtained that shows the Chi2 score of the X quality shown in Eq. (3).

$$Chi2 = \sum_{k=1}^{n} \frac{(O_k - E_k)}{E_k} \tag{2}$$

$$Chi2 = \frac{N(AN - MP)^2}{PM(N - P)(N - M)} \tag{3}$$

3.4.2. Minimum Redundancy Maximum Relevance (MRMR)

The MRMR algorithm is a filter-based feature selection method that operates on two conditions, that is, combines the minimum redundancy and the maximum relevance with class labels (Hanchuan *et al.*, 2005). These conditions are combined by calculating mutual information to obtain the value of relevance and redundancy. MRMR is a discriminant analysis method that selects a subset of features that best represent the entire feature space. Calculation of mutual information between two features is shown in Eq. (4).

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{4}$$

The MRMR algorithm uses the values in the feature vector generated in the dataset. f_i represents the value of i in the feature vector, F_i is an example of a discrete random variable of i. Thus, the mutual information between the i and j attributes is expressed as I(F_i, F_j). Mutual information is used not only between two features, but also for calculating the similarity between a feature and a class. In this case, if the class label vector is expressed as h, the discrete randomness corresponding to the class label is denoted as variant H and the mutual information value between i and class is $I(F_i, H)$ (Gülgezen, 2009). MRMR algorithm consists of combining two algorithms. These algorithms are maximum relevance and minimum redundancy. For a dataset consisting of S features, these algorithms are shown in Eq. (5) and Eq. (6).

The maximum relevance:

$$\max W, W = \frac{1}{|S|} \sum_{F_i \in S} I(F_i, H) \tag{5}$$

The minimum redundancy:

$$\min V, V = \frac{1}{|S|^2} \sum_{F_i \in S} I(F_i, H) \tag{6}$$

The MRMR algorithm combines these two algorithms into two different methods: Mutual Information Difference ($max(V - W)$) and Mutual Information Quotient ($max(V \,/\, W)$). The formulas for these methods are shown in Eq. (7) and Eq. (8).

Mutual Information Difference:

$$max \left[I(F_i, H) - \frac{1}{|S|} \sum_{F_i \in S} I(F_i, F_j) \right] \tag{7}$$

Mutual Information Quotient:

$$max\left[I(F_i,H)/\left(\frac{1}{|S|}\sum_{F_i\in S} I(F_i,F_j)\right)\right] \quad (8)$$

3.4.3. ReliefF

The Relief feature selection algorithm is a method of a filter feature selection. Relief algorithm is firstly defined as a simple, fast and effective feature weighting approach. The output of the Relief algorithm is a weight between -1 (worst) and 1 (best) for each feature, features with a higher positive value indicate features indicating a better predictor result. (Rosario and Thangadurai, 2015). The original Relief algorithm is now rarely used in practice. The ReliefF algorithm is used as the most known and most used Relief-based algorithm (Urbanowicz *et al.*, 2017). "F" in ReliefF refers to the proposed sixth (A to F) algorithm variant. ReliefF algorithm has high efficiency and does not limit the properties of data types. Relief ensures discrete or continuous interest in data clusters. While dealing with multi-class problems, the ReliefF algorithm selects the nearest neighbors from each of the samples in different categories. First, after selecting a random sample X from the training set, it finds the nearest neighbors of sample X and subtracts random nearest neighbor samples from the neighbors in the different classes. The formula in Eq. (9) is used in the ReliefF algorithm to update the weight value of the property.

$$W_f^{i+1} = W_f^i + \sum_{c\neq sınıf(x)} \frac{\frac{p(x)}{1-p(sınıf(x))}\sum_{j=1}^{k} diff_f\left(x, M_j(x)\right)}{m*k} - \frac{\sum_{j=1}^{k} diff_f\left(x, H_j(x)\right)}{m*k} \quad (9)$$

3.4.4. Support Vector Machines-Recursive Feature Elimination (SVM–RFE)

The RFE algorithm is an efficient algorithm for feature selection depending on the specific learning model. SVM is used in the learning method of the RFE attribute selection algorithm. So the method used is called the SVM-RFE. The SVM-RFE algorithm is a wrapper feature selection method. The SVM-RFE algorithm is actually a recursive elimination process. In this iteration, the irrelevant, no-comprehension or noisy qualities are removed in order and the important qualities are kept. The SVM-RFE algorithm is basically composed of the following three steps:

1. The SVM classifier is trained with the current samples and information about SVM capability is obtained. For example, when the linear cadence is used in SVM, weight information of each characteristic is obtained.
2. According to some evaluation criteria, the score of each qualification is calculated.
3. The feature corresponding to the smallest score from the current feature set is removed.

