Academic Platform Journal of Engineering and Science

journal homepage: http://apjes.com/

Application of Grid Search Parameter Optimized Bayesian Logistic Regression Algorithm to Detect Cyberbullying in Turkish Microblog Data

[1]Akın Özçift, [2]Deniz Kılınç,*[3]Fatma Bozyiğit
[1]Manisa Celal Bayar University, Teknoloji Faculty of Hasan Ferdi Turgutlu, Turgutlu, Manisa
[2]Manisa Celal Bayar University, Teknoloji Faculty of Hasan Ferdi Turgutlu, Turgutlu, Manisa
[3]Manisa Celal Bayar University, Teknoloji Faculty of Hasan Ferdi Turgutlu, Turgutlu, Manisa

Research Paper Arrival Date: 12.12.2018 Accepted Date: 24.04.2019

Abstract

There is a huge interaction between users of various social media platforms. This communication produces enormous amount of user data worth to be analyzed from numerous aspects. One of the research area emerging from the user data is a major security issue known as cyberbullying. Since this problem has been recognized as the source of cybercrimes, design of a system to detect cyberbullying attacks/sources through the micro-blog texts is evident. Most of the academic search of this topic has been conducted in English language. The originality of this paper is that we develop an accurate cyberbullying detection system for Turkish language. We used data from Twitter to develop a supervised machine learning model on top of Bayesian Logistic Regression whose parameters are tuned with the use of grid-search algorithm. Since the text data produces a high dimensional training space for machine learning algorithms, we also used Chi-Squared (CH2) feature selection strategy to obtain best subset of features. The optimized version of the proposed algorithm on top of reduced feature dimension has produced an f-measure value of 0.925. Finally, we also compared the results of the proposed algorithm with the frequently used machine learning methods from literature and we provided the corresponding results in related sections.

Keywords: Cyberbullying, Logistic Bayes Regression, Turkish, Machine Learning, Natural Language Processing

Grid Aramayla Optimize Edilmiş Bayes Lojistik Regresyon Algoritmasının Türkçe Mikro Blog Verilerinde Sanal Zorbalık Tespitinde Kullanılması

[1]Akın Özçift, [2]Deniz Kılınç,*[3]Fatma Bozyiğit

Öz

İnternet kullanıcıları ve sosyal medya platformları arasında büyük bir etkileşim vardır. Bu etkileşimin sonucu olarak ortaya çıkan devasa boyutlardaki kullanıcı verileri birçok yönden incelenmeye değerdir. Kullanıcı verilerini baz alarak ortaya çıkan araştırma alanlarından birisi de önemli güvenlik problemlerinden biri olan siber zorbalıktır. Bu sorun, siber suçların kaynağı olarak kabul edildiğinden, mikro-blog metinleri üzerinden siber zorbalık saldırılarını/kaynaklarını tespit etmeyi hedefleyen bir sistemin tasarımı önemli bir konudur. Bu alandaki akademik çalışmaların birçoğu İngilizce dilinde yazılmış metinleri ele almaktadır. Bu çalışmanın özgünlüğü Türkçe metinlerde yer alan sanal zorbalık öğelerini en doğru şekilde tespit edebiliyor olmasıdır. Bu amaçla, Twitter'dan toplanan kullanıcı twitleri üzerinde parametreleri Grid Arama Algoritması ile belirlenen, Bayes Lojistik Regresyon denetimli öğrenme algoritması kullanılmıştır. Metin verilerinin makine öğrenmesi algoritmaları için yüksek boyutlu bir eğitim alanı oluşturması sebebi ile Ki-Kare özellik seçim stratejisi kullanılarak en belirleyici özelliklere karar verilmiştir. Sonuç olarak, çalışmamız özellik sayısının minimum hale getirilmiş versiyonu ile, 0.925'lik bir F-ölçüm değeri üretmiştir. Önerilen yöntemimizin sonuçları literatürde sıkça kullanılan makine öğrenme yöntemleri ile karşılaştırılmış ve ilgili bölümlerde sonuçları paylaşılmıştır.

Anahtar Kelimeler: Sanal Zorbalık, Lojistik Bayes Regresyonu, Tükçe, Makine Öğrenmesi, Doğal Dil İşleme

Corresponding Author: Manisa Celal Bayar Üniversitesi, Hasan Ferdi Turgutlu Teknoloji Fakültesi, Turgutlu/Manisa ftmbozyigit@gmail.com 0236 314 10 10

Doi: 10.21541/apjes.496018

1. INTRODUCTION

The popularity and widespread usage of social networking sites have generated user interactions without geographical location and physical limitations. Any user may be part of a social group and he/she may find opportunities to communicate freely. The result of this interaction is a dynamically growing data which is worthy to be analyzed from different perspectives [1,2]. From cyber-crimes perspective, a few research areas emerging from the mentioned user data are spamming, phishing, malware spread, and cyberbullying [3].

Cyberbullying is defined as "the use of information and communication technology by an individual or a group of users to harass other users" [1,4]. A traditional bully attacks his/her victim before a group that increases the adverse negative effects. In case of cyberbully, the victim is harassed before social groups having enormous number of users. Unfortunately, the social media (e.g., Twitter, Instagram and Facebook) has got innumerable harmful openings from cyber-crimes perspectives [5]. An evaluation of the negative effects of "cyberbullying" highlight that the adverse effect of cyberbullying intensifies with public attacks which is a characteristics of social media [6].

In this context, detection of cyberbullying is an important task to restore the negative results or to prevent the attackers to continue bullying. In other words, being one of the sources of cyber-crimes, design of an intelligent system to discover cyberbullying attacks/sources evolving from social media texts is evident [7].

Intelligent systems are used in numerous domains to automate language processing tasks. In particular, since the user generated data from many social media resources is dynamically increasing in amount, manual investigation of this huge data is impossible. Machine Learning (ML) algorithms are promising solutions to this problem. In the literature, there are many studies conducted on the design of ML systems to detect cyberbullying. However, most of the research is particularly conducted on English language. In this research, we develop an intelligent system to detect cyberbullying attacks on Turkish Twitter data. This work is among the first studies that handles cyberbullying problem through an intelligent system. We therefore first give a brief survey for the most recent English language (or other languages such as Dutch and Spanish) cyberbullying detection systems and then we are going to evaluate Turkish related literature.

The most frequent ML algorithms used in cyberbullying domain are Support Vector Machine (SVM), Naïve Bayes (NB), Random Forests (RF), J48 (Java version of C4.5 algorithm), K-Nearest Neighbour (KNN) and Neural Networks (NN) [8]. One of the recent studies in this domain has been conducted by Cynthia Van Hee et all [9]. The authors made use of SVM algorithm on top of Bag of Words (BOW) model applied to data collected from ask.fm social media [10] in Dutch language. They obtained F-measure value of 55.39%. Another study evaluating a fuzzy-rule based system applied to myspace [11] dataset has produced F-measure value of 91% [12]. In work [13], a NB method has been applied to social media data and the researchers has obtained an average accuracy of 86%. One recent work that has evaluated a NN model on their data and they obtained 87.3%, 89.4% in terms of precision and recall [14]. In their study, Qianjia Huang et. al. [15] used J48 and SVM algorithms and obtained 62.8% and 70.3% F-measures to classify cyberbullying data.

In case of Turkish language based cyberbullying studies there are only one public dataset collected by Bozyigit et. al [16]. The authors have used Twitter as data source and produced a TurkishCyberBulling sample Turkish dataset for cyberbullying detection problem. To the best of our knowledge there are only two studies that use the mentioned dataset to develop a supervised ML model to detect cyberbullying in social media. The first study that uses a newly collected Turkish Twitter dataset to differentiate cyberbullying text from non-cyberbullying text has been conducted in [17]. The researchers have developed a system based on Information Gain (IG) feature ranking method and KNN ML algorithm. The proposed system produces an accuracy of 84% in terms of F-measure. They also have experimented J48, NB and SVM in their study and the mentioned algorithms have produced 54%, 81%, and 74% in terms of F-measure. The second study in Turkish cyberbullying domain has been conducted in [16]. In their study, the authors evaluated a wide range of the algorithms such as SVM, NB, RF, KNN, Bagging, and J48 correspondingly. Before evaluating the algorithms, they applied IG feature ranking to decrease the dimension of feature space while eliminating irrelevant words. After feature selection, the performance of the algorithms in terms of F-measure have been obtained as 91%, 89%, 88%, 88%, 86%, and 73% respectively.