The output of the SVM-RFE algorithm is a feature list that lists the features according to their importance. In the case of using Linear Kernel SVM as the learning method in the algorithm, the weight value (ω) is used to calculate the evaluation criterion (c_i) (Wang *et al.*, 2011).

$$c_i = (\omega_i)^2 \quad (10)$$

In Eq. (10), ω_i is i[th] element of the weight vector (W). If the value of c_i is the smallest value i^{th} feature is removed from the feature set.

4. RESULTS AND DISCUSSIONS

In this study, well-known machine learning based classification algorithms such as SGD, RBF, LR, SVM, and MLP are used separately for cyberbullying detection and the results were compared with each other to determine the most ideal classification algorithm. Phyton programming language is used for feature extraction and feature selection, MATLAB and Weka software packages are used to develop classification algorithms.

Eight different pre-processing combinations have been applied and the results have been tested with the SGD classifier. Because it is known that the SGD classifier gives faster results than the other classifiers, the comparisons are made with the SGD classifier at this step. The effect of the pre-processing methods described in Section 3.2 on the classification performance for each dataset has been investigated. The F_measure (Eq. (11)) values of each pre-processing method were compared using the SGD classifier and the results are presented in Fig. 2. Eq. (12) and Eq. (13) are used for F_measure calculation.

$$F_{measure} = \frac{2*recall*precision}{recall+precision} \quad (11)$$

$$recall = \frac{TruePositive}{TruePositive+FalseNegative} \quad (12)$$

$$precision = \frac{TruePositive}{TruePositive+FalsePositive} \quad (13)$$

The best pre-processing combination for Myspace, Formspring.me and YouTube datasets are 111, 011 and 100 respectively. F_measure values for these combinations are calculated as 0.970, 0.934 and 0.842. As is seen from Fig. 2, the 001 combination of the pre-processing methods achieved the best F_measure score for all datasets. The 001 combination is chosen as the default pre-processor for the rest of the paper. Since the datasets used in this study are unbalanced, the number of positive samples is increased by using the oversampling method. The micro (i.e. the harmonic mean of micro-averaged recall and precision) and macro-averaged F_measure (i.e. the harmonic mean of macro-averaged recall and precision) values obtained using the SGD classifier with 001 pre-processing combination are shown in Table 5. It is seen that when the oversampling method is used, the micro and macro average F_measure values increase and reach the same values. This indicates that our dataset has become a balanced dataset.

Table 5. Effect of oversampling on F_measure classification performance

Datasets	With Oversampling		Without Oversampling	
	Macro Average F_measure	Micro Average F_measure	Macro Average F_measure	Micro Average F_measure
YouTube	0.982	0.982	0.569	0.839
Formspring.me	0.951	0.951	0.707	0.930
Myspace	0.983	0.983	0.953	0.968