In our study, we used Turkish Twitter data to develop a supervised machine learning model on top of Bayesian Logistic Regression whose parameters are tuned with the use of grid-search algorithm. Since the text data produces a high dimensional training space for machine learning algorithms, we have used Chi-Squared (CH2) feature selection strategy to obtain best subset of features. The parameter-tuned/optimized version of the proposed algorithm on top of reduced feature dimension has produced an F-measure value of 92.5% higher than the performance conducted in [16]. Finally, we also compared the results of the proposed algorithm with Naïve Bayes (NB), Support Vector Machine (SVM), C4.5, Random Forest (RF) and we provided the corresponding results in Section 3.

This work provides three main contributions: i) To the best of our knowledge BLR algorithm has been used the first time in cyberbullying domain, ii) Parameter-tuning concept is first time evaluated in this particular topic and finally iii) The

proposed method increases cyberbullying detection accuracy 1.5% compared to SVM algorithm used in [16].

2. MATERIALS AND METHODS

2.1. Dataset

TurkishCyberbulling dataset [16] consisting of 3000 twits, which are marked as cyberbullying and non-cyberbullying, is used in order to test proposed approach. Fifty percent of the twits in the data set are tagged as positive (including cyber bullying) and the other half is tagged negative (without cyber bullying).

2.2. Proposed Architecture

The architecture of the proposed cyberbullying detection system is given in Figure 1. The system consists of four components and the function of each module is explained in related sections.

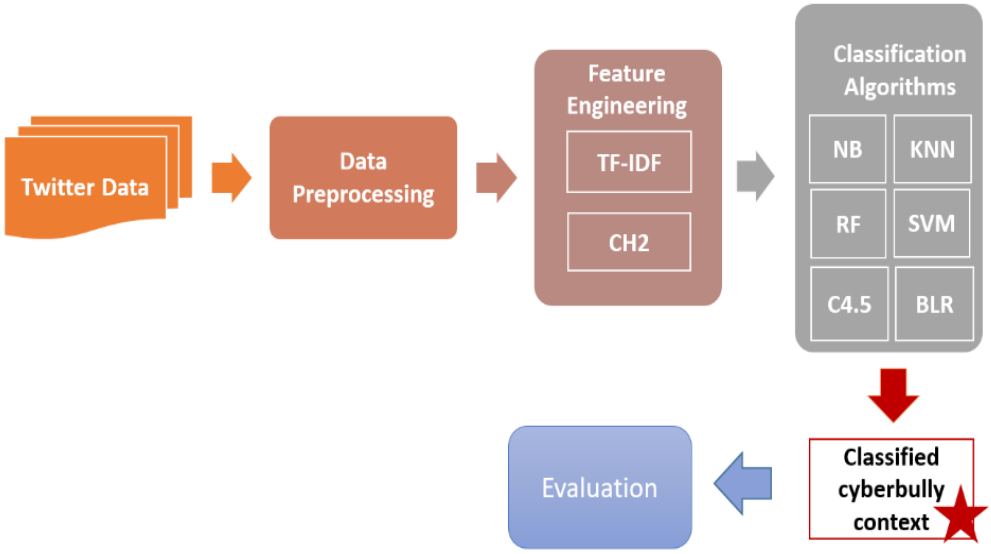

Figure 1. General architecture of the proposed model

As a brief explanation based on Figure 1, the overall system ingests raw Twitter data. Raw data is then pre-processed to increase its quality. Dimension (i.e., 11,534 words) of pre-processed data is then decreased with the use of CH2 strategy and 579 words are obtained. The parameters of BLR algorithm is tuned with the use of grid-search parameter-tuning. As the last step, the prepared dataset is evaluated with tuned BLR algorithm on top of 10-fold Cross Validation (CV). In the following sections, we explain each step of the proposed architecture in detail.

2.3. Data Preprocessing

In this part, a series of preprocessing steps are applied to improve the data set used. Accordingly, non-letter characters, unnecessary website links, and punctuation marks are cleared and all characters are converted to lowercase.

Although there have been studies for the detection and correction of spelling errors for Turkish language, it is seen that they do not perform well. So, we develop an application to normalize tweets including misspelled words. We first create a list of correctly spelled terms related to cyberbullying. Then, we calculate the proximity between the terms in the cyberbullying list and input query. Finally, the misspelled word is normalized considering the alternative correct spelling regarding the value of proximity. The pseudo code of normalization process is shown Figure 2.

Input:
CL: Cyberbulling list
ut: the word in the user tweet
x: the number of characters in the tweet

```
Algorithm:
For each word w in CL Do
        y = number of the characters in the w
        create D matrix including x rows and y columns
        From i=0 To x Do
D [0, i] = i
        End For

        From j=0 To y Do
D [j, 0] = j
        End For

        From i=0 To x Do
        From j=0 To y Do
                If w = ut
                        c=1
                Else
                   D [i, j] = Minimum {{D [i−1, j] + 1,
D [i, j−1]+ 1, D [i−1, j−1] + c}
                End If
        End For
        End For

        distance = D[x-1,y-1]
        max = Maksimum {x,y}
        similarity = (max – distance)/ max
        If similarity > 0.8
                ut = word
        End If
End For
```

Figure 2. Pseudo code of preprocessing

2.4. Data Representation and Feature Engineering

Automated text analysis requires the data to be represented in a suitable form that may be handled by ML algorithms. In this context, Bag of Words (BOW) representation is used to model the text as an unordered collection of its words. In other words, the texts are represented as frequency of the words they contain. The words and their corresponding weights form the mentioned BOW representation. However, the term frequency BOW representation has a "term weight" problem. More clearly, highly frequent terms in Turkish ("şey (thing)", "o (that)", "bu (this)") may dominate the model without containing discriminative information content. One solution to this problem is known as Term Frequency-Inverse Document Frequency (TF-IDF) that rescales word frequency to eliminate domination of irrelevant terms [19]. Having the data pre-processed and represented as TF-IDF BOW model, another key problem arises to be solved before ML methods applied. In particular, another major problem in text mining field is the high dimensional nature of the data. In more clear terms, BOW model representation generates a high dimensional data model having thousands of terms. A major effect of this high dimensional data on ML algorithms is that redundant or irrelevant terms (i.e. features) in feature space reduces accuracy of the algorithms. This problem is solved with the use of various Feature Selection (FS) strategies [20]. The three other benefits of FS are following: (i) better model understandability, (ii) increase in the generalization capability of the model and decrease in over fitting risk and (iii) reduction of computational cost in terms of training and execution time [21]. There are mainly three approaches of FS strategies: (i) Filter Approach. The frequently used methods in this group are IG, CH2, and Correlation Feature filtering (CF). The methods make use of a metric such as correlation, entropy, and mutual information to obtain the most valuable feature subset (terms in text mining domain).

In particular, CH2 filtering approach controls independency between two events. In terms of terms and cyberbullying classes, the filter tests the occurrence of specific word and occurrence of a cyberbullying class to be independent or not. The rank of selected terms is calculated with Equation 1.

$$x^2(D,t,c) = \sum_{et\in\{1,0\}} \sum_{ec\in\{1,0\}} \frac{(N_{etec} - {E_{etec}}^2)}{E_{etec}} \qquad (1)$$

where *et* and *ec* are defined as in Equation 1. *N* is the observed frequency in *D* and *E* the expected frequency.

The other two approaches of FS algorithms are wrappers and embedded strategies. For the details of the two methods the reader is referred to [20] which is an extended survey.

2.5. Baseline Machine Learning Algorithms and Bayesian Logistic Regression Method

After the pre-processing and feature selection steps are utilized, some baseline machine learning algorithms, which are commonly used to classify the textual data, are implemented in the first part of this section Then, BLR algorithm which has been used the first time in cyberbullying domain is executed on the dataset. These methods are described as following.

Naïve Bayes (NB)

It is a frequently used statistics-based supervised learning algorithm based on Bayes' theorem [22]. In NB algorithm, the classification of text documents is implemented by calculating the conditional probabilities on the education dataset. The main advantage of the NB is that it is easy to implement and it performs well on classification problems.

In our study, we experiment Multinomial NB classifier having default value of parameters which provided by scikit-learn library.

Support Vector Machine (SVM)

It is a classification algorithm based on statistical information theory and basic risk minimization. In SVM method, an unlimited number of hyper planes are created to group the samples in the dataset and the most suitable one of these is selected [23]. The advantage of this method is that it can cope with over-fitting. We set the regulation parameter (C) as 1 and kernel as polynomial. Also, the degree of the

polynomial kernel function ('poly') is set as 3 which is default value in Phyton.