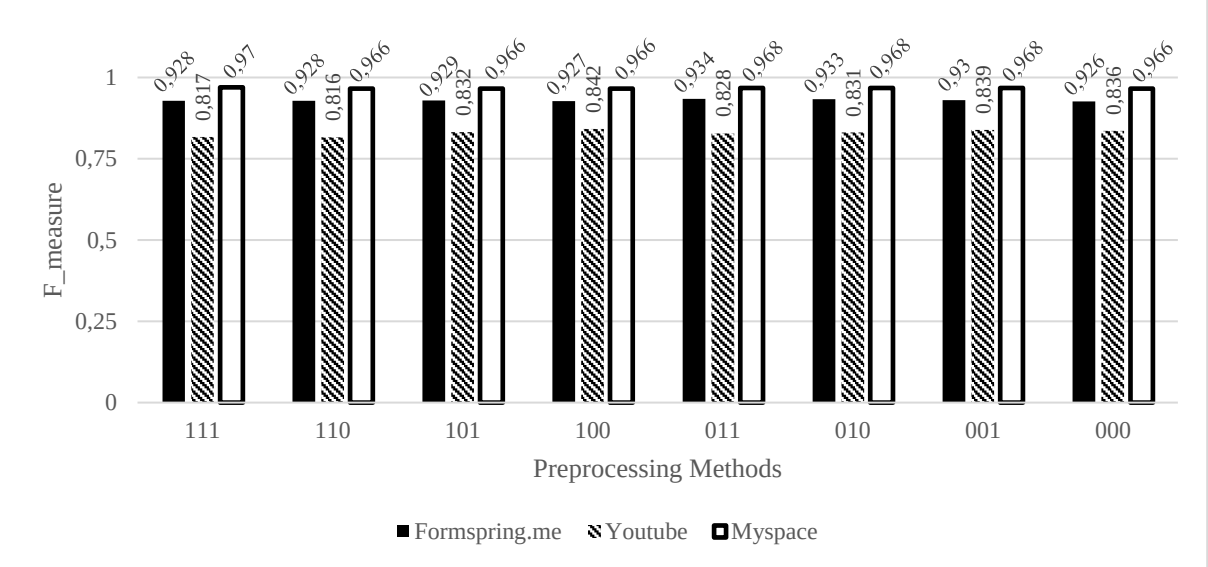

Fig. 2. F_measure comparison of the pre-processing methods

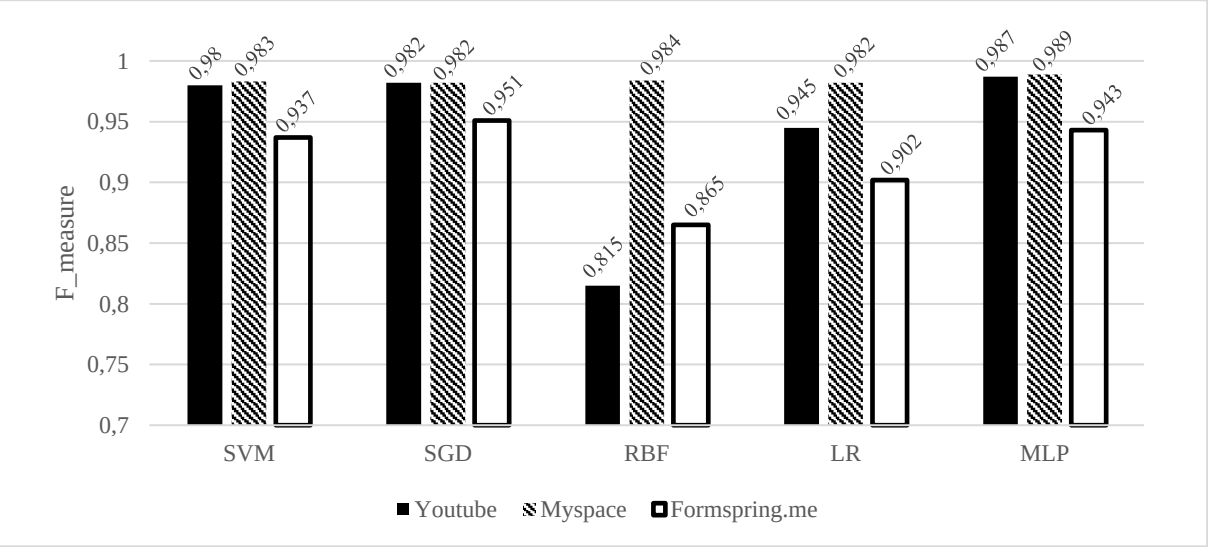

Fig. 3. Performance comparison of SVM, SGD, RBF, LR and MLP classifiers

4.1. Performance Analysis of Classifiers

RBF, SVM, LR, MLP, and SGD classifiers have been tested to have a correct classification method for cyberbullying detection problem. 0.001 df and 001 pre-processing combinations were used in the experiments. A single intermediate layer consisting of 10 neurons was used in the MLP algorithm and the scaled conjugate gradient (*trainscg*) back propagation method was used as the learning method. The classification results are shown in Fig. 3 for all datasets.

MLP and RBF classifiers give the best results for the Myspace dataset. The F_measure values for these classifiers are 0.989 and 0.984 respectively. The best results were obtained in the MLP and SGD classifiers with F_measure values of 0.987 and 0.982 for YouTube dataset; 0.951 and 0.943 for Formspring.me dataset as is presented in Fig. 3. As a result of these experiments, it has been found out that the most suitable classifications for the detection of cyberbullying are SGD and MLP classifiers.

It may be erroneous to evaluate only the classifier's accuracy when comparing classifier performances. The classification speed is also an important parameter as well as the accuracy of classification. Classification times (sec.) of classifiers are shown in Table 6. It can be seen that the LR and SVM algorithms are very slow and the RBF classifier is the fastest algorithm. The SGD and MLP classifiers generally appear to have both a high degree of accuracy and a classification duration at an acceptable level. Given the speed and accuracy of classification at the same time, SGD and MLP algorithms can be considered as the best classifiers.

4.2. The Effects of Feature Selection Algorithms

When you are working on problems where the feature set is too large such as cyberbullying detection problems, the classification time is quite high. In order to decrease the time of classification in such problems, the best sub-features in the feature set are selected. When this method is applied, some important features are lost and the classification performance decreases. Therefore, we have identified the most appropriate feature selection algorithm that preserves the classification performance. The number of features was selected as 50, 100, 250 and 500 using Chi2, ReliefF, SVM-RFE and MRMR algorithms in the experiments and their performances were compared.

Since the comparison results are generally similar, the results obtained from only the YouTube dataset are presented in Fig. 4. The results of all datasets are also summarized in Table 7. In Fig. 4, it can be seen that the best classification accuracy is achieved from the MLP classifier for all feature selection algorithms. The highest F_measure value (i.e. 0.953) is obtained from the SVM-RFE algorithm with 500 attributes.

As is seen from Table 7, the ReliefF algorithm gives lower results than the other feature selection algorithms by means of F_measure. While the MRMR and Chi2 algorithms show accuracy performance with close results, the highest results for all datasets are obtained from the SVM-RFE algorithm. It can be seen that the feature selection paradigm leads to a decrease in classification accuracy. This is because some unselected features have an important role in the training of classification algorithms. However, when the classification time analysis shown in Fig. 5 is examined, it is seen that the feature selection algorithms significantly reduce the classification time. Choosing of 500 as the number of features has the least effect on classification performance and the best results have achieved in terms of time and classification accuracy.

5. CONCLUSION

In this study, a feature-based model is developed on the datasets obtained from YouTube, Formspring.me and Myspace social network platforms. The developed model performs a classification according to whether the comments in the datasets contain cyberbullying. SVM, RBF, MLP, LR, and SGD algorithms are used as classifiers. SGD and MLP classifiers showed better results than other classifiers by giving a F_measure value above 0.930. In order to overcome high classification time problem, MRMR, ReliefF, SVM-RFE and Chi2 algorithms are used as feature selection methods. The SVM-RFE method with 500 selected features reduced the classification time by 20 times for Youtube, 2.5 times for Formspring.me, and 10 times for Myspace datasets while preserving the classification performance.