K Nearest Neighbor (K-NN)

It is an instance-based classification algorithm which does not have a training phase [24]. In the K-NN algorithm, the input consists of the *k* closest neighbor sample with certain tags in the feature space. We set the value of *k* as 1 in our study. The distance between the samples can be calculated using different metrics such as Euclidean, Manhattan, Minkowski, and Hamming. We measured the distance between the neighbor samples calculating Euclidean distance which is the most commonly preferred one. The advantages of this method are that there is no training phase and it can handle noisy data.

C4.5

It creates a decision tree on the current training set and estimates which class the input data belongs to [25]. This method generates a decision tree by selecting the properties have the most discriminating characteristics of the sample items. The advantages are that it can identify distinctive features, make inference for estimation, and succeed in training sets with lost data.

Random Forest (RF)

It is a supervised learning method in which many decision trees are available. First of all, the properties of the samples in the data set are randomly selected and decision trees are created [26]. Then, the training data is designed to form each decision tree. The RF is created by bringing all the trees together. The classification process is carried out by voting of the trees in the RF and the class with the most votes is returned as a result. This classifier can manage large volume data and work efficiently.

In our study, we experiment RF classifier having default value of parameters which provided by scikit-learn library.

Bayesian Logistic Regression (BLR)

The linear logistic regression is a classification model that aims to predict a target attribute considering one or more predictor attributes. Bayesian model has three basic steps as following (i) setting prior probabilities of parameters, (ii) deciding the marginal likelihood of sample data, (iii) and using Bayes theorem to specify the posterior distribution of the parameters. BLR model finds out the non-linear relation between the predictor attributes and the target attribute applying Bayesian model [27]. The following formula calculates the posterior probability of an instance in a specific class with the integration of conventional logistic function.

$$P = \frac{1}{(1 + \exp(b + w_0 \times c + \sum_{i=1}^{n} w_i \times f(a_i)} \qquad (2)$$

where, 'a_i' specifies the predictor attributes, 'c' is the prior log odds ratio the 'b' is bias and w_0-w_i are that weights calculated after training, and the ith attribute ai is utilized to compute the feature $f(a_i)$. In general, the default prior is used as univariate Gaussian having mean '*0*' (zero).

BLR algorithm is implemented using the methods in Weka. Though the algorithm has many parameters, the most crucial ones affecting the performance are *maxIteration*, *priorClass*, and *threshold*. In this study, these parameters are tuned with the use of grid-search parameter-tuning which is a brute force method to estimate the hyperparameters [28]. It works in an iterative way and attempts to find the best set of values for the parameters. The grid points (range) for the parameters are experimentally specified as shown in Table 1.

Table 1. The grid points of grid search parameters

Parameters	Min-Value	Max-Value	Step-Size	# of Steps
maxIteration	10	100	10	10
priorClass	Gauss.	Lap.	1	2

The results of the grid-search are obtained as *threshold =0.5*, *priorClass= Gaussian* and *maxIteration=100*. We run BLR algorithm with default parameters (BR1) and with grid-searched parameters (BR2). The results of the experiments are given in the following section.

3. MATERIALS AND METHODS

In this section, various pre-processing methods, feature extraction and selection processes on TurkishCyberBullying [16] dataset, and then the widely used classification algorithms are applied to determine cyberbullying.

The evaluation results of each machine learning method are obtained with the use of 10-fold cross validation. The results of the classifiers are evaluated with F-measure criterion. Overall results of the experiments are given in Table 2.

Table 2. Evaluation results of the experimented methods

ML Algorithm	Precision	Recall	F-measure
NB	0.742	0.723	0.732
SVM	0.913	0.914	0.913
K-NN	0.875	0.864	0.869
C4.5	0.738	0.725	0.731
RF	0.887	0.879	0.887
BLR1	0.924	0.922	0.922
BLR2	0.929	0.925	0.925

3.1. Evaluation Metrics

F-measure (Fm) metric is calculated based on confusion matrix outcomes. In other words, Fm is calculated with the

use of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) outcomes. A TP is a result where classifier correctly predicts the positive label. And similarly a TN is a result of the classification if the algorithm predicts the negative label correctly. FP is the case where the classifier predicts negative class as positive. The last confusion matrix term, i.e. FN, is the prediction of positive label as negative.