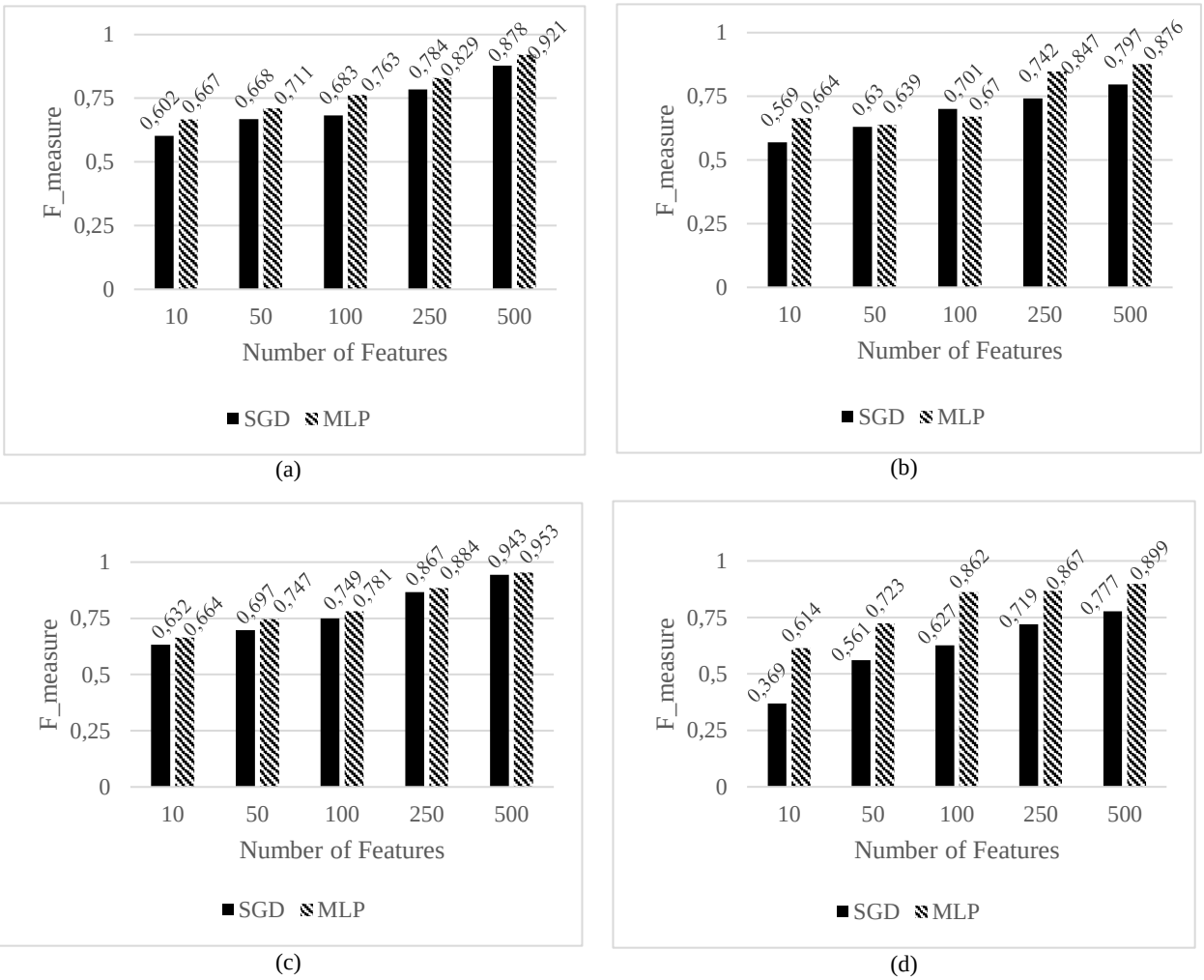

Fig. 4. Performance on SGD and MLP classifiers of the feature selection algorithms for YouTube dataset (a) Chi2; (b) MRMR; (c) SVM-RFE; (d) ReliefF