The precision in terms of TP, FP, TN is calculated with the Equation 3.

$$\textbf{Precision (Pr)} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}} \quad (3)$$

Similarly, recall is calculated with the use of Equation 4.

$$\textbf{Recall (Re)} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}} \quad (4)$$

In order to calculate the accuracy of the proposed model, the harmonic mean of the precision and recall values are obtained and the F-measure is calculated according to the equation given in Equation 5.

$$\textbf{F}_{\textbf{measure}} = \frac{\textbf{2(Pr} \times \textbf{Re)}}{\textbf{Pr} + \textbf{Re}} \quad (5)$$

The best results of the classifiers in detection of cyberbullying are summarized in Figure 3.

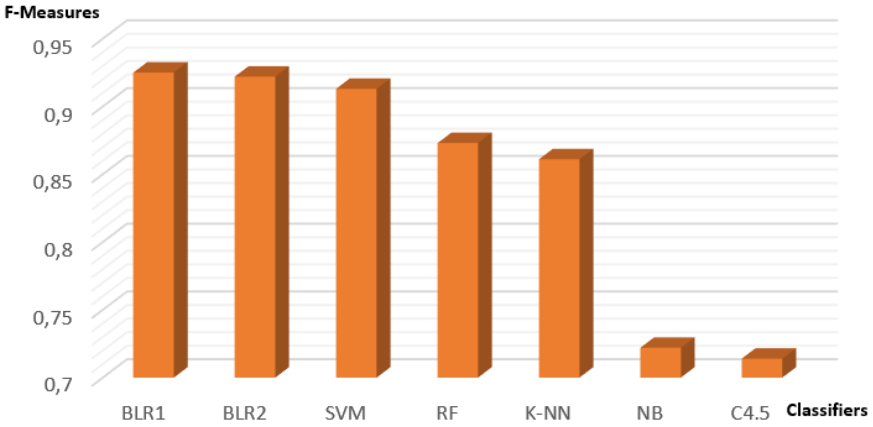

Figure 3. Performance comparision of the experimented methods

Considering the evaluation results of the experimented ML algorithms, it is obviously seen that SVM and RF has better performance scores than the current studies in the literature. Comparision of results is presented in Table 3.

Table 3. Evaluation results of the experimented methods

	[16]	[17]	Our Study
SVM	0.91	0.74	**0.913**
NB	0.89	0.81	0.732
RF	0.88	Not experimented	**0.887**
KNN	0.88	Not experimented	0.869
Bagging	0.86	Not experimented	Not experimented
C4.5	0.73	0.54	0.731
BLR1	Not experimented	Not experimented	0.922
BLR2	Not experimented	Not experimented	0.925

4. CONCLUSION

Social networking applications and corresponding user interactions are the new source of cyber-crimes. Automatic detection of the cyberbullying sources is an important research field. Since the data related to cyberbullying-like risk increases in size, automatic detection of such threads need machine learning algorithms in charge. In this study, a grid-search parameter optimized BLR algorithms is applied to newly collected Turkish cyberbullying dataset and the experimental results have shown that the propped algorithm on top of CH2 feature filtering is precise enough to detect cyberbullying. The result of the optimized BLR is superior to the widely used ML algorithms in the literature. As a future work, we will experiment the combination of various ML algorithms to improve cyberbullying detection performance.

REFERENCES

[1].M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, “Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network”, Computers in Human Behavior, vol. 63, pp. 433–443, 2016. doi: https://doi.org/10.1016/j.chb.2016.05.051

[2].N. Tahmasbi and A. Fuchberger, “Challenges and future directions of automated cyberbullying detection”, in Twenty-fourth Americas Conference on Information Systems, New Orleans, USA, (2018).