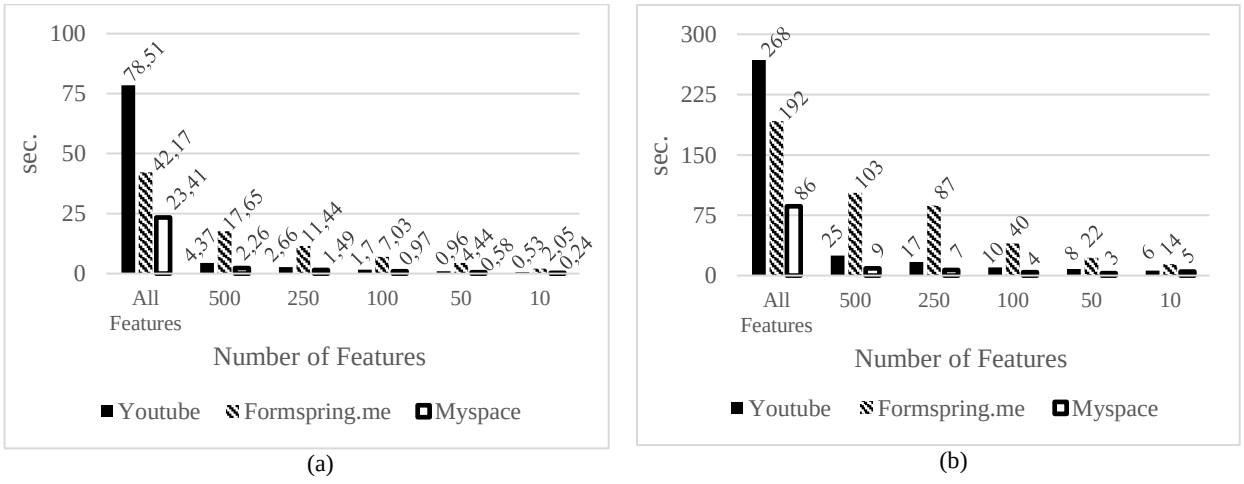

Fig. 5. Comparison of runtime performance of the SVM-RFE algorithm with and without feature selection for all datasets (a) SGD; (b) MLP

Table 7. F_measure results of all experiments

		YouTube		Formspring.me		Myspace	
		SGD	MLP	SGD	MLP	SGD	MLP
MRMR	10	0.566	0.664	0.689	0.779	0.752	0.798
	50	0.630	0.639	0.771	0.802	0.871	0.945
	100	0.701	0.670	0.792	0.821	0.923	0.941
	250	0.742	0.847	0.829	0.837	0.967	0.986
	500	0.797	**0.876**	**0.871**	**0.871**	0.980	**0.988**
ReliefF	10	0.369	0.614	0.362	0.663	0.434	0.653
	50	0.561	0.723	0.419	0.647	0.559	0.755
	100	0.627	0.862	0.464	0.670	0.686	0.771
	250	0.719	0.867	0.561	0.662	0.805	0.832
	500	0.777	**0.899**	0.658	**0.708**	0.853	**0.895**
Chi2	10	0.602	0.667	0.699	0.786	0.640	0.793
	50	0.668	0.711	0.774	0.811	0.870	0.911
	100	0.683	0.763	0.803	0.842	0.908	0.967
	250	0.784	0.829	0.847	0.860	0.969	0.970
	500	0.878	**0.921**	0.878	**0.909**	**0.986**	0.982
SVM-RFE	10	0.632	0.664	0.489	0.760	0.653	0.782
	50	0.724	0.747	0.656	0.776	0.892	0.920
	100	0.749	0.781	0.768	0.824	0.932	0.976
	250	0.867	0.884	0.871	0.837	0.978	0.972
	500	0.943	**0.953**	**0.911**	0.908	**0.988**	0.986
Without any selection		0.982	**0.987**	**0.951**	0.943	0.983	**0.989**

*The best values are emphasized in bold font.

REFERENCES

Al-garadi, M.A., Varathan, K.D. and Ravana, S.D. (2016). "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network" Computers in Human Behavior, Vol. 63, pp. 433–443.

Andreou, E. (2004). "Bully/victim problems and their association with Machiavellianism and self-efficacy in Greek primary school children" British Journal of Educational Psychology, Vol. 74, No. 2, pp. 297–309.

Bing Liu (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, 2nd Ed., Springer Publishing Company, Chicago, USA.

Chakrabarti, S. (2003). *Mining the Web : discovering knowledge from hypertext data*, Morgan Kaufmann, San Francisco, USA.

Dadvar, M. and Jong, F.M.G. de (2012). "Improved cyberbullying detection through personal profiles." International Conference on Cyberbullying, Paris, France.

Dadvar, M., Jong, F.M.G. de, Ordelman, R.J.F. and Trieschnigg, R.B. (2012). "Improved cyberbullying detection using gender information" *12th Dutch-Belgian Information Retrieval Workshop*, Ghent, Belgium, pp 1-6.

Dadvar, M., Trieschnigg, R.B. and Jong, F.M.G. de (2013). "Expert knowledge for automatic detection of bullies in social networks." 2*5th Benelux Conference on Artificial Intelligence*, Delft, Netherlands, pp. 1-7.

Diamanduros, T., Downs, E. and Jenkins, S.J. (2008). "The role of school psychologists in the assessment, prevention, and intervention of cyberbullying" Psychology in the Schools, Vol. 45, No. 8, pp. 693–704.

Dinakar, K., Reichart, R. and Lieberman, H. (2011). "Modeling the Detection of Textual Cyberbullying. The Social Mobile Web", 5th *International AAAI Conference on Weblogs and Social Media*, Barcelona, Catalonia, Spain.