[3].M. Arntfield, "Toward a cybervictimology: Cyberbullying, routine activities theory, and the anti-sociality of social media", Canadian Journal of Communication, vol. 40, pp. 371-388, 2015. doi: https://doi.org/10.22230/cjc.2015v40n3a2863
[4].C. Salmivalli, "Bullying and the peer group: A review", Aggression and Violent Behavior, vol. 15, pp. 112-120, 2010. doi: 10.1016/j.avb.2009.08.007
[5].R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth", Psychological Bulletin, vol. 140, no. 4, pp. 1073-1137, 2014. doi: 10.1037/a0035618
[6].E. Menesini et al., "Cyberbullying definition among adolescents: A comparison across six european countries", Cyberpsychology, Behavior, and Social Networking, vol. 15, no. 9, pp. 455–463, 2012. doi: 10.1089/cyber.2012.0040
[7].K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying", ACM Transactions on Interactive Intelligent Systems, vol. 2, no. 3, pp. 1–30, 2012. doi: 10.1145/2362394.2362400
[8].S. Nadali, M. A. A. Murad, N. M. Sharef, A. Mustapha, and S. Shojaee, "A Review of cyberbullying detection . An overview", in 2013 13th International Conference on Intelligent Systems Design and Applications (ISDA), Kuala Lumpur, Malaysia, 325-330, 2013.
[9].C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, and V. Hoste, "Automatic detection and prevention of cyberbullying", in: International Conference on Human and Social Analytics (HUSO 2015), Julians, Malta, 13-18, Oct. 2015.
[10]. "Ask and Answer", ASKfm. [Online]. Available: https://ask.fm/. [Accessed: 06-Dec-2018].
[11]. "Featured Content on Myspace", Myspace. [Online]. Available: https://myspace.com/discover/featured. [Accessed: 09-Dec-2018].
[12]. B. S. Nandhini and J. I. Sheeba, "Online social network bullying detection using intelligence techniques", Procedia Computer Science, vol. 45, pp. 485–492, 2015. doi: https://doi.org/10.1016/j.procs.2015.03.085
[13]. R. I. Rafiq, H. Hosseinmardi, S.A. Mattson, R. Han, Q. Lv, and S. Mishra, "Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network", Soc. Netw. Anal. Min., vol. 6, no.88, pp. 87-103, 2016. doi: https://doi.org/10.1007/s13278-016-0398-x
[14]. K., A. Sudhir, "A predictive model to detect online cyberbullying", PhD Thesis, Auckland University of Technology, 2015.
[15]. Q. Huang, V. K. Singh, and P. K. Atrey, "Cyberbullying detection using social and textual analysis", in Proceedings of the 3rd International Workshop on Socially-Aware Multimedia , Orlando, Florida, USA, 3-6, 2014.
[16]. Bozyiğit, S. Utku, and E. Nasiboğlu, "Sanal zorbalık içeren sosyal medya mesajlarının tespiti", in 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, 2018. doi: 10.1109/UBMK.2018.8566432.
[17]. S. A. Ozel, E. Sarac, S. Akdemir, and H. Aksu, "Detection of cyberbullying on social media messages in Turkish", in 2nd International Conference on Computer Science and Engineering UBMK'2017, Antalya, Turkey, September , 366–370, 2017.
[18]. F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification", in 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, March 2016.
[19]. S. D. Sarkar, S. Goswami, A. Agarwal, and J. Aktar, "A novel feature selection technique for text classification using Naive Bayes", International scholarly research notices, 2014.
[20]. P. Kumbhar and M. Mali, "A survey on feature selection techniques and classification algorithms for efficient text classification", International Journal of Science and Research (IJSR), vol. 5, no.5, pp. 1267-1275, 2016.
[21]. S. George K, and S. Joseph, "Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature", IOSR Journal of Computer Engineering, vol. 16, no. 1, pp. 34–38, 2014.
[22]. A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification",in Workshop On Learning For Text Categorization, July 1998, pp. 41-48.
[23]. A. Soualhi, K. Medjaher, and N. Zerhouni, "Bearing Health Monitoring Based on Hilbert–Huang Transform, Support Vector Machine, and Regression", IEEE Transactions on Instrumentation and Measurement, vol. 64, no. 1, pp. 52–62, Jan. 2015. doi: 10.1109/TIM.2014.2330494
[24]. H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition", in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006, vol. 2, pp. 2126–2136.
[25]. D. Kılınç, A. Özçift, F. Bozyigit, P. Yıldırım, F. Yücalar, and E. Borandag, "TTC-3600: A new benchmark dataset for Turkish text categorization", Journal of Information Science, vol. 43, no:2, pp. 174-185, 2017.
[26]. P. Yildirim and D. Birant, "The relative performance of deep learning and ensemble learning for textile object classification", 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia and Herzegovina, September 2018, pp. 22-26.
[27]. A. Genkin, D. Lewis, and D. Madigan, "Large-scale Bayesian Logistic Regression for text categorization", Journal Technometrics, pp. 291–304, 2012.
[28]. R. A Thisted. "Elements of statistical computing", London: Chapman and Hall, 1988.