Eşsiz, E.S. (2016). Selecting Optimum Feature Subsets With Nature Inspired Algorithms for Cyberbully Detection, Çukurova University, Adana, Turkey.

Gülgezen, G. (2009). Stable and Accurate Feature Selection, Istanbul Technical University, İstanbul, Turkey.

Hanchuan Peng, H., Fuhui Long, F. and Ding, C. (2005). "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp. 1226–1238.

Hinduja, S. and Patchin, J.W. (2008). "Cyberbullying: An Exploratory Analysis of Factors Related to Offending and Victimization" Deviant Behavior, Taylor & Francis Group, Vol. 29, No. 2, pp. 129–156.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. European Conference on Machine Learning, Springer, Berlin, Heidelberg, pp. 171–182.

Kontostathis, A. and Kontostathis, A. (2009). "ChatCoder: Toward the Tracking and Categorization of Internet Predators". *9th Siam International Conference*

on Data Mining, Nevada, USA.

Kowalski, R.M., Limber, S.P. and McCord, A. (2018). "A developmental approach to cyberbullying: Prevalence and protective factors" Aggression and Violent Behavior, Vol. 45, pp. 20-32.

Langos, C. (2012). "Cyberbullying: The Challenge to Define" Cyberpsychology, Behavior, and Social Networking, Vol. 15, No. 6, pp. 285–289.

Li, T.B.Q. (2005). "Cyber-Harassment: A Study of a New Method for an Old Behavior" Journal of Educational Computing Research, Vol. 32, No. 3, pp. 265–277.

McHugh, M.L. (2013). "The Chi-square test of independence" Biochemia Medica, Medicinska naklada, Vol. 23, No. 2, pp. 143–149.

Nahar, V., Unankard, S., Li, X. and Pang, C. (2012). "Sentiment Analysis for Effective Detection of Cyber Bullying." *14th Asia-Pacific international conference on Web Technologies and Applications*, Springer-Verlag, Beijing, China, pp. 767–774.

Noviantho, Isa, S.M. and Ashianti, L. (2017). "Cyberbullying classification using text mining." *1st International Conference on Informatics and Computational Sciences*, IEEE, Semarang, Indonesia, pp. 241–246.

Ozel, S.A., Sarac, E., Akdemir, S. and Aksu, H. (2017). "Detection of cyberbullying on social media messages in Turkish." *International Conference on Computer Science and Engineering*, IEEE, Antalya, Turkey, pp. 366–370.

Poland, S. (2010). "Cyberbullying Continues to Challenge Educators" District Administration, Vol. 46, No. 5, p. 55.

Rosario, S.F. and Thangadurai, K. (2015). "RELIEF: Feature Selection Approach" International Journal of Innovative Research and Development, Vol. 4, No. 11.

Snakenborg, J., Van Acker, R. and Gable, R.A. (2011). "Cyberbullying: Prevention and Intervention to Protect Our Children and Youth" Preventing School Failure, Vol. 55, No. 2, pp. 88–95.

Urbanowicz, R.J., Meeker, M., LaCava, W., Olson, R.S. and Moore, J.H. (2017). "Relief-Based Feature Selection: Introduction and Review" Journal of Biomedical Informatics, Vol. 85, pp. 189-203.

Wang, J., Shan, G., Duan, X. and Wen, B. (2011). "Improved SVM-RFE feature selection method for multi-SVM classifier." International Conference on Electrical and Control Engineering, IEEE, Yichang, China, pp. 1592–1595.

Wolak, J., Mitchell, K.J. and Finkelhor, D. (2007). "Does Online Harassment Constitute Bullying? An Exploration of Online Harassment by Known Peers and Online-Only Contacts" Journal of Adolescent Health, Vol. 41, No. 6, pp. S51–S58.

Ybarra, M.L., Diener-West, M. and Leaf, P.J. (2007). "Examining the Overlap in Internet Harassment and School Bullying: Implications for School Intervention" Journal of Adolescent Health, Vol. 41, No. 6, pp. S42–S50.

Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A. and Edwar, L. (2009). "Detection of Harassment on Web 2.0.", *Content Analysis in the WEB 2.0*, Madrid, Spain, pp. 1–7.